Using Uncertainty To Interpret Supervised Machine Learning Predictions

Michael C. Darling

University of New Mexico

Follow this and additional works at: https://digitalrepository.unm.edu/ece_etds

Part of the Other Computer Engineering Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Engineering ETDs at UNM Digital Repository. It has been accepted for inclusion in Electrical and Computer Engineering ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact amywinter@unm.edu, Isloane@salud.unm.edu, sarahrk@unm.edu.
This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Don Hush, Chairperson
Using Uncertainty To Interpret Supervised Machine Learning Predictions

by

Michael C. Darling

B.S., Computer Engineering, University of New Mexico, 2012
M.S., Computer Engineering, University of New Mexico, 2015

DISSERTATION

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy Engineering

The University of New Mexico

Albuquerque, New Mexico

December, 2019
Using Uncertainty To Interpret Supervised Machine Learning Predictions

by

Michael C. Darling

B.S., Computer Engineering, University of New Mexico, 2012
M.S., Computer Engineering, University of New Mexico, 2015
Ph.D., Computer Engineering, University of New Mexico, 2019

Abstract

Traditionally, machine learning models are assessed using methods that estimate an average performance against samples drawn from a particular distribution. Examples include the use of cross-validation or hold-out to estimate classification error, F-score, precision, and recall.

While these measures provide valuable information, they do not tell us a model’s certainty relative to particular regions of the input space. Typically there are regions where the model can differentiate between the classes with certainty, and regions where the model is much less certain about its predictions.

In this dissertation we explore numerous approaches for quantifying uncertainty in the individual predictions made by supervised machine learning models. We develop an uncertainty measure we call minimum prediction deviation which can be used to assess the quality of the individual predictions made by supervised two-class classifiers. We show how minimum prediction deviation can be used to differenti-
ate between the samples that a model predicts credibly, and the samples for which further analysis is required.
Contents

List of Figures ix

1 Introduction 1

2 Uncertainty Quantification for Machine Learning 4

2.1 Two-class Supervised Classification Problem .................. 5

2.2 Uncertainty in a Two-class Classifier ......................... 7

2.2.1 Summary of Approach ...................................... 9

2.3 Sources of Uncertainty ....................................... 10

2.4 Uncertainty, Probability, and Error .......................... 12

2.5 Measures of Uncertainty and Their Limitations ............... 12

2.5.1 Confidence Intervals ...................................... 14

2.5.2 Standard Deviation ....................................... 15

2.5.3 Instability ................................................ 16

2.6 Approaches to Uncertainty .................................... 19
## Contents

2.7 Extracting Uncertainty From Machine Learning Pipeline .................................. 21

3 Minimum Prediction Deviation ................................................................................. 24

3.1 Notions of Uncertainty ......................................................................................... 24

3.1.1 Notion 1: Intrinsic Uncertainty ........................................................................ 24

3.1.2 Notion 2: Empirical Uncertainty .................................................................... 26

3.1.3 Notion 3: The significance of uncertainty dictated by the relative sparseness of the data .............................................................. 29

3.2 Data for Comparison .......................................................................................... 29

3.2.1 Synthetic Data .................................................................................................. 31

3.3 Measures for Comparison .................................................................................... 34

3.3.1 Standard Deviation ......................................................................................... 34

3.3.2 Covariate-dependent Confidence Intervals for Logistic Regression ............ 35

3.3.3 Empirical Uncertainty Deviation .................................................................. 37

3.3.4 Summary of Measures .................................................................................... 40

4 Comparison of Uncertainty Measures ................................................................... 41

4.1 Minimum Prediction Deviation with CART ....................................................... 43

4.1.1 I-I Data ......................................................................................................... 43

4.1.2 I-sI Data ....................................................................................................... 46

4.1.3 I-M Data ....................................................................................................... 48

4.2 MPD with Logistic Regression ........................................................................... 50
Contents

4.3 Consistency .................................................. 53
4.4 MPD vs. Standard Deviation ................................. 55
4.5 MPD vs. Confidence Intervals for Logistic Regression .... 58
4.6 MPD with Density Weighting ................................ 60
4.7 MPD in Many Dimensions .................................... 62

5 Minimum Prediction Deviation Applied to URL Analysis .... 66
  5.1 The Data .................................................. 67
  5.2 MPD Analysis ................................................. 68

6 Minimum Prediction Deviation Applied to Pixel Classification .. 73
  6.1 The Data .................................................. 74
  6.2 MPD Analysis ................................................. 75

7 Conclusion .................................................. 79
  7.1 Contributions ............................................... 79
  7.2 Discussion .................................................. 81
  7.3 Future Work ................................................. 83
List of Figures

1.1 Two Gaussian distributions ............................................. 2

2.1 The Machine Learning Lifecycle ................................. 7

2.2 Two classifiers constructed using same learning algorithm and probability distribution .............................. 8

2.3 Two realizations of a classifier possibly with differing predictions ............................................. 8

2.4 Distribution of predictions for n classifier realizations ..................................................... 9

2.5 The steps of the standard machine learning task and their associated sources of uncertainty ......................... 11

2.6 Example probability distributions ..................................... 14

2.7 Distribution over the probability that a URL is malicious ................................................... 15

2.8 Example Stratified Performance Plot ................................. 16

2.9 Two URL prediction distributions with almost the same means and standard deviations .......................... 17

2.10 Mean vs. Label Instability .................................................. 18

2.11 Two URL prediction distributions ...................................... 18
List of Figures

3.1 Example of continuous and sample distributions .................. 25
3.2 Example probability distributions ................................. 26
3.3 Example 2-dimensional data with two Gaussian distributions .... 30
3.4 Example Bayesian Classification. With and Without Uncertainty.. 31
3.5 Example I-sI data set ............................................. 32
3.6 Example I-M data set .............................................. 33
3.7 Examples of Empirical Classification and Uncertainty ............. 39
4.1 Plot of I-I test data with 1000 samples ............................ 43
4.2 3-dimensional Intrinsic and CART-MPD uncertainty plot for I-I test set. .................................................. 44
4.3 I-I test set thresholds plots for intrinsic and MPD-CART uncertainty 45
4.4 Plot of I-sI test data with 1000 samples ............................ 46
4.5 3-dimensional plots of Intrinsic and MPD-CART uncertainty for I-sI test set ..................................................................... 47
4.6 I-sI test set threshold plots for intrinsic and MPD-CART uncertainty 47
4.7 Plot of I-M 1000 sample I-M test set ................................. 48
4.8 Intrinsic and MPD with CART uncertainty plots for 500 sample I-M dataset. .................................................. 49
4.9 I-M data with 500 samples with intrinsic and empirical uncertainty. 50
4.10 3-dimensional uncertainty plots for MPD with LR with I-I, I-M, and I-sI test sets .................................................. 51
List of Figures

4.11 MPD-LR threshold plots for I-I, I-M, and I-sI test sets ... 52
4.12 MPD uncertainty with LR for increasing numbers of samples ... 53
4.13 MPD uncertainty with LR threshold plots for increasing numbers of samples ... 54
4.14 Intrinsic, MPD, and Standard deviation uncertainty heatmaps for I-sI dataset classified with CART ... 55
4.15 Standard deviation heat maps for CART with increasing numbers of samples ... 56
4.16 3-dimensional uncertainty plots for MPD and Standard Deviation with LR model ... 57
4.17 Threshold plots of I-sI data using Standard Deviation with CART and LR models ... 58
4.18 Confidence interval heat maps with LR ... 59
4.19 Confidence interval heat maps with LR ... 60
4.20 MPD-CART with Density Term ... 60
4.21 MPD-LR with Density Term ... 61
4.22 Intrinsic Uncertainty for I-I Data in 4-dimensions ... 63
4.23 Intrinsic Uncertainty for I-sI Data in 4-dimensions ... 64
4.24 Intrinsic Uncertainty for I-M Data in 4-dimensions ... 65

5.1 URL Components ... 68
5.2 Example URL distributions ... 69
List of Figures

5.4 URL Rejection Graph .................................. 70
5.5 Example malicious distributions for three URLs .......... 71
6.1 Multi-Source Imagery .................................. 75
6.2 Results of pixel classification ............................. 77
6.3 Stratified Plots for Pixel Analysis ......................... 78
Chapter 1

Introduction

When a machine learning model makes predictions on a set of data samples, we separate its results into two sets: right and wrong. With this information we are able to calculate the model’s accuracy on that data set, and a host of other measures such as f-score, precision, and recall.

Within the categories of right and wrong predictions, there is always some relative degree of uncertainty: there are samples which are “easy” predictions for the model, and samples for which it is not entirely certain. For example, Figure 1.1 shows two sets of data. If we train a model to find the differences between the red and blue dots, it will be easy for it to identify the samples on the extremes, where there is no overlap between the classes. However, it is obvious that, the region in the middle of the space is more problematic. This kind of overlap will exist in higher dimensional problems which cannot be visualized in the same manner as our simple example.

Assessing uncertainty increases our ability to interpret the validity of a model’s prediction. Though the notion of uncertainty is often conflated with probability, uncertainty and probability are distinct concepts. A probability estimation provides a sample’s relative fit to a label given a model; uncertainty defines the model’s
Chapter 1. Introduction

Figure 1.1: Data generated from two gaussian distributions.

credibility in assessing the sample. A prediction presenting high uncertainty (low model credibility) indicates that alternate, valid interpretations of the data exist and the degree to which the model can distinguish between them.

If we can identify the samples on which a model has certainty, we can use this information to inform subsequent decisions. For example, we could tune the model’s parameters to reduce uncertainty on the most critical class in a given problem. Or, after deployment, we could reject a model’s prediction if its uncertainty is too high.

Consider a classifier charged with protecting a network by identifying malicious websites. Regardless of the model’s overall performance, a single wrong prediction can result in infection or infiltration of the network. If we can assess the classifier’s uncertainty with respect to each prediction, we can retune the model to reduce uncertainty on the class of malicious websites. Another option, after deployment,
Chapter 1. Introduction

would be to determine when uncertainty is significant enough to reject the model’s prediction and send the website for a more computationally intense analysis or, perhaps, human verification.

In this study, we are not interested in developing learning algorithms for the Supervised Classification Problem. Instead we are interested in analyzing the characteristics of a learned classifier, and we wish to do this regardless of the learning algorithm employed. In particular we are interested in characterizing uncertainty in order to quantify the degree to which the end user can have confidence in the predictions made by a supervised machine learning model.

The work presented in this dissertation develops methods for quantifying and interpreting uncertainty in the probabilistic predictions made by two-class supervised classifiers. Specifically, we construct distributions over all possible probability estimates made by a learned classifier and develop a novel measure of uncertainty, which we call minimum prediction deviation (MPD), and compare it to existing measures. This uncertainty measure is covariate-dependent: it is a function of the covariate value $x$. This means that it can be used to identify regions of the input space where the uncertainty is low and predictions can be considered reliable.
Chapter 2

Uncertainty Quantification for Machine Learning

To frame the particulars of our study, before we introduce the new uncertainty measure in Chapter 3, this chapter provides background on fundamental concepts, related studies, and our previous work in interpreting uncertainty of machine learning models.

This project is part of a larger effort to quantify and leverage uncertainty in machine learning. Our guiding hypothesis is that a data-driven uncertainty analysis provides information that is not available from traditional machine learning evaluation methods, and is useful to both machine learning practitioners and decision makers.

The need for this type of work became apparent during attempts to implement large searchable semantic graphs based on remote sensor data. The researchers concluded that in order to fully leverage the objects and patterns found in the data, uncertainty had to be accounted for to fully characterize confidence in their results (Stracuzzi et al., 2015). Subsequent projects applied uncertainty analyses
2.1. TWO-CLASS SUPERVISED CLASSIFICATION PROBLEM

to a number of problem domains including malicious URL detection (Darling and Stracuzzi, 2018), multi-source image analysis (Stracuzzi et al., 2018b), and seismic onset detection (Vollmer et al., 2017; Stracuzzi et al., 2018a).

