COST-SENSITIVE CLASSIFICATION METHODS FOR THE DETECTION OF SMUGGLED NUCLEAR MATERIAL IN CARGO CONTAINERS

A Dissertation

by

JENNIFER BLAIR WEBSTER

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Wolfgang Bangerth
Co-Chair of Committee,	Jean C. Ragusa
Committee Members,	Jim E. Morel
	Mohsen Pourahmadi
Head of Department,	Emil Straube

August 2013

Major Subject: Mathematics

ABSTRACT

Classification problems arise in so many different parts of life – from sorting machine parts to diagnosing a disease. Humans make these classifications utilizing vast amounts of data, filtering observations for useful information, and then making a decision based on a subjective level of cost/risk of classifying objects incorrectly.

This study investigates the translation of the human decision process into a mathematical problem in the context of a border security problem: *How does one find special nuclear material being smuggled inside large cargo crates while balancing the cost of invasively searching suspect containers against the risk of allowing radioactive material to escape detection?* This may be phrased as a classification problem in which one classifies cargo containers into two categories – those containing a smuggled source and those containing only innocuous cargo. This task presents numerous challenges, e.g., the stochastic nature of radiation and the low signal-to-noise ratio caused by background radiation and cargo shielding.

In the course of this work, we will break the analysis of this problem into three major sections – the development of an optimal decision rule, the choice of most useful measurements or features, and the sensitivity of developed algorithms to physical variations. This will include an examination of how accounting for the cost/risk of a decision affects the formulation of our classification problem.

Ultimately, a support vector machine (SVM) framework with F-score feature selection will be developed to provide nearly optimal classification given a constraint on the reliability of detection provided by our algorithm. In particular, this can decrease the fraction of false positives by an order of magnitude over current methods. The proposed method also takes into account the relationship between measurements, whereas current methods deal with detectors independently of one another.

ACKNOWLEDGEMENTS

There are many people that I would like to thank for their assistance in this endeavor. First and foremost is my advisor, Dr. Wolfgang Bangerth. Without his guidance and support, this study would not have been possible. I can not begin to thank him for all the help he has provided.

I would also like to thank several people for their assistance in generating data sets and working with MCNP. Dr. Sunil Chirayath was an extremely valuable resource on campus and generated the initial data sets used to test the algorithms discussed here. Los Alamos National Laboratory, in particular Dr. Avneet Sood and Dr. C.J. Solomon, generously donated computer time and guidance in running MCNP to effectively generate additional data sets as needed.

This project was funded by the Department of Homeland Security Academic Research Initiative Large Grant # 2008-DN-077-ARI018-02/03/04/05. As such, I would very much like to thank both the DHS and the many professors and departments at Texas A&M University who worked to start this effort.

Finally, I would like to thank my family, who put up with much lucubration as I worked with this problem. Thank you to you all for your support.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Classification problems occur in many different aspects of life, be they in the human decision process or in a computational evaluation. Classifying objects involves a complex interplay between the available information, the ultimate goals of the process, and the various consequences of the resulting actions. For instance, when examining information in preparation to making a decision, we analyze the information as a collection and not just as individual measurements, allowing us to utilize correlations between data and improve the accuracy of our decision. Furthermore, we as humans analyze the costs and risks involved in the situation and weight various pieces of information in order to compensate for the varying importance of goals in our decision making process. This could include restricting the likelihood of occurrences of one of the outcomes or weighting the different outcomes to compensate for differences in costs.

In this work, we will discuss the balance of these ideas in the context of detecting smuggled nuclear material entering through US ports. Every year, approximately 40 million shipping containers pass through American ports [54] – this is more than 100,000 crates each day. After the terrorist attacks on September 11, 2001, the border security problem has drawn more attention. This has led to increasing the number of checkpoints at border crossings and more stringent security requirements when flying. It has also facilitated the enactment of several border security laws, including the mandatory screening of maritime cargo as specified in Public Law 110-53 – August 3, 2007: Implementing Recommendations of the 9/11 Commission Act of 2007 [55]. Thus, there is a need for effective, efficient sorting (or classification) algorithms that locate smuggled nuclear material inside of cargo containers.

First, we will discuss the physical and mathematical frameworks necessary for a formal discussion of this source detection problem and some of the challenges it presents. Then, several methods for classification will be considered using statistics, optimization, and machine learning. Each of these methods will be evaluated through examination of the consequences of the classification process as compared with other methods. Extraction of the most relevant information from all of the available information will also be considered and the benefits of data-mining as a precursor to classification will be discussed. Once we have developed the classification and feature selection methods, we will discuss the impact of variations in the problem parameters, including the size and position of source and the shielding effects of cargo, on the effectiveness of the automated numerical methods. Finally,

we will provide some concluding remarks pertaining to the applicability of the classification methods and the challenges that remain to be addressed.

I.1 Basics of the Source Detection Problem

The source detection problem encompasses several physical phenomena – in particular, the radiation produced by the smuggled nuclear material and other background sources, and the physical geometry of the cargo container and its contents. Before we begin to apply mathematical classification frameworks, an understanding of these influences is required. Thus, this section will provide a detailed description of the physical situation that we will examine in terms of radiation physics, the problem geometry, and practical considerations of cargo shipment, as well as some contemporary detection systems.

I.1.1 Radiation Basics

Before we discuss our problem and the challenges it presents, we need to take note of some basic physics that we will make use of here. In general, radioactive materials can emit many different forms of radiation including photons, neutrons, electrons, positrons, and alpha particles. These particles can be absorbed by or scattered off of material that the radiation encounters as it travels. Sometimes these interactions can induce additional radiation. Such interactions occur randomly with a measurable probability determined by the interaction cross sections of the material, which are dependent upon the type of material and the energy of the incident particles. These interactions affect the amount of radiation that will be observed by a detector. In this case, we say that the radiation is *shielded* by the other materials. This effect will play an important role in our studies.

Let us consider a localized source of radioactive material that is surrounded by a vacuum with no other radiation sources present and suppose it emits radiation isotropically, as do all specific materials discussed in this study. If we are given two identical detectors that are both placed at a distance r away from the center of the source material, then each detector will observe, on average, the same amount of radiation from the source. This is because each of the detectors occupies the same solid angle. As we move one of these detectors farther away from the source, the amount of observed radiation will decrease as $1/r^2$ in proportion to its decreasing solid angle. The real world, of course, is more complicated than this hypothetical case, but there is still a strong correlation between distance from the source and a decrease in the amount of source radiation detected. This effect will have a significant impact on our source detection problem as discussed in Sec. I.1.2.

Some of the most prominent photon producing materials are Plutonium, Uranium–233

and Uranium–235. These are collectively referred to as *Special Nuclear Material* (SNM). However, these are not the only materials that produce radiation. Naturally occurring radioactive materials (NORM) such as bananas, potash and concrete all emit radiation with enough strength to potentially mask out the signal from small quantities of SNM. These isotopes and others can be identified by examining their radiation spectra. The strength and energies of photons emitted by a material can narrow down which isotopes have decayed in order to produce the observed radiation. In this study, we will concentrate on the detection of Highly Enriched Uranium (HEU), which is predominantly a mixture of ^{235}U and ^{238}U. Natural Uranium contains less than 0.72 wt. % ^{235}U. In contrast, HEU contains a minimum of 20 wt. % ^{235}U and weapons grade HEU is typically much larger (over 70 wt. %). As the quality of the HEU increases, the percentage of ^{235}U increases and that of ^{238}U decreases. For HEU, the most commonly emitted photons have energies of either 186 keV (produced by the ^{235}U), 766.4keV or 1.001 MeV (produced by the ^{238}U) [44]. The lower energy photons are more easily absorbed and, therefore, less likely to exit the cargo container under observation. Furthermore, those that do survive are not easily distinguishable from the vast amount of background radiation. Thus, this study focuses on the 1 MeV line in order to distinguish HEU from NORM. There are several options for detecting photons available currently. At 1 MeV, High Purity Germanium (HPG) detectors have a 2 keV energy resolution for photons and NaI (Sodium Iodide) detectors have a 20 keV energy resolution window. Therefore, it is possible for researchers to examine fluctuations in the 1 MeV peak and use the methods described in this study to determine whether or not these fluctuations provide enough information in the classification of cargo containers, as described in the next section.

I.1.2 Problem Description

Stated simply, our goal in this study is to use knowledge of the radiation exiting a standard cargo container in order to determine whether or not there is illicit radioactive material within the crate. We must balance the necessity for accurate classification with the costs of implementing any algorithm. To be completely accurate in our classification, we would need to open a container and test each and every material inside to determine which objects are composed of radioactive material. However, this is prohibitively expensive, so we must use more indirect measurements – in this case, the photon count rates for detectors located outside the container.

In the source detection problem for ocean-going cargo crates, we have several physical challenges that need to be kept in mind. A standard cargo container has dimensions of 20 ft long by 8 ft wide by 8.5 ft tall and a total weight limit of about 21,000 kg (depending on the

manufacturer). Within these confines, an SNM source can be placed at any location with any variety of other material that the crate owner specifies. According to the International Atomic Energy Agency, 25 kg of HEU is a significant quantity of interest and enough to cause serious damage if it were to escape detection [23]. This amount of Uranium could be packaged in a space of less than half a cubic foot, which is very small in comparison with the total volume (1360 cubic feet) of the container. The studies in this work concentrate on methods that may be used to locate a 1 kg HEU source inside of a filled cargo container. With such a small source, the radiation produced by the source is masked by background sources and further disguised by the other material in the cargo container that shields some of the radiation that the source produces. HEU is in general harder to detect than other sources as it produces lower energy photons [30, 31].

We will be considering methods here that do not delay the cargo container unduly as it passes through the port in order to minimize the impact of these tests on the commercial transport process. Therefore, our detectors are placed outside of the cargo container along the large vertical sides of the container, much like the radiation portal monitors currently in place [24]. For a localized source placed in the center of the cargo container, the nearest detector will be approximately 4 ft away and radiation may have to pass through a significant amount of other cargo before reaching the detector. As mentioned in the previous section, the distance between the detector and source directly impacts the amount of observed radiation and, thus, our ability to accurately detect the material. On the other hand, we know that the radiation falls off in a predictable way (see Sec. I.1.1), so there are correlations in detector measurements as a result of their spatial relationship. We will make use of these relationships to improve the accuracy of our detection algorithms. The exact characterization of the detector configuration used in this study can be found in Ch. II.

The last major physical piece of the puzzle is the contents of the cargo container. Initially, we only have knowledge of the cargo contents as given in the manifest and we can only consider that information reliable in so far as we trust the person who wrote the description. Since every material interacts differently with radiation, the cargo in the container will affect not only the amount of radiation from the HEU source that is observed, but also the amount of background radiation that reaches the detectors. Without more accurate knowledge of the container contents, it is difficult to determine how radiation interacts with the cargo. This introduces a significant statistical variation to our measurement data, as will be shown in Ch. VI.

I.1.3 Current Detection Methods

There are several detection systems currently in use to detect nuclear material – fixed radiation portal monitors, personal radiation detectors, hand-held gamma and neutron search detectors, and hand-held radio-nuclide identification devices [27]. In most cases, the background radiation levels must be established. The detection equipment is then set to alarm when radiation exceeds four standard deviations above the mean [25]. In order to determine the exact type of nuclear material present, radiation spectra are analyzed in a process known as energy windowing – count rates for various energies of particles are determined and the percentage of counts in each window is compared to determine the type of material [15]. Neither of these methods fully utilize the correlations between spatial detection measurements nor do they control the global error rates of a system of detectors, which will be the focus of our studies.

I.2 Translating Physical Observations into Mathematics

In the previous section, we concentrated on the physical realities of the source detection problem, but a mathematical description of our problem is necessary for rigorous analysis. Two approaches will be used in this study – one using the statistical formalism already available for classification and the other using optimization algorithms. Before we discuss the particular terms that will be required for each individual approach, there are some common concepts that we need to grasp.

First, recall the desired outcome of our research – to decide whether or not a shipping crate contains a smuggled HEU source using only detector measurements obtained from devices outside of the container. There are several ways that we could approach this kind of problem. For example, an *inverse problem* approach would attempt to uniquely determine a complete model of the container contents that most closely matches the given detector measurements. However, this requires extensive computational resources in order to match every single container coming into port and it usually provides a lot of extraneous information [1]. Typically, we do not have enough information to implement this concept in practice. In order to make a decision regarding whether or not an HEU source is in the container, we only need to know if it exists in the crate, not the exact source position nor the nature of all of the other materials inside the container. Therefore, we want to create a decision algorithm that compromises between the computational resources and the accuracy of the solution. One such approach phrases our situation as a classification problem:

Definition I.2.1. A **classification problem** is the problem of assigning a label (from a fixed label set) to each member of a population based on a collection of measurements associated with each object in the population.

I.2.1 Developing a Labeling System for Classification

In order to develop a classification algorithm, we need to identify both a set of labels that we will assign and a set of measurements or features that we will use in order to assign such labels. Then, an algorithm which analyzes a measurement and assigns the label is referred to as a *classifier* or *decision rule*. Usually, the labels are determined by the ultimate purpose of the classification. For instance, in our particular situation, the objects under consideration are the cargo containers, the measurements through which we obtain our knowledge of the cargo container come from the exterior detectors, and we label the container as belonging to one of two categories – crates containing an HEU source and those that do not. A common shorthand used throughout this text will be to call crates containing an HEU source "dangerous" or belonging to set D. Similarly, cargo containers without such a source will be denoted as "safe" or members of set S.

I.2.2 Choice of Feature Space

Now that we have our set of objects and the categories we want to sort into, we will need to define the set of measurements that will be used as a basis for our decision. Before we discuss what this means in terms of our problem, it behooves us to have an abstract definition of the feature or measurement space:

Definition I.2.2. The **feature space**, usually denoted $\mathbb{S}^n$, is the set of all possible measurements or quantities, x, that will be used to make the decision as to which category the item belongs.

Remark I.2.3. The feature space is not unique for any given problem and can be chosen by the researcher to fit whatever criterion he chooses. Furthermore, once a set of measurements is taken, they can be transformed, combined or disregarded to create a new feature space that is more useful in the classification process. We will discuss some of these types of modifications further in the remainder of this section as well as in Ch. V.

It should be noted that good choices for the feature space will drastically affect the algorithm generated and is often the most difficult portion of formulating any classification algorithm [46]. It is wise to choose items that vary as little as possible within each class of objects, but change more obviously between classes. The choice of measurements in this

study is based on experience in the nuclear detection field, where studies have been made to characterize the radiation produced by materials. In this particular classification problem, there is no single measurement that will allow for classification with complete accuracy, except chemical examination of all the materials in the container. Unfortunately, this requires opening and physically searching every container, which is extremely expensive in terms of money, time and man-power and, as such, is infeasible in terms of a real-world approach. Therefore, we will restrict ourselves to non-invasive measurements, which are more indirect and less reliable.

Utilizing independent features is helpful in developing quick and accurate classification methods. Generally, two features are independent if knowledge of the value of one feature gives no information about the value of the other feature. By choosing independent features from among the entire set of available data, one can reduce the number of measurements (the dimensionality n of the feature space $\mathbb{S}^n$) by eliminating redundant information while maintaining the level of accuracy in our decision. A more mathematically rigorous explanation of independence requires probability theory (Def. I.2.5), but as a heuristic example, suppose we use volume (V), mass (m) and density (ρ) to determine whether an object is metal or not. Recalling basic physics, we know that we can relate these three quantities using the formula $\rho = m/V$. From this we can see that knowing any two quantities, will allow us to calculate the third and thus, these features are not independent. However, any two of these features will be independent. This idea extends easily to other fields in that one always wants to determine a set of features which provides all the necessary information to make an accurate classification without duplicating data or adding too much computational complexity.

For the detection of smuggled sources, we have chosen to use 1 MeV photon count rates from a spatially distributed array of detectors. Therefore, each measurement x is actually an n-dimensional vector of count rates where each component corresponds to readings from a specific detector in the array. Frequently, a set of detector readings will be referred to by the labels D and S and, in that case, we really mean that a cargo container producing the given measurements is a member of the specified class.

Since each detector in our scenario is counting the number of photons that hit its surface, the measurement may be viewed as a positive integer. Hence, we can define our feature space as $\mathbb{N}^n \subset \mathbb{R}^n$ (see Sec. I.1). Knowledge of the values which measurements can obtain is not enough information to perform classification accurately. Additional structure on the space must be imposed to denote which values are more likely to be observed. This can be given in the form of probability distributions or random samples from such a distribution, as discussed in Sec. I.2.4.

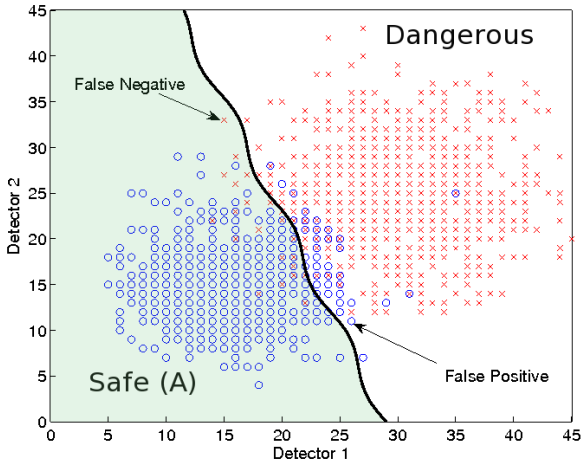


Figure I.1: Shown here is a pictorial representation of the result of the classification process. The red crosses represent measurements, x, that would be obtained by scanning a cargo container that contains a radiation source. The blue circles represent measurements obtained from a harmless cargo container. The decision boundary (black wavy line) is determined by the function $f(\cdot)$ and divides the feature space into two regions – the green space is the set of measurements that will cause an object to be assigned the "safe" label, i.e., where $f(x) = S$, and any other measurement is labeled as "dangerous," where $f(x) = D$. The various types of errors in classification (false positives and false negatives) are also marked here, which will be discussed further in Sec. I.2.5.

I.2.3 Arriving at a Decision Rule

Once we have the feature space and labels defined, we need a way to develop our decision rule – our method of assigning labels. In our particular problem, we want to sort objects into two categories based on detector measurements, which can be done by dividing the feature space into two disjoint pieces – one region where measurements are more likely to have been obtained by measuring a safe cargo container and the complement of this region containing measurements from a container with a source. Thus, our task is to find the appropriate boundary between the two regions in $\mathbb{S}^n$. This can be denoted by a function $f : \mathbb{S}^n \to \{S, D\}$, as demonstrated in Fig. I.1. Ideally, the optimal boundary would be found in a solution space containing all possible functions that partition the feature space into two sets. However, this space is quite large so the solution space of the classifier is often restricted to linear or piecewise linear functions. The ultimate classification of objects will be given by testing a measurement x from that container to determine its membership in one of the regions of feature space.

There are several choices for methods, usually referred to as *machine learning methods*, for developing the final classification rule. Such methods include the classical theory of statistics, optimization methods or support vector machines. Several of these methods will be discussed in the following pages. By common consent of the community, a good decision algorithm, $f(x)$, is one that

- Makes as few errors in classification as possible (see Sec. I.2.5).
- Is an accurate predictor for any object and not just those used to develop the method.
- Balances the relative costs of different kinds of mistakes in classification, when necessary, (see Sec. I.2.5).
- Provides a (relatively) simple rule for making a decision on a single point in feature space.
- Can be generated or adapted to new information quickly.

Upon implementation, various methods must also be compared to find the optimal classification method, as depicted in Sec. I.3. We will address these points for each method discussed in this study.

Generally, algorithms to create classification rules are lumped into one of four categories based on what information we have about the distribution of measurements (see Sec. I.2.4): Analytical, Supervised, Unsupervised and Reinforced. The first requires the analytical form of the distribution of measurements or a good approximation to it and then uses this along with prespecified formulas to create the classification rule. The other three categories use a set of random samples drawn from the distributions of measurements in order to do the same. If it is a supervised algorithm, then the user provides a data set containing pairs of data points from the feature space and the appropriate label under which they should be classified. Unsupervised algorithms require only the data points from the feature space, not the labels, and use clustering of the points to make the classification rule. Similarly, the reinforcement algorithms require only the data points and a starting classification rule. These points are then processed by the algorithm and a user decides whether the classification of that point is correct or incorrect. This information is then used to adapt the algorithm so that this point in feature space is more accurately classified in future trials.

I.2.4 Probabilities of Measurement

Before using the categories and feature space to create a decision rule, we need a way to discuss all of the variations and uncertainties that arise in this problem. As mentioned in Sec. I.1, the production, scattering and absorption of photons are all random processes. Therefore, if we take measurements, x, from the same cargo container and detector set up multiple times, then the photon count rates may differ each time. To account for this, we must obtain some information about how likely it is that a set of measurements can be linked to a specific type of object (a safe or dangerous container). We can do this by analyzing the probability distributions of the measurements. This information can be given as an analytical form of the distribution of each population or as a set of sample measurements.

In the analytical setting, we can use the common language of probability and statistics to discuss the variations of our measurements. For any measurement, x, we can denote how likely we are to observe such a measurement by the *probability density function* or pdf, $p(x)$. We can also talk about how likely we are to observe a specific phenomenon, such as $x \in A$, for which we will use the notation, $P(A) = \sum_{x \in A} p(x)$, where the capitalization indicates that we are referring to the probability of the event $x \in A$ and no longer concerned with a specific point x. We can further delineate our data by recalling that our problem considers two disjoint populations, S and D, that completely cover the sample space from which we have measurements x. This will give us additional information to incorporate into our probability distribution in the form of conditional distributions:

Definition I.2.4. If A and B are events in a sample space S and $P(B) > 0$, then the **conditional probability** of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{I.1}$$

In words, this conditional probability is the probability of both events occurring as scaled by the probability of the prior knowledge that B occurs.

We would like to know how likely the object is of a specified type given that we observe a specific measurement x, i.e., $p(S|x)$ and $p(D|x)$. To determine these quantities straight from Def. I.2.4 would require knowledge about how likely we are to obtain every single possible measurement in order to determine the denominator of Def. I.2.4. This would be extremely difficult information to gather since our feature space could have a very large dimension if we have a large detector array. However, it is relatively easy to measure the conditional probabilities $p(x|S)$ and $p(x|D)$, which indicate how likely it is to obtain the

measurement x when observing an object from the specified population. Similarly, it is relatively easy to estimate the prior probabilities $P(S)$ and $P(D)$, which give the fraction of objects from each population present in the sample we are studying. From the quantities we do know, we can use Bayes' Rule to calculate the probability that a container with measurement x belongs to population S or population D:

$$\begin{aligned} p(S|x) &= \frac{p(x|S)P(S)}{p(x|S)P(S)+p(x|D)P(D)} \\ p(D|x) &= \frac{p(x|D)P(D)}{p(x|S)P(S)+p(x|D)P(D)} = 1 - p(S|x) \end{aligned} \tag{I.2}$$

With this information, we can use our measurements to determine which scenario is more likely and create our classification algorithm. It is important to note that this formula, and any decision rule that follows from it, can be extended to any collection of populations, but for the purposes of this research we will concentrate on developing a decision rule for two disjoint populations.

With the definition of conditional probabilities, we can also be more specific about independence of events as suggested in Sec. I.2.2. While we will not make use of this relation directly, we will discuss the algorithmic effects of having redundant or useless information in Ch. V. Recall that we are looking for measurements or events that give us no information about one another, i.e., knowing the value of one variable does not provide any information about the other variable. Mathematically, this can be stated:

Definition I.2.5. Two events A and B are independent if $P(A|B) = P(A)$.

While there is a surfeit of information on using probability to make statements about the likelihood of a decision being correct, this does not always help once the complexities of the real world are introduced. For instance, it is often difficult to know the exact distributions of our measurements because so many different variables can affect the outcome. Therefore, we can use a large collection of sample measurements to approximate the necessary distributions to develop our classification methods. The challenges of obtaining a large enough sample to ensure accuracy will be discussed further in Sec. I.6.

I.2.5 Comparing Outcomes of Classification – Cost vs. Risk

In addition to the basic statistics language, we need a common language with which to evaluate the effectiveness of our algorithms. In our classification framework, accuracy of the algorithm is naturally determined by the comparison of the true state of an object in our population and the labeled outcome produced by our algorithm. There are four

Table I.1: The four possible outcomes of a binary classification are determined by comparison of the true label and the label given by our classification algorithm in a confusion matrix. This allows us to examine the consequences of different types of mistakes when we discuss the effectiveness of our algorithm.

Label \ Truth	Class S	Class D
Class S	Correct	False Negative Source Not Detected
Class D	False Positive Unnecessary Search	Correct

outcomes that we need to consider in this binary classification problem, as expressed in Table I.1. An accurate classification occurs when the true state of the object and the label coincide. Mistakes or misclassifications occur when the label assigned by our algorithm is not the same as the true label. This leads to two different kinds of mistakes (see Fig. I.1) – a *false positive* where our algorithm thinks that there is a radiation source in the crate when it is really harmless cargo and a *false negative* where our algorithm does not detect the present source.

Each of these outcomes has a different cost associated with the decision. For every false positive or false alarm that occurs, we must investigate the contents more thoroughly and this will cost both time and money. Each false negative has allowed nuclear material into the country and possibly led to the destruction of a major city. While we can calculate the monetary cost of invasively searching each false alarm, it is more difficult to arrive at a single value that incorporates all the property damage, loss of human capital, economic impact and societal pressures that arise from the loss of a city. Thus, the translation of ideas into a numeric cost for classification is not a trivial matter.

One way to consider this cost conundrum is by looking at global expected error rates of our system of detectors and algorithms. If we examine Fig. I.1, we can determine that a certain percentage of our sample measurements are false positives and another set are false negatives. Therefore, instead of asking a researcher to assign an arbitrary and subjective cost value to misclassifications, it is often asked if we can bound a particular error type and thus guarantee a certain reliability in the detection process. The Neyman-Pearson Lemma stated below gives exactly such a constraint.

Theorem 1 (Neyman-Pearson Lemma [39])**.** *Suppose we are labeling objects as belonging to one of two populations: S and D. Given knowledge of the conditional distributions of measurements for each population and the proportion of each population in the overall con-*

text, then the likelihood-ratio test which gives the label S to an object having measurement x with a specified false negative rate α is

$$\Lambda(x) = \frac{L(D|x)}{L(S|x)} = \frac{p(x|D)P(D)}{p(x|S)P(S)} \leq \eta \tag{I.3}$$

where η is a constant chosen such that $P(\Lambda(x) \leq \eta|D) = \alpha$

Remark I.2.6. In terms of our decision rule $f(\cdot)$ discussed previously, this statement of the Neyman-Pearson Lemma provides the decision rule:

$$f(x) = \begin{cases} S & \text{when } \Lambda(x) \leq \eta \\ D & \text{otherwise} \end{cases} \tag{I.4}$$

The constraint $P(\Lambda(x) \leq \eta|D) = \alpha$ mathematically states that the expected probability of mislabeling a container with a source (should correctly belong to class D) as a safe container S is α. However, the usual method for enforcing the false negative constraint involves integration (analytic or numeric) of the probability distributions over the set of all x for which $\Lambda(x) \leq \eta$ and some sort of search method that will locate the appropriate value of η. As stated in Sec. I.6, this can lead to problems in higher dimensions.

Remark I.2.7. This formulation can be generalized to classification problems with any number of classes and multiple constraints on various types of misclassifications [17]. To the best of my knowledge, this still has only been implemented for cases where the dimension of x is low due to computational complexity.

I.3 Classification by Bayes' Risk Minimization

There are many ways to go about classifying populations, as mentioned above. If exact information about the conditional distributions of these populations, the probability of each object type in the overall population and the costs of various classification actions (misclassification costs, for example) are known or can be approximated with some accuracy, then an optimal classification rule can be determined. This classification method is called the Bayes' Rule for Cost Minimization or the Bayes' Minimal Risk Decision.

This method is based on the fact that it is usually easiest in an experimental situation to control the type of an object rather than the exact measurement, x. Consider an experiment to determine the gender of a person based on height measurements. It is relatively easy to find many people that are clearly male or clearly female and then measure each of those people's height. It is much harder to find a sufficient sample of people of any gender that

are exactly the same height for every possible height measurement.

These distributions, where we control either the class or the measurement and allow the other event to vary, are called conditional probabilities, as mentioned in Sec. I.2.4, and can be transformed using Bayes' Rule (I.2) from functions that have the class controlled by the researcher into functions that presume that the measurement was the controlled quantity. As we can see in the following theorem, we can construct a decision rule that minimizes the expected cost of misclassification using analytic information about the probability distributions.