2.1 Two-class Supervised Classification Problem

We review the standard two-class supervised classification problem. Those familiar with the problem may wish to skip this section.

First we assume that the pattern formation process produces feature vectors \( x \) of the form \( x = (x_1, x_2, \ldots, x_d) \), where each \( x_i \) is a real-valued component and \( d \) is the number of features. In the two-class problem the class labels are denoted by the variable \( \omega \) which takes the values \( \omega_1 = 1 \) or \( \omega_2 = 2 \). Our probabilistic model assumes that labeled data points \((x, \omega)\) are generated as samples from probability functions \( P_1p(x|\omega_1) \) and \( P_2p(x|\omega_2) \) where \( p(x|\omega) \) is the density for class \( \omega \) and \( P_\omega \) is the (prior) class probability. The class probabilities \((P_1, P_2)\) satisfy \( P_1 = 1 - P_2 \) and represent the probability that a sample is generated from class 1 or 2 respectively. Equivalently they represent the fraction of samples generated from class 1 or 2 over the long run.

For the purposes of this dissertation a pattern classifier is a real-valued function \( q \) that produces a value in the range \([0, 1]\) that represents an estimate of the (posterior) probability \( p(\omega_1|x) \):

\[
q(x) = \hat{p}(\omega_1|x) \approx p(\omega_1|x) \quad (2.1)
\]

This function assigns a label

\[
T(q(x)) = \begin{cases} 
1, & q_1(x) > 0.5 \\
2, & q_1(x) \leq 0.5
\end{cases}
\]
2.1. TWO-CLASS SUPERVISED CLASSIFICATION PROBLEM

to every point \( \mathbf{x} \in \mathbb{R}^d \). A classifier \( q \) commits an error for sample \((\mathbf{x}, \omega)\) when \( T(q(\mathbf{x})) \neq \omega \). The goal is to construct a classifier whose expected error is as small as possible. If we define the \textit{indicator} function
\[
I(a) = \begin{cases} 
1, & a \text{ is true} \\
0, & a \text{ is false}
\end{cases}
\]
then the expected error \( e \) (also called the \textit{error rate}) can be written
\[
e(q) = E[I(T(q(\mathbf{x})) \neq \omega)]
\]
The minimum error rate over all possible functions is defined
\[
e^* := \min_{q \forall} e(q).
\]
This is the so-called \textit{Bayes error} and it represents a lower bound on the error that can be achieved. The classifier \( q^*(\mathbf{x}) = p(\omega_1|\mathbf{x}) = \frac{P_1p(\mathbf{x}|\omega_1)}{P_1p(\mathbf{x}|\omega_1) + P_2p(\mathbf{x}|\omega_2)} \) achieves the Bayes error.

**Example 1.** Consider a problem where the class densities are multivariate Gaussian of the form
\[
p(\mathbf{x}|\omega) = (2\pi)^{-d/2} |\Sigma_\omega|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu_\omega)^T \Sigma_\omega^2 (\mathbf{x} - \mu_\omega) \right\} \quad \omega \in \{\omega_1, \omega_2\}
\]
where the class means and covariances are given by
\[
\mu_\omega = E[\mathbf{x}|\omega] \\
\Sigma_\omega = E[(\mathbf{x} - \mu_\omega)(\mathbf{x} - \mu_\omega)^T|\omega].
\]
If we know the means, covariances, and prior probabilities then we can implement the optimal classifier by using Bayes’ rule
\[
p(\omega_1|\mathbf{x}) = \frac{P_1p(\mathbf{x}|\omega_1)}{P_1p(\mathbf{x}|\omega_1) + P_2p(\mathbf{x}|\omega_2)} \quad (2.2)
\]
and substitute into (2.1). But in practice, even when the distributions are known to be Gaussian, we almost never have complete knowledge of the means \((\mu_1, \mu_2)\), covariances \((\Sigma_1, \Sigma_2)\), and class probabilities \((P_1, P_2)\).
2.2. **UNCERTAINTY IN A TWO-CLASS CLASSIFIER**

Figure 2.1: The Machine Learning Lifecycle: (a) A model is trained, (b) the model’s performance is based on its predictions on test data, (c) after the deployment, the model is used to predict the true nature of unseen data.

One of the most common ways to compensate for incomplete distribution knowledge is to use **empirical data** to “fill in” the missing information. In particular we can often obtain a collection of labeled data samples \((x_1, \omega(1)), \ldots, (x_m, \omega(m))\) that can be used to help design the classifier. The process of using labeled data samples to help design the classifier is called **learning**, and the labeled data samples are called **training** data. In summary, the **supervised classification problem** can be stated as follows.

**Definition 1. The Supervised Classification Problem:** Given a collection, \(((x_1, \omega(1)), \ldots, (x_m, \omega(m)))\), of independent and identically distributed samples from a density \(p\), and incomplete knowledge of \(p\), determine a classifier \(q\) whose error \(e(q)\) is as close to \(e^*\) as possible.

### 2.2 Uncertainty in a Two-class Classifier

Consider a classifier, \(q\), trained using some learning algorithm and data set (Figure 2.1a). We can measure the classifier’s performance relative to a set of test data (Figure 2.1b). If we use \(q\) to classify an unknown sample, \(x\) (Figure 2.1c), the only knowledge we have regarding \(q\)’s efficacy in predicting \(x\), is our measures of its performance on the test data set as a whole.
2.2. **UNCERTAINTY IN A TWO-CLASS CLASSIFIER**

![Diagram](image)

Figure 2.2: Two classifiers constructed using same learning algorithm and probability distribution.

![Diagram](image)

Figure 2.3: Two realizations of a classifier possibly with differing predictions.

Theoretically, if we can measure $q$’s uncertainty with respect to the particular input, we would have some insight into the reliability of its prediction. If we construct two realizations of $q$, we have classifiers, $q_1$ and $q_2$, designed with different (equal size) training sets, both from the same distribution $p$ (Figure 2.2). If we use $q_1$ and $q_2$ to predict the same unseen sample, $x$ (Figure 2.3), and their outputs differ, then $q$’s prediction is not reliable. Intuitively, any $x$ where $|q_1(x) - q_2(x)|$ is large suggests a high degree of uncertainty.

Suppose we generate $n$ classifiers in the same way (Figure 2.4). If the number of
2.2. UNCERTAINTY IN A TWO-CLASS CLASSIFIER

Figure 2.4: Distribution over $n$ classifier realizations of the estimated probability that the unseen sample belongs to class $i$.

training samples is large, then the classifiers will be similar, but not identical. Even when trained on a large amount of data, there may be regions of the input space where the classifiers give very different predictions. Any $x$ where there is a high degree of variability in the classifier outputs, suggests a high degree of uncertainty.

There are many ways to quantize this variability. For example we might estimate the standard deviation of the classifier outputs, Another, closely related, option is to calculate covariate-dependent confidence intervals, which we describe in 3.3.2. These measures of variability also relate to an empirical estimate of a well-known stability measure which we describe in Section 2.5.3.

2.2.1 Summary of Approach

The experiments presented in this dissertation are conducted with Classification and Regression Trees (CART: Breiman et al., 1984) and Logistic Regression (LR) models implemented in the scikit-learn python package (Pedregosa et al., 2011). We define bootstrap sampling, then summarize our approach to creating distributions over a
Bootstrap Sampling

Given a random sample $X = (x_1, x_2, ..., x_n)$ from an unknown probability distribution $F$, we estimate the sampling distribution of random variable $R(X, F)$ on the basis of the observed data, $X$. To do so, we sample from $X$ with replacement to obtain a sampled set $X^*$ from the pool of observed data (Efron, 1979).

Estimating Uncertainty

The procedure is summarized as follows. First, obtain $S$ bootstrap samples by sampling the data $X = x_1, ..., x_n$ with replacement and obtaining $n$ data points. Denote the bootstrap samples $X^{*1}, X^{*2}, ..., X^{*S}$. Then, for each of the $S$ bootstrap samples, fit a learning model and obtain probability $p_i$ estimates for candidate labels $y_i$ for each $X^*$. The $S$ values of $p_i$ provide a probability distribution for each candidate label for each sample.

Those familiar with Ensemble learning will notice that this approach to quantifying uncertainty is similar to bagging (Breiman, 1996). While our approach is similar in its use of sampling, it examines the distribution of the classifiers’ predictions rather than aggregating their results.

2.3 Sources of Uncertainty

The classic machine learning problem maps observed data to unobservable properties of interest. Inverse uncertainty quantification (represented by the outer box in Figure 2.5) combines sources of variability in order to define and characterize the
2.3. SOURCES OF UNCERTAINTY

Figure 2.5: The steps of the standard machine learning task and their associated sources of uncertainty. The solution uncertainty is the union of uncertainties associated with measurements, regularization, model-form, and inference.

range of a system’s possible behavior. The problem is well-studied for cases in which an equation-based model is known (see Smith, 2014, for example).

The switch to statistical and stochastic models, however, raises new questions about uncertainty quantification methods. The major difference between classical inverse problems and machine learning is that the model in Figure 2.5c would traditionally be composed of a set of theoretical equations that define the mapping from observations to unobservables instead of an induced statistical model.

Figure 2.5 illustrates that uncertainty can arise from every step of the machine learning pipeline. Data collection introduces measurement errors and questions of data sufficiency. Which and how much data should be collected? How long before recollection is necessary? How the data is processed can lead to biases which may or may not be desirable. The choice of learning algorithm and the parameterization of the subsequent model is a source of variability. Some models require sampling to
make inferences. The model’s interface is also a source of uncertainty: the presenta-
tion of the outcome can influence a user’s perception of its meaning or validity.

Note that the goal of the work presented in this paper is to quantify and inter-
pret the uncertainty in the final outcomes of a supervised machine learning process
without investigating its sources. Identifying the uncertainties of each step in the
pipeline and understanding how they propagate are subjects of further study.

2.4 Uncertainty, Probability, and Error

Though not unrelated, we define error, probability, and uncertainty as distinct con-
cepts. Given a fixed model, probability measures the likelihood or belief in an inferred
outcome, uncertainty characterizes the variability in outcomes, and error describes
differences between predicted outcomes and actual observations. The sources of un-
certainty throughout the machine learning process give rise to a variety of possible
models, each with its own sets of probability estimates and errors.

2.5 Measures of Uncertainty and Their Limita-
tions

Supervised machine learning models are evaluated by techniques that measure their
accuracy in several forms, such as confusion matrices, ROC curves, and F-Scores.
These accuracy-based measurements indicate the model’s performance when distin-
guishing samples from a set of test data. Cross-validation techniques are also used
to estimate performance on unseen examples under the assumption that they are
drawn from the same distribution as the test data.
2.5. MEASURES OF UNCERTAINTY AND THEIR LIMITATIONS

Some approaches estimate uncertainty by computing confidence intervals to estimate the variance in cross-validation and ROC curves (see LeDell et al., 2015, for example). However, none of these methods are covariate-dependent, and though they distinguish between type 1 and type 2 errors, they do not distinguish between the different reasons for these errors: a model may consistently fail on certain areas of the input space, or it may be inconsistent in its response to samples with certain feature values (which results in it effectively flipping coins to determine a prediction).

Our approach is to directly evaluate a learned model’s variability for each input. This is important because, a model may display low uncertainty on some areas of the input space, and be highly variable in other regions. Once we obtain a distribution over possible estimates of a sample, there are a number of ways to measure its spread.

In the remainder of this section, we examine the benefits and limitations of some measures for interpreting the uncertainty in classifier predictions. To do so, we use two types of examples; one abstract (Figure 2.6) and one from the domain of identifying URLs from malicious websites (Figures 2.7 and 2.8).

The distribution of an estimated probability can take on a number of forms such as the examples in Figure 2.6. The distribution in panel 6.3a is clustered tightly around its mean suggesting that it is a stable approximation of the true probability that a sample, \( x \), belongs to the class, \( \omega_1: p(\omega_1|x) \).

Though the curves in 2.6b and 2.6c have the same means as the distribution in panel a, their shapes do not give the same indication of stability. The bimodal curve in 2.6b suggests that at least two plausible interpretations of \( x \) exists. Figure 2.6c shows a more uniform distribution, which implies that the model’s response to \( x \) is highly sensitive to the particular sampling of the input data.

Figure 2.7 gives an example from one of our application problems: a distribution over the estimated probability that a given URL belongs to a malicious website. This
2.5. MEASURES OF UNCERTAINTY AND THEIR LIMITATIONS

![Figure 2.6: Possible distributions of $q$.](image)

distribution is constructed by bootstrapping the data and training a decision tree for each sampled data set. For each URL in the testing data, each model outputs its estimated probability of the URL being malicious. These probability values are recorded into histograms.

For illustrative purposes, we create Kernel Density Estimates of the histograms where we choose a bandwidth (smoothing parameter) value that creates a continuous distribution without smoothing out the salient features of the distributions. The characteristics of these distributions can then be used to interpret the model’s response to each input.

### 2.5.1 Confidence Intervals

A typical approach for quantifying uncertainty is the calculation of confidence intervals, which we formally define for this problem in section 3.3.2. For the three example distributions of Figure 2.6, confidence intervals would allow us to distinguish the distribution in panel a from the other two but would not show the difference between the distributions in panels b and c.

A drawback of confidence interval analyses is they require the assumption of a known distribution such as normal. For example, if we were to calculate a confidence
2.5. MEASURES OF UNCERTAINTY AND THEIR LIMITATIONS

interval for the prediction distribution of the URL in Figure 2.7, we would have to assume an ideal form to its distribution which is clearly not valid here.

![Figure 2.7: Distribution over the probability that a URL is malicious.](image)

2.5.2 Standard Deviation

As a measure of variability, a distribution’s standard deviation provides a useful measure of uncertainty. This is seen in Figure 2.8 where we plot means versus standard deviations of the distributions generated when predicting URLs in a data set. The bars of various colors annotate the accuracy of the classifiers’ predictions on the samples whose standard deviation fall within their specified range.

For example, the green bar contains 1626 samples which have standard deviations between 0 and 0.1 and have been classified with an accuracy of 97.7%; the red bar contains 25 samples which have standard deviations between 0.4 and 0.5 and have been classified with an accuracy of 66.3%.

Standard deviation, though providing a useful measure of uncertainty, does not provide enough nuance to identify all pertinent characteristics of the distributions. For the three example distributions of Figure 2.6, the standard deviation of the distribution in panel a would distinguish it from the other two. However, the distributions in panel b and c have the same standard deviation value.

The two URL distributions in Figure 2.9 have almost the same means (0.49 and
2.5. MEASURES OF UNCERTAINTY AND THEIR LIMITATIONS

Figure 2.8: The left panel shows the distribution of predicting the specified URL. The right panel plots the means and standard deviations of the prediction distributions for all samples. The color bars annotate the number of samples within each standard deviation range and the average accuracy with which they were classified.

0.5) and standard deviations (0.29 and 0.3). Though both URLs represent malicious websites, the URL represented in 2.9b was classified correctly in 81% of 100 trials, while the other URL in 2.9a was classified correctly in only 53 trials.

Since we are interested in the uncertainty surrounding classification, we are not just interested in the variances or standard deviations of the distributions, but also in their deviation from 0 or 1. If we examine the distribution in panel (b), we see a large portion of its mass closer to 1, therefore we might have increased confidence of a malicious prediction. The distribution in panel (a) is more evenly spread with its most dense region around 50% (this can be measured with an analysis of highest density regions).

2.5.3 Instability

The closer a sample is to the decision boundary, the more likely that small changes in the input data or parameters will change its predicted class. This instability is
2.5. MEASURES OF UNCERTAINTY AND THEIR LIMITATIONS

![Figure 2.9: Two URL prediction distributions with almost the same means and standard deviations.](image)

Illustrated in Figure 2.10 where we define label instability for binary classification as the complement of the absolute difference between the number of times a sample classified as positive or negative divided by the total number of classifications:

\[
\text{instability}(x) = 1 - \frac{|\#pos - \#neg|}{\#pos + \#neg}
\]  

(2.3)

Figure 2.10 shows the instability values plotted against the standard deviations of the distributions for all the URLs in a test set. The URLs with higher standard deviation values tend to have more variation in their assigned labels through 100 trials.

While label instability captures an important aspect of classification uncertainty, it does not distinguish between the distributions in Figure 2.11. Both distributions have label instability values of 0.96 meaning the URLs they represent were classified as malicious and benign almost the exact same number of times through 100 trials. However, the label stability measure does not capture the the fact that the distribution in panel (b) has a higher amount of variance 2.11b.
2.5. MEASURES OF UNCERTAINTY AND THEIR LIMITATIONS

Figure 2.10: Label Instability: URL prediction distributions with larger standard deviations, have more instability in their label assignment.

Figure 2.11: Two URL prediction distributions with the same measure of label instability.

Given the limitations of the measures seen in this section. In Chapter 3, we develop a measure we call *minimum prediction deviation*, that captures pertinent information for interpreting the distributions over a model’s estimates in a classification context.
2.6 Approaches to Uncertainty

Our methods are solely based on probability theory and we leave investigation of alternative approaches to future work. This section provides a brief overview of uncertainty quantification and methods.

Though the study of probability theory dates back to the 16th century, its use in scientific applications generally began in the 20th century. Traditionally, scientists viewed uncertainty as undesirable and sought to develop theories of nature based in precision. Einstein’s sentiment that “God does not play dice with the universe” summarizes the prevailing wisdom prior to the development of quantum mechanics.

In the early twentieth century physicists required statistical methods to model the behavior of microscopic particles since their exact behaviors are often unknowable. Since then, the scientific community has increased the use of probability theory to describe the uncertainties associated with models containing imprecise or random elements (Booker and Ross, 2011).

Probability theory is generally divided into two interpretations. In the Frequentist interpretation, probabilities represent long run frequencies of events; the true properties about parameters, which represent information about an event or system of interest, are revealed in the long run.

In the Bayesian interpretation, probabilities represent our knowledge about a parameter. Our knowledge without observing data is represented by a prior distribution. We update what we know about the parameter as we observe data, which is represented by a likelihood function. The posterior distribution represents our updated knowledge about the parameter after observing data (Bickel and Doksum, 2001; Samaniego, 2010; Gelman et al., 2014). Since the posterior is a probability distribution, it is used to quantify uncertainty about an event occurring. The ad-
2.6. APPROACHES TO UNCERTAINTY

Vantage of this approach is its ability to model uncertainty in events for which long term frequencies have not been or cannot be measured (Murphy, 2012).

Starting with Lotfi Zadeh’s introduction of fuzzy sets in 1965, alternative theories to classical probability have emerged for describing uncertainty. In his seminal paper on *Fuzzy Logic Theory*, Zadeh reasoned that in order to handle many types of systems realistically, we need approaches that do not place undue importance on precision (Zadeh, 1973). In order to describe physical systems, Fuzzy logic theory uses, in addition to numerical variables, “linguistic” variables to characterize simple relations between variables with fuzzy conditional statements and complex relations with fuzzy algorithms. Uncertainty is represented by a certainty factor which is calculated using fuzzy, rather than precise, quantifiers (Zadeh, 1983).

The *Dempster-Shafer Theory of Evidence* is a framework for combining evidence based on two ideas. First, it defines propositions and assigns them intervals containing degrees of belief and plausibility. Secondly, it uses rules to combine beliefs (Luger, 2008). The size of a belief/plausibility interval represents confidence in a hypothesis. If there is no evidence for a hypothesis, its belief-plausibility interval will be [0,1]; the interval begins to shrink if evidence for the hypothesis is uncovered.

Quantitative *Possibility Theory*, based on Fuzzy Set theory, was introduced by Zadeh (1978) and extended by Dubois and Prade (1988). Quantitative possibility theory is an uncertainty framework designed to distinguish between uncertainty due to variability of observations and uncertainty due to incomplete data. Rather than using one set function as in probability theory, possibility theory uses two: maxitive, and minitive, which are also called possibility and necessity measures respectively. Possibility defines the extent to which an event is consistent with what is known and necessity specifies to what extent an event is implied by current knowledge. The difference between the two measures constitutes the uncertainty of an event occurring (Dubois and Prade, 2007).
2.7. EXTRACTING UNCERTAINTY FROM MACHINE LEARNING PIPELINE

Random Set Theory models data with random subsets instead of conventional vectors. It is used for types of data that cannot be represented as points. This allows for the construction of priors and likelihood functions that are capable of modeling a wide range of phenomena and has been used for information fusion of imperfect data (Khaleghi et al., 2013).

Imprecise Probability Theory, encompasses a variety of approaches that generalize traditional probability theory in order to model uncertainty and partial ignorance. Walley (2000) works toward the unification of imprecise probability theories as he argues that many theories work well for particular applications, but are insufficient for representing all types of uncertainty.

2.7 Extracting Uncertainty From Machine Learning Pipeline

We include a selection of papers which address uncertainty in aspects of the machine learning pipeline. Whereas our methods are generally model agnostic, with the exception of Sun (2015), the following methods address uncertainty for particular approaches.

Nix and Weigend (1994) estimate uncertainty in a function estimator based on feed-forward neural networks by analyzing the statistical properties of the model errors when reproducing the observed data.

Heskes (1997) estimate prediction intervals in neural networks containing limited data and address uncertainty in their estimator with confidence intervals.

2.7. EXTRACTING UNCERTAINTY FROM MACHINE LEARNING PIPELINE


Glasmachers and Igel (2008) address model-form uncertainty in support vector machines when choosing kernels and regularization parameters.


Muhlbaier et al. (2005) use a modified softmax equation with the classifier weights output from an ensemble to estimate confidence in a prediction.

Shrestha and Solomatine (2006) estimate prediction uncertainty as two quantiles of the distribution of errors. Using fuzzy c-means clustering, they partition the input space in terms of the samples with similar model errors.


Gal (2016) obtain uncertainty estimates in deep learning. They tie approximate Bayesian inference to dropout and other stochastic regularization techniques and assess approximations empirically.

Mentch and Hooker (2016) through confidence intervals and hypothesis tests, quantify the uncertainty of predictions generated by supervised learning ensemble methods. This averages over models built on subsamples of the training data and demonstrates that the resulting estimators take the form of a U-statistic. As such, predictions for individual feature vectors are asymptotically normal, which allows for the construction of confidence intervals around predictions. While this is a mathematically rigorous method for uncertainty
2.7. *EXTRACTING UNCERTAINTY FROM MACHINE LEARNING PIPELINE*

quantification in the domain of supervised learning, asymptotic normal distributions of predictions do not capture the distributions’ intricacies. We show, in our example problems, that these intricacies can have critical ramifications for decision-making.
Chapter 3

Minimum Prediction Deviation

3.1 Notions of Uncertainty

We introduce a novel measure of uncertainty, which quantifies three notions of uncertainty for supervised classification:

1. **Intrinsic**: The uncertainty of classification even when the probabilistic structure is completely known.


3. **Density**: The significance of uncertainty relative to the density of the data.

3.1.1 Notion 1: Intrinsic Uncertainty

Suppose we wish to classify samples from two continuous distributions whose forms we know exactly (Figure 3.1a). This is an optimal two-class problem. The per-
3.1. **NOTIONS OF UNCERTAINTY**

formance of a solution for separating the two data populations is measured by its classification error rate.

The vertical line set at $x = 0$ represents the optimal boundary for distinguishing the two distributions in terms of minimal error rate. If we draw a sample, $x^*$, from any point along the horizontal axis, it will have a probability, $p(\omega_1|x^*) = 1 - p(\omega_2|x^*)$, of belonging to the $\omega_1$ class. There is no uncertainty about these probability values. Likewise, there is no uncertainty in the optimality of the boundary line. In this idealized scenario, the only aspect that is uncertain is the exact label of $x^*$.