Theorem 2 (Bayes' Rule for Minimization of Cost of Misclassification [28])**.**
Let us consider a set of features $\mathbb{S}^n$ obtained from two populations, S and D, with population conditional distributions $p(x|S)$ and $p(x|D)$, respectively, for $x \in \mathbb{S}^n \subseteq \mathbb{R}^n$. Also, suppose $P(S)$ and $P(D)$ are the prior probabilities of encountering each respective population and $P(S)+P(D) = 1$, i.e., all items that produce readings in the feature space must be classified as belonging to exactly one of the two populations. Further, let the cost of misclassifying a point from the S population as a member of population D be given by $c_{D|S}$, and $c_{S|D}$ be the cost of classifying a point from the D population as coming from the S population. Then, for a given point in the feature space, x, the classification rule that minimizes the expected cost of misclassification is:

$$x \text{ is classified as belonging to } S \iff x \in \left\{ x \in \mathbb{S}^n \mid c_{S|D} p(x|D) P(D) \leq c_{D|S} p(x|S) P(S) \right\} \tag{I.1}$$

Proof:

Suppose that a point x in feature space is classified as belonging to population S if it is a member of a region, A, of the feature space, $\mathbb{S}^n$, and is classified as a member of population D otherwise. We want to define the region A so that the expected cost of misclassification, ECM, is minimized, i.e., the total number of mistakes made by using our decision rule as weighted by the costs of making these mistakes and their likelihood of occurring is low. Using membership of x in A as a template for classification, we can write the probability of producing a false positive as:

$$P(D|S) = \sum_{x \in A^C} p(x|S) P(S) = P(S) - P(S) \sum_{x \in A} p(x|S) \tag{I.2}$$

Similarly, the probability of obtaining a false negative is given by:

$$P(S|D) = \sum_{x \in A} p(x|D)P(D) \tag{I.3}$$

Given that the fixed cost of misclassifying a false positive is given by $c_{D|S}$ whereas the fixed cost of misclassifying a false negative is given by $c_{S|D}$, the total expected cost of misclassification, ECM, is given by:

$$\begin{aligned} \text{ECM} &= c_{S|D}P(S|D) + c_{D|S}P(D|S) \\ &= c_{S|D}\sum_{x \in A} p(x|D)P(D) + c_{D|S}\sum_{x \in A^C} p(x|S)P(S) \\ &= c_{S|D}\sum_{x \in A} p(x|D)P(D) + c_{D|S}P(S) - c_{D|S}\sum_{x \in A} p(x|S)P(S) \\ &= c_{D|S}P(S) + \sum_{x \in A}\Big[c_{S|D}P(D)p(x|D) - c_{D|S}P(S)p(x|S)\Big] \end{aligned} \tag{I.4}$$

This is clearly minimal when $\sum\limits_{x \in A}\Big[c_{S|D}p(x|D)P(D) - c_{D|S}p(x|S)P(S)\Big]$ is minimal, since the only adjustable quantity is the region A. Since all of the probabilities and costs for each individual x are positive, this implies that the sum is minimal if we define A as:

$$\begin{aligned} A &= \Big\{x \in \mathbb{S}^n \mid c_{S|D}p(x|D)P(D) - c_{D|S}p(x|S)P(S) \leq 0\Big\} \\ &= \Big\{x \in \mathbb{S}^n \mid c_{S|D}p(x|D)P(D) \leq c_{D|S}p(x|S)P(S)\Big\} \end{aligned} \tag{I.5}$$

Thus, to minimize the expected cost of misclassification, an object having a measurement x will be classified as safe if x belongs to the set A as defined above.

■

Remark I.3.1. Under the further assumption that $P(S)$, $c_{D|S}$ and $p(x|D)$ are strictly greater than zero for all points x in the feature space, then

$$A = \left\{x \in \mathbb{S}^n \;\middle|\; \frac{c_{S|D}P(D)}{c_{D|S}P(S)} \leq \frac{p(x|S)}{p(x|D)}\right\} \tag{I.6}$$

This demonstrates that the absolute cost of each individual error need not be known nor do the probability distributions need to be normalized for this analysis to be valid. Only the relative costs of the two error types and the proportionality of the two cargo container types are required. Further, by treating the cost ratio or indeed the whole left side of the inequality as an unknown, we can use this framework to impose constraints on the expected total misclassification rate or even the individual expected rates of false negatives and false

positives through the Neyman-Pearson Lemma (Thm. 1). This will be the foundation for all of the methods discussed in Ch. III.

Remark I.3.2. In this proof of the theorem, the conditional probability distributions were assumed to be discrete distributions with countably many possible outcomes. However, as one can see in [28], this need not be the case as the basic principles still hold for continuous distributions mutatis mutandis.

Remark I.3.3. One could generalize this proof and assign a different cost to each specific measurement x, making the costs $c_{S|D} = c_{S|D}(x)$ and $c_{D|S} = c_{D|S}(x)$, respectively. However, it does add an extra term to the sum as follows:

$$\begin{aligned} \text{ECM} &= \sum_{x \in A} \left[c_{S|D}(x) p(x|D) P(D) + \frac{c_{D|S}(x) P(S)}{\text{vol}(A)} - c_{D|S}(x) p(x|S) P(S) \right] \\ &= \sum_{x \in A} \left[c_{S|D}(x) p(x|D) P(D) + c_{D|S}(x) P(S) \left(\text{vol}(A)^{-1} - p(x|S) \right) \right] \end{aligned} \tag{I.7}$$

where $\text{vol}(A)$ is the volume of the region or the number of elements in the region for continuous and discrete distributions, respectively. Thus giving the decision rule: x is a measurement classified as coming from a container of type S if and only if

$$x \in A := \left\{ x \in \mathbb{S}^n \mid c_{S|D}(x) p(x|D) P(D) \leq c_{D|S}(x) \left((\text{vol}(A))^{-1} - p(x|S) \right) P(S) \right\} \tag{I.8}$$

Unfortunately, this makes the definition of A self-referencing, which means that, for a given set of costs and probability distributions, a numerical solution could be obtained through careful iteration, but a closed analytical form is not possible.

I.4 Optimization

Optimization algorithms have been used to find the equivalent of the Bayes' Rule solution when the distributions are not completely known, but there are several concerns that need to be addressed [49]. Generally, an optimization problem is composed of three parts

1. an objective or cost function,
2. a set of controllable inputs,
3. and a set of constraints.

The goal is to either minimize or maximize the objective function by adjusting the controllable input variables, which may be restricted as a result of the constraints. The objective function is generally given as a real-valued function describing a physical requirement as

a function of several parameters that are either fixed by the physical system or adjustable in order to reach the physical goal. The objective function may also contain other terms called *penalty terms*, which improve performance or enhance certain features of the feasible input space. Constraint functions may be used to limit the feasible values of the adjustable parameters and often correspond to a physical constraint as well.

Once the objective function, constraints and variable parameters have been defined, there are many methods one can use in order to find the extrema of the system, including gradient descent and interior point search methods. The choice of method is determined largely by the characteristics of the objective function. For instance, combinatorial or graph theoretic methods are needed if the controllable parameters take on only discrete values. There are a variety of software packages available to solve such problems, including Opt++ [36], TAO [38], and MATLAB [35].

I.4.1 Overview of the Optimization Formulations

Such formulations are quite common in science and engineering when one seeks to minimize the entropy or maximize the power output of a system by changing pressure or temperature. Constraints in this situation might be something like a function limiting the metal stresses as a result of the adjusted quantities in order to ensure safety of the system. In our problem, these concepts are naturally extended – our objective function and constraints will relate the overall costs of each type of misclassification to parameters that will describe a function that partitions our feature space into two regions as mentioned in Sec. I.2.3. We will discuss two formulations utilizing optimization to solve the cost-sensitive classification problem in Sec. IV.1 and IV.2.

I.4.1.1 Shape Optimization

The first formulation we will be discussing is a variation of *shape optimization*, where a variety of parameters describe a region and are adjusted incrementally to achieve a specified goal. This is a common situation in engineering disciplines, where the basic design of a part is given and then evolved until it reaches a shape providing for optimal performance [20]. In our case as described in Sec. IV.1, we will define the basic shape of our decision boundary and modify it until we minimize the expected probability of obtaining a false positive subject to the constraint that there is no more than a set expected probability of false negatives.

I.4.1.2 Support Vector Machines

Support Vector Machines (SVM) are a specialized optimization method for kernel based classification rule development, which will be further described in Sec. IV.2. The goal of SVM algorithms is to find the hyperplane separating two populations which minimizes misclassifications. Studies have shown that decision rules produced in this manner approach the Bayes' Optimal Rule in the space of such restricted classifiers as the sample size increases [9, 21, 34, 50]. In the linear case, SVM methods optimize over both the normal and distance to the origin of a plane separating the two populations, i.e., SVM algorithms seek to find an optimal separating hyperplane to partition the feature space, $\mathbb{S}^n$, according to the labels S and D. There are algorithms that use a kernel such as an n^{th} degree polynomial or radial basis function in order to transform the high dimensional feature space into a lower dimensional space where optimization algorithms can work more effectively. The hyperplane is found in this transformed space and then pushed back to the original feature space to perform classification. The choice of kernel is initially provided by the researcher, so it requires expert judgment and experimentation to choose the appropriate function.

I.4.2 Challenges in Our Particular Problems

In both of these formulations, there are several quirks, described in more detail in Sec. I.6, that impact the effectiveness of the optimization framework. For example, both of these formulations are created based on the assumption that we are dealing with sample data of integer values. This means that any function we construct will be a non-smooth approximation that can have large flat regions with no information about where the true optima actually lie. Mollifiers have been used and tested in this situation, but this can introduce an error as an artifact of the smoothing. Alternatively, different objective functions, such as the Perceptron Criterion function or distance penalty functions, can be used to assure a smoother gradient. The Perceptron Criterion function, more commonly called the least squares error, is frequently used in many scientific disciplines for just this reason [14, p227,235]. More information about how we dealt with this challenge can be found in Sec. IV.1.

Local optima can also hamper the effectiveness of optimization algorithms and can occur as real features of the underlying distributions or as artifacts from inadequate sampling of the measurement space. According to Miller et al. [37], the common way to help ensure that your algorithm finds global optima instead of settling at local optima is to take numerous random starting points and choose the best results from the collection. Rose [45] also mentions the same problems with optimization surfaces created from probability

distributions, but suggests using stochastic gradient techniques in addition to the repeated starting point approach in order to deal with the local optima. Stochastic gradient techniques take advantage of the fact that the objective function $J(w)$ can be written as the sum of differentiable functions of a single parameter, $J_i(w)$, where each function contributes information about the relation of i^{th} sample of the population and the parameter. This allows the step size in the optimization algorithm to be adjusted based on the effect of one sample at a time. These challenges make the process of choosing an appropriate objective function and constraints especially important in our analysis.

I.5 Other Methods of Classification and Challenges

There are many other methods of classification available at this time, though we will not discuss their application in this work. Two of the most common types of formulations are clustering methods and discriminant analysis. We will give two short examples of such methods here.

I.5.1 k-Nearest Neighbor Classification

The k-Nearest Neighbor classification algorithm is one of the simplest methods of machine learning available [14]. In this case, one begins with an initial sample of labeled data, called the *training set*. Assignment of a label to a measurement $\mathbf{x}$ is done by analyzing the k points from the training set which lie closest to the new point and using the class label which has a majority of points among the k analyzed. One can incorporate the cost-risk analysis that we are interested in by weighting the votes of the points appropriately, although this might require adjustment of the weights if one is trying to control the cost in the Neyman-Pearson sense.

Despite the benefits of the simplicity of this algorithm, there are two drawbacks that make this method inconvenient for our source detection problem. First, when measurements from the two classes differ by only a small amount, the distributions overlap and the algorithm can have difficulty classifying points that lie near the decision boundary. In fact, based on variations in the initial training data, one could frequently return measurements for which no classification can be made as there are an equal number of points from each class in the k-nearest neighboring points. Since we can not guarantee that the radiation measurements will be well separated for all those containers with a source and those without, we would not have as much control over the total expected number of containers that would be invasively searched. Both the containers that are unable to be classified as well as all the containers classified as containing a source would impact this secondary scan. Secondly,

this method requires a large number of points to accurately characterize and sample the feature space, which we have mentioned brings the Curse of Dimensionality into effect.

I.5.2 Multiple Discriminant Analysis

Multiple Discriminant Analysis [14, p117-124] is an area of classification that deals with some of the problems of high dimensional spaces by projecting the data into a lower dimension space and then separating the data. These methods do not necessarily correspond with the Bayes' Rule, but they do allow for an optimal solution to be chosen within the projection space, which may have been difficult in the higher dimension space. However, the Bayes' Rule and the Fisher Linear Discriminant method produce the same rule in the case where both the conditional distributions, $p(x|S)$ and $p(x|D)$, are Gaussian and have equal covariance matrices. The Fisher Linear Discriminant method finds the vector $\mathbf{w}$ in feature space for which the functional

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{I.1}$$

is maximal independent of the magnitude of $\mathbf{w}$, where $\mathbf{S}_B$ and $\mathbf{S}_W$ are scatter matrices with the following definitions

$$\begin{aligned} \mathbf{S}_B &= (\mathbf{m}_S - \mathbf{m}_D)(\mathbf{m}_S - \mathbf{m}_D)^T \\ \mathbf{S}_W &= \sum_{x \in S} (x - \mathbf{m}_S)(x - \mathbf{m}_S)^T + \sum_{x \in D} (x - \mathbf{m}_D)(x - \mathbf{m}_D)^T \end{aligned}$$

where $\mathbf{m}_S$ is the mean of population S, $\mathbf{m}_D$ is the mean of population D, and each sum is over all the points in the appropriate training set. Multiples of this vector $\mathbf{w}$ form the projection space and then further optimization would be necessary to determine the optimal decision rules in the new space. Physically, $\mathbf{w}^T \mathbf{S}_B \mathbf{w}$ represents the squared difference of the population means in the projection space where $\mathbf{w}^T \mathbf{S}_W \mathbf{w}$ is the sum of the population variances in the projection space. This method can be extended to multiple class situations, but the manner in which this is currently stated does not take into account a difference in the cost of misclassification. The most common method to incorporate the disparate risks of such misclassification is to weight the points in each class and adjust these weights to control the global error.

I.6 The Curse of Dimensionality

As stated in Sec. I.2.2, we are utilizing an array of photon detectors to make our classifications of cargo containers. This, of course, gives us a large data array of discrete

integer-valued measurements. Furthermore, as in Sec. I.2.4, the real world rarely provides complete information about the distribution of such measurements. All of these quirks require that we keep the character of our feature space and the type of available information in mind when we create a decision rule in order to effectively combat their influence on the accuracy of the machine learning methods, e.g., challenges of dealing with large data arrays, overfitting due to inadequate sampling, non-smooth functions as a result of discretization. Each of these problems will be addressed more fully as they impact each of the algorithms we discuss, but ultimately, these problems are greatly influenced by the size of the feature space that we are working with and the *Curse of Dimensionality*.

The Curse of Dimensionality is a well-known problem in algorithm implementation for many areas of mathematics. Briefly, this problem refers to the fact that techniques used to analyze data in one or two dimensions become inefficient as the number of dimensions in the problem and thus, the complexity of the problem grows. The container classification problem under discussion suffers from the Curse of Dimensionality as it requires numerical integration of or sampling from high dimensional spaces, both of which increase the complexity of the problem exponentially. The rule of thumb in the areas of numerical integration or sampling is that when using "brute force" methods, like simple numerical quadrature rules or uniform random sampling, roughly one power of ten is needed for every dimension in order to get any real kind of accuracy. This type of problem only worsens as the volume of the region under investigation grows, e.g., dealing with functions on the unit n-dimensional sphere is more tractable than dealing with all of $\mathbb{R}^n$, but is still more difficult than studying the unit circle in $\mathbb{R}^2$. Current classification methods are not exempt from this effect and most methods break down in higher dimensions [14]. For instance, when working with sample data in high dimensions, a classification method may be tailored to work extremely well with initially provided data, but fail to detect patterns in additional data. This situation is known as *overfitting* and is caused when sample data does not provide information about the entire feature space, as is frequently the case in higher dimensions.

In classification, the most common method for combating this curse is to limit the dimension of the problem as much as possible before classification is performed. One such method is to transform measured variables $x \in \mathbb{S}^n$ to a lower dimensional space, $y = g(x) \in \mathbb{R}^m$ and then use the computed value, y, to make the actual classification, e.g., using density to make a decision instead of the mass and volume separately. In our particular case, we will use feature selection methods prior to classification in order to reduce the 320 initial measurements (Ch. II) to 30 equally spaced detectors and then to 2–6 useful measurements (Ch. V). Choosing independent features instead of multiple, correlated pieces of data can

also reduce the dimensionality. Continuing our example, it does not make sense to use density, volume and mass to make a decision since given any two of these measurements, one can compute the third exactly. This means that we have duplicated some of the information and thus, the third measurement will likely not change the classification of an item obtained using only the other two measurements.

A third general class of methods to reduce dimensionality is to treat each measurement separately in order to arrive at classification. Many current classification methods use a decision tree to incorporate data from multiple tests in order to lessen the effects of dimensionality. For example, a single test could be performed and the decision to label or continue testing made. Then, things that could not be classified by the first test alone would undergo a second test and a similar decision process implemented. This could continue for as many tests as desired until total classification is achieved. Therefore, each item tested undergoes a single "labeling" step, but may undergo many tests and the number of tests it undergoes may differ for each item of interest. In this decision tree framework, each decision is made independently of all other decisions in the process. This method is similar to that currently in practice for radiation detection in cargo containers. Another possible way to deal with data from multiple tests is to have an item undergo every test and develop the thresholds for each test independently of all other data. This means that each item will undergo every test available, but its classification may be based on the results of a single test. Of course, we would like to perform on average as few tests as possible in order to classify an item in order to save time and money required for such tests. This requires a significance ordering of the tests and then some sort of estimate of the overall rate of misclassification as a result of the rates of misclassification for each individual test [51]. Since thresholds for each test are determined independently of the others in these examples, we can think of this as a "box" threshold, which is an n-dimensional hyper-rectangle contained in our sample space $\mathbb{S}^n$ with one vertex on the origin. Ultimately, these tests are not optimal because these types of decisions discard the correlation between different tests [51]. For instance, we may have a situation where several tests could be close to the decision boundaries for the individual tests, but none of the readings by themselves are enough to change the assigned label. However, when taken as a whole, the set of measurements may change the classification. This research will continue to explore the use of sensor data as an ensemble instead of a chain of separate readings.

I.7 Comparing Classifiers – Receiver Operating Characteristics

With so many different methods for generating our decision rule, one needs a way to compare the accuracy and precision of the varying methods. For a fixed false negative

Table I.2: The confusion matrix below is utilized by ROC curves to measure the accuracy and sensitivities of classification algorithms by analyzing the outcome of labeling a data set.

		True class	
		D	S
Labeled class	D	**T**rue **P**ositives	**F**alse **P**ositives
	S	**F**alse **N**egatives	**T**rue **N**egatives
Column Totals		**P**	**N**

rate, i.e., a fixed percentage of dangerous cargo containers escaping detection, we can compare the false positive rates of two classifiers to determine which method requires us to search through fewer safe containers. The natural choice for the better classifier will save time and money by opening fewer containers. However, it would be beneficial to choose the more effective classifier, irrespective of the particular false negative rate as the cost of such mistakes may change with the political and social climate. Thus, we need a way to discuss the abilities of a classifier for a wide range of false negative rates.

Receiver Operating Characteristics (ROCs) are a commonly used tool in evaluating the effectiveness of classifiers. We begin with a set of test data containing **P** samples from the class of containers with a source and **N** samples from containers with harmless cargo. For each binary classifier, the number of true positives, false positives, true negatives and false negatives for this data set may be calculated as given in Table I.2. A classifier is then mapped into the ROC space $[0, 1] \times [0, 1]$ by determining the true positive or *recall* rate and the false positive rate:

$$\text{TP (Recall) rate} = \frac{TP}{\mathbf{P}} \qquad \text{FP rate} = \frac{FP}{\mathbf{N}} \tag{I.1}$$

By comparing the various false positive rate – recall rate pairs of the classifiers, we can analyze the relative trade-offs between the benefits (correctly finding the nuclear material) and the costs (opening containers unnecessarily). We can also use this information to determine the accuracy and precision of the various classification methods:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Accuracy} = \frac{TP + TN}{\mathbf{P} + \mathbf{N}} \tag{I.2}$$

The cost–benefit comparison is easy to understand through a visualization of the ROC space [16], as seen in Fig. I.2. A perfect classifier will correctly label all points, i.e., there will be no false positives or false negatives. Thus, the False Positive Rate will be 0 and

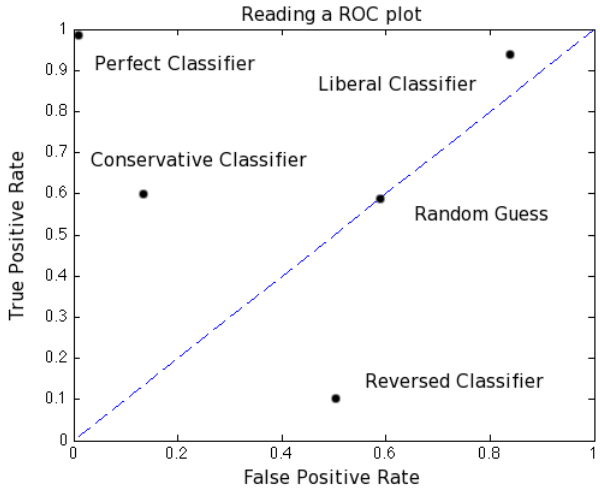


Figure I.2: Examples of the positions of 5 different decision rules and their corresponding positions on a Receiver Operating Characteristic Plot are illustrated. For a more complete description of the features of this graph, see Sec. I.7.

the True Positive Rate will be 1, corresponding the the upper left hand corner of the ROC plot. The closer a classification method lies to this corner of the graph, the better its performance. Conversely, the worst classifier possible would incorrectly label every point, reversing the False Positive Rate and True Positive Rate values, and be plotted in the lower right corner of ROC space. One can note that this worst case classifier could be corrected (and turned into a perfect classifier) by switching the labeling system so that all previously labeled safe containers were now labeled as dangerous and vice versa. This situation presents a *reversed classifier*, where the labeling system need only be reversed to improve its performance and reflecting its (TP rate, FP rate) mapping across the diagonal line of the ROC space. Thus, points lying on this diagonal can not be easily improved without further information provided and are the equivalent of a "random guess" strategy when assigning labels.

A ROC plot also distinguishes between liberal and conservative classifiers. A *liberal classifier* is one that will assign a label with very little evidence and, as a result, they cause more frequent false alarms. A *conservative classifier* requires more stringent proof before declaring that something belongs to class D, which forces the false alarm rate to stay low, but often causes the true positive rate to be low as well. As a consequence, liberal classifiers have a higher false positive rate than conservative ones, but generally have a higher true positive rate as well. These positions are also illustrated in Fig. I.2.

Every point in the ROC space gives the performance of a classifier with a specific false

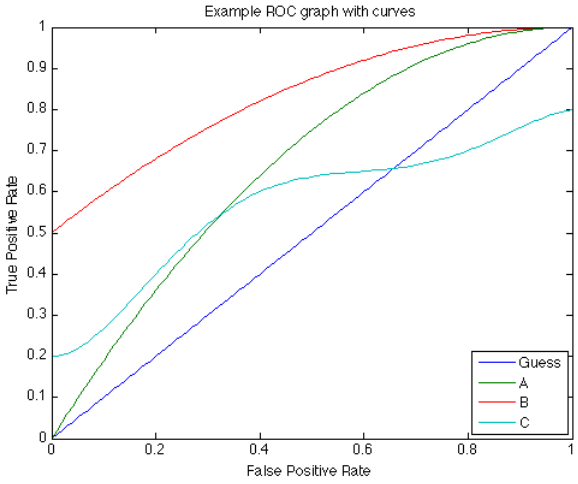


Figure I.3: Classification families can be represented as curves on a ROC plot. A single point denotes a decision rule with a specific false positive rate, but by adjusting user-chosen parameters in the decision rule, one can generate a family of classifiers of varying error rates created by essentially the same type of algorithm. Here, Algorithm B is consistently better than the other two algorithms, regardless of the false positive rate. Algorithms A and C, however, are better for different ranges of classifiers (C is the better of the two when classifying conservatively, and A is better for the more liberal settings).

positive rate. Often, our methods for creating a decision rule introduce a set of parameters, e.g., radiation thresholds for individual detectors, which can be used to tune the performance of the classifier to meet a variety of specified error rates. In doing so, we obtain entire families of classifiers corresponding to curves in ROC space. One classifier is considered "better" than another if its ROC curve is consistently greater than the other, e.g., Classifiers B and A in Fig. I.3. This may only happen for a set range of False Positive rates, e.g., Classifiers A and C in Fig. I.3.

Additionally, ROC classifiers allow us to examine the cost ratio depicted in Theorem 2. Classifiers with the same expected cost ratio will fall on a line with a slope, m, proportional to the ratio of the costs of misclassifications and the probability of encountering a specific class:

$$m = \frac{c_{S|D}p(S)}{c_{D|S}p(D)} \tag{I.3}$$

As shown in Tortorella [52], this can be utilized to help select cost ratios to control global error rates in addition to the analysis already presented by an ROC curve.

CHAPTER II

COMMON TEST CASES

As was mentioned in Sec. I.2.2, any classification method requires some knowledge of the distribution of the measurements in the feature space, be this through an exact distribution or a random sampling of possible scenarios. The following sets of simulated measurements will be used for the development and testing of the methods discussed in this work. The algorithms themselves are completely independent of any specific properties of the distributions. The methods in Ch. III require that the distributions be known to the researcher, but do not make any specific demands about the types of distributions involved. The algorithms in Ch. IV do not require analytic knowledge of the distributions – only sample measurements – so, to be consistent in our comparison of methods, we will use samples drawn from the same distributions as used in the analytic case.

Each measurement set given here depends on two separate physical parameters – the material inside the container and the detector configuration – and a choice of distribution. In this chapter, we will discuss the general process for simulating the data with a choice of distribution and then go through the specific material configurations used. Finally, we will consider the various detector configurations used in these tests. Typically, we will compare measurements with and without an internal HEU source for a single material configuration and detector array. This results in a pair of distributions (for the analytic methods) or sample sets (for the methods in Ch. IV) labeled Safe and Dangerous, respectively, for use in the classification of containers.

II.1 Simulation of the Data Sets

In order to train the classification algorithms, an exemplar set of typical measurements is needed. This could be obtained either through physical experimentation or through computer simulation. There are drawbacks to both methods of data compilation. For instance, physical experimentation can be prohibitively expensive as we would need many detectors, sets of physical cargo and a radioactive source as well as the personnel and physical space to actually take detector readings. On the other hand, computer simulations require a well tested and developed code to handle the radiation transport. Also, most computational codes give only an average count rate for each detector as the final output and ignore the fluctuation from measurement to measurement that is seen in the physical experiment.

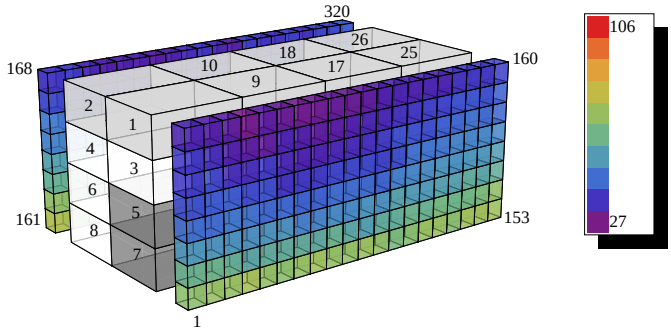


Figure II.1: The geometry of the cargo container and detectors used in the MCNP deck. There are 160 detectors completely covering either side of the cargo container and completely covering the surface of the container. The coloring on the side indicates the mean count rate seen by the given detector in a 30 second interval, as computed by MCNP.

To balance the shortcomings of both the physical experimentation and computer simulation, the data sets used in the development of our classification methods are created using a two step process. First, for each cargo container scenario under examination, the Monte Carlo n-Particle software package (MCNP)[56] is used to simulate the statistical mean particle count rates for each detector in the array for a specified time period. This involves specifying the geometry of the cargo container and the two arrays of detectors, as depicted in Fig. II.1. We will further discuss how the cargo in the container is modeled in Sec. II.2 and describe each scenario under consideration here in detail in Appendix A.