![Figure 3.1](image)

Figure 3.1: (a) Two continuous distributions; (b) probability that $x^*$ belongs to $\omega_1$ as a function of the input space; (c) multiple estimates, $q_1(x)$, of the probability that $x$ belongs to $\omega_1$ as a function of the input space

We can calculate the conditional probability of any $x$ belonging to either distribution using Bayes’ formula (Equation 2.2). Figure 3.1b shows the conditional probability of a sample belonging to $\omega_1$ as a function of the input space.

The uncertainty of any input’s correct label assignment is highest when $p(\omega_1|x) = 0.5$. Equivalently, there is less uncertainty when $p(\omega_1|x)$ is close to 0 or 1. There are many ways to quantify this notion, but one possibility is:

$$U_p(x) = \min(|p_1 - 0|, |p_1 - 1|)$$

(3.1)

where, for convenience, $p_1 := p(\omega_1|x)$. Equation 3.1 quantifies intrinsic classification uncertainty. Note that $U_p$ satisfies $0 \leq U_p \leq 0.5$ and increases as uncertainty in-
3.1. NOTIONS OF UNCERTAINTY

creases.

3.1.2 Notion 2: Empirical Uncertainty

Suppose we are unaware of the true distributions of \( \omega_1 \) and \( \omega_2 \). Let \( \hat{\omega}_1 \) and \( \hat{\omega}_2 \) be distributions estimated by sampling from \( \omega_1 \) and \( \omega_2 \). Then \( \hat{p}(\omega_1|x) \) is the estimated probability that a sample, \( x \), belongs to \( \omega_1 \). For notation convenience, let \( q(x) := \hat{p}(\omega_1|x) \).

If we use a learning method to produce the estimate \( q(x) \) from the sample data, the output will vary depending on particular realizations of \( \hat{\omega}_1 \) and \( \hat{\omega}_2 \). This uncertainty, illustrated in Figure 3.1c, is due to the finite number of training samples, and the variability inherent in a learning method.

Suppose we repeat the sampling process an infinite number of times and obtain as many estimates of the probability that \( x \) belongs to \( \omega_1 \). The probability density function of \( q(x) \) could take on a number of forms. We again examine the example distributions seen in Figure 3.2.

We remind the reader of their interpretations: though the three distributions have the same mean, panel 6.3a indicates a stable approximation of the true probability
3.1. NOTIONS OF UNCERTAINTY

\( p(\omega_1|\mathbf{x}) \), panel b suggests that at least two plausible classifications of \( \mathbf{x} \) exists, and panel c implies that the model’s output on \( \mathbf{x} \) is highly sensitive to the particular sample of training data.

As we show in Chapter 2, traditional measures such as confidence intervals and standard deviation do not capture the differences between these three distributions. We develop our new measure as means to distinguish the pertinent characteristics of a distribution for classification purposes.

Since we are interested in defining the uncertainty surrounding a classification, we wish to capture the notion that the uncertainty of classifying \( \mathbf{x} \) is high when the distribution of its probability estimate, \( q(\mathbf{x}) \), is concentrated near 0.5 and low when the probability mass is clustered close to 0 or 1: when a large majority of the classifiers designed with data from the input distribution produce the same classification for \( \mathbf{x} \).

A distribution’s standard deviation is defined as: \( \sigma(q(\mathbf{x})) = \sqrt{E[(q(\mathbf{x}) - \bar{q}(\mathbf{x}))^2]} \): the square root of the expectation of the squared deviation of a random variable from its mean. Since our goal is to determine the uncertainty in the classification of \( \mathbf{x} \) based on the distribution of \( q(\mathbf{x}) \), we modify the standard deviation equation so that it quantifies the deviation of \( q(\mathbf{x}) \) from 0 and 1:

\[
    u_0(\mathbf{x}) = \sqrt{E[(q(\mathbf{x}) - 0)^2]} \quad (3.2)
\]
\[
    u_1(\mathbf{x}) = \sqrt{E[(q(\mathbf{x}) - 1)^2]} \quad (3.3)
\]

If \( u_0(\mathbf{x}) < u_1(\mathbf{x}) \), the distribution’s mass is clustered closer to 0; if \( u_1(\mathbf{x}) < u_0(\mathbf{x}) \), the distribution’s mass is clustered closer to 1. Thus we define a measure of empirical uncertainty we call minimum prediction deviation as:

\[
    U_q(\mathbf{x}) = \min[u_0(\mathbf{x}), u_1(\mathbf{x})] \quad (3.4)
\]
3.1. NOTIONS OF UNCERTAINTY

$U_q$ increases as uncertainty increases.

Equation 3.4, simultaneously quantifies the first two notions of uncertainty: the uncertainty inherent in optimal classification and the uncertainty of estimated outcomes produced by a learned classifier. Note that if the learning method is consistent ($q_1(x) \to p(\omega_1|x)$), as the number of samples goes to infinity, then the uncertainty in equation 3.4 converges to equation 3.1.

From the perspective of classification, the form of $q(x)$, has three extreme cases of interest. The first case is when $U_q = 0$: all of the probability mass is at 1 or 0; it is obvious that these cases convey the highest confidence in the classification of $x$. The second case is when $U_q(x) = 0.5$: all of the probability mass is at 0.5 and we can infer, confidently, that the classifiers are uncertain about this input and are effectively flipping coins when assigning a classification output.

The third case, the highest level of uncertainty, occurs when exactly half of the probability mass is at 0 and the other half is at 1 where, a simple calculation shows, $U_q(x) = \frac{1}{\sqrt{2}} \approx 0.707$. Since half the classifiers give the complete opposite probability estimate to the other half it is intuitively clear that the uncertainty here should be higher than the second case, where, at least, all of the classifiers agree on the estimate.

Note that the range of the minimum prediction deviation, $U_q$, is $0 \leq U_q(x) \leq \frac{1}{\sqrt{2}}$, which is larger than the range of the intrinsic uncertainty $0 \leq U_p \leq 0.5$. We avoid scaling these two measures to the same range since this would lead to a situation where $U_q(x)$ no longer converges to $U_p(x)$ for consistent learning methods.
3.2 DATA FOR COMPARISON

3.1.3 Notion 3: The significance of uncertainty dictated by the relative sparseness of the data

Let $B(x, D)$ be the set of points that lie within a ball of radius $D$ of point $x$ (where the “shape” of the ball is determined by a user-specified metric). We then calculate the probability mass $P(B(x, D))$ of the data points that fall within $B(x, D)$ and update equation 3.4 with two alternatives:

$$U_q(x) = \min[u_0(x), u_1(x)]P(B(x, D))$$ (3.5)

$$U_q(x) = \min[u_0(x), u_1(x)](1 - P(B(x, D)))$$ (3.6)

Equation 3.5 places less emphasis on low density regions and Equation 3.6 places more emphasis on low density regions. For example, we may want to place less emphasis on low density regions simply because we expect to see fewer data samples from these regions and are therefore less concerned about uncertain decisions for these samples.

On the other hand in anomaly detection problems we may want to place more emphasis on low density regions because we are more concerned about the uncertainty of decisions in these regions.

3.2 Data for Comparison

Notion 1, intrinsic uncertainty, can be seen when we classify data samples using Bayes’ rule (see section 2.1). When the likelihood and prior for a random sample are known, we are able to calculate the true posterior probability that the sample belongs to a particular class.
3.2. DATA FOR COMPARISON

Figure 3.3: Example I-I dataset with 1000 samples from each distribution.

For example, Figure 3.3 shows random samples generated by two gaussian distributions in two dimensions. If we classify each sample by their maximum posterior class probability (Equation 2.2), we obtain the results seen in Figure 3.4a. In this example, we obtain the error rate of 15.5% is $e^*$: the Bayes, or optimal, error rate.

Intrinsic uncertainty can be calculated when the probabilistic structure of a problem is completely known. We call Notion 1 intrinsic uncertainty since when all aspects of the problem are known, the only uncertainty is the exact label of a random sample. The region where the samples from the two distributions overlap contain samples whose posterior probabilities are near 50%. If we set a threshold such that we classify as uncertain the samples whose whose posterior probability is in the set $[0.3,0.7]$, which corresponds to an intrinsic threshold of 0.3, we get the results seen in Figure 3.4b.
3.2. DATA FOR COMPARISON

In most real-world problems, the distributions that underlie a data set, though at times can be estimated, are often unknowable and, therefore, the true posterior probabilities and intrinsic uncertainty cannot be calculated. Learning methods can be used to induce models that estimate posterior class probabilities. This process introduces model-form uncertainty to the estimated outcomes—hence the need for Notions 2 and 3.

To understand the characteristics and implications of our estimates and measurements of empirical uncertainty, we generate synthetic data so that we can calculate both true posteriors and estimated probabilities of the samples so that we can compare their intrinsic uncertainty to their empirical uncertainty estimates.

![Figure 3.4: Example Bayesian Classification. Without (a) and With Uncertainty (b).](a) (b)

3.2.1 Synthetic Data

We use three types of synthetic data sets which, for short-hand purposes, we call I-I, I-sI, and I-M. Figures 3.3, 3.5, and 3.6, show example realizations of these types
3.2. DATA FOR COMPARISON

![Figure 3.5: Example I-sI data set](image)

of data in 2-dimensions with 2000 datapoints where the blue dots are labeled as generated from distribution 1, and the red dots are labeled as distribution 2.

The I-I data sets contain two gaussian distributions with differing means and identical covariances which are the identity matrix (see Example 1 in section 2.1). Figure 3.3 shows 2-dimensional realization of this type of data with 2000 samples.

The overlap in the data ensures some amount of error. However, the learning algorithms we will use, Classification and Regression Trees (CART) (Breiman, 2017) and Logistic Regression (LR) (Nelder and Wedderburn, 1972), are expected to discriminate this type of data well. However, these two approaches use significantly different models and learning methods to produce posterior probability estimates and so we expect their uncertainty profiles to differ.
3.2. DATA FOR COMPARISON

Figure 3.6: Example I-M data set with 1000 samples from each distribution.

The I-sI data, seen in Figure 3.5 contain two gaussian distributions with the same mean but with differing covariance matrices. Distribution 1 uses the identity matrix for its covariance. Distribution 2’s covariance is the identity matrix multiplied by some scalar value (in this example the red distribution’s covariance is the identify matrix multiplied by 8).

We use I-sI data to explore cases where CART is expected to provide much better classification performance than LR, but the expected uncertainty profiles are difficult to predict ahead of time.

Figure 3.6 shows an example realization of what we call the I-M data set. Whereas distribution 2 has the identity matrix for its covariance, distribution 1 is generated from a mixture of two gaussians (it is not easy to distinguish the two gaussians visually).
3.3. MEASURES FOR COMPARISON

Models built with LR are expected to provide better classification performance on I-M data than the I-sI cases, but not as well as on I-I data. CART is expected to provide better classification for I-M data than LR when the training set is sufficiently large, but once again the uncertainty profiles are difficult to predict ahead of time (for both methods).

3.3 Measures for Comparison

To compare MPD with existing measures, we use Standard Deviation with CART and LR and the Covariate-dependent Confidence Interval (CI) for LR. We compare to these measures as they are widely used and MPD is derived by modifying the standard deviation.

In this section, we define the learning methods, standard deviation and derive the CIs. In Chapter 4, we will conduct empirical experiments using these measures of uncertainty on a variety of datasets when comparing to MPD.

3.3.1 Standard Deviation

Standard deviation defines a distribution’s spread from its mean. In the context of classification, a high standard deviation implies that a model’s output on a particular input is highly sensitive to the particular sampling of the training data.

We define uncertainty defined by standard deviation as:

\[ U_{std}(x) = \sigma(q(x)) = \sqrt{E[(q(x) - \bar{q}(x))^2]} \] (3.7)

where \( q(x) \) can be the distribution over probability estimates from a CART or LR model. \( U_{std} \) satisfies \( 0 \leq U_{std} \leq 0.5 \) and increases as uncertainty increases. The
3.3. MEASURES FOR COMPARISON

maximum value of 0.5 occurs when half of the probability mass is at 0 and the other half is at 1.

In section 2.5.2, we show that standard deviation can be a useful measure of uncertainty that shows some correlation to a model’s accuracy. In the next chapter, we compare standard deviation to our new measure.

3.3.2 Covariate-dependent Confidence Intervals for Logistic Regression

Logistic regression uses a weighted linear combination of the input vector to produce an estimated class probability using the logistic function:

\[ \hat{p}(\omega_1|x) = \pi(\hat{\beta}^T \tilde{x}) \]

where \( \pi \) is the logistic function:

\[ \pi(\theta) = \frac{e^\theta}{1+e^\theta}, \]

\( \beta \) is the coefficient vector:

\[ \hat{\beta}^T = \left[ \hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_d \right], \]

and \( \tilde{x} \) is the augmented data vector:

\[ \tilde{x}^T = (1, x^T). \]

The coefficients are estimated from the training data using maximum-likelihood estimation.

Derivation of Covariate-dependent Confidence Interval:

This derivation draws from Hauck (1983); Li (2018); STATISTICA (2013); Dybowski and Roberts (2001); Geyer (2007); Kerns (2017).
3.3. MEASURES FOR COMPARISON

The *Fisher information matrix* for this problem is:

\[
I(\hat{\beta}) = \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T \pi(\hat{\beta}^T \tilde{x}_i) \left( 1 - \pi(\hat{\beta}^T \tilde{x}_i) \right)
\]

where \( \tilde{x}_i \tilde{x}_i^T \) is a \((d + 1) \times (d + 1)\) vector outer product matrix and \( \pi(\hat{\beta}^T \tilde{x}_i) \left( 1 - \pi(\hat{\beta}^T \tilde{x}_i) \right) \) is a scalar.

The (asymptotic) estimated covariance matrix of the estimated coefficients is given by:

\[
\hat{\Sigma}_\beta = \left[ I(\hat{\beta}) \right]^{-1}
\]

The *standard error* at covariate value \( x \) is

\[
\hat{s}_x = \sqrt{\tilde{x}^T \hat{\Sigma}_\beta \tilde{x}}
\]

We define

\[
l^-(\tilde{x}) = \hat{\beta}^T \tilde{x} - z_{\alpha/2} \hat{s}_x
\]

\[
l^+(\tilde{x}) = \hat{\beta}^T \tilde{x} + z_{\alpha/2} \hat{s}_x
\]

where \( z_{\alpha/2} \) is the \( 1 - \alpha/2 \) quantile of the standard normal distribution.

For example if \( \alpha = 0.05 \) then \( z_{0.025} = 1.96 \). Thus, the 100(1 - \( \alpha \))% confidence interval for predicted probability value \( \hat{\pi}(\tilde{x}) \) at covariate \( x \) is:

\[
\hat{\pi}(\tilde{x}) = \left[ \pi(l^-)(\tilde{x}), \, \pi(l^+)(\tilde{x}) \right]
\]

Under the, somewhat technical, assumptions described in Hauck (1983); Li (2018); STATISTICA (2013); Dybowski and Roberts (2001); Geyer (2007); Kerns (2017), the confidence interval above is an interval that contains the true posterior probability with probability at least \( 1 - \alpha \). Thus, if the confidence interval is large then there is significant uncertainty in the posterior probability estimate.
3.3. **MEASURES FOR COMPARISON**

We define the confidence interval uncertainty to be

\[ U_{CI}(x) = \pi(l^+(\tilde{x})) - \pi(l^-(\tilde{x})) \]  

and declare a data sample as uncertain if \( U_{CI} \) is greater than some threshold: \( U_{CI}(x) > t \).

Uncertainty defined by covariate-dependent confidence intervals satisfies \( 0 \leq U_{CI} \leq 1 \).

### 3.3.3 Empirical Uncertainty Deviation

In ensuing experiments, where we compute uncertainty based on estimates of class probabilities, we compare these empirical assessments of uncertainty to the intrinsic rates. This is done by first, setting a threshold and identifying the samples which are intrinsically uncertain with a filter function using Equation 3.1:

\[
U_{int}(x) = \begin{cases} 
0 & \text{if } U_p(x) \leq \text{threshold}_{intrinsic} \\
1 & \text{if } U_p(x) > \text{threshold}_{intrinsic}
\end{cases}
\]  

(3.9)

As an example, the black, uncertain, samples in Figure 3.4b are defined using Equation 3.9 with a threshold of 0.3. After applying the intrinsic filter function, we then have a set of samples that are uncertain per Equation 3.9: \( \mathcal{I} = \{ x | U_{int}(x) = 1 \} \).

We define the *intrinsic uncertainty rate* as the percentage of samples, classified with Bayes' rule, found to be uncertain for a given threshold:

\[
R_I = \frac{\sum_{i}^{n} U_{int}(x_i)}{n}
\]  

(3.10)

where \( n \) is the number of samples in the data set.
3.3. MEASURES FOR COMPARISON

In the empirical setting, after estimating the posterior probabilities with a learning method we apply the filter function using an empirical estimate of uncertainty:

\[
U_{emp}(x) = \begin{cases} 
0 & \text{if } U(x) < \text{threshold}_{emp} \\ 
1 & \text{if } U(x) > \text{threshold}_{emp} 
\end{cases}
\]  

(3.11)

where \( U(x) \) could be any of our empirical estimates of uncertainty. The range of the threshold is dictated by the range of the empirical uncertainty measure used (MPD Standard Deviation, or CIs). Uncertainty is highest in the regions where different classes of data overlap: when the samples’ true posterior probability is close to 0.5. Setting thresholds at higher values results in a smaller uncertainty region; lower thresholds result in larger uncertainty regions.

As an example, Figure 3.7 shows samples that are first classified with logistic regression and decision tree models, then defined to be uncertain, or not, by equation 3.11 with equation 3.4 as the uncertainty measure using a threshold of 0.2.

Figure 3.7a shows that filtering the data using MPD with logistic regression results in an uncertainty region defined by two lines equidistant from the linear decision boundary. In Figure 3.7b, the results of filtering the data using MPD with CART exhibits more variability at the out edges of the overlap region.

In the next chapter we show examples of filtering the data with confidence intervals and standard deviation, and compare the results to filtering the same data sets with minimum prediction deviation.

After applying the empirical filter function, we then have a set of samples that are uncertain per Equation 3.11: \( \mathcal{E} = \{x | U_{emp}(x) = 1\} \).

The thresholds for the empirical and intrinsic filter functions are free parameters. However, as a means to compare the estimated uncertainty to the intrinsic, we lower
3.3. MEASURES FOR COMPARISON

Figure 3.7: Examples of Empirical Classification and Uncertainty for Logistic Regression (left panel) and CART (right panel) given a threshold of 0.2.

the empirical threshold until the set of empirically uncertain samples contains the set of intrinsically uncertain samples: \( \mathcal{I} \subseteq \mathcal{E} \).

After setting the empirical threshold, we find the empirical uncertainty rate:

\[
R_E = \frac{\sum_{i} U_{emp}(x_i)}{n}
\]  

(3.12)

To compare measures of uncertainty, we find the difference between intrinsic and empirical uncertainty rates and call this quantity *empirical uncertainty deviation*:

\[
D_E = R_E - R_I
\]  

(3.13)

In the next chapter, we compare the deviations calculated from three different measures of uncertainty to see how their empirical uncertainty rates compare to the intrinsic uncertainty rate as we increase the number of samples in a data set. If a particular learning method is a consistent estimator of the posterior probability, then we expect \( D_E \) to approach 0 as the number of training samples goes to infinity.
3.3. MEASURES FOR COMPARISON

<table>
<thead>
<tr>
<th>Measure</th>
<th>Equation Number</th>
<th>Range</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic Uncertainty</td>
<td>3.1</td>
<td>[0, 0.5]</td>
<td>Define uncertainty of sample class probability.</td>
</tr>
<tr>
<td>Minimum Prediction Deviation (MPD)</td>
<td>3.4</td>
<td>[0, 1/\sqrt{2}]</td>
<td>Estimate uncertainty of sample classification.</td>
</tr>
<tr>
<td>MPD with density term</td>
<td>3.5 and 3.6</td>
<td>[0, 1/\sqrt{2}]</td>
<td>Estimate uncertainty of sample classification then scale by density defined by ball of some radius.</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>3.7</td>
<td>[0, 0.5]</td>
<td>Define variability of a distribution.</td>
</tr>
<tr>
<td>Covariate-dependent Confidence Intervals (CI)</td>
<td>3.8</td>
<td>[0,1]</td>
<td>Define confidence of sample class probability estimate.</td>
</tr>
<tr>
<td>Intrinsic and Empirical Uncertainty Rates (R_I \text{ and } R_E)</td>
<td>3.10 and 3.12</td>
<td>[0,1]</td>
<td>Find number of samples classified as uncertain for a given threshold.</td>
</tr>
<tr>
<td>Empirical Uncertainty Deviation (D_E)</td>
<td>3.13</td>
<td>[0,1 - (R_I)]</td>
<td>Find difference between empirical and intrinsic uncertainty rates for a given threshold.</td>
</tr>
</tbody>
</table>

Table 3.1: Summary of measures defined in this chapter.

3.3.4 Summary of Measures

In this chapter we define a measure of intrinsic uncertainty for a dataset, five measures of empirical uncertainty for distributions over estimated probabilities, the rates of dataset uncertainty for particular thresholds, and empirical uncertainty deviation which compares rates of empirical uncertainty to intrinsic uncertainty. We summarize these measures and their ranges in Table 3.1.
Chapter 4

Comparison of Uncertainty Measures

In this chapter, we compare the results of defining uncertainty with minimum prediction deviation, standard deviation, and covariate-dependent confidence intervals. We use three test data sets: one each of I-I, I-sI, and I-M with 1000 samples (Figures 4.1, 4.4, and 4.7 respectively).

We first calculate the intrinsic uncertainty values for each sample in the data (in section 3.1.1 we define intrinsic uncertainty to be $U_p(x) = \min(|p_1 - 0|, |p_1 - 1|)$). We show 3-dimensional plots with uncertainty on the vertical axis against the samples’ 2-dimensional locations where the samples are shaded by their level of uncertainty (see Figure 4.2 for example). We also plot such 3-dimensional graphs with empirical uncertainty measures. These graphs provide useful illustrations of the behavior of a given classifier method.

As described in section 3.3.3, the method we use to compare the uncertainty measures is to
Chapter 4. Comparison of Uncertainty Measures

1. set a threshold on intrinsic uncertainty,

2. find the set of samples whose intrinsic uncertainty values are above the threshold,

3. calculate the empirical uncertainty values of the data using one of the empirical measures,

4. decrease the threshold on the empirical uncertainty values until the set of empirically uncertain samples contains the set of intrinsically uncertain samples.

5. Finally, we compute the difference between the number of samples which are deemed empirically uncertain to the number which intrinsically uncertain samples. We call this difference uncertainty deviation, $D_E$.

The experiments in this chapter all use an intrinsic threshold of 0.4.

We expect the empirical uncertainty measures to capture the intrinsic uncertainty, which stems from overlap of the data classes, and the uncertainty of the learning process, what we call the implementation uncertainty. Therefore, not only do we expect to see higher uncertainty with empirical uncertainty measures, we desire to. If an empirical uncertainty measure shows less uncertainty than intrinsic uncertainty, we can conclude that it is not adequately capturing the two types of uncertainty. The key idea is not to eliminate uncertainty, as desirable as that may seem, rather, we wish to measure uncertainty accurately so that we can interpret the reliability of a model’s predictions.

Though this chapter only shows results for data in 2 dimensions, these results generalize to data in higher dimensions.
4.1 Minimum Prediction Deviation with CART

To gain an intuitive understanding of MPD with CART, we calculate both the intrinsic and MPD uncertainty for each type of dataset (I-I, I-sI, and I-M).