Once these mean values are determined, the random fluctuation that is seen in physical experimentation is introduced by letting the count rate for each detector vary according to a Poisson distribution about the specified mean. Many experiments have verified that the detection of gamma radiation behaves according to a Poisson distribution. Furthermore, when considered over time scales that are small in comparison to the half-life decay of the radioactive source, gamma radiation emission and detection is a *Poisson process* since it holds to the Poisson postulates:

Theorem 3 (The Poisson Postulates [5])**.** *For each $t \geq 0$, let N_t be an integer-valued random variable with the following properties. (Think of N_t as denoting the number of arrivals in the time period from time 0 to time t.)*

(i) $N_0 = 0$
(start with no arrivals)

(ii) $s < t \implies N_s$ *and* $N_t - N_s$ *are independent*
(arrivals in disjoint time periods are independent)

(iii) N_s *and* $N_{t+s} - N_t$ *are identically distributed*
(number of arrivals depends only on the period length)

(iv) $\lim_{t \to 0} \frac{P(N_t = 1)}{t} = \lambda$
(arrival probability proportional to period length, if length is small)

(v) $\lim_{t \to 0} \frac{P(N_t > 1)}{t} = 0$
(no simultaneous arrivals)

If conditions (i)-(v) hold, then N_t *follows a Poisson distribution with mean* λt*.*

Using the flux rates calculated by MCNP for each container, we can determine the average number of counts for each detector in a 30 second interval, λ_i, and reintroduce the variation one would see in experimental observations according to a Poisson distribution as follows:

$$p_i(x_i) = e^{-\lambda_i} \frac{\lambda_i^{x_i}}{x_i!} \tag{II.1}$$

where λ_i is the mean value for the i^{th} detector and x_i is the detector measurement observes at that same detector. In the development of methods which require exact knowledge of the distribution of measurements (see Ch. III), the multivariate Poisson distribution is used where the probability of obtaining any specific set of measurements from the detector array is the product of the probability distributions for each detector

$$p\left(\mathbf{x}\right) = \prod_i p_i(x_i) = \prod_i e^{-\lambda_i} \frac{\lambda_i^{x_i}}{x_i!} \tag{II.2}$$

A random sample from this multivariate distribution is used in the methods in Ch. IV so that methods with exact knowledge and with collected data can be more easily compared.

Definition II.1.1. A **random sample** is a number of independent observations from the same probability distribution. This involves selecting N samples from the distributions in such a way that any such choice of samples has an equal change of being selected.

Fig. II.2 demonstrates the readings that one might expect to obtain experimentally if the radiation escaping from a single cargo container was measured several different times. Each point on this graph represents a single possible measurement from two separate detectors.

II.2 Cargo, Sources and Background

As stated in Sec. I.1, the material placed between a source and a detector will prevent some of the source radiation from reaching a detector and some of these materials will emit their

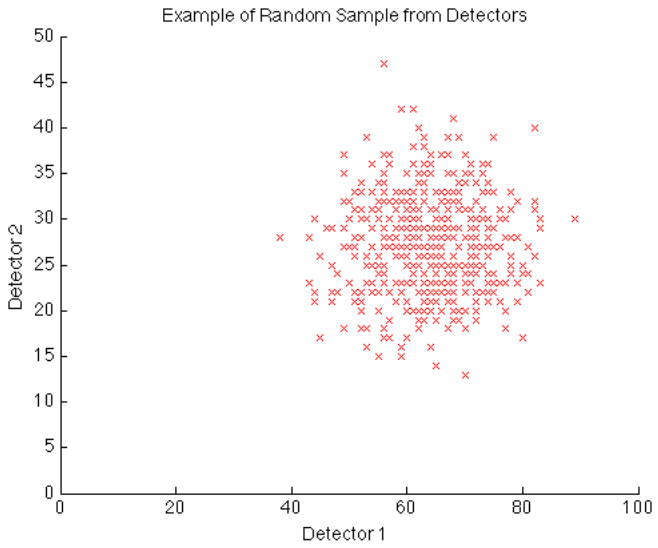


Figure II.2: After determining the mean count rates for each detector, the variation one would observe in reality is reintroduced by drawing a random sample from a Poisson distribution. Here, we depict a random sample of 500 measurements where the means of two of the detectors are $\lambda_1 = 64$ and $\lambda_2 = 27$. One can see that the sample is grouped more tightly near the given means and spreads in the same manner as the variance for the Poisson distribution.

own radiation signatures. Further, background radiation and unknown source positions can influence our ability to detect illicit materials. Any algorithm that we develop needs to be flexible enough to recognize these situations and still accurately detect smuggled HEU. Accordingly, the data sets used to test the algorithms have varying materials in the cargo container, background source strength and internal source positions, which will be outlined below. Details on the exact combinations and placement of materials used in the MCNP simulations can be found in Appendix A.

This work concentrates on 5 different cargo loads designated L1–L5. These scenarios are used to test the influence of materials placed within the cargo container on our ability to detect a 1 kg HEU source. We have divided the interior of the cargo container into 32 brick-shaped blocks of equal volume as depicted in Fig. II.3, which we then fill with various materials. L1 is treated as a base case and contains only light density (under 1 g/cm^3) materials – wood, cotton, and plastic. L2 tests the effects of density variations by replacing some of the light density material close to the source with iron, which has a density of 7.8 g/cm^3. L3 and L4 substitute different amounts of concrete (2.4 g/cm^3) for the iron in L2 providing an internal source as well as a density variation. Concrete is an

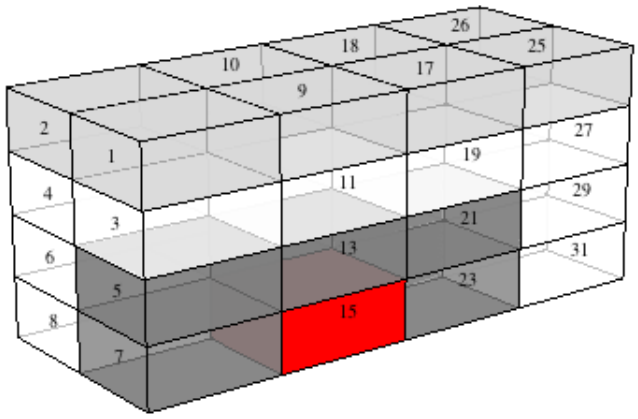


Figure II.3: To test the effects of variations in cargo on the effectiveness of the algorithms, we have divided the container into 32 brick-shaped blocks, which are filled in our simulations with different combinations of materials. Depicted here is an L1 loading scheme where the different shades of the blocks correspond to different density materials, the lightest being white and the most dense being dark gray. The red box is filled with cotton and the S1 source is centered in the long direction and on the back wall of this box.

Table II.1: Summary of the material combinations used and the properties that are tested in each loading scenario.

Loading	Properties
L1	base case – light density material
L2	some medium density material present
L3	light density material with a NORM source (concrete)
L4	smallest internal NORM source, light density materials
L5	all material is light density NORM (potash)

example of NORM containing large percentages of ^{40}K, which has a 1.46 MeV gamma line. These photons can scatter off of other materials in the container and lose energy, e.g., the photons may reach the detectors with an energy of 1 MeV. This phenomenon can affect our detection problem by obscuring the radiation peak attributed to the HEU source in two ways. It can add to the background in such a way as to hide the localized peak coming from the smuggled material or it can be so localized that it creates a false signature for the algorithm. The difference between these two scenarios is the amount of concrete present – L3 has roughly twice the amount of concrete than is present in L4. L5 is filled with only a typical potash fertilizer (2.66 g/cm^3). Potash is a NORM and contains ^{40}K, as did the concrete of the L3 and L4 scenarios. As such, these scenarios result in elevated background measurements. The salient details can be found in Table II.1.

Table II.2: Several source combinations were used in the test sets, designated below.

Source Designation	Position and type
S0	base case – the only source is the concrete slab on which the container rests
S0a	internal NORM sources (part of the background)
S1	HEU source located 7.5 ft back and 2.5 ft in from the side wall on the bottom of the container
S2	HEU source located 7.5 ft back and 2.5 ft in from the side wall and 4.5 ft from the bottom of the container

In addition to the internal source provided by the NORM discussed above, each data set includes background radiation produced by a 30.0 cm thick concrete slab placed under the cargo container and extending past the detectors in all directions. This source is orders of magnitude greater than that produced by either the internal NORM sources or the 1 kg HEU source under consideration. This HEU source is placed in one of two locations either directly on the floor of the container or suspended about halfway between the container floor and ceiling. In both cases, the source is placed well away from either end of the container so that greater amounts of material block the emitted radiation from the detectors. Thus, we have various designations for different source positions, as given in Table II.2. Since the background radiation from the concrete slab is greater near the bottom of the container, the radiation emitted by the HEU source on the bottom of the container makes up a smaller percentage of the detector count rate than the source suspended in the middle of the container, making it more difficult to detect.

For most of the development of these algorithms, we will work with cargo loading L1 and the S1 source, but with the addition of an inch thick steel shield in between the concrete background source and the cargo container. This particular configuration is referred to as Data Set A. The reason for working with this particular data set is twofold. Firstly, the presence of the steel shielding cuts down on the amount of background radiation shown in proportion to that of the local internal source for which we are searching. While this is rather expensive to implement in practice, one could adopt such a shield in order to improve the detection probabilities. Additionally, Data Set A gives us independent data with which to develop and test the classification algorithms. This will help us avoid the overfitting problem mentioned in Sec. I.6.

II.3 Detector Configurations

The MCNP simulations involve the placement of 320 detectors on the long sides of the container as shown in Figs. II.1 and II.4. This detector configuration will be referred to as the original detector set. To ensure that the algorithms developed are independent of the physical detector configuration and the magnitudes of the count rate, there are three other detector sets that are made by combining and taking subsets of the original data set. The original detector configuration without steel shielding for the L1 and L2 material loadings, see Sec. II.2, have mean count rates of anywhere from 50 to 100 photons in a 30 second period, depending on the position of the detector. The detector subsets considered here are the 30 detector subset (30Det), the column totaled subset (CTDet), and the 4 by 4 summed subset (4×4Det), which are discussed further in the following subsections. The S1 HEU source is closest to detectors 57 and 58 as numbered in the original configuration, Fig. II.4, and the S2 source is nearest to detector 61.

II.3.1 30Det: 30 Detector Subset

This detector set uses detectors of the same size as those in the original detector set, but, instead of having detector surfaces completely covering the side of the cargo container, the detectors are spaced out along the container wall. The original detectors chosen for the subset are depicted in Fig. II.4 and it should be noted that these 30 detectors come from both sides of the cargo container. When referring to specific detectors, the numbering system shown in Fig. II.5a will be used. In this subset, the S1 source position is closest to detector position 23. The rest of the diagrams in Fig. II.5 give examples of the average count rate on each detector over a 30 second period for two cargo containers with no internal source and one cargo container with an S1 source.

II.3.2 CTDet: Column Totaled Detectors

This data set involves detectors that are 8 times the area of the original set and are obtained by adding up all of the individual detector readings in one column of Fig. II.4 to obtain the new measurements. Typical detectors in this configuration with L1 or L2 material settings have a mean count rate of approximately 500-600 counts in 30 seconds. Fig. II.6 gives some examples of the average measurements observed if this detector configuration is used for a few of our test configurations. It should be noted that when referring to individual detectors, the numbering is sequential from left to right.

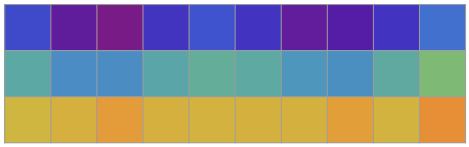
8	16	24	32	40	48	56	64	72	80	88	96	104	112	120	128	136	144	152	160
7	15	23	31	39	47	55	63	71	79	87	95	103	111	119	127	135	143	151	159
6	14	22	30	38	46	54	62	70	78	86	94	102	110	118	126	134	142	150	158
5	13	21	29	37	45	53	61	69	77	85	93	101	109	117	125	133	141	149	157
4	12	20	28	36	44	52	60	68	76	84	92	100	108	116	124	132	140	148	156
3	11	19	27	35	43	51	59	67	75	83	91	99	107	115	123	131	139	147	155
2	10	18	26	34	42	50	58	66	74	82	90	98	106	114	122	130	138	146	154
1	9	17	25	33	41	49	57	65	73	81	89	97	105	113	121	129	137	145	153

168	176	184	192	200	208	216	224	232	240	248	256	264	272	280	288	296	304	312	320
167	175	183	191	199	207	215	223	231	239	247	255	263	271	279	287	295	303	311	319
166	174	182	190	198	206	214	222	230	238	246	254	262	270	278	286	294	302	310	318
165	173	181	189	197	205	213	221	229	237	245	253	261	269	277	285	293	301	309	317
164	172	180	188	196	204	212	220	228	236	244	252	260	268	276	284	292	300	308	316
163	171	179	187	195	203	211	219	227	235	243	251	259	267	275	283	291	299	307	315
162	170	178	186	194	202	210	218	226	234	242	250	258	266	274	282	290	298	306	314
161	169	177	185	193	201	209	217	225	233	241	249	257	265	273	281	289	297	305	313

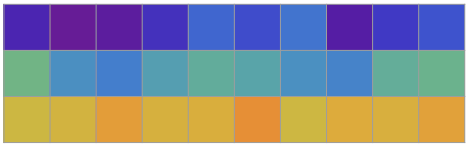
Figure II.4: The original MCNP detector geometry consists of 320 detectors spread over two sides of the container and numbered as given here. The top half of the numbers denotes the detectors nearest to the S1/S2 sources and the bottom half denotes the far side of the container. In order to allow for ease of analysis, we consider a 30 detector space subset of the original detectors, 30Det, as marked by the highlighted red detectors in the original array.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30

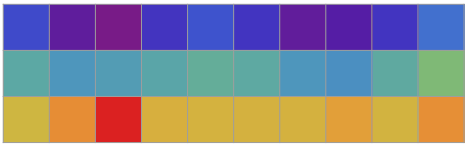
(a) Numbering



(b) L1, S0



(c) L2, S0



(d) L1, S1

Figure II.5: This is a depiction of measurements from the 30 detector space subset of the original detectors, 30Det, and the numbering system of detectors. It should be noted that the detectors in the first five columns come from the side of the container nearest the source and the last five come from the far side. When referring to specific detectors from this subset, we will use the numbering system given in Fig. II.5a. For three different scenarios (two with background only, S0, and on with a smuggled source), we depict the average count rate expected for each of the 30 detectors. Purple and blue detectors indicate low count rates and the count rate increases across the color spectrum until the highest average count rates are indicated in red.

(a) L1, S0

(b) L2, S0

(c) L1, S1

Figure II.6: These figures are examples of average count rates for the column totaled subset of the original detectors, CTDet, from both sides of the container with 20 column detectors apiece. The first 20 detectors correspond to the side nearest the source. Detectors with the lowest count rate are purple and the highest are red, as is usual. The first two figures contain no internal source and the last one has a smuggled source located behind detector 8. One can see that the detectors near the edges of the container see more radiation in general than do those near the interior, leading to the series of yellow detectors in the middle and at each end of the row in each subfigure above. This increase in radiation is a result of two factors – (1) some of the least dense material in each of these scenarios is located near the ends of the container and (2) the concrete slab from which the background radiation comes extends beyond the container bounds which allows some radiation from outside the cargo container to interact with the detectors on the edge of the array.

8	16	24	32	40	48	56	64	72	80	88	96	104	112	120	128	136	144	152	160
7	15	23	31	39	47	55	63	71	79	87	95	103	111	119	127	135	143	151	159
6	14	22	30	38	46	54	62	70	78	86	94	102	110	118	126	134	142	150	158
5	13	21	29	37	45	53	61	69	77	85	93	101	109	117	125	133	141	149	157
4	12	20	28	36	44	52	60	68	76	84	92	100	108	116	124	132	140	148	156
3	11	19	27	35	43	51	59	67	75	83	91	99	107	115	123	131	139	147	155
2	10	18	26	34	42	50	58	66	74	82	90	98	106	114	122	130	138	146	154
1	9	17	25	33	41	49	57	65	73	81	89	97	105	113	121	129	137	145	153

168	176	184	192	200	208	216	224	232	240	248	256	264	272	280	288	296	304	312	320
167	175	183	191	199	207	215	223	231	239	247	255	263	271	279	287	295	303	311	319
166	174	182	190	198	206	214	222	230	238	246	254	262	270	278	286	294	302	310	318
165	173	181	189	197	205	213	221	229	237	245	253	261	269	277	285	293	301	309	317
164	172	180	188	196	204	212	220	228	236	244	252	260	268	276	284	292	300	308	316
163	171	179	187	195	203	211	219	227	235	243	251	259	267	275	283	291	299	307	315
162	170	178	186	194	202	210	218	226	234	242	250	258	266	274	282	290	298	306	314
161	169	177	185	193	201	209	217	225	233	241	249	257	265	273	281	289	297	305	313

Figure II.7: Using the original MCNP configuration of 320 detectors placed on two sides of the containers, we create the 4 by 4 detector subset, 4×4Det, by adding squares of detectors as given in the checkerboard pattern above. This creates two rows of detectors where each detector is 16 times the area of the original detectors.

II.3.3 4×4Det: Four by Four Summed Detectors

This data set involves detectors that are 16 times the size of the original set and are obtained by adding up blocks of 16 detectors from the original set in 4 by 4 squares, as shown in Fig. II.7. Typical detectors in this configuration for the L1 or L2 material settings have a mean count rate of approximately 800 counts for the top row or 1300 counts for the bottom row of detectors in 30 seconds. Fig. II.8 gives some of examples of measurements taken if this detector configuration is used. The numbering system is shown in Fig. II.8a and the S1 source is behind detector 12 and close to the border with detector 13.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

(a) Numbering

(b) L1, S0

(c) L2, S0

(d) L1, S1

Figure II.8: Examples of the average count rate in the 4 by 4 detector subset, 4×4Det, which cover both sides of the cargo container with an equal number of detectors. The first five columns of detectors correspond to the container side near the source positions and the last five to the far side. Fig. II.8a gives the native numbering scheme that will be used to refer to specific detectors here. The rest of the diagrams give the average count rate for each detector in three different scenarios – two different loading schemes with only background and one with a source behind the second detector from the right in the bottom row (S1 scenario).

CHAPTER III

CLASSIFIERS BASED ON BAYES' RULE

Now that the basic elements of classification have been defined, we can discuss the implementation and effectiveness of several different algorithms. In this chapter, our discussion will be confined to two techniques based on Bayes' Rule and the Neyman-Pearson Lemma from Sec. I.3 and I.2.5. In these methods, the necessary probability distributions are assumed to be completely known to the researcher, see Sec. II.1. However, the costs of various misclassifications are unknown. As always in our problem, the goal in classifier development is to define a subset of the feature space, A, where a container is given the label S if the corresponding measurement $x \in A$. We also desire that this definition of the region A minimizes the total fraction or cost of misclassifications made by the algorithm. While working on this problem, we need to keep in mind two goals:

1. The classifier allows no more than a prescribed fraction, α, of containers with HEU to escape detection: $\sum_{x \in A} p(x|D) \leq \alpha$.
2. Among the set of all classifiers for which the first condition holds, the total number of false positives, $\sum_{x \in A^C} p(x|S)$, must be minimal.

We can further interpret these goals in terms of the ROC plots described in Sec. I.7 by examining the relations of these quantities to the true positive (TP) and false positive (FP) rates:

$$\text{TP rate} = 1 - \sum_{x \in A} p(x|D) \qquad \text{FP rate} = \sum_{x \in A^C} p(x|S) \tag{III.1}$$

On a ROC plot, a classifier satisfying the above conditions would be no more than α from the top of the plot and as close to the upper left corner as possible. We should note that we could have easily switched these two conditions, i.e., we could specify the allowable fraction of false alarms and then minimize the expected number of false negatives. This would not drastically affect the implementation of the algorithm, requiring only that the label of the distributions be exchanged.

Two methods, called the Box Threshold Method and the Analytic Bayes' Optimal Decision Method, will be the subject of this chapter. The Box Method (Sec. III.1) uses information about individual detectors to determine the decision rule and then progresses to utilizing all the measurements for actual classification. This is the most basic of our classification

methods and, as such, will be used as a standard by which to judge the effectiveness of all other algorithms. The Bayes' Method uses information about all of the detectors simultaneously to both compute the decision rule as well as make the final classifications. The second method allows correlations in the data to influence the ultimate classification process, as will be shown in Sec. III.2.

III.1 Box Threshold Method

The first algorithm we will discuss is what we will call the Box Threshold Method. The end goal of this method is, for a given set of n measurements, to define an n-dimensional box in the feature space, $\mathbb{S}^n$, such that the total expected misclassification error is minimal and the false negative rate is equal to the specified rate α. Any container that produces a measurement x that lies within this n-dimensional box will be classified as safe. To the best of the author's knowledge, this method, where every detector has an alarm threshold that is independent of the other detectors, is similar to that currently in practice in border security checks.

In this section, we assume that we are given exact probability distributions for the readings from each detector (see II.1) for each type of container, so this algorithm takes advantage of our knowledge of the exact form of the Poisson distribution. The basic algorithm is as follows:

1. Specify an individual detector false negative rate, α_{ind}. This requires that every single detector has the same false alarm rate when taken separately.

2. Using the Neyman-Pearson Criterion, we enforce the constraint on the false negative rate by finding a value c_i such that, for the i^{th} detector of a container with an HEU source having mean $\lambda_{i,D}$,

$$P(x \leq c_i|D) = \sum_{x=0}^{c_i} \frac{\lambda_{i,D}^x}{x!} e^{-\lambda_{i,D}} = \alpha_{ind} \tag{III.1}$$

3. Determine individual thresholds, $t_i \leq c_i$, by minimizing either the expected cost of misclassification or the expected false positive rate, as chosen by the researcher and

specified in the respective equations below:

$$\begin{aligned} \text{ECM}(t_i) &= P(x > t_i|S) + P(x \leq t_i|D) \\ &= 1 - \sum_{x=0}^{t_i} \frac{\lambda_{i,S}^x}{x!} e^{-\lambda_{i,S}} + \sum_{x=0}^{t_i} \frac{\lambda_{i,D}^x}{x!} e^{-\lambda_{i,D}} \qquad \text{(III.2)} \\ P(x > t_i|S) &= 1 - \sum_{x=0}^{t_i} \frac{\lambda_{i,S}^x}{x!} e^{-\lambda_{i,S}} \qquad \text{(III.3)} \end{aligned}$$

4. Arrange these values t_i into the vector $\mathbf{t}$, then calculate the total expected false negative rate $\widetilde{\alpha}$, which in this case can be found by computing:

$$\widetilde{\alpha} := P(S|D) = P(\mathbf{x} \leq \mathbf{t}|D) = \prod_i \sum_{x_i=0}^{t_i} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i} \qquad \text{(III.4)}$$

5. Repeat steps 1-4 adjusting α_{ind} until $\widetilde{\alpha}$ is as close as possible to α, the desired expected false negative rate as given by the researcher. In this particular instance, adjustment of α_{ind} will be made using the bisection method where the function of which we want to find the root is $f(\alpha_{ind}) = \widetilde{\alpha}\,(\alpha_{ind}) - \alpha$.

Then, $\mathbf{x} \leq \mathbf{t}$ is the decision rule that classifies an object which produces a set of measurements $\mathbf{x}$ as safe (without a source). As a result, we have guaranteed that the false negative rate, $P(S|D)$, is less than or equal to the specified level α.

Remark III.1.1. Current methods of radiation detection usually use a constant threshold for every detector in the system, but as we are interested in how correlations in measurements affect the classification process, we will allow each detector threshold to vary individually. Further investigation exploring this difference in approaches is warranted.

Remark III.1.2. If one chooses to minimize the total expected cost of misclassification as in (III.2), it may not be possible to exactly achieve the specified false negative rate α. Even though the false positive rate will continue to decrease as the threshold increases, it may not offset the increase in the overall misclassification rate since the false negative rate will increase as the threshold increases.

If we choose to minimize the false alarm rate as in (III.3), we can further utilize the known structure of the distribution functions to simplify our computations. Using the fact that the cumulative distribution function of any one dimensional distribution increases as the argument x goes to infinity, we know that the false positive rate given by $P(x > t_i|S)$ will decrease as the threshold value increases. More simply, the false positive rate for a one

dimensional function is a monotonically decreasing function. Therefore, with this objective function, we can develop the thresholds using only information about the distribution of measurements from containers containing HEU, $p(x|D)$.

In this particular case, we are guaranteed that the distributions of the individual detector readings are independent Poisson distributions, which means that the multivariate distribution can be expressed as the product of the distributions for each detector:

$$P(S|D) = P(\mathbf{x} \leq \mathbf{t}|D) = \prod_i \sum_{x_i=0}^{t_i} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i} = \prod_i P(x \leq t_i|D) \tag{III.5}$$

As a result, the need for iteration of the first three steps of the above algorithm could have been eliminated by taking the n^{th} root of the desired false negative rate α and then finding the threshold t_i satisfying (III.1) where $\alpha_{ind} = \sqrt[n]{\alpha}$. By leaving the iteration steps in the algorithm, this process could easily be extended to the situation where the exact distribution in unknown, but adequate samples of the distributions are present. Then, a step 0 would need to be included in the algorithm where the sample set would be projected onto each detector space to get the set of approximated 1D distributions, see Fig. III.1. This would also require that, instead of using the analytic forms of the distribution, the calculation of the individual detector false negative rates and the total false negative rates would have to be done by numerical integration using the sample points.

III.1.1 Choice of Objective Function

For initial testing, we examine the minimization of both the expected cost of misclassification and the false positives with the constraint on the total false negative rate. The 30 detector subset described above (Sec. II.3.1) was used to decrease the computational requirements and simplify the analysis. For each of the thresholds, t_i, only information from a single detector was used to determine the optimal threshold, although the training set was labeled as to whether a point was generated as a result of measuring a container with a source or without. For these tests, the respective objective functions were implemented using an approximation method. We began with multi-dimensional sample data consisting of N_S samples of measurements from safe cargo containers and N_D samples from dangerous containers. Then, we worked component-wise with the sample data and binned it appropriately to approximate the distributions of measurements for each detector as discussed previously. Instead of using the explicit analytic form of the objective functions

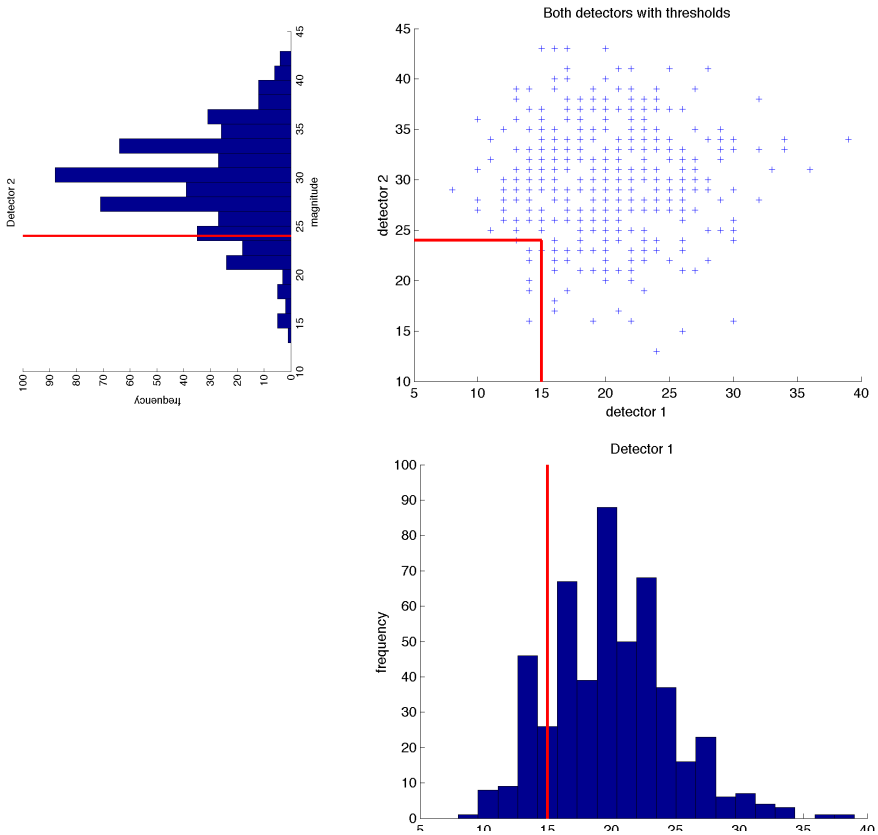


Figure III.1: Illustration of the Box Method implementation for a 2 detector system with minimization of the false positive rate. Assume that we are given many samples of two detector measurements that we might observe when scanning a specific kind of container that contains a HEU source (the scatter plot at the top right). First, we project these samples into the 1D spaces for the set of all detector readings, essentially looking at measurements from one detector at a time while completely ignoring the other measurements, as given by the frequency plots along each access of the scatter plot. Next, we use the Neyman-Pearson criterion to determine a threshold with a specified false negative percentage, α_{ind}, which are denoted by the red line in each frequency plot. Finally, we use these individual thresholds to create a rectangular area in the original feature space, which corresponds to the box created by the axes and the two red lines in the scatter plot. The interior of this rectangle is designated as the region A and points that fall inside this region will be classified as belonging to safe cargo containers. Given this defined region we can count the total fraction of false negatives that fall inside this region to get the total false negative rate, $\widetilde{\alpha}$ and iteration can be used to adjust the region until $\widetilde{\alpha} = \alpha$.

as in (III.2) and (III.3), we approximate the false positive and negative rates by:

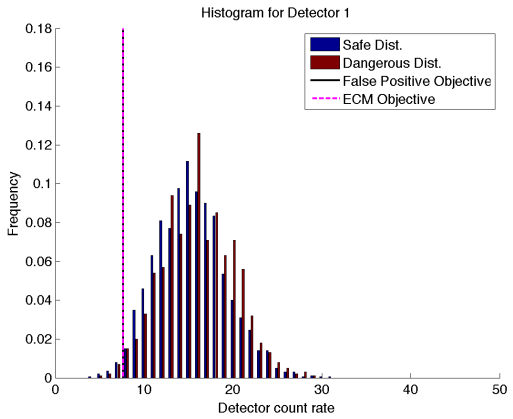
$$P(x_i > t_i|S) \approx \frac{1}{N_S} \sum_{x_i > t_i} 1 \qquad P(x_i \leq t_i|D) \approx \frac{1}{N_D} \sum_{x_i \leq t_i} 1 \tag{III.6}$$

Once these functions were implemented, the optimal thresholds were determined using a trust region Hessian based algorithm from MATLAB called *fmincon*. However, as was mentioned in Sec. I.4.2, this optimization algorithm requires a smooth objective function to operate effectively. Thus, a mollifier was employed to reduce the step function nature of this objective. As a result, points far outside the boundary take values near one, those well inside the boundary are close to zero and points near the boundary interpolate between the two extremes.