For each type of data, we show the raw test data set, 3-dimensional maps of intrinsic and empirical uncertainty, and we compare MPD uncertainty to the thresholded intrinsic uncertainty in terms of empirical uncertainty deviation, $D_E$.

4.1.1 I-I Data

Figure 4.1: Plot of I-I test data with 1000 samples.

Figure 4.1 shows the I-I test data with 1000 samples. When classified optimally, the overlap between the data sets results in a Bayes error rate of approximately 0.145. Though uncertainty is expected to be lower here than with the other types of
4.1. **MINIMUM PREDICTION DEVIATION WITH CART**

Figure 4.2: (a) Intrinsic and (b) MPD with CART uncertainty heatmap plots of I-I test set; darker colors signify higher levels of uncertainty.

datasets, we expect it to be somewhat high due to the finite number of samples—especially with a CART model.

Since the distributions have the same covariance and different means, we expect most learning methods to be able to find reasonable approximations of the optimal linear decision surface. On this data, the CART classifier has an average error rate of approximately 0.180.

Figure 4.2 shows 3D plots of the same I-I dataset with (a) intrinsic and (b) MPD uncertainty on the vertical axis. As expected, uncertainty is highest in the areas with the most overlap. Whereas the intrinsic uncertainty continuously increases as the samples approach the area of overlap. Plot (b) shows the empirical uncertainty a little more spread out in the space.

Since there is a finite number of samples and only two dimensions, the tree does not have much room to grow. Therefore, there are few probability values possible as the trees have a depth of 5.32 with 10.62 leaves on average. In this example,
4.1. MINIMUM PREDICTION DEVIATION WITH CART

Figure 4.3: Threshold plots for (a) Intrinsic and (b) MPD with CART uncertainty for I-I test set with an intrinsic uncertainty threshold of 0.4.

there are 500 samples in the training dataset. If we increase the number of training samples, we see the empirical uncertainty values tend toward the intrinsic values.

Figure 4.3a shows the results of classifying the data optimally with Bayes rule and filtering the data with the intrinsic filter (Equation 3.9) using a threshold of 0.4. The intrinsic uncertainty rate, $R_I$ is 0.088.

Figure 4.3b shows has an empirical uncertainty rate, $R_E$ of 0.355 resulting in an empirical uncertainty deviation of $D_E = 0.355 - 0.088 = 0.267$. Due to the finite number of samples, and variation of the learning method, the empirical uncertainty is significantly higher than the intrinsic uncertainty. As expected with CART, there is high uncertainty in sparsely populated areas of the overlap regions.
4.1. MINIMUM PREDICTION DEVIATION WITH CART

Figure 4.4: Plot of I-sI test data with 1000 samples.

4.1.2 I-sI Data

Figure 4.4 shows the I-sI test set with 1000 samples. When classified optimally, the Bayes error rate is approximately 0.197. The optimal decision boundary has a circular shape which linear classifiers, such as logistic regression, are unable to approximate.

The flexibility of a CART model allows it to perform better on this data than a linear model, in this case the CART classifiers have an average error rate of approximately 0.235.

Figure 4.5 shows a 3-dimensional plot of the data with uncertainty on the vertical axis for (a) intrinsic, and (b) MPD uncertainty computed from a CART-based model using a training data set with 500 samples. The MPD uncertainty values seem to be capturing the intrinsic uncertainty fairly well, though show higher uncertainty
4.1. **MINIMUM PREDICTION DEVIATION WITH CART**

Figure 4.5: (a) Intrinsic and (b) MPD with CART uncertainty 3-dimensional plots for I-sI test set where darker samples signify higher uncertainty values.

Figure 4.6: Threshold plots for (a) Intrinsic and (b) MPD with CART uncertainty for I-sI test set with an intrinsic uncertainty threshold of 0.4.

Figure 4.6a shows a threshold plot for intrinsic uncertainty with a threshold of 0.4 where $R_I$ is 0.082. Figure 4.6b shows an empirical uncertainty rate, $R_E$ of 0.589.
resulting in an empirical uncertainty deviation of $D_E = 0.589 - 0.082 = 0.507$.

The 2-dimensional plots in Figure 4.6b are somewhat misleading. Since the samples are plotted in the order: dataset 1, dataset 2, uncertain, the uncertain region covers over some nuance. In the 1-dimensional plots, we can see that the empirical uncertainty regions have some resemblance to their intrinsic counterparts.

### 4.1.3 I-M Data

Figure 4.7 shows the I-M test data with 1000 samples. With optimal classification, the Bayes error rate is approximately 0.170. The CART model performs well on this type of data, achieving an error rate of approximately 0.185.

Figure 4.8 shows intrinsic uncertainty in panel (a) and empirical uncertainty in
4.1. MINIMUM PREDICTION DEVIATION WITH CART

Figure 4.8: (a) Intrinsic and (b) MPD with CART uncertainty plots for 500 sample I-M dataset. The samples are shaded such that darker colors signifies higher uncertainty.

Panel (b) calculated with CART estimated probability distributions where the MPD values seem to be fairly capturing the intrinsic uncertainty.

Figure 4.9a shows the results of classifying the data optimally with Bayes rule and filtering the data with an intrinsic threshold of 0.4. The intrinsic uncertainty rate, $R_I$ is 0.126.

Figure 4.9b shows the uncertainty when classifying with CART, the uncertainty rate, $R_E$ is 0.442 resulting in an empirical uncertainty deviation of $D_E = 0.442 - 0.126 = 0.316$. The nonzero deviance suggests that in addition to capturing the intrinsic uncertainty well, MPD also captures the implementation uncertainty that appears predominately in the overlap region.
4.2. MPD WITH LOGISTIC REGRESSION

Figure 4.9: (a) Intrinsic and (b) MPD with CART uncertainty for 500 sample I-M dataset with a intrinsic uncertainty threshold of 0.4.

4.2 MPD with Logistic Regression

Up to this point, we show MPD computed from CART model probability estimates. In this section we show MPD based on estimated probabilities from Logistic Regression models.

We choose to experiment logistic regression for three reasons. First, the closest comparison to MPD is the covariate-dependent confidence interval (CI), and there is a well known CI result for LR (described in section 3.3.2); secondly, calculating MPD from LR probability estimates provides an informative comparison to calculating MPD with CART since LR uses a fundamentally different learning method and decision boundary than CART; Finally, pairing LR with an I-I dataset is an ideal case where MPD uncertainty converges to intrinsic uncertainty as the number of training samples goes to infinity, which we show in section 4.3.

While an LR classifier performs near-optimally on an I-I dataset, its accuracy decreases on I-M and, especially, I-sI datasets since their circular decision boundaries
4.2. **MPD WITH LOGISTIC REGRESSION**

are poorly approximated by the linear model.

Figure 4.10 shows 3-dimensional plots with MPD on the vertical axis for (a) I-I, (b) I-M, and (c) I-sI test sets using LR models each trained with 500 sample training sets.

The MPD uncertainty values with LR are very close to the intrinsic uncertainty of the I-I test set, this is because LR fits this data well. However, the LR model’s fit to the other two data sets is sub-optimal, which can be seen if we compare the MPD uncertainty plots in Figures 4.10b and 4.10c to their intrinsic counter parts (Figures 4.5a and 4.8a respectively).

Figure 4.10c shows high values of uncertainty for all samples where the range is between 0.35 and 0.5. Since LR is a linear classifier, we would expect it to choose a line that best separates the two classes of data and show lower uncertainty in the outer regions of the space. However, instead, the model is producing posterior estimates close to 0.5 for all samples which results in an error rate close to 0.5.

While there exists linear classifier design methods that will achieve lower classification error and smaller uncertainty in some regions of the space, the MPD results...
4.3. **CONSISTENCY**

show very clearly that linear classifiers produced by LR are essentially useless here.

![MPD threshold plots for LR models](image)

(a) (b) (c)

Figure 4.11: MPD threshold plots for LR models trained on 500 sample data sets for the (a) I-I, (b) I-M, and (c) I-sI test sets

Figure 4.11 shows the same datasets filtered with an intrinsic threshold of 0.4.

For the I-I data set, the LR model’s uncertainty rate is 0.150 which is close to the intrinsic uncertainty rate of 0.082. The LR models’ average classification error is approximately 0.144 which means that it achieves the Bayes error rate of approximately 0.144.

The average error rate of an LR model accuracy for the I-M dataset is approximately 0.183 where the Bayes rate is approximately 0.172. However, despite the high accuracy, the uncertainty rate seen in 4.11b is 0.796. This suggests, as expected, that the model is not the best fit to the data.

For the I-sI data set, the LR models’ average error rate is approximately 0.495 with an uncertainty rate of 0.807.
4.3. CONSISTENCY

4.3 Consistency

Since empirical uncertainty partially stems from a lack of sufficient data, under particular circumstances, namely a learning method that is a consistent estimator of posterior probabilities trained on a data set oriented ideally for its approach, empirical uncertainty as defined by MPD (Equation 3.4) will converge to the intrinsic uncertainty quantified by Equation 3.1 with an infinite number of training samples.

To illustrate this, Figure 4.12 shows I-I test data where its MPD uncertainty values are calculated from LR models trained with (a) 500, (b) 1,000, and (c) 10,000 samples.

If we examine the plot of these data’s intrinsic uncertainty seen in Figure 4.2a and compare to Figure 4.12, in (a), there are clearly differences in the samples’ intrinsic and empirical uncertainty values. These differences decrease with higher numbers of training samples in (b). With 10,000 training samples, 4.12c looks nearly identical to Figure 4.2a.

Figure 4.13 shows threshold plots of MPD using LR models trained on (a) 50, (b)
4.3. CONSISTENCY

Figure 4.13: MPD threshold plots computed with LE models trained on (a) 50, (b) 500, and (c) 5,000 samples. The empirical uncertainty deviates from the intrinsic by rates of (a) 0.277, (b) 0.102, and (c) 0.014.

500, and (c) 5,000 samples. If we compare to this dataset’s intrinsic threshold plot in Figure 4.3a, we can see that there is a significant difference between the uncertainty regions in the intrinsic uncertainty in (a), whereas there is virtually no difference between the regions in (c).

In this example, we set the intrinsic uncertainty threshold at 0.4. As described in section 3.3.3, we compare the thresholded uncertainties in terms of empirical deviance, $D_E$. The empirical threshold also starts at 0.4 but is decreased until the set of intrinsically uncertain samples are contained within the set of empirically uncertain samples, we then find the difference in the uncertainty rates between the two sets.

In (a), the empirical uncertainty rate is 0.457, the intrinsic uncertainty rate for this dataset is 0.088: $D_E = 0.457 - 0.088 = 0.369$. The deviance value decreases to 0.032 in (b) and 0.006 in (c). With 5,000 training samples, minimum prediction deviation with logistic regression finds only 6 samples of the I-I test set to be uncertain beyond the baseline defined by intrinsic uncertainty.
4.4. **MPD VS. STANDARD DEVIATION**

![Figure 4.14: (a) Intrinsic, (b) Minimum Prediction Deviation, and (c) Standard deviation uncertainty for I-sI test set. MPD and standard deviation values based on CART model trained on 500 sample training set.](image)

4.4 **MPD vs. Standard Deviation**

In section 3.3.1, we define uncertainty with the standard deviation of samples’ estimated probability distribution. When a posterior distribution has a large standard deviation, we can infer high uncertainty in the probability estimate.

A majority of the samples in the I-sI test set have relatively high values of intrinsic uncertainty since all of the samples from distribution 2 are completely contained within the space of distribution 1 (Figure 4.14a). Therefore, if we use a model with relatively high inherent variability, such as CART, to estimate the posteriors of samples from this data set, we will obtain distributions with a fair amount of variation.