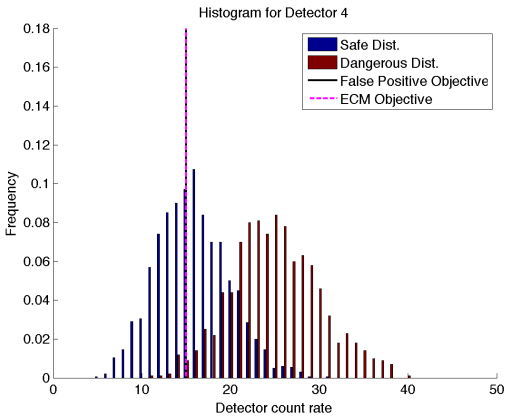
We can see examples of the distributions of measurements for four detectors after we have projected from the 30 dimensional space into each of the individual detector spaces in Fig. III.2. For most of the detectors, the distribution of measurements from a container with and without a source are indistinguishable as in Fig. III.2a. This is a result of the behavior of radiation described in Sec. I.1 and the small size of the source. Detectors nearer the source will have more separated distributions, as in the remaining figures in Fig. III.2, and will provide a better opportunity to classify containers correctly. For this particular container loading (Data Set A) and detector combination (30Det), only two detectors can actually see enough radiation from the source to trigger/create a reasonable threshold alarm – Detectors 22 and 23.

This distinction of two detector classes is more easily shown in Table III.1, where we compare the false positive and false negative rates for thresholds developed with each objective function. As one can see, for most of these detectors with the specified false negative rate, the false positive rate is over 90%. This means that our classification method based on these detectors is so conservative that we will have false alarms on most of our containers. This is not a favorable outcome because it requires that too many containers be needlessly opened, adding a large cost to our algorithm in the field. Looking at the frequency plots for these detectors in Fig. III.2, it is easy to see why such a classification is developed as the graphs of the two distributions overlap considerably in such cases.

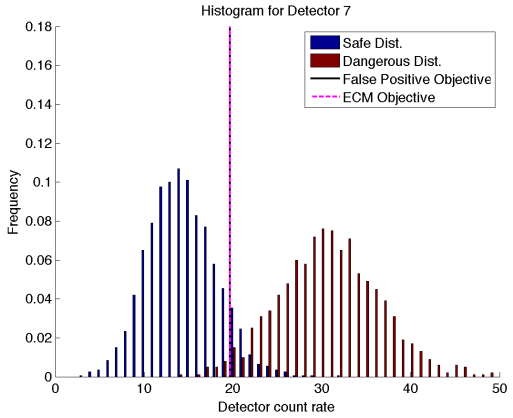
We can better understand the information provided by this data and the thresholds developed by analyzing the sum of the false negative and false positive rates $\beta = P(x_i > t_i|S) + P(x_i \leq t_i|D)$ given in Table III.1, which must take on a value in $[0, 2]$. For most of the detectors here, β obtains a value near 1. However, for detectors 22 and 23, this sum is significantly less than 1. This suggests a way to distinguish those detectors with informa-



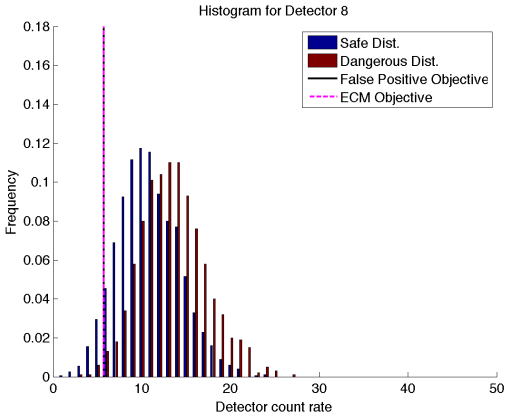
(a) Histogram for Detector 21



(b) Histogram for Detector 22



(c) Histogram for Detector 23



(d) Histogram for Detector 13

Figure III.2: Sequentially generated thresholds were produced using both the false positive (solid black line) and expected cost of misclassification (dashed pink line) objective functions according to the procedure described previously in Sec. III.1. The blue bars show the distribution of the safe measurements and the red bars show the dangerous measurements. In all of the pictured cases, both methods provide the same threshold and this is a common occurrence in this algorithm. Further analysis of these figures can be found in Sec. III.1.1. These measurements were generated using Data Set A (Sec. II.2).

Table III.1: Expanding on Fig. III.2, this table provides a list of the error rates in the sequential determination of detector thresholds for both the false positive and ECM objective functions. Using the two different objective functions described in Sec. III.1, we use the information from Data Set A (Sec. II.2) to determine detector thresholds and calculate the false positive and false negative rates for each objective function. From the data below, we can see that for most detectors, the false positives rates are so high that using only one detector to make the decision is not worth the cost of implementation, even with control over the false negative rate for each detector.

	False Positive Optimization			ECM Optimization		
Detector	Threshold	False Pos.	False Neg.	Threshold	False Pos.	False Neg.
1	3.27	0.94	0.05	3.27	0.94	0.05
2	0.37	0.96	0.05	0.00	0.96	0.04
3	1.11	0.95	0.05	0.00	0.96	0.03
4	2.07	0.95	0.05	0.00	0.97	0.03
5	3.46	0.95	0.05	0.00	0.97	0.02
6	3.27	0.95	0.05	0.00	0.97	0.02
7	3.16	0.94	0.05	3.16	0.94	0.05
8	2.27	0.95	0.05	0.00	0.97	0.03
9	1.26	0.96	0.05	0.00	0.96	0.03
10	2.60	0.95	0.05	0.00	0.97	0.02
11	5.50	0.95	0.05	0.00	0.98	0.02
12	5.57	0.93	0.05	5.57	0.93	0.05
13	5.68	0.88	0.05	5.68	0.88	0.05
14	3.28	0.95	0.05	0.00	0.97	0.02
15	3.62	0.95	0.05	0.00	0.97	0.02
16	5.99	0.93	0.05	5.99	0.93	0.05
17	4.86	0.94	0.05	4.86	0.94	0.05
18	5.77	0.94	0.05	5.77	0.94	0.05
19	3.70	0.95	0.05	0.00	0.97	0.02
20	3.64	0.95	0.05	0.00	0.97	0.02
21	7.66	0.94	0.05	7.66	0.94	0.05
22	14.99	0.55	0.05	14.99	0.55	0.05
23	19.64	0.12	0.05	19.64	0.12	0.05
24	5.37	0.95	0.05	4.26	0.96	0.03
25	5.97	0.94	0.05	5.97	0.94	0.05
26	5.76	0.95	0.05	0.00	0.98	0.02
27	3.99	0.95	0.05	0.00	0.97	0.02
28	5.95	0.94	0.05	5.95	0.94	0.05
29	6.77	0.93	0.05	6.77	0.93	0.05
30	6.84	0.94	0.05	6.84	0.94	0.05

tion about the source from those without. This sum also tells us whether the threshold we have calculated is useful in classification. When β is close to or greater than 1, then using such a detector singly will misclassify the larger portion of our cargo containers.

The final thing that should be noted when analyzing the information given in Table III.1 is the difference between the two objective functions – false positive minimization and expected cost of misclassification (ECM) minimization. The major distinction between the two methods is that the false positive minimization will always achieve the individual false negative rate α_{ind} for every detector, while the ECM method may not achieve this constraint for all detectors. In particular, for detectors where the two distributions overlap significantly, we can see that $P(x_i \leq t_i|S) \approx P(x_i \leq t_i|D)$, so

$$\text{ECM}(t_i) = P(x_i > t_i|S) + P(x_i \leq t_i|D) \approx P(x_i > t_i|S) + P(x_i \leq t_i|S) = 1 \qquad \text{(III.7)}$$

With such a flat objective function for all possible choices of threshold, t_i, the standard optimization algorithms will not find a global minimum of use in classification. Several studies were undertaken where the false negative probability $\alpha = P(x_i \leq t_i|D)$ was specified at a variety of levels. In each case, those detectors with separation in the distributions of the two classes produced the same threshold through minimization of either the false positive rate or the expected cost of misclassification. Furthermore, the specified false negative rate for each such detector, α_{ind}, was achieved with both methods. However, the global false negative rate $P(\mathbf{x} \leq \mathbf{t}|D)$, which is calculated using information about all of the detectors, was more difficult to control when using the ECM objective function as a result of the flat functions for most detectors described above. Thus, while the threshold for each individual detector was increased iteratively in our algorithm, there was not enough information to motivate a change in the global false negative rate. Thus, for the rest of this work and in comparison with other algorithms, we will minimize the false positive rate to simplify the calculations and provide greater control over the global false negative rates.

III.1.2 Challenges in Controlling the Global False Negative Rate

Every time that the Box Method is implemented to determine detector thresholds, we need to calculate the global false negative rate:

$$P(S|D) = P(\mathbf{x} \leq \mathbf{t}|D) = \prod_i \sum_{x_i=0}^{t_i} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i} \qquad \text{(III.8)}$$

which may involve summation over a possibly high dimensional feature space. Thus, controlling global error rates can be computationally expensive, as stated in Sec. I.6. Addition-

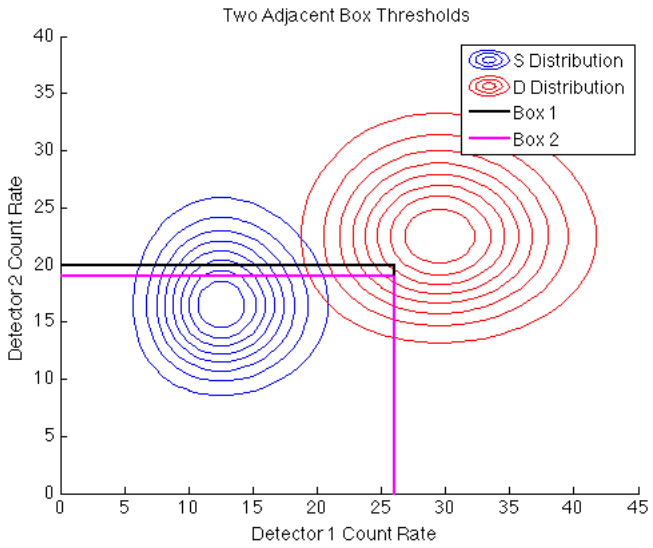


Figure III.3: There are several challenges in adjusting the box threshold to have a specified false negative rate. Here we have two different box thresholds in a 2 dimensional measurement space. By adjusting the detector 2 threshold by 1 count, we can generate two boxes – Box 1 (black line) has a false negative rate of 0.0498 and Box 2 has a rate of 0.0369. We can not further adjust this particular threshold to get any false negative rate in between these two values because of the discrete nature of the Poisson distributions.

ally, the discrete nature of the Poisson distribution can present computational challenges. Due to the fact that we are dealing with a rectangular space, every time we adjust the boundaries of the region, we add a number of points to the labeled safe region, which means that we can not specify the false negative rate exactly. For instance, in Fig. III.3, we have two boxes in two dimensional feature space, where we have changed the detector 2 threshold by 1 count. The inner box has a false negative rate of $\alpha = 0.369$ and the outer box has $\alpha = 0.498$. Since the photon count rate measurements lie at integer points, no values of α between 0.369 and 0.498 can be achieved by varying the threshold of detector 2.

We will further discuss the performance of the Box Method in Sec. III.3, when it is compared to the Bayes' Optimal Decision Method.

III.2 Analytic Bayes' Optimal Decision Method

The other option for a classification method using analytic distribution information that will be discussed in this study is the Bayes' Optimal Decision Method. Unlike in the previous method, all the detectors that will ultimately be used to make the classification

will also be used simultaneously to arrive at the decision rule. This will allow correlations between the measurements to be exploited in order to produce more accurate classifications. In fact, this method will achieve the optimal decision rule. However, it may not be possible to achieve this in practice as we are not usually provided with complete analytic knowledge of the distributions of measurements.

Following the procedure of Sec. I.3, we would like to find the region of feature space A which minimizes the total expected cost of misclassification (ECM)

$$\begin{aligned} \text{ECM} &= c_{S|D}P(S|D) + c_{D|S}P(D|S) \\ &= c_{S|D}\sum_{x\in A} p(x|D)P(D) + c_{D|S}\sum_{x\in A^C} p(x|S)P(S) \end{aligned} \tag{III.1}$$

subject to the constraint that the false negative rate $P(S|D)$ must be no more than α, i.e., $\sum_{x\in A} p(x|D)P(D) \leq \alpha$. This will be done utilizing the likelihood ratio formulation where a cargo container is classified as safe if its measurement, x, satisfies the inequality

$$\frac{c_{S|D}P(D)}{c_{D|S}P(S)} \leq \frac{p(x|S)}{p(x|D)} \tag{III.2}$$

where $c_{S|D}$ is the cost of a false negative, $c_{D|S}$ is the cost of a false positive, $P(\cdot)$ is the fraction of containers of the specified class, and $p(x|\cdot)$ is the conditional distribution of measurements of the given class.

As discussed in Sec. I.2.5, determining the exact costs of each type of misclassification and even the correct proportions of containers in the overall population can be exceedingly difficult. Therefore, we will treat the left hand side of (III.2) as a constant and utilize the Neyman-Pearson Lemma (Theorem 1) to enforce our constraint. Thus, our problem becomes to determine a value for the cost ratio η such that a container is labeled safe if

$$\eta \leq \Lambda(x) = \frac{p(x|S)}{p(x|D)} \tag{III.3}$$

where η is chosen so that $P(\Lambda(x) \geq \eta|D) \leq \alpha$. In this fashion, we will trade estimation of the misclassification costs in the original cost minimization formulation for control of the global false negative rate.

Remark III.2.1. It is important to note that we could have specified the overall false positive rate that we would find acceptable instead of the false negative rate and this would not substantially change the overall algorithm. The general statement of the Neyman-Pearson Lemma makes this possible by changing the labeling system and a few inequalities.

III.2.1 Determination of Cost Ratio

One way to determine the cost ratio is through a root finding approach. We begin by assuming that there is a value η_0 such that the safe region is defined by $\frac{p(x|S)}{p(x|D)} \geq \eta_0$ and has the desired false negative rate $\alpha = P(\Lambda(x) \geq \eta_0|D)$. Then for any other choice of η, we can determine the associated false negative rate, $\alpha_\eta = P(\Lambda(x) \geq \eta|D)$. One can then adjust η until α_η is as close as possible to the specified level α by finding the roots of

$$f(\eta) = \alpha_\eta - \alpha = P(\Lambda(x) \geq \eta|D) - \alpha. \tag{III.4}$$

This turns the problem of finding thresholds for multiple detectors into an effectively one dimensional problem.

Furthermore, we can notice that $f(\eta)$ is a monotonically decreasing function as a result of the properties of probability distributions. One of the basic properties of probability distributions is monotonicity, i.e., if $A \subseteq B$, then $P(A) \leq P(B)$. Thus, if $\eta < \beta$, then this implies that $\{x|\Lambda(x) \geq \beta\} \subset \{x|\Lambda(x) \geq \eta\}$ and hence

$$P(\Lambda(x) \geq \beta|D) \leq P(\Lambda(x) \geq \eta|D). \tag{III.5}$$

Combining this monotonicity with the normalization of probability measures, we can see that, as $\eta \to 0$, $P(\Lambda(x) \geq \eta|D) \to 1$ and for η large, $P(\Lambda(x) \geq \eta|D)$ will be near 0, so $f(\eta)$ will range from $1 - \alpha$ to $-\alpha$ as η grows. From the monotonicity property, we can also conclude that there is an interval $[a, b]$ of finite length such that $f(\eta)$ is close to 0 for all $\eta \in [a, b]$. This function may not actually attain 0 because we are dealing with the discrete Poisson distribution rather than a continuous distribution. Measurements drawn from the Poisson distribution are either outside of the set $\{x|\Lambda(x) \geq \eta\}$ or inside and as such the summation over this set has a step like increase when a new point is included. This is similar to the problem mentioned in Sec. III.1.2 for the Box Threshold Method. Likewise, the discrete nature of the distribution means that we may not be able to choose one particular value of η as the root of the function $f(\eta)$ since, by definition, neither distribution varies smoothly.

We will use a standard bisection method to locate roots of $f(\eta)$ to within a reasonable tolerance, thereby determining an appropriate value for η_0. Because we are generally working in high dimensions, we will use stochastic integration techniques to approximate $P(\Lambda(x) \geq \eta|D)$ instead of actually performing the summation:

$$P(\Lambda(x) \geq \eta|D) = \sum_{\{x|\Lambda(x) \geq \eta\}} p(x|D). \tag{III.6}$$

This means that we also avoid having to parameterize the boundary of the region and avoid some of the computational problems that can occur in high dimensional integration.

As a result of this treatment of the cost ratio, we can make use of our calculated values for η and known information to estimate the dollar cost of a false negative $c_{S|D}$ for a particular implementation of the algorithm. For example, we know that there are roughly 11 million containers entering US ports each year [53] and according to the International Atomic Energy Agency (IAEA), there have been 2331 confirmed incidents involving illicit trafficking and other such unauthorized activities involving nuclear material in the period from 1993 to 2012, 16 of which have involved "unauthorized possession" of HEU or Plutonium [22]. Supposing that all of these events were to occur by using cargo containers to smuggle such material into American ports, we have an average of 0.8 events per year, giving us the proportions of each container type in our population as $P(D) \approx 7.3 \times 10^{-8}$ and $P(S) = 0.99999993$. We can further estimate the cost of physically searching the cargo container needlessly, $c_{D|S}$, by assuming that it will take 8 man hours to perform the search and the average dock worker is paid \$14 per hour. Giving these workers a hazard pay of \$25 per hour, we can assume that the cost of such a needless search is $c_{D|S} = \$200$. Thus, the cost of allowing nuclear material to escape detection by our algorithm is

$$c_{S|D} = \eta \, c_{D|S} \, \frac{P(S)}{P(D)} \approx 2.7 \times 10^{9} \eta \tag{III.7}$$

Therefore, if we found that a 5% false negative rate gave a value of $\eta = 0.1$, then our algorithm has assigned a dollar value of around 274 million to the destruction of a city. In comparison, Hurricane Katrina cost roughly 108 billion dollars in property damage and destruction [29]. It should be noted that the estimate of $P(D)$ used here is different from the true average value as a result of our assumptions about the IAEA statistics. The events recorded by the IAEA are international statistics, not just those events occurring at American ports. Furthermore, 25 kg of HEU are required by the IAEA before a "significant quantity" of material is obtained. The IAEA defines a *significant quantity* to be "the approximate amount of nuclear material for which the possibility of manufacturing a nuclear explosive device cannot be excluded" [23]. Based on this fact, 25 crates, each containing 1 kg of HEU, would need to be smuggled into the country in a relatively short time period. If we assume that all of the material must make it into the country within a single year, the proportions of each container type become: $P(D) = 2.3 \times 10^{-6}$ and $P(S) = 0.999997$. Performing the calculation in the same manner as before, $c_{S|D} \approx 8.8 \times 10^{7} \eta$, which decreases the cost of allowing a single container to escape detection significantly ($\approx \$9 \times 10^{6}$). Since the total smuggled source requires 25 crates, this translates to a cost of approximately \$220 million for the entire 25 kg of HEU.

We can also reverse our procedure and determine the equivalent cost of searching every container, $c_{D|S}$, if we suppose that the destruction of a city through nuclear material is equivalent to the impact of this hurricane. In this case,

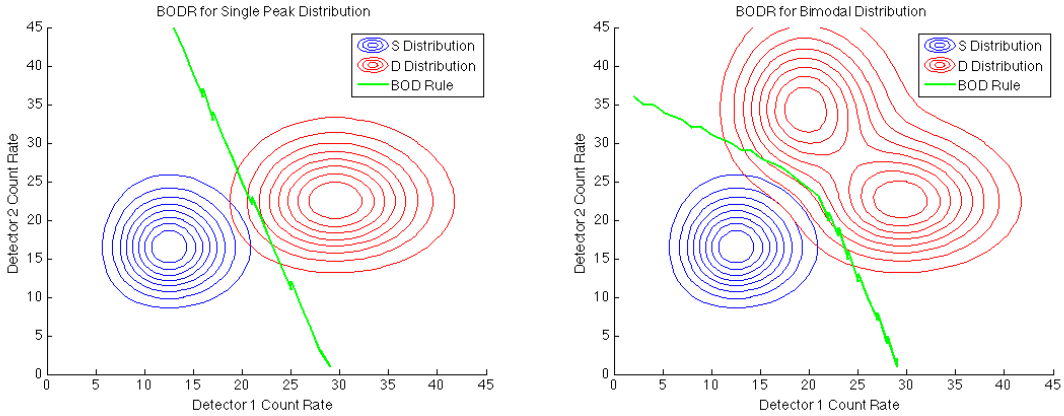
$$c_{D|S} = \frac{c_{S|D}}{\eta} \frac{P(D)}{P(S)} \approx \frac{7.8 \times 10^3}{\eta} \tag{III.8}$$

With the same 5% false negative rate and $\eta = 0.1$ as in the previous analysis, this computation suggests that hand searching a single container is worth on the order of 78 thousand dollars. In reality, the cost of searching containers is far less than this and leads one to conclude that searching containers is worth the cost in return for a large decrease in risk. Furthermore, these calculations suggest that it might be more appropriate to control the overall false alarm rates of the system if one wants to control the cost of the entire system since the general population contains many more safe containers than those with a source. However, we will continue to constrain the false negative rate in this study for consistency.

III.2.2 Initial Tests of the Bayes' Optimal Method

The actual shape of the region defined by specifying $\Lambda(x) \geq \eta$ is highly dependent on both the character of the distributions used to determine the value η and the global false negative rate α. For example, tests were completed using two detectors where the distribution of measurements for containers with a source, $p(x|D)$, changed from a single Poisson distribution to a bimodal distribution that is the sum of two Poisson distributions. Both tests had the same distribution for safe measurements, $p(x|S)$, which is a two-dimensional Poisson distribution with mean $(13.0, 17.0)$. In the first test (Fig. III.4a), the mean of the dangerous distribution was placed at $(30.0, 23.0)$. In the second test (Fig. III.4b), the means of the two dangerous distributions are $(30.0, 23.0)$ and $(25.0, 35.0)$. Given the same false negative rate $\alpha = 0.05$, we can see that the two curves generated differ substantially as the bimodal distribution has forced the boundary of the Bayes' Optimal Decision Region to bend. As a result, accurate and complete characterization of the distribution of measurements is necessary in order to develop the most accurate classification algorithms.

Alternatively, we can study the region described by the Bayes Optimal Decision rule for a fixed distribution as the percentage of false negatives is varied. Using the bimodal test distribution as in the previous discussion, tests were completed allowing only the desired false negative rate α to vary. In this particular case, it appears that the boundaries vary along two vectors that are linked to the difference in means of the two distributions and the parameter that changes is the distance from each major boundary portion to the origin, as seen in Fig. III.5. More generally, the overall shape of the region, A, appears to be



(a) Single peak distribution and boundary

(b) Bimodal distribution and boundary

Figure III.4: Given here are the Bayes' Optimal Decision Boundaries for two different distributions as determined by a root finding method. This simple example shows how the character of the distributions influences the shape of the region A and emphasizes the need for an accurate characterization of the entire feature space for the most complete classification.

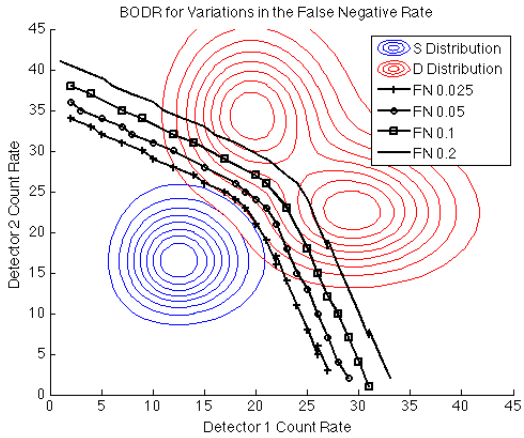


Figure III.5: Effects of varying the allowable false negative rate on the shape of the Bayes' Optimal Decision Region. As the level α increases, the boundary moves further away from the origin, but the overall shape remains constant, suggesting that the shape is controlled by the character of the distributions alone.

Table III.2: Variations in the cost ratio η and false positive rate as the false negative constraint is changed for a fixed distribution, corresponding to the curves in Fig. III.5.

False Negatives	η	False Positives
0.025	1.3580	0.0307
0.05	0.3629	0.0121
0.1	0.0736	0.0039
0.2	0.0115	0.0006

controlled by the character of the distributions used to generate the decision boundary and the false negative level α controls the size of the region. This could prove useful in updating algorithms in practice as one large detailed study could be made to understand the character of the region, which is then adjusted using more naive methods to obtain specific false negative rates as requested.

It is interesting to note the trade off in error types caused by adjusting the decision boundary. Table III.2 shows the computed values of η and the false negative and false positive rates for each of the optimal decision regions in Fig. III.5. As the percentage of false negatives doubles, the false positive rate decreases by an order of magnitude. This may not be true for the realistic distribution of measurements, but it will allow us to discuss the effects of the disparate numbers of safe and dangerous containers in our population. In particular, with the nearly 40 million safe containers each year passing through ports, decreasing the false positive rate corresponds to a significant decrease in the number of highly scrutinized containers. This translates to a reduction in cost for scanning systems and hence must be evaluated when determining the acceptable risk of undetected smuggled nuclear material.

III.3 Comparison of Analytic Algorithms

There are several metrics through which we can compare our classification algorithms, including:

1. expected false positive rate $FP(\alpha)$ for a fixed false negative rate, α,
2. total probability of error $E(\alpha) = P(S|D)P(D) + P(D|S)P(S)$ for a fixed false negative rate; in this calculation, we will use the proportions of each type of container found in Sec. III.2.1,
3. the trade-offs in these values as a result of varying the false negative rate.

In this section, we will discuss these methods with the assistance of ROC curves, as described in Sec. I.7 for the Box Threshold and Bayes' Optimal Decision Rule methods.

Table III.3: For a fixed false negative rate of $\alpha = 0.0668$, we can compare the false positive and total error rates of the Box Threshold and Bayes' Optimal methods. In the calculation of the total probability of error, we set $P(D) = 5.8275 \times 10^{-5}$ and $P(S) = 0.999942$. For both metrics, the Bayes' Method performs two orders of magnitude better than the Box Method.

	False Positive $FP(\alpha)$	Total Error $E(\alpha)$
Box	0.1378	0.137796
BODR	0.0033	0.003304

These methods will also serve as benchmarks by which we can judge the effectiveness of the algorithms described in Ch. IV.

Before we begin the analysis, it is important to note that, for any of these comparison metrics, one must integrate or sum over a portion of the feature space, which can be high dimensional. This is notoriously difficult (see Sec. I.6), so we will be approximating these quantities using Monte Carlo integration techniques. For example,

$$P(x \in A|D) = \int_A p(x|D)\, dx = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_A(x_i) \tag{III.1}$$

where x_i are a collection of N samples drawn from the distribution $p(x|D)$ and $\mathbb{I}_A(\cdot)$ is the standard indicator function returning 1 if $x \in A$ and 0 otherwise. This approximation of the integral is well known to converge as $1/\sqrt{N}$. Similar results hold for the case of discrete distributions where the integral in (III.1) is replaced by a summation.

Let us begin our analysis by examining the two detector scenario depicted in Fig. III.6, where the mean of the safe distribution is at $(13, 17)$ and the mean of the dangerous distribution at $(25, 30)$. We begin by specifying the false negative rate as $\alpha = 0.0668$ and following the procedures outlined in this chapter determine both the Box Thresholds and Bayes' Optimal Decision Rule for this data. From here, we can compare the false positive and expected cost of misclassification for each method as seen in Table III.3. For this particular level α and distribution pair, we can see that the Bayes' Method outperforms the Box Threshold method in both the false positive and total probability of error. One should also note that because there are so many more safe containers in the population, the false positive rate is a good approximation for the overall error probability, so in the future, we will just use this quantity as a measure of algorithm effectiveness.

Examining this single false negative level does not provide a complete picture of the classification method. We will discuss the sensitivity of the two algorithms to physical changes

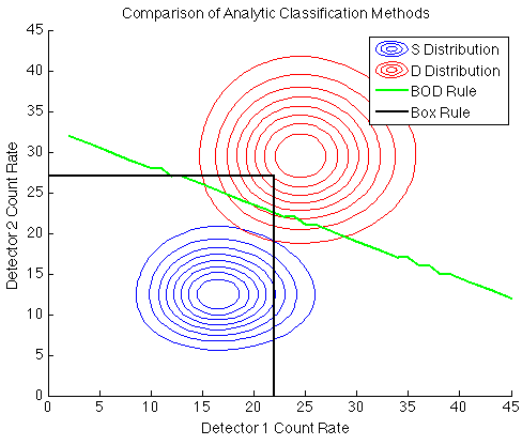
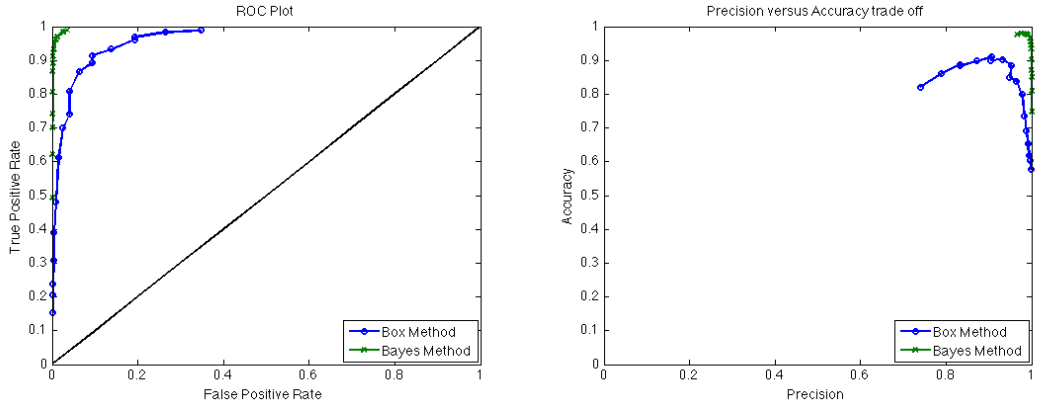


Figure III.6: Pictorial comparison of the regions generated by the Box Threshold and Bayes' Optimal methods with a two dimensional feature space. The means of the distributions are $(17, 13)$ and $(25, 30)$. The cost ratio for the Bayes' Method was $\eta = 0.0911$.

in the cargo and source in Ch. VI, but here we will discuss how variations in the false negative constraint affect the performance of the two algorithms. First, we will use ROC curves in order to analyze the performance of the two algorithms. As in Sec. I.7, this tool examines the recall (true positive) rate versus the false positive rate. We will begin with a sample set of measurements containing equal proportions of measurements from both safe and dangerous container types. This is not true in the realistic population setting, but it will be computationally useful in this analysis. For each of a range of false negative rates, the false negative and true positive values were calculated as shown in Fig. III.7a. As one can see for each tested value of our constraint, the Bayes' Optimal Method outperforms the Box Method since all of the points lie closer to the upper left corner of the plot. It should be noted that this plot has been normalized in such a way that if one wanted to examine the changes in false positive rate as a function of the false negative constraint, Fig. III.7a would simply be reflected across the line $y = 0.5$. In Fig. III.7b, one can see that varying the false negative level can affect both the accuracy and the precision of the algorithms. However, the Bayes' Method produces more precise, accurate classifications than the Box Method.

We can perform a similar analysis for higher dimensional spaces to investigate the effects of increasing the available information on our algorithm. In Fig. III.8, a 6 detector array, which includes the two detectors analyzed previously, is used in order to perform the calculations for the two plots as described above. These detectors were chosen from the 30Det subset by a feature selection method, discussed in Ch. V, to be the most useful in

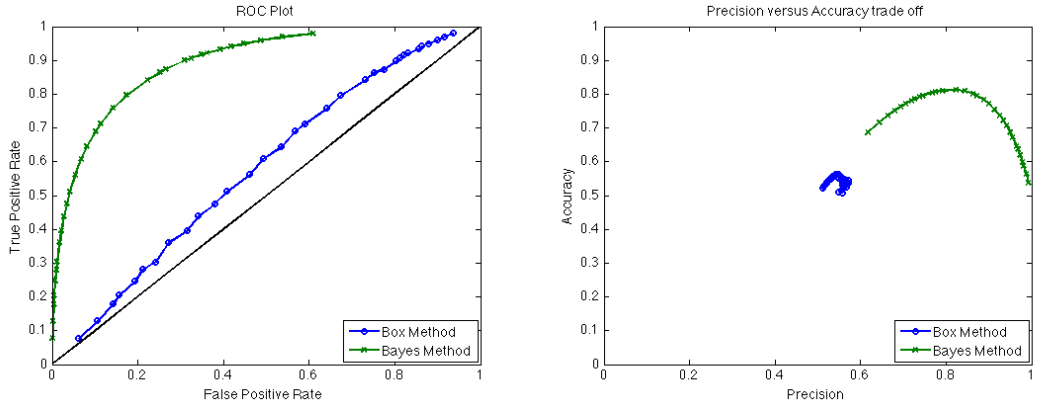


(a) ROC Plot

(b) Accuracy v. Precision Plot

Figure III.7: Depicted here for a two dimensional feature space are the ROC and Accuracy curves for comparing classification methods irrespective of a specific desired false negative constraint, α. The left plot is a standard ROC plot analyzing the relative trade off of cost versus benefit. In this case as well, the Bayes' Optimal Method is a more perfect classifier than the Box Method. The right plot shows that the Bayes' Method is both more precise and more accurate than the Box Method and that changing the false negative constraint does affect both of these quantities. The means of the distributions are $(17, 13)$ and $(25, 30)$, as before.

classification as the two distributions are the most well separated of the available information. As one can see in the diagrams, the ROC curve for the Bayes' Method is once again closer to the corner indicating a superior classifier while the Box Method is little better than a random guess in this higher dimensional space. Similarly, the Bayes' Method is still more accurate than the Box Method. Unfortunately in this case, increasing the number of features available in order to make the classification appears to decrease our accuracy as compared to the 2 detector scenario and increased the false positive rate for any fixed false negative rate. This phenomenon is explained by a combination of the Curse of Dimensionality and the introduction of redundant data. We have added computational complexity by increasing the dimension of our array without adding much more information about the nature of the source detection problem. This is reflected in the fact that measurements from any of the 4 new detectors are almost indistinguishable between classes, when taken singly.



(a) ROC Plot

(b) Accuracy v. Precision Plot

Figure III.8: Shown here for a 6 dimensional feature space are the standard ROC and Accuracy/Precision plots. The ROC curve for the Bayes' Method is once again closer to the corner indicating a more perfect classifier while the Box Method is little better than a random guess in this higher dimensional space. The Bayes' Method is still more accurate and more precise than the Box algorithm.

CHAPTER IV

OPTIMAL CLASSIFICATION WITH SAMPLE DATA

Perhaps, one of the more natural ways to consider this classification problem is as an optimization problem with an objective function and constraints. This chapter examines two different classification methods that incorporate sample measurements as opposed to analytical information about the distributions of measurements within the framework of optimization. We will first consider an approach involving a naive optimization formulation and then move to a support vector machine implementation. The classical optimization formulation, while more intuitively understandable than the support vector machine approach, provides several examples of the challenges of implementing classification methods with discrete distributions, which we will go over in depth in Sec. IV.1.2.

In both these cases, we will use sample data pulled from the same distributions and cargo loading scenarios used in the previous chapter, but the classification algorithm will be developed with no knowledge of the underlying distributions. Instead of two distributions that are referred to as either "safe" or "dangerous" depending on whether they describe the possible measurements from an innocuous container or one containing an HEU source, we have two sets of example measurements with the same labeling system. This is all explained in detail in Ch. II, including how the example measurements are generated.

IV.1 Misclassification Minimization as an Optimization Problem

In the previous section (Sec. III.3), we noticed that because our population of containers has far more harmless containers than those with sources, we can use minimization of the false positive rate with the false negative rate constrained to obtain an almost identical result. This was, in fact, the first approach taken to analyze this detection problem. However, as mentioned in the literature search (Sec. I.4), this approach is not very effective in two dimensions let alone in higher dimensions. We will first discuss the construction of the objective and constraint functions from sample measurements. Then, we will examine a few challenges of practical implementation and some of the initial results obtained with this optimization formulation.

IV.1.1 Construction of the Optimization Problem

Here we will begin with a collection of samples – N_S of which are obtained by observing containers of type S and N_D associated with class D. For simplicity when discussing

this formulation, we will abuse the notation $x \in S$ to denote a measurement x from this being obtained by observing a container of type S, and similarly for measurements from dangerous containers, D.

As always, our goal for any classification approach to this problem is to minimize the expected number of misclassifications while restricting the expected number of false negatives (the percentage of containers containing an HEU source that escape detection) to be less than a specified percentage. In a similar manner to the methods from Ch. III, the optimization algorithm will ultimately define a region of the measurement space, $A \subset \mathbb{S}^n$, and a container will be labeled as "safe" if its associated features lie within this region, i.e., $x \in A$.

Our first task is to create the objective function – a function that defines the expected probability of obtaining a false positive as a function of the region A. We can utilize the same principles that make Monte Carlo integration possible in order to define the percentage of false positives produced by the region A as follows:

$$f(A) = P(S|D) = \frac{\int\limits_{A^C} p(x|S)\, dx}{\int\limits_{\mathbb{S}^n} p(x|S)\, dx} \approx \frac{1}{N_S} \sum_{x \in S} \mathbb{I}_{A^C}(x) \tag{IV.1}$$

where $p(x|S)$ is the conditional distribution for safe measurements, possibly unknown, from which the N_S sample measurements in the "safe" set, S, are drawn. Simply put, this function determines the fraction of points from our random sample of safe measurements which lie outside of the region A.

Similarly, the number of false negatives, which we are constraining in order to produce a reliable algorithm, can be denoted by:

$$g(A) = P(D|S) = \frac{\int\limits_{A} p(x|D)\, dx}{\int\limits_{\mathbb{S}^n} p(x|D)\, dx} \approx \frac{1}{N_D} \sum_{x \in D} \mathbb{I}_{A}(x) \tag{IV.2}$$

where $p(x|D)$ is the distribution of "dangerous" measurements, A is as before, D is the set of example measurements from "dangerous" containers and N_D is the number of points in the set D.

This gives us the optimization formulation:

$$\min_A f(A) \quad \text{subject to } [\alpha - g(A)] \geq 0 \tag{IV.3}$$

where α is a maximal allowable false negative percentage specified by the user. We are searching for a region A that will satisfy these conditions. However, the set of all possible such subsets of the measurement space $\mathbb{S}^n = \mathbb{N}^n \subset \mathbb{R}^n$ is infinite, so for the sake of decreasing the computational complexity of the problem, we will limit our discussion to regions that can be described by the intersection of half spaces. This will be further restricted so that the user specifies the number of half spaces and the normal of the bounding hyperplanes. To designate a set of M hyperplanes, one requires M n-dimensional vectors, $\mathbf{v}_i$, which denote the normal vectors for each plane, and magnitudes, $c_i \in \mathbb{R}$, which are used to specify the distance from the origin to the plane along the corresponding vector. In other words, the region A is defined by a set of vectors $\mathbf{V} = \{\mathbf{v}_i\}$ and a point from the set:

$$\mathcal{X}_{\mathbf{V}} := \{\mathbf{c} \in \mathbb{R}^n \mid \mathbf{v}_i \cdot x \leq c_i \; \forall x \in S \;\text{ with }\; \mathbf{v}_i \text{ fixed for } i = 1, \ldots, m\} \tag{IV.4}$$

As a result of this description of the space $\mathcal{X}_{\mathbf{V}}$, our minimization problem is really to choose $\mathbf{c} \in \mathcal{X}_{\mathbf{V}}$ such that we have satisfied (IV.3). The classification rule that results will guarantee that any point, x, that is in the safe region, A, must satisfy all of the following equations:

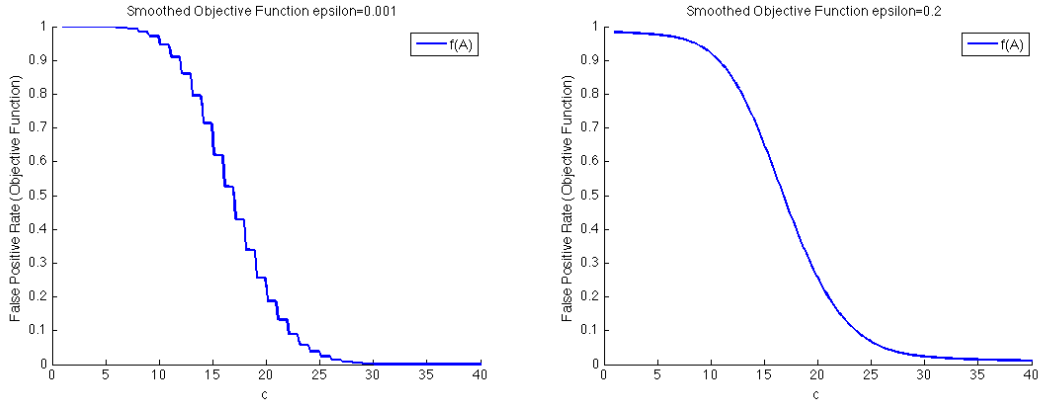
$$\mathbf{v}_i \cdot x \leq c_i \quad i = 1, ..., M \tag{IV.5}$$

for $\mathbf{V}$ fixed by the researcher and $\mathbf{c}$ chosen as a result of the optimization approach with provided training data.

IV.1.2 Implementation of the Optimization

Since the point of this study is the discussion of cost-sensitive classification algorithms, not the study of the inner workings of standard optimization techniques, we examined the effectiveness of several different optimization software packages – TAO, OPT++ and MATLAB. TAO and OPT++ had several limitations on the types of allowable constraints. In this problem, this ultimately involved a mixture of linear and non-linear inequality constraints as discussed here as well as in Sec. IV.1.1, thereby rendering TAO and OPT++ ineffective. As such, MATLAB was chosen for analysis purposes as it is a robust software package with greater flexibility in the constraints. Other challenges with the Naive Optimization formulation (Sec. IV.1.1) will be discussed in this section. Ultimately, the methods for overcoming the difficulties associated with this approach will be in vain, since the initial choice of vectors, $\mathbf{v}_i$, have a much larger impact on the efficiency of the developed classification rule.

There are two issues with our original construction of the problem that need to be addressed. First, we need to examine the minimization of real-valued functions that only



(a) Smoothed $\epsilon = 0.001$ (b) Smoothed $\epsilon = 0.2$

Figure IV.1: Using the mollified versions of our objective function, we have many versions of a one dimensional optimization function for our problem, two of which are depicted here. Using (IV.8), the overall shape of the function is preserved, but the flat regions that can create false local optima are minimized. As $\epsilon \to 0$, the smoothed objective function approaches the original function given in (IV.1).

change their value at a discrete set of points. Recall from (IV.5), we are optimizing over $\mathbf{c} \in \mathcal{X}_{\mathbf{V}}$. Most standard optimization methods for determining such quantities require information about the partial derivatives of the objective function with respect to these variables. Such data is frequently determined by numerical approximation using changes in the function value. However, in our problem, the objective function, $f(A)$, only changes when a new point in the integer lattice becomes a member of the region A as a result of the summation of indicator functions, $\mathbb{I}_A(x)$. If δA is too small here, then $f(A) = f(A+\delta A)$ and the derivative approximation will result in $f'(A) = 0$. Thus, the optimization algorithm will not be given any useful derivative information and terminate with a false optimum, when in fact the function is simply very flat in that region, see Fig. IV.1a. This is a well known problem in the realm of optimization and the most common method for alleviating the computational challenge is the use of mollifiers, which smooth away these false optima as in Fig. IV.1b.

Definition IV.1.1. We can define a family of **mollifiers** $\phi_\epsilon = \epsilon^{-1}\phi(x/\epsilon)$ in terms of convolution with a function T by requiring that:

1. $\lim_{\epsilon \to 0} T * \phi_\epsilon = T$ (the original function can be recovered)
2. $\text{supp}\,(T * \phi_\epsilon) \subset \text{supp}\,(T) \uplus \text{supp}\,(\phi_\epsilon)$ (the support of the smoothed function does not deviate too much from that of the original function)

We begin, in our particular case, to construct our smoothed function by noticing that the statement $x \in A$ is equivalent to stating that $x \cdot \mathbf{v}_i \leq c_i$ for all i as shown in (IV.5) and so the indicator function on A can be decomposed as:

$$\mathbb{I}_A(x) = \prod_{i=1}^{M} \mathbb{I}_{x \cdot \mathbf{v}_i \leq c_i}(x) \tag{IV.6}$$

Thus, we can generate a smoothing function for the whole indicator by smoothing each of the component indicator functions. We can write a single indicator function as the shifted Heaviside function, $H(c_i - x \cdot \mathbf{v}_i)$. It can be shown that convolving $\phi(x) = \left[\pi\left(1 + (c_i - x \cdot \mathbf{v}_i)^2\right)\right]^{-1}$ with this Heaviside function satisfies all of the above properties over a sufficiently large interval on the real line and provides the smoothed indicator function:

$$t_i(x) = \frac{1}{\pi} \arctan\left(\frac{c_i - x \cdot \mathbf{v}_i}{\sigma\epsilon}\right) + \frac{1}{2} \tag{IV.7}$$

Therefore, we can use the smoothed approximation of the indicator function:

$$\mathbb{I}_A(x) \approx t_A(x) = \prod_{i=1}^{M} t_i(x) = \prod_{i=1}^{M} \left(\frac{1}{\pi} \arctan\left(\frac{c_i - x \cdot \mathbf{v}_i}{\sigma\epsilon}\right) + \frac{1}{2}\right) \tag{IV.8}$$

where c_i and $\mathbf{v}_i$ are the same as in the half space description above, σ is the square root of the variance of the distribution of safe points, and ϵ allows for the adjustment of the smoothness of the function. As ϵ tends to zero, this function, $t_A(x)$, becomes more like the indicator function function $\mathbb{I}_A(x)$. At the boundary of the region, the function varies smoothly and monotonically between 0 and 1. By adjusting the smoothing parameter, ϵ, one can mitigate the step function nature of the objective and constraint functions as shown in Fig. IV.1b. In order to approximate $\mathbb{I}_{A^C}$ as is needed in the objective function, a similar calculation can be performed to show that:

$$\mathbb{I}_{A^C}(x) \approx t_{A^C}(x) = \prod_{i=1}^{M} \left(-\frac{1}{\pi} \arctan\left(\frac{c_i - x \cdot \mathbf{v}_i}{\sigma\epsilon}\right) + \frac{1}{2}\right) \tag{IV.9}$$

The next problem we will discuss is dealing with an objective function that contains large flat regions, as do most probability functions. Since we are again reconstructing the true objective function from sample data, there may be large regions of the feature space which contain no sample measurements and thus no useful information about the function. In addition, we are also working with probability distributions with a range limited to the values in $[0, 1]$. Even when using analytic knowledge of such distributions, cumulative probability distributions like the false positive rate have large regions of their domain

where the function changes very little. This can be seen in Fig. IV.1 for values of the coefficient $c \geq 30$. Therefore, we need to further modify our objective function to account for these small changes in slope, which will be done here by using $\ln(f(A))$ giving us the modified objective function:

$$\min_{\mathbf{c}\in\mathcal{X}_{\mathbf{V}}} \left\{ \ln\left(\frac{1}{N_S} \sum_{x\in S} t_{A^C}(x) \right) \right\} \quad \text{subject to} \quad \left[\alpha - \frac{1}{N_D} \sum_{x\in D} t_A(x) \right] \geq 0 \tag{IV.10}$$

There are several methods for solving constrained optimization problems numerically, including quadratic penalty methods, Lagrange multiplier formulations, and active set methods [41]. In this study, MATLAB's *fmincon* function from the Optimization Toolbox is used to solve the optimization problem with the specified constraints. This function constructs and solves a Lagrange multiplier formulation of the user provided problem.

Unfortunately, as a result of these modifications to make the problem numerically tractable, several false optima have been introduced. From a common sense and physical point of view, we know that for large count rates (even on a single detector), our algorithm should classify the measurement as dangerous, but that is not necessarily reflected in our choice of objective function, as we do not have sample information from the entire space of dangerous measurements. To ensure this classification occurs with this formulation, either more points from the dangerous distribution are needed so that information about the whole domain of the distribution is available or one must add a penalty term for large values of the coefficients, $\|\mathbf{c}\|$. This particular phenomenon can also be managed by the choosing the bounding hyperplanes so that not all normals are parallel to a detector axis, which is the method we chose here.

Furthermore, our smoothed objective function exhibits behavior that is not present in the original formulation. We consider the case where our objective function is a given by the product of two indicator functions where the basis vectors are not orthogonal:

$$f(c) = \sum_{x\in S} \mathbb{I}_{(x\cdot\mathbf{v}_1 \leq c_1)}(x) \mathbb{I}_{(x\cdot\mathbf{v}_2 \leq c_2)}(x) \,, \tag{IV.11}$$

where $\mathbf{v}_1 = \left[\frac{1}{2}, \frac{\sqrt{3}}{2}\right]$ and $\mathbf{v}_2 = \left[\frac{\sqrt{3}}{2}, \frac{1}{2}\right]$. Then, for a fixed value of one coefficient c_1, the other coefficient c_2 can be increased to a point where $\{x|\ x\cdot\mathbf{v}_1 \leq c_1\} \subseteq \{x|\ x\cdot\mathbf{v}_2 \leq c_2\}$ and increasing c_2 further should not change the objective function either, and yet, with our smoothed formulation, it does. As a result, changing the starting value in numerical optimization software can change the solution, which is one reason why starting from a number of randomly selected points is suggested in the literature (Sec. I.4).

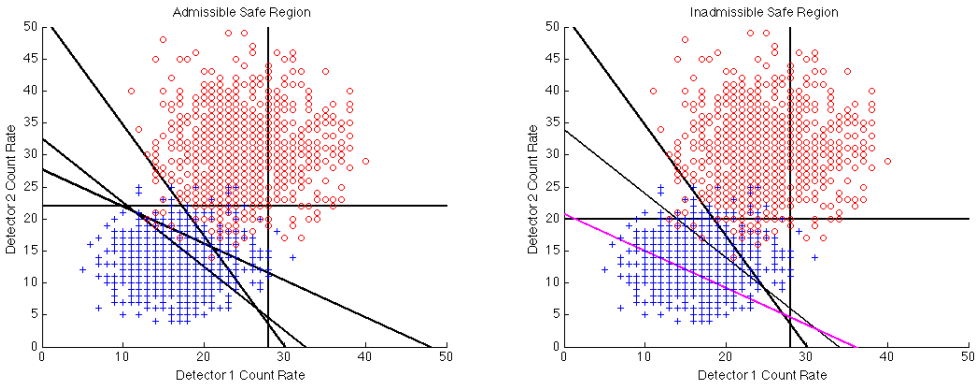
There are three other ways in which we could deal with these false optima. The first is to gain more example measurements with which to build the objective function and, thus gain more complete information about the true objective function. While we are generating our own sample measurements, this may be possible, but it is usually infeasible to do so in a real world situation. Therefore, we will ignore this option. The next method is to introduce a distance penalty term to the objective function. This modification will penalize safe region choices where points from the dangerous distribution lie farther inside the region by adding the distance from each point to the boundary, for instance by a distance function like:

$$h(A) = \sum_{x \in D} \widetilde{u}(x) = \sum_{x \in D} \min_i \left(\frac{c_i - x \cdot \mathbf{v}_i}{\sigma \epsilon} \right) \tag{IV.12}$$

where σ and ϵ are scaling parameters based on the standard deviation of the points from population D and the smoothing parameter discussed previously. When adding this term into the objective function, one must weight it with an adjustable parameter, $1/\mu$, that will serve as a Lagrange multiplier which a researcher must vary by hand. As μ becomes large, the objective function will behave just as (IV.10) and, as μ becomes small, it will more heavily penalize regions with higher false negative rates. This distance penalty method works most effectively when there are points spread throughout the feature space, so that in places where the conventional objective formulation changes very little, the distance penalty can still effect the function. As mentioned previously, this is not necessarily the case in our situation, so although this method was tested initially, it was abandoned in the final formulation.

The final method, and the one we will use in this particular formulation, is a system of linear inequality constraints that force all of the hyperplanes to contain at least one point on the boundary of the safe region. Suppose we examine five hyperplanes described by the pairs $(c_i, \mathbf{v}_i)$ in a two dimensional feature space, which, in concert with the x and y axes, describe the convex hull that is the safe region. Then, as in Fig. IV.2, one can see that the main difference between an admissible and an inadmissible hull is the fact that an inadmissible hull has one line, j, where there is no point on this line x_j that satisfies all of the i inequalities $x_j \cdot \mathbf{v}_i \leq c_i$. One can test this property by verifying that no line can be removed without altering the convex hull. Unfortunately, this method only limits where the false minima of the function may occur and does not remove them. Therefore, starting the numerical optimization from a variety of different points is still advisable in order to find the global minima.

One should note that, as stated in Sec. I.4 and Sec. I.6, all of these problems become more pronounced in higher dimensional feature spaces and as the number of optimization



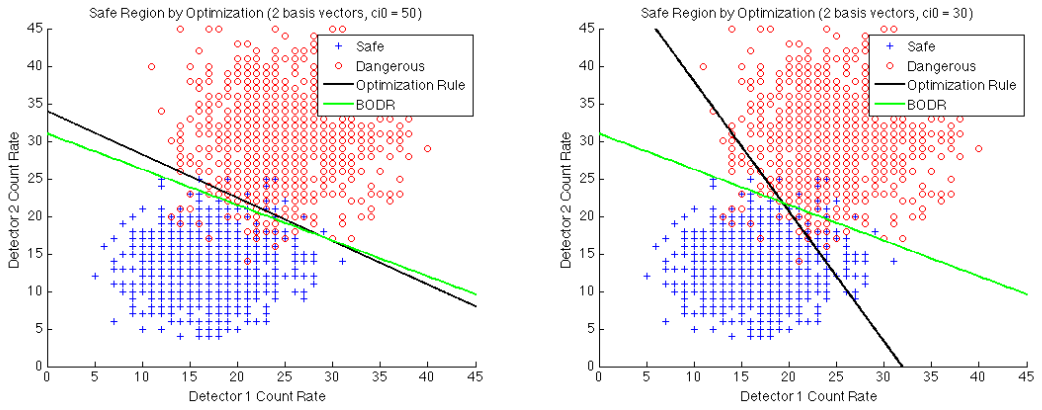
(a) Admissible Region (b) Inadmissible Region

Figure IV.2: These figures show two examples of the safe region, A, made up of the intersection of 5 halfspaces – one of which is admissible and the other is not. The pink line in the figure on the right is the one that causes problems in this formulation because now only four of the five lines adjoin the safe region, A.

variables, M is increased. A larger and more descriptive sample set would be required in order to make effective use of this formulation. Thus, this method was discarded in the final comparison of algorithms.

IV.1.3 Initial Results and Comparison to Bayes' Optimal Decision Rule

Before abandoning the naive optimization method, we will look more closely at some of the results for a two dimensional feature space. As one can see in Table IV.1 and Figs. IV.3 and IV.4, the outcomes of the optimization method are heavily dependent on the choice of hyperplane normals and the starting point of the optimization algorithm, even in this low dimensional feature space. In this case, the dominant hyperplane in the description of the safe region changes as a result of a change in the initial starting point for the optimization algorithm.



(a) Safe Region, $A - c_{i0} = 50$

(b) Safe Region, $A - c_{i0} = 30$

Figure IV.3: Here, we compare the classification rules developed using the Bayes' Optimal and the Naive Optimization methods with normal vectors $\left[1/2, \sqrt{3}/2\right]$ and $\left[\sqrt{3}/2, 1/2\right]$ and two choices for starting point $\mathbf{c}_{i0} = 30$ and $\mathbf{c}_{i0} = 50$. The means of the distributions are $(17, 13)$ and $(25, 30)$. Both regions have a false negative rate of 0.05, for which the Bayes' method has a cost ratio of $\eta = 0.1734$. The associated false positive rates are given in Table IV.1.

Table IV.1: False positive rates for the various methods shown in Fig. IV.3 are calculated here using a common sample set of measurements rather than analytic knowledge of the distributions. The Bayes' Optimal Decision Rule has a false positive rate of 0.005, if calculated analytically. While it appears that the false positive rate decreases for our Naive Optimization routine with $c_0 = 50$, this is actually a response to the fact that the Naive method deals with incomplete sample data and not the analytic distribution as in the Bayes' Optimal method.