In this case, standard deviation will yield some insight into the model’s reliability for each sample. However, if we compare Figure 4.14c to Figure 4.14b, the samples’ MPD values are generally closer to their intrinsic values than their standard deviations. This is likely due to the fact that standard deviation measures variability in a prediction without accounting for its proximity to one class or the other.
4.4. MPD VS. STANDARD DEVIATION

The utility of using standard deviation as an uncertainty metric is diminished with a larger sized training set. Figure 4.15 shows samples’ standard deviation values that result from training sets with (a) 1,000, and (b) 10,000 samples. The reduction in values partially results from an increase in the CART model’s complexity: as we increase the training set size, the tree grows to a larger depth which means that the samples are more likely to be classified the same way on every trial. However, even if we limit the depth of the tree, the increase in the training set size still reduces the variability of the predictions.

If use a model with relatively low variability, such as LR, the standard deviations of the posterior distributions will also be extremely narrow regardless of the size of the training set and the model’s fit to the data. Figure 4.16 shows the data’s (a) MPD and (b) standard deviation values calculated from an LR model. The range of the standard deviation values in Figure 4.16b is [0, 4.77e−8].

The LR model is estimating nearly the same probability values on every trial—which is expected from a highly biased model. Figure 4.16a highlights a utility of
4.4. MPD VS. STANDARD DEVIATION

Figure 4.16: 3-dimensional uncertainty plots for MPD and Standard Deviation with LR

MPD in that it, correctly, shows high uncertainty even when there is little variation in the model’s estimates.

Figure 4.17 shows threshold plots of I-sI data using Standard Deviation with (a) CART and (a) LR models. The 2-dimensional plots in Figure 4.17a lack the nuance that can be seen in the 1-dimensional plots: since the uncertain samples are plotted last, they cover the samples that are classified as certain. Still, however, a large portion of the space is deemed uncertain, whereas the intrinsic uncertainty region is relatively small for this type of data.

Figure 4.17b shows that since the standard deviation values of the posterior distributions from LR models are extremely small, in order to cover the intrinsically uncertain space, all of the samples have to be deemed uncertain.
4.5. **MPD VS. CONFIDENCE INTERVALS FOR LOGISTIC REGRESSION**

Figure 4.17: Threshold plots of I-sI data using Standard Deviation with (a) CART and (a) LR models

![Threshold plots of I-sI data using Standard Deviation with (a) CART and (a) LR models](image)

(a) ![Threshold plots of I-sI data using Standard Deviation with (a) CART and (a) LR models](image)

(b) ![Threshold plots of I-sI data using Standard Deviation with (a) CART and (a) LR models](image)

**4.5 MPD vs. Confidence Intervals for Logistic Regression**

Figure 4.18 shows uncertainty plots of covariate-dependent confidence intervals calculated for logistic regression models for (a) I-I, (b) I-M, and (c) I-sI test sets.

Figure 4.19 shows confidence interval threshold plots for (a) I-I, (b) I-M, and (c) I-sI datasets. In section 3.3.2, we define uncertainty with confidence intervals to be defined by the width of the interval. Therefore, uncertainty will be lower in more densely populated regions of the space, even where the class probability values are close to 0.5. This is seen in Figure 4.18 where the highest levels of uncertainty are in the outer regions of the distributions.

The empirical, CI-based, uncertainty rates for the three data sets are 0.378, 0.700, and 0.594 respectively. Which means that they deviate from the intrinsic uncertainty by rates of 0.290, 0.574, and 0.512 respectively.
4.6. **MPD WITH DENSITY WEIGHTING**

![Confidence Interval Heat Maps](image)

Figure 4.18: Confidence interval heat maps with LR for (a) I-I, (b) I-M, and (c) I-sI datasets with 500 samples each.

The fact that the CI-based empirical uncertainty deviation values are higher for the I-M data than the I-sI dataset is notable since the LR models’ average error rates on the two datasets are approximately 0.183 for the I-M samples, and 0.495 for the I-sI data.

Since CIs are designed to measure the confidence in a probability estimate, they do not speak to the classifier’s ability to distinguish between the two classes of data. On the same set of I-sI data with the same intrinsic threshold, pairing LR with MPD results in an uncertainty rate of 1.0. MPD allows us to, rightly, interpret that LR is a poor fit for this data. This conclusion is harder to infer if we use confidence intervals as our uncertainty measure.

For the I-I data which, as we have seen, LR fits very well, MPD with LR achieves an uncertainty rate of .102 which is closer to this dataset’s intrinsic rate of 0.088 than the CI-based uncertainty rate of 0.290.
4.6. **MPD WITH DENSITY WEIGHTING**

![Images of confidence interval threshold plots](a) (b) (c)

Figure 4.19: Confidence interval threshold plots with LR for (a) I-I, (b) I-M, and (c) I-sl datasets with 500 samples each.

### 4.6 MPD with Density Weighting

In section 3.1.3, we define MPD multiplied by the probability mass $P(B(x, D))$ of the data points that fall within a radius $D$ of the sample in question, $x$. Equation 3.5, $U_q(x) = \min[u_0(x), u_1(x)]P(B(x, D))$, is designed to place less emphasis on sparse

![Images of 3-D uncertainty plots](a) (b)

Figure 4.20: 3-D uncertainty plot for MPD with CART with Density Term for 1000 samples with (a) uncertainty multiplied by the samples’ density value, (b) multiplied by the inverse of the samples’ density value.
4.6. **MPD WITH DENSITY WEIGHTING**

Figure 4.21: 3-D uncertainty plot for LR with CART with Density Term for 1000 samples with (a) uncertainty multiplied by the samples’ density value, (b) multiplied by the inverse of the samples’ density value.

regions, and Equation 3.6, \( U_q(x) = \min[u_0(x), u_1(x)](1 - P(B(x, D))) \) is designed to place more emphases on regions with lower density.

Figures 4.20 and 4.21 show 3-dimensional uncertainty plots where \( P(B(x, D)) \) is a circle with a radius such that each sample’s density value reflects all samples in the dataset.

Figure 4.20 shows the I-I test set classified with CART with uncertainty defined by (a) Equation 3.5, and (b) Equation 3.6. Both equations reduce most of the samples’ values of uncertainty compared to MPD without the density term. Figure 4.20a shows the uncertainty region concentrated closer to the overlap regions and uncertainty suppressed in the outer regions compared to the uncertainty seen in Figure 4.2. Figure 4.20b shows higher uncertainty in the sparsely populated overlap regions while retaining the major uncertainty region in the densely populated area.

These effects are more pronounced in Figure 4.21, where we define uncertainty by (a) Equation 3.5 and (b) Equation 3.6 with an LR model on the I-sI data set.
4.7. MPD IN MANY DIMENSIONS

Since the LR models do not fit the I-sI data well, uncertainty is high throughout the space. However, adding the density terms deemphasize in (a), and emphasize in (b), the outer regions of the data.

4.7 MPD in Many Dimensions

Thus far, all of our examples have been performed on 2-dimensional datasets. In the next two chapters we apply uncertainty analyses to problems with 4 and 87 features.

In this section we show the results of measuring uncertainty in for synthetic data 2, 4, 6, 8, and 10 dimensional spaces. For each type of data (I-I, I-sI, and I-M) and dimensionality we train and test on datasets with 1000 samples using both CART and LR models.

We do not show graphical representations of all the experiments since data is difficult to visualize in more than 2-dimensions. However, to provide some examples, Figures 4.22, 4.23, and 4.24 show threshold plots of the intrinsic uncertainty for each type of data in 4-dimensions. These provide limited utility. In these higher dimensional spaces, it is difficult to clearly discern the regions of the three types of data.

In this section, as with the rest of the Chapter, for all experiments, we set the intrinsic uncertainty threshold at 0.4 and measure the intrinsic uncertainty rate ($R_I$), we then allow the MPD thresholds to decrease until the set of intrinsically uncertain samples are contained with the set of empirically uncertain and measure the empirical uncertainty rates ($R_E$). We report results in Tables 4.1, 4.2, and 4.3.
4.7. MPD IN MANY DIMENSIONS

Figure 4.22: Intrinsic Uncertainty for I-I Data in 4-dimensions with an intrinsic uncertainty threshold of 0.4

<table>
<thead>
<tr>
<th>Number of Dimensions</th>
<th>Bayes Error Rate</th>
<th>Intrinsic Unc. Rate $R_I$</th>
<th>CART Error Rate</th>
<th>CART-MPD Error Rate $R_E$</th>
<th>LR Error Rate</th>
<th>LR-MPD Error Rate $R_E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.124</td>
<td>0.069</td>
<td>0.172</td>
<td>0.350</td>
<td>0.125</td>
<td>0.144</td>
</tr>
<tr>
<td>4</td>
<td>0.174</td>
<td>0.102</td>
<td>0.178</td>
<td>0.797</td>
<td>0.178</td>
<td>0.125</td>
</tr>
<tr>
<td>6</td>
<td>0.148</td>
<td>0.076</td>
<td>0.298</td>
<td>0.771</td>
<td>0.157</td>
<td>0.189</td>
</tr>
<tr>
<td>8</td>
<td>0.140</td>
<td>0.087</td>
<td>0.330</td>
<td>0.830</td>
<td>0.144</td>
<td>0.216</td>
</tr>
<tr>
<td>10</td>
<td>0.159</td>
<td>0.098</td>
<td>0.361</td>
<td>0.834</td>
<td>0.170</td>
<td>0.167</td>
</tr>
</tbody>
</table>

Table 4.1: Rates of bayes error, intrinsic uncertainty, CART error, CART with MPD uncertainty, LR error, LR with MPD uncertainty for I-I datasets in many dimensions with an intrinsic uncertainty threshold of 0.4.
Figure 4.23: Intrinsic Uncertainty for I-sI Data in 4-dimensions with an intrinsic uncertainty threshold of 0.4

<table>
<thead>
<tr>
<th>Number of Dimensions</th>
<th>Bayes Error Rate</th>
<th>Intrinsic Unc. $R_I$</th>
<th>CART Error Rate</th>
<th>CART-MPD Error Rate</th>
<th>LR Error Rate</th>
<th>LR-MPD Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.172</td>
<td>0.087</td>
<td>0.31</td>
<td>0.91</td>
<td>0.507</td>
<td>0.896</td>
</tr>
<tr>
<td>4</td>
<td>0.153</td>
<td>0.076</td>
<td>0.346</td>
<td>0.603</td>
<td>0.453</td>
<td>0.934</td>
</tr>
<tr>
<td>6</td>
<td>0.161</td>
<td>0.090</td>
<td>0.401</td>
<td>0.950</td>
<td>0.501</td>
<td>0.956</td>
</tr>
<tr>
<td>8</td>
<td>0.159</td>
<td>0.083</td>
<td>0.400</td>
<td>0.969</td>
<td>0.494</td>
<td>0.983</td>
</tr>
<tr>
<td>10</td>
<td>0.164</td>
<td>0.082</td>
<td>0.422</td>
<td>0.988</td>
<td>0.491</td>
<td>0.931</td>
</tr>
</tbody>
</table>

Table 4.2: Rates of bayes error, intrinsic uncertainty, CART error, CART with MPD uncertainty, LR error, LR with MPD uncertainty for I-sI datasets increasing dimensions with an intrinsic uncertainty threshold of 0.4.
4.7. MPD IN MANY DIMENSIONS

Figure 4.24: Intrinsic Uncertainty for I-M Data in 4-dimensions with an intrinsic uncertainty threshold of 0.4

<table>
<thead>
<tr>
<th>Number of Dimensions</th>
<th>Bayes Error Rate</th>
<th>Intrinsic Unc. $R_I$</th>
<th>CART Error Rate $R_E$</th>
<th>CART-MPD Error Rate $R_E$</th>
<th>LR Error Rate</th>
<th>LR-MPD Error Rate $R_E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.163</td>
<td>0.136</td>
<td>0.189</td>
<td>0.862</td>
<td>0.165</td>
<td>0.834</td>
</tr>
<tr>
<td>4</td>
<td>0.151</td>
<td>0.059</td>
<td>0.233</td>
<td>0.822</td>
<td>0.200</td>
<td>0.835</td>
</tr>
<tr>
<td>6</td>
<td>0.164</td>
<td>0.055</td>
<td>0.278</td>
<td>0.854</td>
<td>0.208</td>
<td>0.734</td>
</tr>
<tr>
<td>8</td>
<td>0.151</td>
<td>0.042</td>
<td>0.283</td>
<td>0.679</td>
<td>0.222</td>
<td>0.647</td>
</tr>
<tr>
<td>10</td>
<td>0.157</td>
<td>0.040</td>
<td>0.298</td>
<td>0.743</td>
<td>0.260</td>
<td>0.822</td>
</tr>
</tbody>
</table>

Table 4.3: Rates of bayes error, intrinsic uncertainty, CART error, CART with MPD uncertainty, LR error, LR with MPD uncertainty for I-M datasets in many dimensions with an intrinsic uncertainty threshold of 0.4.
Chapter 5

Minimum Prediction Deviation
Applied to URL Analysis

In Chapter 3, we define minimum prediction deviation as a means to capture the uncertainty in a classifier’s predictions. Once the uncertainty is defined, this information has a number of potential uses. In this chapter, we return to the problem of malicious URL classification and examine using MPD as a measure of confidence in a model’s predictions.