Method	False Positive Rate
Bayes' Optimal	0.0140
Optimization – $\mathbf{c}_{i0} = 50$	0.0120
Optimization – $\mathbf{c}_{i0} = 30$	0.0570

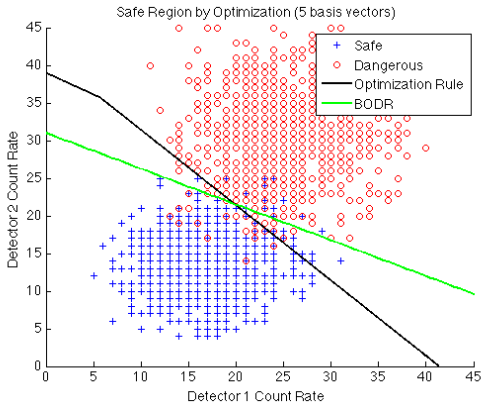


Figure IV.4: Here, we compare the classification rules developed using the Naive Optimization and the Bayes Optimal methods with 5 basis vectors. In this case, $[1, 0]$, $[0, 1]$, $[\sqrt{2}/2, \sqrt{2}/2]$, $[1/2, \sqrt{3}/2]$ and $[\sqrt{3}/2, 1/2]$ are used. Both regions have a false negative rate of 0.03380, for which the Bayes' method has a cost ratio of $\eta = 0.514984$.

IV.2 Support Vector Machine Implementation

As we saw in the previous section, one of the major drawbacks of the Naive Optimization formulation of our detection problem is that the researcher must choose the normals of the separating hyperplane *a priori*. This choice combined with a choice of starting point for the numerical optimization algorithm can drastically affect the accuracy of the classification rule produced. Support Vector Machines (SVM) are a learning method that takes these problems into account by allowing both the normal direction and the distance to the origin to vary when choosing a single hyperplane that separates the two populations. This section will discuss the most basic SVM formulation before moving onto one of the cost sensitive versions of the method and the initial implementation and results for our problem.

IV.2.1 Basics of Support Vector Machines

Support Vector Machines are a specialized type of classification formulation first stated by Cortes and Vapnik in [9]. It can be stated as a problem where the labels of each sample reading are known (supervised), the labels for some of the samples are unknown (semi-supervised) or no knowledge of the class is tied to the samples (unsupervised). The most common and tested formulation is the supervised learning example discussed here. Consider a set of n samples (y_i, x_i) where x_i is the vector of measurements on which the decision is based and y_i is either 1 or -1 and corresponds to the labeling of the two classes we are trying to distinguish. There are two steps to any support vector formulation

– transformation of the vectors of measurements, x_i, to the Hilbert space on which the decision function will operate and then utilization of quadratic programming methods to find the optimally separating hyperplane between the two classes of samples in this space.

The first step is referred to as the kernel application, in which a sample is transformed from one space to another by means of a researcher-chosen function, $\mathbf{\Phi}(x)$. The purpose of this step is to manipulate the data into a space where there is a clear separation between the two groups under study. It can also be used to reduce the dimension of the measurement space and improve the efficacy of numerical optimization programs. These transformed points are used in the optimization problem as part of a inner product between the normal decision vector in the transformed space $\mathbf{w} = \mathbf{\Phi}(\widetilde{\mathbf{w}})$ and a transformed sample measurement – $\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x})$, where $\widetilde{\mathbf{w}}$ is in the same space as x. Thus, it is more common to see this inner product treated as a single function called the kernel:

$$k(\widetilde{\mathbf{w}}, x) := \mathbf{w} \cdot \mathbf{\Phi}(x). \tag{IV.1}$$

Commonly utilized kernels include linear transformations, the Gaussian radial basis function, higher degree polynomial functions and hyperbolic tangent functions.

After completing the kernel application step, a quadratic programming method chooses the optimal separating hyperplane by determining the normal vector $\mathbf{w}$ and associated distance b that maximizes the margin, the distance between the decision boundary and the nearest x_i from each class, on either side of the separating hyperplane, as shown in Fig. IV.5. In a similar manner to the Naive Optimization approach of Sec. IV.1, the $|b|/\|\mathbf{w}\|_2$ is the distance from the hyperplane to the origin along the normal $\mathbf{w}$. The optimal hyperplane is found through solving the constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i \left(\mathbf{w} \cdot \mathbf{\Phi}(x_i) + b\right) \geq 1, \text{ for } i = 1, \ldots, n \end{aligned} \tag{IV.2}$$

In practice, one often introduces slack variables ξ_i for each constraint in order to reduce the sensitivity of the method to outliers and to deal with the case where the distributions of measurements as represented by the sample population overlap significantly. This results in one of two equivalent formulations: the C-SVM method [9] or the ν-SVM [47] formulation.

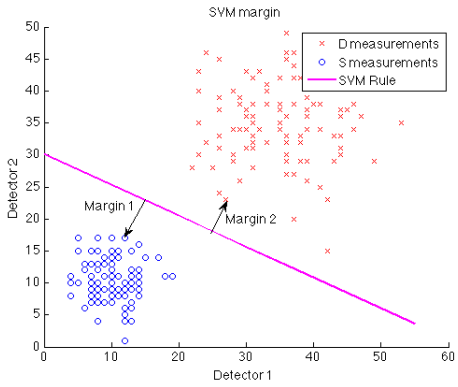


Figure IV.5: SVM methods choose the separating hyperplane that maximizes the margins between the two classes of points. In the original formulation, the two margins have equal size $1/\|w\|$. The ν-SVM formulation allows some of the points from the training method to be inside of these margins or to be misclassified. This can be necessary, especially when the two sets of points S and D are not easily separable.

The ν-SVM formulation is shown here:

$$\begin{aligned} \min_{\mathbf{w},b,\rho} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \nu\rho \\ \text{s.t.} \quad & y_i\left(\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b\right) \geq \rho - \xi_i, \text{ for } i = 1,\ldots,n \\ & \xi_i \geq 0 \text{ for } i = 1,\ldots,n \\ & \rho \geq 0 \end{aligned} \tag{IV.3}$$

where $\nu \in [0,1]$ is an upper bound on the fraction of margin errors produced by the algorithm, i.e., the fraction of sample points which are incorrectly classified. We have chosen the ν-SVM formulation here because the limited domain for ν makes tuning the SVM for optimal learning easier than the unbounded approach in the original slack variable formulation, C-SVM. As an additional feature, the parameter ν bounds the margin errors and thus provides an estimate on the total probability of error (both false positives and false negatives) on the training set.

In order to make this method cost-sensitive as in the other methods in this study, we will utilize the 2ν-SVM formulation proposed in [7, 11], which relies on the introduction of a

single parameter $\gamma \in (0,1)$ that gives the trade off in the two error types:

$$\begin{aligned} \min_{\mathbf{w},b,\rho} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{\gamma}{n}\sum_{i\in I^+}\xi_i + \frac{1-\gamma}{n}\sum_{i\in I^-}\xi_i - \nu\rho \\ \text{s.t.} \quad & y_i\left(\mathbf{w}\cdot\mathbf{\Phi}(x_i)+b\right) \geq \rho - \xi_i, \text{ for } i = 1,\ldots,n \\ & \xi_i \geq 0 \text{ for } i = 1,\ldots,n \\ & \rho \geq 0 \end{aligned} \tag{IV.4}$$

where $I^+ = \{i : y_i = 1\}$ ($n_+ = |I^+|$) and $I^- = \{i : y_i = -1\}$ ($n_- = |I^-|$). As further shown in [7, 12], the choice of ν and γ can provide bounds on the fraction of margin errors of each type, ν_+ and ν_-:

$$\nu_+ = \frac{\nu n}{2\gamma n_+} \qquad\qquad \nu_- = \frac{\nu n}{2(1-\gamma)n_-} \tag{IV.5}$$

In practice, one usually works with the dual form of (IV.4):

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{\Phi}(x_i)\cdot\mathbf{\Phi}(x_j) \qquad \text{(IV.6)} \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq \frac{\gamma}{n}, \text{ for } i \in I^+ \\ & 0 \leq \alpha_i \leq \frac{1-\gamma}{n}, \text{ for } i \in I^- \\ & \sum_{i=1}^{n}\alpha_i y_i = 0, \;\; \sum_{i=1}^{n}\alpha_i \geq \nu \qquad \text{(IV.7)} \end{aligned}$$

where α_i are the Lagrange multipliers associated with each constraint. There are restrictions on the choices of ν and γ for the feasibility of the dual problem [12]. A search over this feasible space is necessary to determine the combination (ν, γ) so that the false negative rate is bounded by the specified rate. One can recover our original solution $(\mathbf{w}, b)$ with

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{\Phi}(x_i) \qquad \text{and} \qquad b = \frac{1}{N_{SV}}\sum_{i=1}^{N_{SV}} \mathbf{w}\cdot\mathbf{\Phi}(x_i) - y_i \tag{IV.8}$$

where N_{SV} is the number of test points $(y_i, \mathbf{x_i})$ such that the associated Lagrange multiplier is non-zero. These $\mathbf{x_i}$ are called the *support vectors* and they are the sample vectors, which influence the solution $\{\mathbf{w}, b\}$. The optimization problem can be sped up if the support vectors are known before starting, since the corresponding constraints are the most likely to be violated by changing the position of the decision boundary.

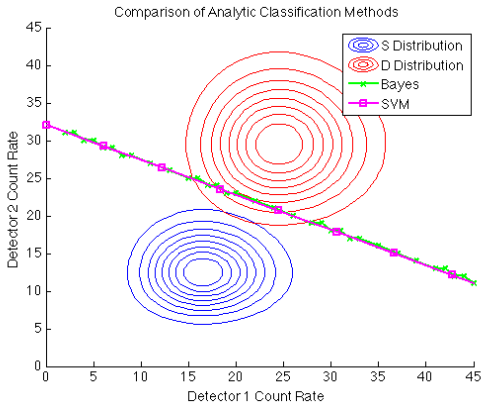
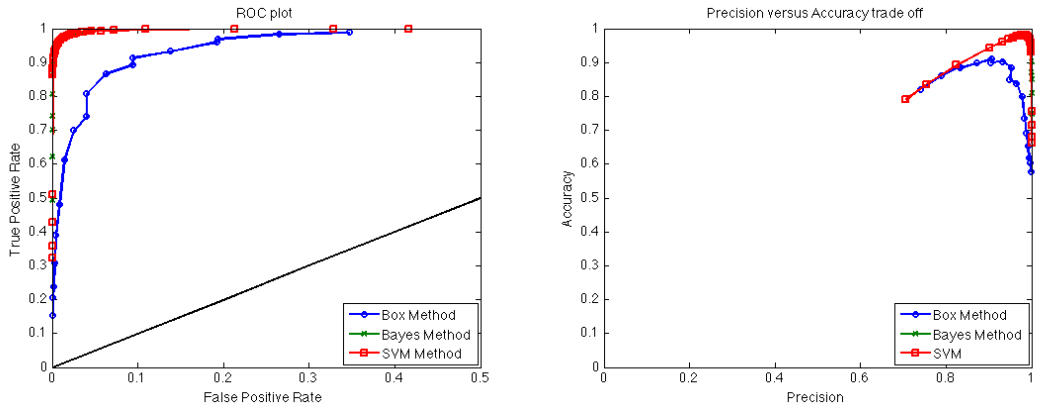


Figure IV.6: Depicted here for a two dimensional feature space are the decision boundaries created by the Bayes' Optimal and SVM methods. The two boundaries are identical for a false negative rate of $\alpha = 0.05$ with identical false positive rates 0.0055. The means of the distributions are $(17, 13)$ and $(25, 30)$, as before.

IV.2.2 Implementation and Comparison of SVM Trials

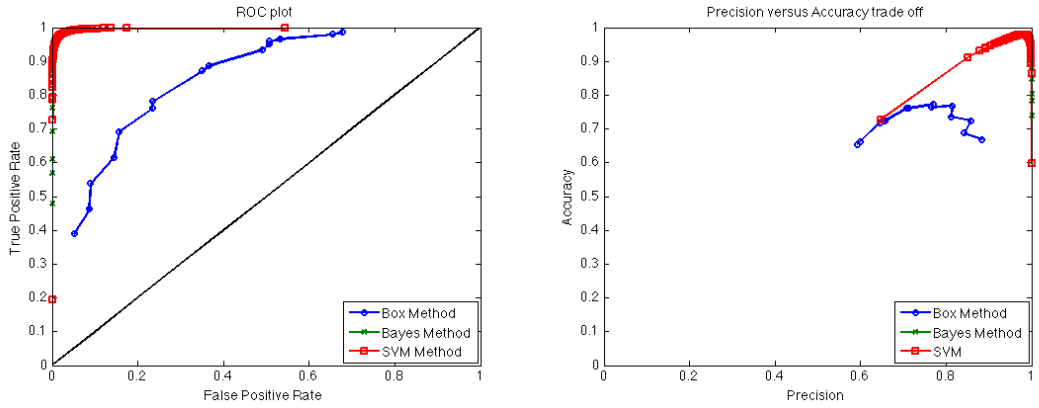
There are many pieces of software that need to work together for the support vector machine framework to operate effectively. Therefore, we used the LIBSVM software package [6] with the modifications for the 2ν-SVM formulation as given in [10]. In order to bound the false negative rate, a simple search over the feasible region is performed to gain a basic understanding of how the false negative rate changes and then localized to find a specific false negative rate in a similar fashion to the coordinate descent method described in [12]. The largest difference is that our method varies ν and γ where Davenport's coordinate descent method varies ν_+ and ν_-.

As one can see, in Figs. IV.6 and IV.7, for the two detector case, the SVM method is an improvement on the Box Threshold Method and matches well with the Bayes' Optimal solution in the space of restricted classifiers [33]. By this, we mean that, in the space of hyperplane decision boundaries, the SVM method will choose the Bayes' Optimal solution, which may differ from that described in Sec. III.2 because the likelihood ratio test does not necessarily generate a planar boundary. In this case, this may provide a skewed sense of the efficacy of the SVM as compared with the Box Method since the Bayes' Optimal solution is linear here. The kernel choice can effect the outcome as much as the sample points provided for training and there is a danger of overfitting in higher dimensions for all of the methods, as can be seen with the Box Method in Fig. IV.8.



(a) ROC Plot (b) Accuracy v. Precision Plot

Figure IV.7: As before, we can use the ROC and Accuracy curves for comparing all three classification methods (Box, Bayes' and SVM) irrespective of a specific desired false negative constraint, α for a two dimensional feature space. In this case as well, the SVM method which takes into account the correlations in measurements improves upon the classification provided by the Box Method. The means of the distributions are $(17, 13)$ and $(25, 30)$, as before.



(a) ROC Plot (b) Accuracy v. Precision Plot

Figure IV.8: Adding features until we are working with a 4 dimensional feature space, we can utilize the ROC and Accuracy curves for comparing classification methods irrespective of a specific desired false negative constraint, α. The SVM method performs almost as well as the optimal choice. However, we are starting to add information about features that do not differ between safe and dangerous containers. Therefore, the overall classification performance has begun to decline (as seen by the Box curve) and we are in danger of overfitting the SVM method to our sample points since they do not provide information about the greater volume of the feature space.

CHAPTER V

THE IMPORTANCE OF FEATURE SELECTION

One of the most important questions in the process of creating a classification method is "What measurements should we use in order to produce the best decision algorithm?" As mentioned in Sec. I.1, there are several choices for the type of measurements one can utilize, including comparison of the gross count rates of photons exiting the container and analysis of the energy spectrum of this radiation. Even if we restrict our physical measurement to the 1 MeV photon count rates discussed throughout this study, we have already seen that not every detector will perceive enough information from a source to significantly distinguish it from the background noise (Sec. III.1). Additionally, enforcing the constraint on the percentage of false negatives produced by each algorithm, requires integration over a potentially high dimensional space. As stated in Sec. I.6, this can create a myriad of problems with computation of the optimal decision rules. Good feature selection can mitigate such problems and lead to classification with minimal information.

As a result, feature selection algorithms are the topic of much discussion in several fields [2–4, 13, 19, 26, 40, 42, 43, 46]. There are three main categories of feature selection algorithms – filters, wrappers and embedded methods. Filter methods apply knowledge external to the classification framework, but intrinsic to the samples under consideration to judge the value of the subsets like the correlations between various features. Wrapper methods use the performance of the learning machine itself as an evaluation metric for the various feature subsets. This results in a set of nested loops of evaluations – first to determine the features under consideration and then to find the optimal decision rule. Embedded feature selection methods choose the feature subset and the optimal boundary simultaneously by adding more variables to the optimization problem [18]. Wrapper and embedded methods are much more computationally expensive than filter methods, but they can lead to more accurate classifications. Thus, we will focus our examination on filter methods that can be applied to any of the methods we have discussed previously.

Some studies also incorporate normalization and ranking methods as preprocessing steps before feature selection occurs to improve upon the effectiveness of the algorithm [32]. This is a way to incorporate external knowledge of the measurement types into the feature selection process, but it does have drawbacks – namely, it is very specific to the type of measurement given and the assumptions made about the initial conditions. This section will look at methods for both feature selection and data normalization that may be

employed for the source detection problem.

V.1 Filter and Ranking Methods

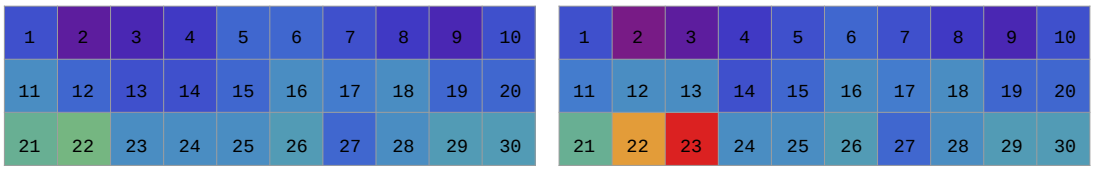
As was mentioned previously, filter methods usually focus on either analysis of the standard statistical properties such as the mean, variance and correlations of the samples. However, they can also involve external knowledge of the physical properties of measurements or the utilization of an information theoretical approach. In this section, we will discuss three different methods for feature selection, two of which will be implemented.

The first two filter methods discussed here give a ranking of each feature on its own merits and do not generally give a feel for which subset of features will work best in an ensemble classification setting. The best two features on their own may in fact give redundant information when used together. Duch advocates combating this redundancy by using the ranking capabilities of the filter in combination with the subset selection of a wrapper by adding or subtracting features in the order they were ranked when evaluating their effectiveness with the classification method [13]. However, this method requires choosing a single classification method and ensuring that a sufficient description of the feature space has already been achieved. We will discuss this later in the section.

The introductory work on these filter methods uses the 30Det detector subset from Sec. II.3.1 for Data Set A (Sec. II.2). We have avoided the full detector array since, as determined previously, utilizing 15 or 30, let alone the 320 detectors calculated with MCNP, is infeasible computationally with MATLAB. This smaller subset makes it more intuitively obvious when determining which detectors will be most helpful in making a classification. It is important to note that distance from the ground (which is a major source of background radiation) is a major factor in how much radiation is seen by a specific detector, as can be seen in Fig. V.1.

V.1.1 External Knowledge Filtration

As one can see in Fig. V.1, the cargo in the container and the distance from both the background source provided by the concrete and the internal smuggled source affect the mean count rates for each detector. In an ideal world, knowledge of the interior material of every cargo container would be available and could potentially be used to simulate the radiation pattern created by a source and the expected count rates for these detectors. However, this is not the case, as was explained in Sec. I.1. Often, the classification method and thus, the feature selection algorithm must make a decision with only the information from one example of measurements belonging to a specific container configuration. Ultimately, we



(a) Measurements without HEU source

(b) Measurements with HEU source

Figure V.1: We will begin our discussion of feature selection by using the 30 detector subset for Data Set A. This data set will be used as a starting place in order to discuss our feature selection algorithms. One can intuitively tell that detectors 22 and 23 contain information about our source and should be most useful for classification.

would like to eliminate the necessity to have knowledge of the exact cargo loading scenario of the container as well. From initial simulations, the average background count rates vary as a function of the distance from each detector to the ground or major background source contributions. If we take the average of all detectors at a fixed height, then any detector that deviates greatly from this average is more likely to contain information about whether or not there is a source in the container.

Of course, now we need to determine what we mean by "deviates greatly." We have several options for this. First, we could use the sample deviation of all detectors at a specific height in the array

$$\sigma_h^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x}_h)^2, \quad \text{where} \quad \overline{x}_h = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{V.1}$$

and the x_i are the count rates for each detector at height h as a basis for comparison. As a second option, since the detection of radiation is a Poisson process and Poisson distributions have a variance equal to the mean, one could consider using the square root of the mean as the standard deviation $\sigma_h = \sqrt{\overline{x}_h}$. This choice is based on the assumption that the dominant source of radiation is the background and that it travels evenly throughout the container so that no streaming effects are present.

Once one of these options is chosen, the algorithm claims that any detector reading of fixed height that is greater than one deviation, σ_h, above the mean, $\overline{x}_h$, is highly likely to contain information about the smuggled source. One standard deviation is chosen because the high detector readings that we want to isolate for later use have been incorporated into the mean value previously calculated. A σ_h threshold set too high in this method will not designate any detectors for further investigation. One gains further information by simply ranking

each of the detectors based on the distance in standard deviations from the observed count rate x_i to the mean count rate $\overline{x}_h$. Using the first option for σ_h and the arrays described previously, only two detectors on the bottom row (23 and 22) were determined to deviate sufficiently from the background, while the others were ranked at approximately the same level. However, the lower thresholds of the second option designate detectors 23, 22 and 6 as markedly different from the average count rates.

While this particular method of feature selection does allow the researcher to incorporate information about the shielding effects of the cargo into the decision making process, there are some substantial drawbacks that must be addressed. For instance, this method assumed that the radiation due to the concrete background traveled in a smooth fashion throughout the container and, thus, there were no substantial differences in detector readings of similar height. This is not always the case, as will be shown in Sec. VI.3.

Additionally, as discussed in Sec. I.1, we know that detectors closer to the source should observe more radiation from the internal smuggled point source. Unfortunately, with this method for either choice of sigma, we do not account for this phenomenon. Because we know the location of the source in this particular study, we know that an optimal feature selection method should choose detector 23 first and then it should rank detectors that are physically near this location as the next most likely to contain information about the internal source. However, this algorithm does not do so and is highly susceptible to fluctuations that can occur in the background radiation levels. For instance, detector 6 has a higher count rate than other detectors of similar height, but it is due to the fact that it is physically on the edge of a container and observes more radiation from the background unshielded by cargo and not its proximity to the internal source. However, according to our second option for σ_h, this is one of the more promising detectors for the purposes of our decision.

Another major drawback of this method of feature selection is that it requires a distinction in the height of the detectors in order to work effectively. If we instead used the column averaged detectors produced in Sec. II.3.2, this particular method chooses detectors 40, 39, 38 , 19, 16, 20, and 8 as the ranking for the top seven detectors to use in classification. However, the only detector that is actually near the source and is seeing radiation different from what one would expect in a container with harmless cargo is Detector 8. The rest of the detectors have larger count rates because they are near the physical edges of the cargo container and thus observe more radiation from outside of the cargo container than the other detectors.

V.1.2 Mutual Information and Maximum Relevant, Minimum Redundancy Methods

Another filter method which looked promising in initial studies is an information theoretical approach. This method looks at the correlation between a feature and the class types through the mutual information framework [8, 46]. Usually, this is done through knowledge of the joint and partial distributions of the feature space and the two classes in order to calculate an information score [19] for each test feature such as:

$$I(i) = \int_{x_i} \sum_{y \in \{S,D\}} p(X = x_i, Y = y) \log \frac{p(X = x_i, Y = y)}{p(X = x_i)P(Y = y)}. \tag{V.2}$$

This method requires accurate knowledge of the probability distributions of measurements for each container type. Furthermore, depending on which distributions are available, it may require an additional integration step to obtain the individual distributions for each feature within the label classes. Once this is completed, the problem becomes a series of one dimensional integration problems with the highest information scores corresponding to the features most likely to make accurate classifications. Because the source of our problem may move and thus, cause the relevant features to change with each container under discussion, it is difficult to get enough information about a particular container in order to estimate the appropriate distribution, making this an impractical method for implementation in the field.

There are many variations on this theme that counter some of the drawbacks mentioned previously. For instance, the Shannon Entropy for each feature can be used to overcome some of the error and slowness caused by the binning required to approximate the probability distributions in this method. However, to get the most accurate approximation, the entropy must be calculated for multiple combinations of features and then combined in a similar manner to derivatives in a Taylor expansion of a function [8]. While this makes the approximations more accurate it does not negate the need for large amounts of sample points. Furthermore, as with the other filtration methods discussed in this section, this kind of feature selection does not take correlations between features into account. However, this method can be extended to take these relations into account with the Maximum Relevance – Minimum Redundancy framework, which uses an information score between two features in addition to the score between each feature and the class labeling [42]. This allows features to be chosen that correlate well to certain classes while avoiding duplicate information. Even though these methods show a great improvement in classification accuracy in their respective studies, the necessity for large numbers of points in order to approximate distributions and choose relevant features makes this impractical in our

particular problem. Therefore, we will ignore this method in further studies.

V.1.3 F-Score Testing

With all of the assumptions required by the previous methods and the misidentifications of useful detectors in even these simplified test cases, one can see that it will be ineffective in practical application. Thus, we will move on to a feature selection method that relies less on the physical characteristics of the measurements. In doing so, we will utilize measurements from a larger set of samples than the single sample in the previous method. In this case, we will analyze the class separation in each variable through the following formulation to calculate an F-Score [13, pp. 315-324] for each feature:

$$F(i) = \frac{\left(\overline{\mathbf{x}}_i^+ - \overline{\mathbf{x}}_i\right)^2 + \left(\overline{\mathbf{x}}_i^- - \overline{\mathbf{x}}_i\right)^2}{\frac{1}{n^+-1}\sum\limits_{k\in n^+}(x_{i,k} - \overline{\mathbf{x}}_i^+)^2 + \frac{1}{n^--1}\sum\limits_{k\in n^-}(x_{i,k} - \overline{\mathbf{x}}_i^-)^2} \tag{V.3}$$

where there are n^+ "safe" sample measurements, n^- "dangerous" sample measurements, $\overline{\mathbf{x}}_i^+$ is the mean of the sample population of "safe" measurements, $\overline{\mathbf{x}}_i^-$ is the mean of the "dangerous" sample population, $\overline{\mathbf{x}}_i$ is the mean for all samples in the training population and $x_{i,k}$ is the i^{th} component of the k^{th} sample of the population in question. This provides a look at the average separation of the conditional means of the two populations as scaled by the average variance of the two populations. The features with higher F-Scores have greater separation between classes and thus, for a fixed number of features, a higher chance of minimizing the expected cost of misclassification. Also, higher F-Scores are linked to lower variances in each class, which concentrates the bulk of the readings in a more localized region of the feature space.

Working with the same 30 detector subset used in the external knowledge filter, we have evaluated the F-Score for each detector in the set, Table V.1 and Fig. V.2. Detectors 23, 22 and 13 are more obviously well separated and localized than the rest of the features. The next set of potentially effective detectors is detectors 2, 3, 5, 11 and 12, which have F-Scores of less than one tenth the size of the first group. The rest all have an F-Score less than one hundredth the size of the F-Scores in the first two categories.

This selection method is much more successful in choosing detectors that are physically nearer to the true source location than the external knowledge feature selection method and its effectiveness is not diminished with physical variations in the container provided there are multiple measurements of a single configuration with and without a source. In reality, we will not have multiple samples with which to make such decisions. However, this

Table V.1: An F-Score for each detector in the 30 detector subset can be calculated, as shown here.

0.0000	0.0061	0.0051	0.0000	0.0025	0.0000	0.0000	0.0000	0.0000	0.0000
0.0015	0.0071	0.0158	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0230	0.1333	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

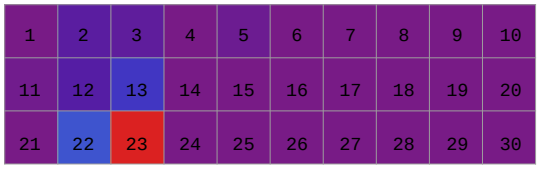


Figure V.2: We can represent the F-score graphically for 30 detector subsets. Here, we see that only three, possibly four detectors, have sufficiently greater F-scores to warrant further scrutiny.

method can operate effectively with a large sample of "safe" measurements and a single example of a "dangerous" or unknown sample, making it more practical for implementation, provided the measurement distributions do not overlap significantly. However, if only one measurement of the container is used, then the F-Score will be sensitive to detectors that deviate greatly from the mean detector count rate and may incorrectly identify the most important features if the distributions of the measurements overlap significantly.

There are situations in which this method can choose relevant features incorrectly. As a first example, if we compare cargo loadings that differ substantially even without an internal source present, the F-Score method will choose detector measurements that deviate the most between the compared container types, as will be shown in Ch. VI. This identification does not indicate the presence of a source, only differences. Similarly, if the calculated F-Scores are relatively constant throughout the feature space, then there is no information provided by this method. This can indicate that either the container has no internal smuggled source or that there is a distributed source throughout the container that masks the signal of the point source effectively. Without sufficient samples over the entirety of the feature space and with blind application of the procedure, there is no guarantee of correct identification of relevant features and it may even introduce false positives before classification begins. Despite these drawbacks, this was the most effective feature selection method tested and the one that we continue to use in further tests.

Using the F-Score as a ranking method, we can apply the subset selection method suggested

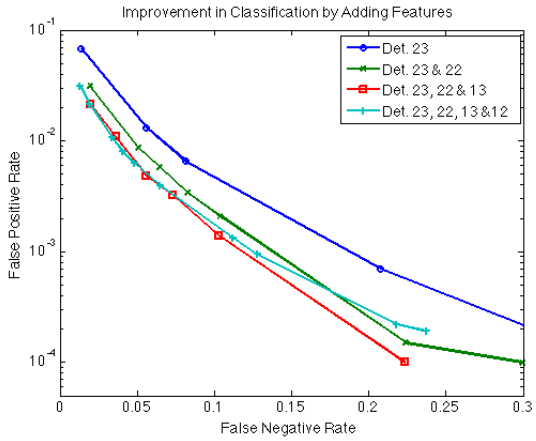


Figure V.3: Increasing the number of features used in the decision making process improves the accuracy of classification methods, but only as long as the new features contain information about the source. Even when developing classification rule via the Bayes' Optimal Decision framework, as shown in this figure, the addition of further features decreases the false positive rate, but only up to a point. Since the radiation source is small, only three features in the 30Det set contain significant amounts of radiation from the smuggled source.

by Duch [13] to determine the number of features that will allow us to make the most accurate classification possible. A subset to size k is created by choosing the features with the k highest F-Scores. As shown in Fig. V.3, for a fixed false negative rate, one can create decision rules with between 1 and 4 detectors. Initially, adding information provided by other detectors improves the false positive rate while maintaining the same level of reliability of detection. However, after a point (in this case three detectors), adding detectors only increases the computational complexity of the problem without decreasing the false positive rates. This phenomenon is due to the fact that our source is physically so small, that only a few detectors are near enough to observe a statistically significant amount of radiation, as seen in Fig. V.2. It should be noted that the number of useful features may vary with the detector array or measurement type. In particular, if we examine the 320 detector array, there are at least a dozen detectors that see an average of 15 photons more than the background radiation. In the 30Det Scenario used for initial testing, only a single detector, Detector 23, observed that much of an increase in observable radiation levels. The improvement in false positive rates for the 320 detector array was not tested due to the computational limitations of MATLAB, which can not draw enough points in the high dimensional space for the stochastic integration methods to converge in a reasonable time period.

V.2 Normalization

Normalization of features before attempting feature selection or classification is a common theme in many areas [14, 19, 32]. The goal here is to remove noise and fluctuations between measurement types so that their deviations are more easily compared while maintaining the overall characteristics of the distribution of measurements. Thus, instead of classification using the raw measurement vectors, x_i, one uses the rescaled measurements, $\widetilde{x}_i$. Many classification algorithms, including the SVM method from Sec. IV.2, use distances between points as a criterion for weighting their effect on the solution. Thus, if we consider two features, f_1 and f_2, which are known to be equally useful in classification and there is a difference in their average magnitudes, i.e. $f_1 >> f_2$ on average, these algorithms will favor variations in the feature with the higher magnitude, f_1, when determining the solution.

There is no standard method for performing such modifications to the data since it usually combines expert knowledge of the behavior of the measurements initially obtained with statistical properties of the sample. Each feature in the space is rescaled independently of the others and scaling is consistent no matter which category (S or D) from which the point is drawn. In many cases, a z-score method is used, in which the j^{th} feature is centered about the corresponding mean from the safe distribution $\overline{x}_S^j$, or its approximation, and then scaled by (an approximation of) the same distribution's standard deviation σ_S^j, i.e., $\widetilde{x}_i^j = \frac{x_i^j - \overline{x}_S^j}{\sigma_S^j}$. The intent behind this is to force the measurements to take on characteristics of a normal distribution with mean 0 and standard deviation 1, $\mathcal{N}(0,1)$, which works well with systems that have underlying normal distributions. Then, feature selection and classification work by measuring the deviation of the second distribution from this standard.

Another common method is to use Min-Max Normalization which scales the data so that the minimum measurement for the j^{th} feature, m^j, and the maximum reading, M^j, correspond to 0 and 1, respectively. This can be done using the rescaling:

$$\widetilde{x}_i^j = \frac{x_i^j - m^j}{M^j - m^j} \tag{V.1}$$

The Min-Max Method is sensitive to outliers in the data, but it does preserve the relationships among the original sample points in the rescaled sample.

In our particular study, we examined several methods for rescaling the photon count rate data, as depicted in Fig. V.4. With knowledge of the physical system, one can make several observations. First, getting a sample that completely describes the entire feature space for every detector is difficult, so having accurate knowledge of the global minimum and maximum value of each detector is not possible. For instance, the maximum count rate

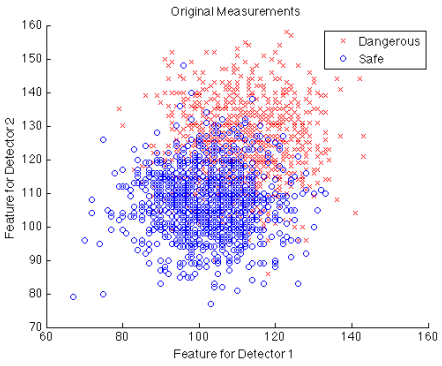
could be extremely large for sources placed immediately adjacent to a detector, but this type of source position may not be included in the training set, making Min-Max scaling impractical here. Secondly, in our situation, the average count rate for a detector is roughly correlated with the height of the detector from the concrete slab, as noted in Sec. V.1.1. In order to avoid overfitting to a single cargo loading, we normalized each detector using the average count rate for detectors of a fixed height, rather than for a specific detector. Thirdly, since the detection of photons is a Poisson process, as mentioned in Sec. II.1, the measurements from each detector roughly obey a Poisson distribution and thus, we can approximate the standard deviation by the square root of the sample mean. However, this also means that this scaling method may spread out portions of the sample with smaller count rates. Finally, in reality, there are many more examples of harmless cargo containers than of dangerous measurements, so most realistic data will be available for the safe distribution.

It is important to note that none of the analytic classification methods discussed in Ch. III will be affected by any of these linear rescaling methods. Since the exact probability of obtaining each point is known, we can associate the same probability to its rescaled point and thus, any region with a fixed false negative probability in the rescaled features will contain the same points as a region in the unscaled feature space with the same false negative probability. The rescaling will affect the performance of the SVM method. As all of these rescaling methods change the average magnitude of any given feature, the problem of unfairly weighting one feature over another is removed.

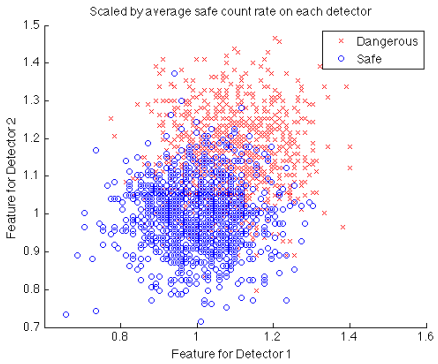
After testing the SVM method with each of these normalization schemes, except for the Min-Max scaling determined previously to be unreliable for physical reasons, we found that the z-score rescaling using height-dependent means and variances, $\overline{x}_h$ and $\sqrt{\overline{x}_h}$, proved to give nearly the optimal solution in multiple tested cases with our typical 1 MeV photon readings:

$$\widetilde{x}_i^j = \frac{x_i^j - \overline{x}_h}{\sqrt{\overline{x}_h}} \tag{V.2}$$

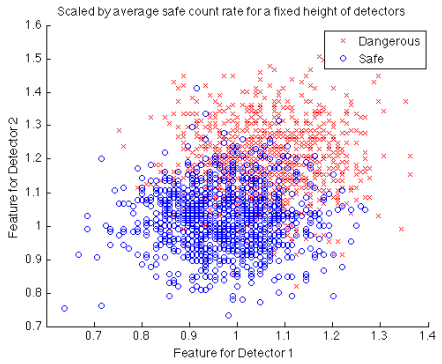
However, this method of rescaling has two major drawbacks in the general setting. First, the mean and standard deviation choices used here are tied to the physical scenarios in our test set where a set of detectors of fixed height have roughly the same average count rate. If one were to incorporate other types of measurements or intelligence data into the decision process, then these features would need to be normalized in a different fashion based on their specific physical properties. Secondly, this form of normalization does not compensate for internal sources created by harmless radiative sources like concrete. These extra internal sources force the centers of the distribution functions for some of the detectors



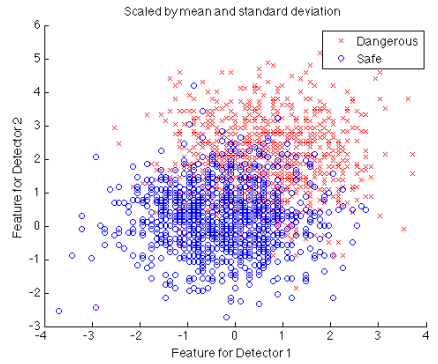
(a) Original measurements



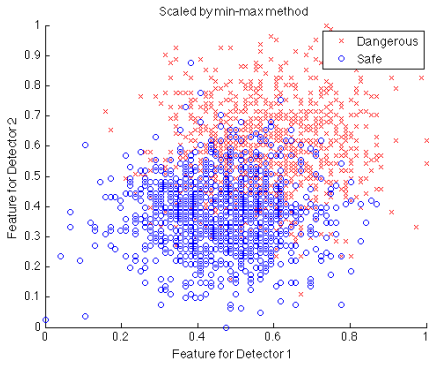
(b) Scaled by individual mean $\frac{x_i}{\overline{x}_i}$



(c) Scaled by mean for similar height h, $\frac{x_i}{\overline{x}_h}$



(d) Scaled by height based z-score $\frac{x_i - \overline{x}_h}{\sqrt{\overline{x}_h}}$



(e) Scaled by Min-Max $\frac{x_i^j - m^j}{M^j - m^j}$

Figure V.4: There are a variety of ways to rescale data so that classification methods using Euclidean distances between points do not unduly weight features based on their magnitude. Shown here are a small sample of these methods that were considered in this study that rescale data according to various statistical properties of the sample. Based on the modeling of detector readings by Poisson distributions used throughout this study, we will approximate the standard deviation of the sample by the square root of the mean.

from safe containers off of the origin, which could lead to false positives when classifying an unknown container type with a previously developed rule assuming no such sources. If the effects of harmless internal sources on the detector measurements are known, then the mean for each detector $\overline{x}_i$ can be used instead of the average for a given height $\overline{x}_h$ to improve the classification accuracy. Unfortunately, this requires more accurate information about a particular cargo configuration than we may commonly expect. This consequence will be discussed further in Ch. VI. For the more homogeneous cargo scenarios discussed in this study, the height-dependent z-score normalization works most consistently and will be used in the rest of this study.

CHAPTER VI

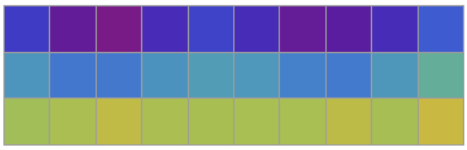
METHODOLOGY SENSITIVITY TO PHYSICAL VARIATIONS

Now that we have examined the development of methods for classification, it is natural to ask how sensitive these methods are to physical variations such as source size, source position, and cargo loading.

VI.1 Variations in Source Size

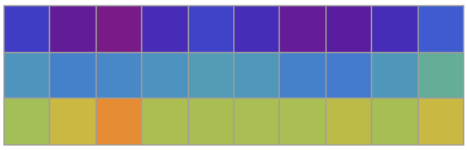
In our first study, we will examine the effects of source strength on the efficacy of the algorithms previously discussed. Let us consider the case where the cargo loading, L1, and source position, S1, are constant and the source strength is the only thing that varies, as depicted in Fig. VI.1. We first need to determine if the feature selection method will choose the same detectors in each case. Using the F-score method in Table VI.1, which was the most flexible of the feature selection methods discussed in Ch. V, we can see that Detectors 23, 22, 13 and 12 are consistently chosen as the most disparate features in the 30Det set. If one considers other loading scenarios, it is important for this particular study that the only parameter that differs between the safe and dangerous containers is the source size. Therefore, harmless internal sources, such as the concrete blocks from the L3 scenario, are present in both the safe and dangerous example measurements, so the efficacy of the feature selection method extends to these other scenarios in a similar fashion to the L1 scenario.

Next, we will consider the performance of classifiers for three source strengths through the development of the classification algorithm with information concerning only the 1 kg source and then testing its effectiveness in the classification of the other source strengths. As one can see from Fig. VI.2a, the SVM method provides a good approximation of the optimal decision rule for any source strength, above that of the Box Method. Furthermore, one can see that using the 1 kg rule with a fixed false positive rate, any of the sources are detected more effectively than randomly guessing, but it is unlikely that a half strength source will be detected. As there are a far greater numbers of harmless cargo containers scanned each year, controlling the false positive rate may prove desirable. Unfortunately, for the lowest false positive rates, most of the sources of size 1 kg or less will pass through undetected. However, this study is somewhat misleading, since we were considering methods developed by constraining the false negative rates in order to make statements about the reliability of detection of smuggled sources. Using the developed rule with a fixed



(a) Background

(b) Half Strength 1 kg HEU source



(c) 1 kg HEU source

(d) Double Strength 1 kg HEU source

Figure VI.1: Using the 30Det (Sec. II.3.1) scenario with the L1 loading and S1 source position, we have varied the source size from half the strength of a 1 kg source to twice the strength. As one can see, the relevant detectors do not change among the source strength variations, but the mean count rates are noticeably different. Purple detectors indicate low count rates and the highest average count rates are indicated in red.

Table VI.1: F-score Test for comparative source strengths of 0.5, 1, and 2 times a 1 kg source. For each case, detectors 23, 22, 13 and 12 are consistently an order of magnitude larger than any of the other detectors.

(a) Half Strength Source

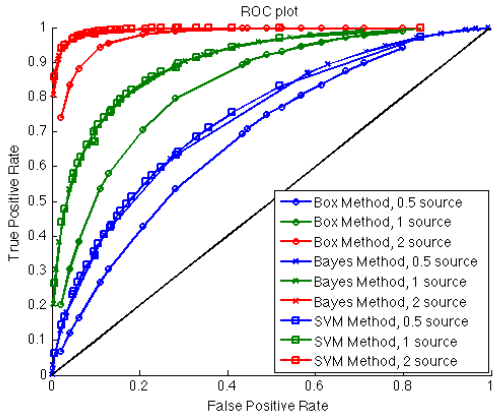
3.2e-07	3.1e-05	3.8e-06	1.7e-06	3.2e-05	4.4e-05	5.5e-07	1.9e-05	2.4e-05	4.0e-08
1.1e-06	1.1e-04	2.1e-04	2.6e-05	7.9e-08	4.5e-07	4.5e-06	3.3e-06	3.6e-06	4.3e-07
3.1e-06	3.4e-04	1.7e-03	5.6e-06	1.7e-05	4.0e-06	5.7e-06	7.9e-08	3.5e-05	5.6e-07

(b) Full Strength Source

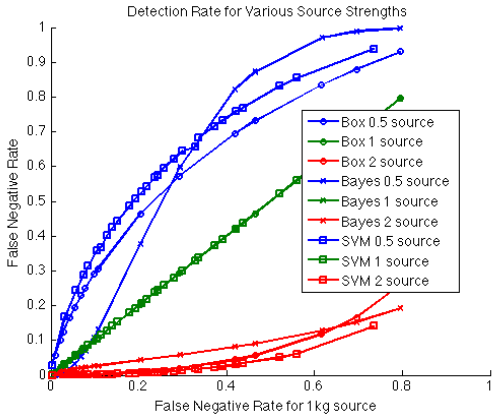
5.6e-09	6.0e-05	3.4e-08	2.6e-09	1.3e-05	1.4e-06	2.9e-08	3.6e-08	1.7e-06	6.8e-06
5.5e-08	1.8e-04	5.9e-04	2.2e-05	2.4e-05	2.2e-08	1.2e-05	8.0e-06	2.6e-06	3.5e-08
4.5e-06	2.1e-03	5.9e-03	2.8e-06	2.0e-06	1.7e-06	1.5e-07	7.4e-06	1.7e-05	1.5e-06

(c) Double Strength Source

5.2e-06	1.1e-06	1.4e-06	1.0e-06	3.6e-05	3.3e-07	2.6e-05	1.0e-07	1.3e-07	3.6e-06
1.4e-05	1.1e-03	2.7e-03	3.6e-06	3.7e-06	1.4e-05	1.5e-05	2.1e-05	1.2e-05	2.5e-06
3.8e-06	5.5e-03	2.0e-02	1.3e-05	1.3e-06	7.2e-07	1.3e-06	3.4e-06	7.0e-08	7.4e-07



(a) ROC Plot for varying source strengths

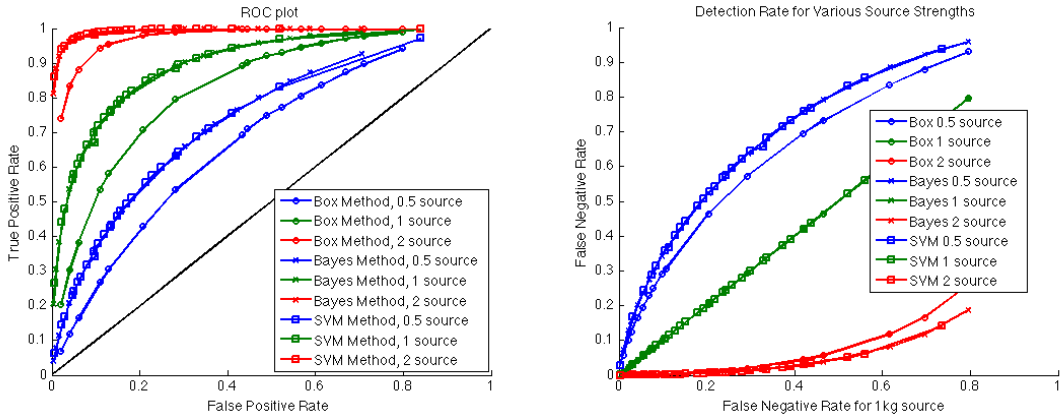


(b) Detection Rate for varying source strengths

Figure VI.2: Using Detectors 23 and 22, as designated by the F-score method for all source strengths, a classification rule using information about only the 1 kg source was developed with each of the three methods – Box, Bayes' Optimal and SVM. Blue curves represent the effectiveness for the half strength source, green for the 1 kg source, and red for the double strength source. Circles show rates developed using the Box Method, $\times$'s correspond to the Bayes' Optimal Method, and squares the SVM method. Here, the Bayes' Optimal Rule defined as a test on the likelihood ratio $p(x|S)/p(x|D_i)$ depends on the source size as the rule assumes exact knowledge about the distribution $p(x|D_i)$ of measurements for each source size D_i. For a more complete interpretation of this information, please see the main text of this section.

detection rate (false negative rate), the double strength source is regularly detected by all methods, as demonstrated in Fig. VI.2b. However, as expected, the half strength source, though still detectable, is almost twice as likely as the 1 kg source to be misclassified and escape detection.

One should note that there are two ways in which one could consider the Bayes' Decision Rule for the 0.5 kg and 2 kg in this situation. First, as shown in VI.2, one could assume that one has exact knowledge of the distributions of measurements for all source sizes, in which case these can be used in the calculation of the Bayes' Decision Rule. Thus, for a fixed likelihood ratio η generate by a specified rate of detection for a 1 kg source, both the false positive and false negative rates will vary. This does not occur in the evaluation of other methods, which have a constant false positive rate. It does however give the optimal choice of decision rule for a source of the specified size. In essence, this changing distribution type actually changes the nature of the classification rule. This brings us to the second option for the Bayes' Decision Rule, which uses the same probability distributions for classification of all source sizes, as given in Fig. VI.3. This ensures that the major characteristics of the



(a) ROC Plot for varying source strengths

(b) Detection Rate for varying source strengths

Figure VI.3: In a similar fashion to Fig. VI.2, each of the classification methods is tested for its effectiveness in detecting different size sources while ensuring that the decision rules are unchanged. Here, the Bayes' Optimal Rule defined as a test on the likelihood ratio $p(x|S)/p(x|D)$ is fixed as the rule assumes each dangerous point follows the distribution $p(x|D)$ for a 1 kg source sample. For a more complete interpretation of this information, please see the main text of this section.

classification rule do not change and gives a more realistic picture of the classification of these alternative source sizes. With this constancy in all of the decision rules, the Bayes' Optimal Decision Rule and the SVM solution are in agreement for any fixed false negative rate. As shown in Fig. VI.3, rules developed for a 1 kg source will reliably detect the double strength source. The half strength source is far less likely to be detected than the 1 kg source, but the decision rule produced by the 1 kg source does work better than randomly guessing at the containers.

Instead of developing the classification rule from data about the single source, one could use example measurements from multiple source sizes in order to create the decision rule. In this case, the detection rule is most heavily influenced by points from the smallest source, as they are the most similar to those of a safe container and, thus, most likely to be misclassified. This suggests that the signal to noise ratio, or rather the separation between the examples of safe and dangerous measurements, plays an important role in our ability to accurately detect the smuggled material.

(a) L1, background only

(b) L1, S1 source

(c) L1, S2 source

Figure VI.4: Using the 30Det scenario with the L1 and source positions S1 and S2, one can see the effects of the change in source position on the average detector count rates. Purple and blue detectors indicate low count rates and the count rate increases across the color spectrum until the highest average count rates are indicated in red. It should be noted that the detectors in the first five columns come from one side of the container and the last five come from the opposite side.

VI.2 Source Position Variations

We can also study the susceptibility of our algorithm to a change in the source position. While maintaining the cargo loading and source size constant, we can change the position of the source from S1 to S2, which corresponds to a change in the height of the radiating object from the floor of the container. Furthermore, it changes the type of material surrounding the HEU source from plastic with a density of 0.99 g/cm^3 to wood of 0.5 g/cm^3. This change in density of the surrounding material allows significantly more radiation to escape the container, making it far easier for the classification process to occur correctly. As shown in Fig. VI.4, the source position definitely affects which features are most useful in the course of classification, with 23 being most useful in the detection of the S1 source and detector 13 most useful for the S2 source. In this case, the F-score feature selection method accurately locates the relevant detectors for each position type, as shown in Table VI.2.

Application of any of the normalization methods suggested in Sec. V.2 does not affect the selection of the relevant features either for this L1 loading scenario. However, a careful choice of normalization can be beneficial to the classification process, irrespective of the method used. Let us additionally use the F-score as a ranking method and compare detectors of similar rank, i.e., compare measurements from detector 23 on an S1 source with those from detector 3 on an S2 source. Then, one can use the z-score to normalize each

Table VI.2: Testing the sensitivity of feature selection to source position. The F-score method correctly identifies detectors 23, 22, 13 and 12 as being the most useful for classification with an S1 source, provided that multiple examples of the dangerous measurements are used. For the S2 source location, detectors 3, 13, 12 and 2 as most useful. The S2 source requires far fewer example measurements for the correct choices since (1) it is surrounded by less dense material which allows a greater portion of the radiation to escape the container and (2) the location is farther from the background source making the signal to noise ratio lower.

(a) F-scores for S1 Source

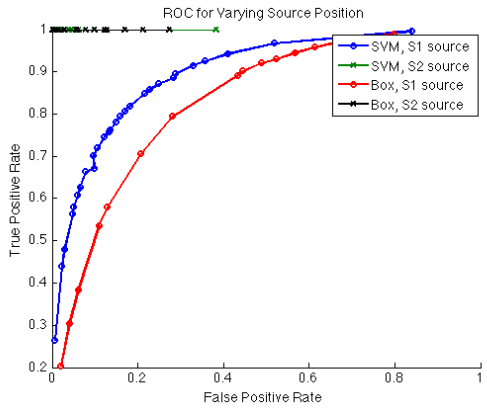
5.6e-09	6.0e-05	3.4e-08	2.6e-09	1.3e-05	1.4e-06	2.9e-08	3.6e-08	1.7e-06	6.8e-06
5.5e-08	1.8e-04	5.9e-04	2.2e-05	2.4e-05	2.2e-08	1.2e-05	8.0e-06	2.6e-06	3.5e-08
4.5e-06	2.1e-03	5.9e-03	2.8e-06	2.0e-06	1.7e-06	1.5e-07	7.4e-06	1.7e-05	1.5e-06

(b) F-scores for S2 Source

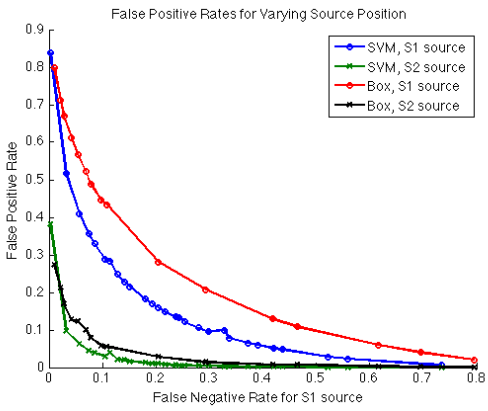
3.2e-03	1.2e-01	2.3e-01	3.3e-03	4.4e-04	1.3e-03	2.1e-04	7.1e-04	1.3e-03	1.8e-04
1.5e-03	1.3e-01	2.3e-01	4.4e-04	3.5e-05	9.3e-04	2.8e-07	1.2e-05	3.3e-05	8.5e-07
5.9e-06	5.1e-05	5.7e-04	2.1e-06	4.2e-05	2.6e-05	1.7e-04	7.0e-06	2.0e-05	2.6e-05

measurement according to the average count rate of detectors of similar height, $\overline{x}_h$, and use the classification rule developed with an S1 source to attempt detection of the S2 source, as depicted in Fig. VI.5. As one can see, this decision rule is an almost perfect classifier for the S2 source even though no knowledge of this particular source was used in the development. However, this classification rule does not provide a constant false positive rate for all source positions. Since we are normalizing detectors by the average count rate for a group of detectors instead of each detector individually, the distribution of measurements for individual detectors will not be perfectly centered on the origin. Therefore, if a classification rule is developed using detectors with above average background radiation for their height (i.e., $x_i > \overline{x}_h$), then other detectors will have a lower false positive rate than predicted in the development stages. On the other hand, false positive rates could be higher than usual if $x_i < \overline{x}_h$ in the development stage. The choice of normalization helps us avoid overfitting the method to a particular container type, but at the same time it loses some of the consistency that normalizing to measurements from each detector individually.

Unfortunately, this choice of normalization may not allow us to perform the same sort of near perfect classification of sources if harmless internal sources are present in the cargo container. In the L1 scenario, the average count rate for each detector of fixed height was nearly constant. For the L3 and L4 cargo loadings, the presence of natural internal sources causes variations among the average count rate, as shown in Figs. VI.6c and VI.6e. As before, we should note that the change in source position for these two scenarios again causes



(a) ROC Plot for varying source strengths



(b) False Positive Rates for Decision Rules developed with an S1 source of specified False Negative rate

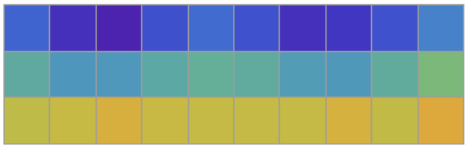
Figure VI.5: Using normalization by the z-score method with the mean being a function of height $\widetilde{x}_i = (x_i - \overline{x}_h) / \sqrt{\overline{x}_h}$, classification rules developed for the source position S1 can be used to classify the measurements from an S2 source in the case of an L1 loading. As we can see in Fig. VI.5a, the S2 source is almost perfectly classified by any of the decision rules developed for the S1 source. Fig. VI.5b allows us to analyze the effectiveness of the normalization scheme. We notice that for the S2 source position, the safe distribution means are slightly lower than average for detectors of that height, which reduces the false positive rate when classifying sources in this position. This may not always be the case, especially if the classification rule is developed for detectors with below average background radiation for their height. For more information, see the main text of the section.

a change in the density of the material surrounding the S2 position. If one normalizes using height averages as in the L1 scenario, then the means of the normalized safe measurement distributions are not as near to 0 as before and classification of the S2 source for the L3 loading is not as good as in the L1 scenario. We will discuss this problem further in the next section.

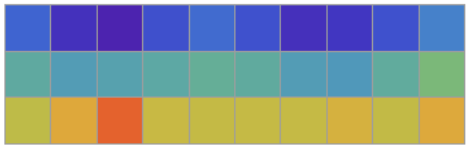
VI.3 Cargo Loading Sensitivity

Our final sensitivity study will analyze the effects of the cargo content of the container. As can be seen in Fig. VI.6, modifications in the cargo can cause significant variations in the mean count rates for each container. Each of the classification, feature selection and normalization methods themselves are ignorant of the cargo loading, detecting only deviations from the normal – in particular, once a decision rule is developed, classification of any provided point will proceed under the assumption that enough information was provided to do so accurately. Therefore, if one develops the classification, feature selection and normalization rules with only information about the radiation emanating from the L1 loading, then all future comparisons will be made with this basis for the radiation levels one would expect to see from either a safe or dangerous container.

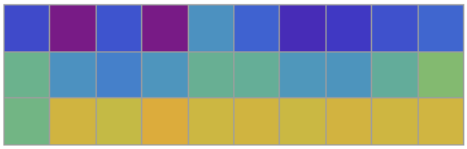
Both the feature selection and the normalization will emphasize deviations from the considered normal and any such deviations from this will show as a potential source. Let us first suppose that the only examples of safe containers used during the development of the classification methods are of L1 type (Fig. VI.6a). Then, when one attempts to compare a container of L3 type without any internal source (Fig. VI.6c), the F-score method designates detectors 4, 21, 2 and 3 as having the most deviation between the two container instances. We are fortunate here that the mean count rates for the L3 no source case are generally lower than or comparable to those of the L1 scenario, resulting in a lower false positive rate. Unfortunately, as a result of the lower background radiation from the L3 cargo, the average count rate for detectors near an HEU source are less than we would expect and more importantly they deviate less from the average than other detectors with depressed background (like detector 21). Thus, the F-score method as it stands does not make a distinction between measurements from an L3 container with and without a smuggled source in comparison to the L1 scenario, selecting detectors 4, 21, 2 and 3 in both cases. If, however, we add the additional requirement that not only must the F-score be large, but that the mean of the suspected dangerous container be larger than that of the means of the safe container for each detector, then this mitigates the problem somewhat. This modified F-score method selects detectors 3, 22, 5 and 23 when an HEU source is present and 3, 5, 24 and 11 otherwise.



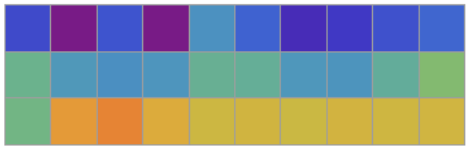
(a) L1, background



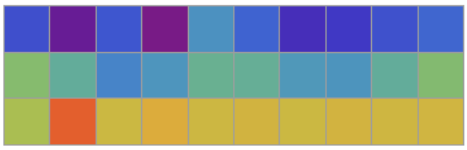
(b) L1, S1 source



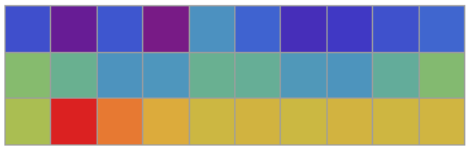
(c) L3 no internal source, background



(d) L3 no internal source, S1 source

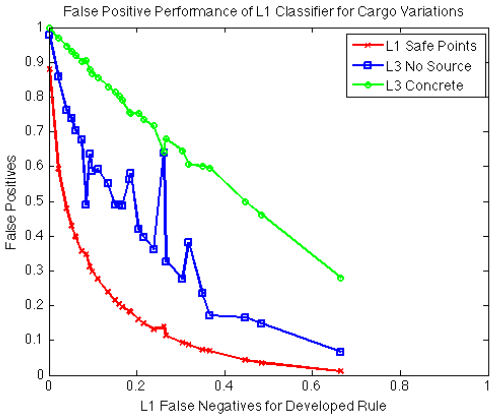


(e) L3 concrete internal source, background

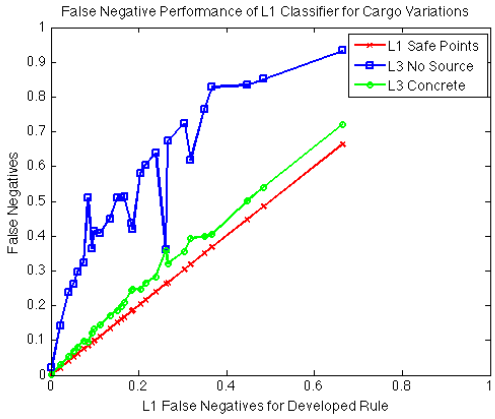


(f) L3 concrete internal source, S1 source

Figure VI.6: Using the 30Det scenario with the L1 and L3 loadings and S1 source position, one can see that a choice of background will matter in the feature selection methods. In each of these figures, the same color scale is used to denote the photon count rates. Comparing the background and source for a single cargo container, one can accurately determine the most useful detectors for classification (Detectors 23, 22, 13, and 12). However, if one were to compare the various backgrounds in Figs. VI.6a, VI.6c, and VI.6e, it is possible to locate false positives based on the variations in harmless internal sources and cargo density. Purple and blue detectors indicate low count rates and the count rate increases across the color spectrum until the highest average count rates are indicated in red.



(a) False Positives for background only



(b) False Negatives for S1 source position

Figure VI.7: The starting background assumptions play a major role in the accuracy of classification as well. If one develops a classification methodology (here an SVM classifier, with feature selection and count rate minimum, normalization by z-score with height dependent means) with evidence of only one container type (L1 in these figures), then all decisions made within this framework are in fact measuring deviations from the considered normal. For a fixed false negative rate used in the development of the classification rule, the rate of both error types has increased as a result of changing container types in the implementation phase. This suggests that the classification has been overfitted to one particular cargo scenario. For a more complete interpretation of this information, see the main text of this section.

As one can see from Fig. VI.7, both the rate of false positives and false negatives from the L3 cargo loading are larger than those of the L1 container type utilizing either classification method, although only the SVM is pictured here. However, this is almost certainly a case of overfitting the feature selection method, as emphasized by the non-smooth variation of the L3 error rates in comparison to the L1 error rates. If the roles of the cargo distributions had been reversed and we were measuring deviations from the L3 scenario, it is likely that false positives will be produced due to the differences in measurements of detectors 4 and 21, as identified by the F-score test. These detectors have larger means in the L1 background scenario than the L3 scenario, but they are still examples of harmless cargo and should not be flagged for further study.

If we add a single internal concrete source adjacent to detectors 21 and 22 in the L3 scenario, then this again changes the likelihood of accurate classification. For instance, the F-score test designates detectors 4, 22, 3 and 2 as useful in classification, of which detectors 22 and 3 have higher than normal count rates. Therefore, when applying the classification rule, the chance of a false positive is higher than predicted in the development of the classification

rule, as seen in Fig. VI.7. Also, the elevation in background radiation caused by the internal concrete source increases the likelihood of correctly identifying the smuggled HEU source. This is a misleading identification though, since the F-score method identifies detector 22 as most useful in classification rather than detector 23 as one would expect. The increase in radiation observed by detector 22 is caused in large part by the internal concrete source and the smuggled source presents a much smaller portion of the signal.

Even from these few examples, it is clear that the self-shielding of the cargo within the container as well as the presence of harmless internal sources plays a significant role in the performance of any classification method. The accuracy of our classification methods is tied to what we consider normal levels of radiation and the deviations from this standard. Thus, we can conclude that in order to gain the improvement in performance shown by the support vector machine methods, one must have accurate information about the expected levels of radiation and variations in the cargo.

CHAPTER VII

CONCLUSIONS

The classification problem for the detection of smuggled HEU has many facets that must be addressed – the development of an optimal classification rule that balances the costs and risks associated with a decision rule, the selection of features containing the most information about the source, normalization of data so that measurements for each feature are comparable, and the treatment of physical changes in the additional cargo and their effect on the classification process. The current Box Methods ignore correlations in measurements that can improve classification by decreasing the likelihood of a false positive by an order of magnitude, under the right conditions. Since far more harmless containers pass through ports each year, reductions in the false positives produced by detection algorithms can mean real reductions in monetary costs.

VII.1 Summary of Results

In this study, we have developed a cost-sensitive SVM framework with F-Score feature selection and z-score normalization by height that can achieve nearly optimal classification. It allows the researcher to control the reliability of detection of our algorithm (the expected false negative rate) without requiring exact knowledge of the cost of various outcomes or analytic information about the distribution of measurements, as required by the Bayes' Optimal Decision Rule. The SVM method developed here is insensitive to source position and size, provided that the background radiation and additional cargo meet certain conditions. First, the background radiation must be fairly regular to ensure that the signal from the source is not obscured by fluctuations in the background. In our study, we used one of the simplest possible background characterizations – a single concrete slab. It is well established that there are other sources of naturally occurring radioactive material and their presence may vary between screening sights at different ports. Each of the classification and feature selection methods are really measuring deviations from what the researcher designates as expected levels of radiation. Thus, accurate characterization of the background levels are absolutely necessary. Any variations in background may require modifications to the normalization method to maintain the effectiveness of the algorithms. In a similar fashion, the cargo in the container with the potential smuggled source can not have any large variations in density or unknown, harmless internal sources. As we discussed in Ch. VI, such variations cause too large of a deviation for a single classification rule to encompass all of the variations. Finally, for the decision rule to work effectively

for all source positions, it must be developed using the "worst case scenario" – a minimal source strength placed in a position as far from detectors as possible with large background contribution and strong shielding. These restrictions influence the minimal signal to noise ratio that will be acceptable for accurate classification.

VII.2 Possible Future Improvements

As with most projects, investigating the intricacies of the source detection problem has only spawned more ideas with which to attack the problem. For instance, it is clear that utilizing correlations in measurements is a key feature to improve the likelihood of detection. There are several signal separation techniques used in pattern recognition that may prove useful in locating features that are influenced most heavily by a point source of smuggled material. It would also be interesting to investigate the use of radiographs (x-rays) to estimate the likely fluctuations in background as a result of density variations. As noted in Ch. VI, these fluctuations have a large influence on the effectiveness of the algorithms and are difficult to characterize when classification methods only measure deviations from the average. Finally, the support vector machine framework could be expanded to a real time learning algorithm that would be able to account more easily for small variations in background radiation that can occur daily.

REFERENCES

[1] M. Allmaras, D. P. Darrow, Y. Hristova, G. Kanschat and P. Kuchment, *Detecting Small Low Emission Radiating Sources*, Inverse Problems and Imaging, 7 (2013), pp. 47–79.

[2] P. Bermejo, L. de la Ossa, J. A. Gàmez and J. M. Puerta, *Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking*, Knowledge-Based Systems, 25 (2012), pp. 35–44.

[3] A. L. Blum and P. Langley, *Selection of relevant features and examples in machine learning*, Artificial Intelligence, 97 (1997), pp. 245–271.

[4] L. Bruzzone and C. Persello, *A Novel Approach to the Selection of Spatially Invariant Features for the Classification of Hyperspectral Images With Improved Generalization Capability*, IEEE Trans. Geoscience and Remote Sensing, 47 (2009), pp. 3180–3191.

[5] G. Casella and R. L. Berger, *Statistical Inference*, Thomson Learning, Inc., Pacific Grove, CA, 2 ed., 2002.

[6] C.-S. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, 2 (2011), pp. 27:1–27:27.

[7] H.-G. Chew, R. E. Bogner and C.-C. Lim, *Dual-ν support vector machine with error rate and training size biasing*, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001, Proceedings (ICASSP '01), vol. 2, 2001, pp. 1269–1272.

[8] T. W. Chow and D. Huang, *Estimating Optimal Feature Subsets Using Efficient Estimation of High-Dimensional Mutual Information*, IEEE Trans. Neural Networks, 16 (2005), pp. 213–224.

[9] C. Cortes and V. Vapnik, *Support-Vector Networks*, Machine Learning, 20 (1995), pp. 273–297.

[10] M. A. Davenport, *Neyman-Pearson SVMs: Controlling Error Rates with Cost-Sensitive Support Vector Machines*, `http://www.ece.rice.edu/~md/np_svm.php`, 2012.

[11] M. A. Davenport, R. G. Baraniuk and C. D. Scott, *Controlling False Alarms with Support Vector Machines*, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006, Proceedings (ICASSP '06), vol. V, 2006, pp. 589–592.

[12] M. A. Davenport, R. G. Baraniuk and C. D. Scott, *Tuning Support Vector Machines for Minimax and Neyman-Pearson Classification*, IEEE Trans. Pattern Analysis and Machine Intelligence, 32 (2010), pp. 1888–1898.

[13] W. Duch, *Filter Methods*, Feature Extraction, Foundations and Applications, ed. by I. Guyon, M. Nikravesh, S. R. Gunn and L. A. Zadeh, vol. 207 of Studies in Fuzziness and Soft Computing, Springer, New York, NY, 2006, ch. 3, pp. 89–117.

[14] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, John Wiley and Sons, Inc., New York, NY, 2 ed., 2001.

[15] J. Ely, R. Kouzes, J. Schweppe, E. Siciliano, D. Strachan, *et al.*, *The use of energy windowing to discriminate SNM from NORM in radiation portal monitors*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 560 (2006), pp. 373–387.

[16] T. Fawcett, *An introduction to ROC analysis*, Pattern Recognition Letters, 27 (2006), pp. 861–874.

[17] E. Grall-Maës and P. Beauseroy, *Optimal Decision Rule with Class-Selective Rejection and Performance Constraints*, IEEE Trans. Pattern Analysis and Machine Intelligence, 31 (2009), pp. 2073–2082.

[18] J. Guo, *Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis*, Biostatistics, 11 (2010), pp. 599–608.

[19] I. Guyon and A. Elisseeff, *An Introduction to Variable and Feature Selection*, J. Mach. Learn. Res. 3 (2003), pp. 1157–1182.

[20] J. Haslinger and R. Mäkinen, *Introduction to shape optimization: theory, approximation, and computation*, Advances in Design and Control, Society for Industrial and Applied Mathematics, Philidelphia, PA, 2003.

[21] C.-C. Hsu, K.-S. Wang and S.-H. Chang, *Bayesian decision theory for support vector machines: Imbalance measurement and feature optimization*, Expert Systems with Applications, 38 (2011), pp. 4698–4704.

[22] IAEA Incident and Trafficking Database (ITDB), *Incidents of nuclear and other radioactive material out of regulatory control: 2013 Fact Sheet*, `http://www-ns.iaea.org/downloads/security/itdb-fact-sheet.pdf`, 2013.

[23] International Atomic Energy Agency, *IAEA Safegaurds Glossary*, vol. 3, IAEA International Nuclear Verification Series, IAEA, Vienna, 2001.

[24] International Atomic Energy Agency, *Detection of radioactive materials at borders*, IAEA-TECDOC 1312, International Atomic Energy Agency, Vienna, Sept. 2002.

[25] International Atomic Energy Agency, *Combating Illicit Trafficking in Nuclear and Other Radioactive Material*, vol. 6, IAEA Nuclear Security Series, 2007.

[26] A. G. Janecek, W. N. Gansterer, M. A. Demel and G. F. Ecker, *On the Relationship Between Feature Selection and Classification Accuracy*, J. Mach. Learn. Res.: Workshop and Conference Proceedings, 4 (2008), pp. 90–105.

[27] K. Jarman, C. Scherrer, L. Smith, L. Chilton, K. Anderson, *et al.*, *Indirect estimation of radioactivity in containerized cargo*, Radiation Measurements, 46 (2011), pp. 10–20.

[28] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Pearson Prentice Hall, Upper Saddle River, NJ, 6 ed., 2007.

[29] R. D. Knabb, J. R. Rhome and D. P. Brown, *Tropical Cyclone Report: Hurricane Katrina: 23-30 August 2005*, `http://www.nhc.noaa.gov/pdf/TCR-AL122005_Katrina.pdf`, Sept. 2011.

[30] R. Kouzes, *Challenges for interdiction of nuclear threats at borders*, First International Conference on Advancements in Nuclear Instrumentation Measurement Methods and their Applications (ANIMMA), June 2009, pp. 1–3.

[31] R. Kouzes, *Neutron and gamma ray detection for border security applications*, First International Nuclear & Renewable Energy Conference (INREC), 2010, Mar. 2010, pp. 1–3.

[32] M.-L. T. Lee and G. A. Whitmore, *Intensity-Dependent Normalization in Microarray Analysis: A Note of Concern*, Bernoulli, 10 (2004), pp. 943–949.

[33] Y. Lin, *Support Vector Machines and the Bayes Rule in Classification*, Data Mining and Knowledge Discovery, 6 (2002), pp. 259–275.

[34] Y. Lin, Y. Lee and G. Wahba, *Support Vector Machines for Classification in Nonstandard Situations*, Machine Learning, 46 (2002), pp. 191–202.

[35] MATLAB, *version 7.12.0635 (R2011a)*, The MathWorks Inc., Natick, MA, 2011.

[36] J. Meza, P. Hough, P. Williams and R. Oliva, *Opt++: An Object-Oriented Nonlinear Optimization Library (version 2.4)*, Software: `https://software.sandia.gov/opt++/`, Sandia National Laboratory, Albuquerque, NM, 2007.

[37] D. Miller, A. V. Rao, K. Rose and A. Gersho, *A Global Optimization Technique for Statistical Classifier Design*, IEEE Trans. Signal Process. 44 (1996), pp. 3108–3122.

[38] T. Munson, J. Sarich, S. Wild, S. Benson and L. C. McInnes, *TAO 2.0 Users Manual*, tech. rep. ANL/MCS-TM-322, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 2012.

[39] J. Neyman and E. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypothesis*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 231 (1933), pp. 289–337.

[40] R. Nilsson, J. M. Peña, J. Björkegren and J. Tegnèr, *Consistent Feature Selection for Pattern Recognition in Polynomial Time*, J. Mach. Learn. Res. 8 (2007), pp. 589–612.

[41] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, NY, 2nd ed., 2006.

[42] H. Peng, F. Long and C. Ding, *Feature selection based on mutual infromation criteria of max-dependency, max-relevance, and min-redundancy*, IEEE Trans. Pattern Analysis and Machine Intelligence, 27 (2005), pp. 1226–1238.

[43] A. Rakotomamonjy, *Variable Selection Using SVM-based Criterion*, J. Mach. Learn. Res. 3 (2003), pp. 1357–1370.

[44] D. Reilly, N. Ensslin, H. Smith Jr. and S. Kreiner, *Passive Nondestructive Assay of Nuclear Materials*, URL: `http://www.lanl.gov/orgs/n/n1/panda/`, LA-UR-90-732, Los Alamos National Laboratory, Los Alamos, NM, Mar. 1991.

[45] K. Rose, *Deterministic Annealing for Clustering, Compression, Classification, Regression and Related Optimization Problems*, Proceedings of the IEEE, 86 (1998), pp. 2210–2239.

[46] C. Sakar and O. Kursun, *A method for combining mutual information and canonical correlation analysis: Predictive Mutual Information and its use in feature selection*, Expert Systems with Applications, 39 (2012), pp. 3333–3344.

[47] B. Scholköpf, A. J. Smola, R. C. Williamson and P. L. Bartlett, *New Support Vector Algorithms*, Neural Computation, 12 (2000), pp. 1207–1245.

[48] A. Sonzogni, *Interactive Chart of Nuclides*, Software: `http://www.nndc.bnl.gov/chart`, National Nuclear Data Center, Brookhaven National Laboratory, Upton, NY, Jan. 2012.

[49] S. Sra, S. Nowozin and S. J. Wright, Eds., *Optimization for Machine Learning*, MIT Press, Cambridge, MA, 2012.

[50] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, New York, NY, 2008.

[51] J. D. Storey, *The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing*, J.R. Statis. Soc. B, 69 (2007), pp. 347–368.

[52] F. Tortorella, *A ROC-based reject rule for dichotomizers*, Pattern Recognition Letters, 26 (2005), pp. 167–180.

[53] U.S. Department of Homeland Security: Customs and Border Patrol, *How Cargo Flows Securely to the U.S.*, http://www.cbp.gov/linkhandler/cgov/trade/cargo_security/cargo_control/cargo_flow_map.ctt/cargo_flow_map.pdf, 2008.

[54] U.S. Department of Transportation, Research and Innovative Technology Administration, Bureau of Transportation Statistics, *America's Container Ports: Linking Markets at Home and Abroad.* http://www.bts.gov/publications/americas_container_ports/2011/html/long_term_trends.html, Washington, DC., 2011.

[55] U.S. Senate and House of Representatives, *Implementing Recommendations of the 9/11 Commission Act of 2007*, Public Law 110-53, http://www.gpo.gov/fdsys/pkg/PLAW-110publ53/pdf/PLAW-110publ53.pdf, Washington, DC., Aug. 2007.

[56] X-5 Monte Carlo Team, *MCNP - A general Monte Carlo n-particle transport code, Version 5*, Report LA-UR-03-1987, Los Alamos National Laboratory, Los Alamos, NM, 2003.

APPENDIX A

DETAILS OF MCNP INPUT DECKS

The data used in this study was generated through extensive use of MCNP. Each testing set is given a 3 number designation, e.g. B1S1L1, which corresponds to the box drawing scenario, the source position and the material loading scenario, respectively. The meaning of each of these numbers is given in the following subsections and the results in following section.

A.1 Problem Geometry - Box Drawing Scenario

The 20 ft dry cargo container has dimensions 5.898 m by 2.352 m by 2.394 m in the MCNP input deck created by Dr. Sunil Chirayath. The box drawing scenario has 4 boxes in the x-direction with a width of 1.5235 m, 2 in the y-direction with a width of 1.217 m, and 4 in the z-direction with a width of 0.646625 m. The boxes are enumerated by rows in the z direction from the top of the cargo container to the bottom, as shown in Fig. A.1. The HEU source is in box 15 for the S1 source position and in box 11 for the S2 source.

A.2 Problem Materials - Loading Scenarios

In accordance with the MCNP modeling scenarios developed by Dr. Chirayath and the Smuggled HEU Interdiction through Enhanced anaLysis and Detectors (SHIELD) team, the highest density materials being nearest to the S1 source box (box 15) and the density decreases as the distance to the source increases. This is so that the greatest shielding covers most of the source and makes the detection problem harder. This principle will be used in loading the first box drawing scenario as well to continue to make the problem more difficult. The container has a weight limit of 21,630 kg, which causes the iron and concrete boxes to be reduced in either density or percentage of box filled. For consistency with the other portions of this project, the only materials that will be used to fill the boxes in this round of tests are wood, plastic, cotton, concrete, iron and potash. The density of these materials is given in Table A.1.

Potash is a fertilizer, usually labeled with the formula, NPK. However, many different fertilizers can be referred to by this label with chemical formulas including but not limited to Potassium Nitrate, Potassium Chloride, Potassium Sulfate and Potassium-Sodium Nitrate mixtures. Typical densities are in the range 1-3 g/cm^3. The exact formula and density may vary between companies. One of the more common potash fertilizers is sulfate of potash,

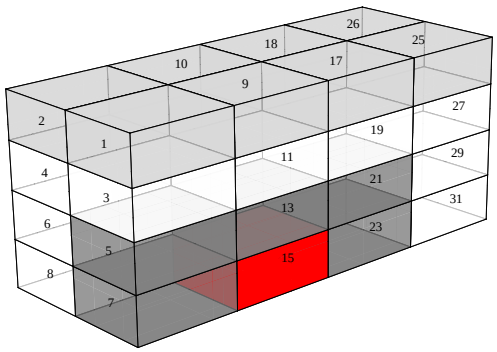


Figure A.1: Depiction of the box drawing scenario where the red box contains the source and density is given by darkness of the gray – the darker the box, the more dense the material.

Table A.1: Densities of materials used to fill containers for testing.

Material	Density (g/cm^3)
Wood	0.5
Plastic	0.91
Cotton	0.03
Concrete	2.35
Iron	7.8
Potash	2.6

K_2SO_4, which is used in this study.

There are 5 loading scenarios that are common to all sections of this project. The highest density materials are placed in the boxes closest to the source in box 15.

A.3 Problem Sources - Background, NORM and HEU sources

Two different source positions have been devised – one resting on the bottom of the cargo container towards the center of the container in the x-direction (S1) and the other in the middle of the container in the z-direction (S2). The S1 position is in loading box 15 and the S2 position is in loading box 27. These two positions are chosen for several reasons. First, they are towards the center of the container, which means that more material is between each detector and the source and each detector covers a smaller solid angle of radiation paths and, therefore, each detector sees less radiation from the source. Next, the placement along the bottom of the container hides the source radiation in the greater

Table A.2: Proportions of boxes containing the given materials in the considered test schemes. Each scenario has one box filled with the HEU and plastic.

Scenario No.	Plastic	Wood	Cotton	Iron	Concrete	Potash
L1	8	8	15	0	0	0
L2	7	7	15	2 (50% full)	0	0
L3	6	6	17	0	2 (100% full)	0
L4	6	6	17	0	2 (40% full)	0
L5	0	0	0	0	0	31

radiation coming from the background concrete. The higher source placement is present to verify that we are not over-compensating for the low source to signal ratio given by the first scenario. S0 will be the designation for MCNP runs with only an external concrete source. In the case where the background includes both the external concrete source and an internal concrete or potash source, there will be additional runs (designated S0a) which calculate the contribution from only the internal source. The total background source, in this case, will be calculated by adding the internal (S0a) and external (S0) contributions.

The background and norm sources that are present in our problem are concrete and potash. The active nuclide in both these materials is Potassium-40, which most commonly produces a 1.46 MeV photon, with an intensity of 10.67% production for every disintegration of the nuclide (numbers from [48] obtained on 2/1/12). To put this in an MCNP deck, we need to calculate the mass of Potassium in the given volume of the substance and use a distributive source. The number of photons can be given by:

$$\begin{aligned}
\frac{\text{1.46 MeV photons/sec}}{\text{g } K^{39}} &= \left(\text{abundance of } K^{40}\right)\left(\text{specific activity of } K^{40}\right) \qquad \text{(A.1)}\\
&\quad \times \left(\frac{\text{dps}}{\text{Ci}}\right)\left(\frac{\text{1.46 MeV photons}}{\text{disintegration}}\right)\\
&= \left(\frac{0.000117 \text{ g } K^{40}}{\text{g } K^{39}}\right)\left(\frac{7.1\times10^{-6} \text{ Ci}}{\text{g } K^{40}}\right) \qquad \text{(A.2)}\\
&\quad \times \left(\frac{3.7\times10^{10}\text{dps}}{\text{Ci}}\right)\left(\frac{0.1067 \text{ photons}}{\text{disintegration}}\right)\\
&= 3.2795 \ \frac{\text{1.4MeV photons/sec}}{\text{g } K^{39}} \qquad \text{(A.3)}
\end{aligned}$$

Using this fact and the composition of concrete and potash (K_2SO_4) found in the MCNP decks, one can calculate that, in the background concrete of volume 30 cm by 1524 cm by 1012 cm with density 2.4 g/cm^3, the total production rate of 1.46 MeV photons is 6.9×10^6

photons per second. Similarly, we can calculate the contributions of alternate internal sources, like potash and concrete, by first calculating the volume of the material in the interior. For example, a 100% filled box of concrete has a volume of 1.2×10^6 cm^3 and a 40% filled box has a volume of 4.8×10^5 cm^3. Using the same chemical composition of concrete as in the background source, we can see that each box percentage generates 1.46 MeV photons at a rate of 1.8×10^5 photons per second and 7.1×10^4 photons per second, respectively. If there are two such boxes in a given scenario, the total internal source is twice what is mentioned here.

In the potash scenario, one can calculate the internal source in a similar fashion. If we use K_2SO_4 and the density of 2.66 g/cm^3, we can calculate that one standard box produces 1.46 MeV photons at a rate of 4.621×10^6 photons per second. Only 14 of the 32 boxes can be filled at this density before the weight limit of the cargo container is reached. Potash varies greatly in density depending on type, country of origin and composition. This is the composition of muriate of potash, a common fertilizer type in the mid-20th century. Other compositions may be lighter and, thus, fill the containers more completely.

The HEU source was 1 kg of HEU (70 wt% ^{235}U and 30 wt% ^{238}U), which produces roughly 33 photons per gram per second (22 of which are of 1 MeV), as designated by Dr. Chirayath's initial input deck. This source description was unchanged to conform with the MCNP used by other portions of the DHS project.

A.4 Variance Reduction in the MCNP Runs

Due to the long length scales of this problem, importance weighting must be used in MCNP to counteract the loss of particles due to absorption. For the wood, plastic and cotton filled cargo containers, the importance must be multiplied by a power of two every time it crosses a boundary since roughly half of the particles are lost in each cell. This weighting must be changed for higher density materials. Note that the Uranium shells present in the original input file are much smaller than the material boxes and still have an importance weighting of powers of two.

The importance weighting alone is not enough to guarantee that the MCNP runs will converge and produce the average count rates for each detector. Therefore, we utilized the weight window generation feature of MCNP to help reduce the variation in measurements and aid the computational process [56, Vol. I, Ch. 2, Sec. 7].