For domains, such as cyber security, where every classification is potentially critical, it is important to have the ability to assess the credibility of a model’s output. In the case of a classifier charged with protecting a network by identifying malicious URLs, regardless of the model’s overall performance, it will occasionally misclassify some examples, risking infection and infiltration of the network.

Obtaining and characterizing a model’s uncertainty with respect to each sample provides a means to differentiate between the predictions that are more and less reliable. This knowledge can be used to send the samples presenting higher uncertainty for an alternative, perhaps more computationally intense, analysis—or human veri-
5.1. **THE DATA**

This section adds an uncertainty analysis to the URL classification system presented by Darling et al. (2015).

In the context of URLs, the distributions of features change rapidly as new websites are introduced on a daily basis — especially in the set of malicious URLs since attackers, in order to evade detection, change domains often. Boostrapping the data set allows us to assess the range of possible outputs given an input—creating the variety necessary for constructing an effective estimation of uncertainty. After using the bootstrap method to generate the sampled data sets and training models for each, we obtain the probabilistic distributions of the candidate labels for each URL and measure these distributions with minimum prediction deviation.

### 5.1 The Data

The data set contains 127,684 URLs of which half are labeled as *malicious*; the other half are labeled as *benign*. The URLs are separated into three parts (see Figure 5.1), and are then quantified into 87 features including:

- 16 Language features: we create a n-gram-frequency-based language model based on the set of Benign URLs and use it to draw features for all training data.
- 10 Length features: including lengths of URL, host name, and path.
- 29 Counting features: including number of delimiters, numbers, and letters.
- 15 Pattern features: including case changes, consecutive occurrences of a character, and most frequent token.
- 5 Binary features: including ‘.com’ out of place, IP address for a host name, and presence of black-list words.
5.2. **MPD Analysis**

- 12 Ratio features: including vowel to consonant ratio, digit to letter ratio, and ratio of the sections of the URL compared to the overall length.

For our example task presented here, we partition the data such that 90% is used for training the model, and the remaining 10% is withheld for the testing set.

![Figure 5.1: URL Components.](https://www.facebook.com/help/cookies/?ref=sitefooter)

**5.2 MPD Analysis**

We demonstrate the application of minimum prediction deviation, as defined by Equation 3.4, to the URL classification system. We use the same approach to building the model where we bootstrap the data and train a classifier for each sampling of the data. We use CART based classifiers since decision trees have proven to be effective in this problem space (Darling et al., 2015)

After the bootstrap process, for each URL, we obtain distributions over the estimated probabilities that a URL is malicious or benign. Figure 5.2 shows the estimates for an example URL.

In these two-class problems, the distributions are mirror images since the probability estimates collected during each training iteration add up to 1. Therefore, we can glean most of the relevant information from one of the two distributions, so, for most of the analysis we can discard the benign distribution and examine the distribution over the sample’s malicious estimate. However it can be instructive to look at the mean of the distribution of estimating sample’s true label, as is done in Figure 5.3.
5.2. MPD ANALYSIS

Figure 5.2: Distributions over the estimated probabilities that the example URL is malicious or benign.

Figure 5.3 shows for each example, the mean of the mode’s estimate of its true label plotted against its MPD value. Samples whose true value is malicious are

Figure 5.3: The prediction distributions’ mean plotted against prediction uncertainty overlaid with accuracy at each interval.
5.2. **MPD ANALYSIS**

![Graphical representation of Table 5.1: the number of samples above uncertainty threshold and the accuracy of the remaining samples.](image)

represented by the red plus symbol; truly benign samples are symbolized by blue underscores. The color bands annotate the number of samples contained within the range that each spans, as well as the model’s accuracy on that group of samples.

We choose to plot MPD against mean, since, the mean of the distribution represents our baseline decision boundary: if a sample’s distribution over its estimate of being malicious is greater than 0.5, we classify the sample as malicious. Using this criterion, the classifier’s overall accuracy on the testing set of 12,538 URLs is approximately 0.946. This plot shows on which samples the model is 99.6% accurate (green band) and the examples where the model’s accuracy is significantly lower: the model’s accuracy is only 0.639 for the samples within the red band.

Figure 5.3 shows that MPD can be used to inform decision making based on a model’s predictions. MPD can be a measure of confidence in the model’s outcomes since predictions presenting higher MPD values are less likely to be correctly classified.

Suppose we filter the predictions such that we will reject the samples presenting
5.2. **MPD ANALYSIS**

<table>
<thead>
<tr>
<th>MPD Threshold</th>
<th>Accuracy of Remaining URLs</th>
<th>Number of URLs Rejected</th>
<th>Number of URLs Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.946</td>
<td>0</td>
<td>12,538</td>
</tr>
<tr>
<td>0.6</td>
<td>0.947</td>
<td>26</td>
<td>12,512</td>
</tr>
<tr>
<td>0.5</td>
<td>0.954</td>
<td>275</td>
<td>12,263</td>
</tr>
<tr>
<td>0.4</td>
<td>0.970</td>
<td>940</td>
<td>11,598</td>
</tr>
<tr>
<td>0.3</td>
<td>0.985</td>
<td>1661</td>
<td>10,877</td>
</tr>
<tr>
<td>0.2</td>
<td>0.995</td>
<td>2845</td>
<td>9693</td>
</tr>
<tr>
<td>0.1</td>
<td>0.999</td>
<td>4507</td>
<td>8031</td>
</tr>
</tbody>
</table>

*Table 5.1: At each specified threshold, the accuracy of the remaining samples, the number of samples whose MPD value is above (rejected), and below the threshold (remaining).*

MPD values above some threshold. The rejected samples could be sent for an alternative analysis or human verification. Figure 5.4 and Table 5.1 show the improvement of the model’s classification accuracy, as we remove the the samples presenting MPD values above a threshold.

A majority of the URLs present low uncertainty such as the example in Figure 5.5a which has a mean of almost 1 and an MPD value of approximately 0.003. There are however a number of distributions that look similar to Figure 5.5b with a mean of 0.54 and an MPD value of 0.472. Figure 5.5c, with a mean of 0.53, shows the URL whose estimate obtained the highest MPD value of approximately 0.672.
5.2. **MPD ANALYSIS**

which is close to the worst possible value of $\frac{1}{\sqrt{2}}$ (approximately 0.707).

Though the distributions in panels (b) and (c) have similar means, they have significantly different MPD values as they represent two distinct cases. Figure 5.5b has a low variance which meant that it is a relatively stable estimation showing that the URL is not well described by either label given the model. Figure 5.5c fits, under different conditions, both labels extremely well for a given bootstrap sampling.

The subject of improving the model to reduce uncertainty in our estimates, is reserved for further study. However, the URL analysis shows that the ability to separate the samples on which the model is unable to predict reliably, allows us to have increased confidence in the model’s predictions on the remaining samples.

Classifying malicious websites solely on the lexical features of their URLs is the most lightweight approach to this problem. Other approaches include time-consuming queries to blacklist websites and domain registration information.

Even more time consuming, are the “dynamic” approaches which actually execute the page’s content. One possibility for marrying the speed of a static approach, to the accuracy of a dynamic system, would be to use the lightweight model as a gatekeeper: accept the predictions of the static model, and send the websites it finds uncertain to a more thorough, time-consuming system.
Chapter 6

Minimum Prediction Deviation
Applied to Pixel Classification

The previous chapter describes a example for using the uncertainty information measured by minimum prediction deviation to a deployed system. We set a threshold on the MPD values of the classifier’s prediction and are then able to triage URLs into two categories: uncertain and certain.

In this chapter we apply an uncertainty analysis to pixel classification. We choose this domain in order to apply our methods to a problem with characteristics unlike those of the URL problem. Whereas the URL classifier uses 87 features, from the pixel data we only derive 4. Furthermore, this problem allows for visual inspection of uncertainty as we shade the pixels of an image by their MPD values.

For our second example task, we map urban tree canopy using aerial imagery. Organizations, such as the U.S. Department of Agriculture, commission studies to determine the amount of tree coverage and help communities improve tree cover over time (USDA, 2019).
6.1. **THE DATA**

The traditional approach for determining the amount of tree cover in aerial imagery is pixel-based classification, of which there are two approaches: pixel-based hard and pixel-based soft classification. A pixel-based hard classification assumes that each pixel contains only one type of class; a pixel-based soft approach produces a probabilistic assessment of each possible class (Chen et al., 2018).

This problem is less critical than URL classification. In cyber security problems, every prediction is potentially critical, which is clearly not the case here: the level of acceptable uncertainty is relatively high. However, understanding uncertainty in our outcomes can still help us to improve the analysis. For example, if too many pixels an image are above a uncertainty-defined threshold, we could determine that we need to collect more training data, perhaps from a different type of sensor. Another option, like in the URL example, would be to send images containing high levels of uncertainty for an alternative analysis or human verification.

The experiments seen in this section build on work by Stracuzzi et al. (2017, 2018a,b). Whereas in the previous studies classified multiple objects in each image, here we restrict the problem to two-classes: each pixel is classified based on its probability of belonging to a tree or not.

### 6.1 The Data

Figure 6.1 shows optical and lidar images of similar scenes taken from a stretch of Philadelphia’s Schuylkill river. Each image contains some combination of trees and other objects such as grass, water, buildings, and boats.

The first row of Figure 6.1 shows the optical imaging of the regions where each pixel contains red, green, and blue values scaled from 0 to 1. The middle row shows the same regions imaged with lidar, also scaled from 0 to 1, which has been
6.2. MPD ANALYSIS

Figure 6.1: Optical (first row), lidar (middle row), and ground-truth labels (last row) of five similar regions. The labels are represented as: green for trees and gray for everything else.

preprocessed into height maps where lighter pixels indicate taller objects.

The third row contains hand-drawn segment labels where red represents trees, and the pixels with all other objects are blue. We follow O’Neil-Dunne et al. (2013) by combining each pixel’s R, G, B, and height values into four-dimensional vectors. We will refer to the scenes by their column letter (for example, we will refer to the first scene as “scene a”).

6.2 MPD Analysis

We demonstrate the application of minimum prediction deviation, as defined by Equation 3.4, to the pixel classification system. Following Falk et al. (2015), we use
6.2. MPD ANALYSIS

the CART decision tree algorithm.

We use the same approach to model building as in the previous two chapters. For each experiment, we combine four out of the 5 scenes to produce a training set and then test on the 5th scene, and rotate so each image is tested. For example, to test on the scene in column (e), we train on the scenes from columns a-d. Each training set undergoes 100 trials of bootstrapping, a realization of the classifier is induced for each bootstrap sampling, and each classifier then produces estimates for each pixel in the test image.

Figure 6.2 shows the classification results. Row 1 plots each pixel’s most likely class per the model (red for tree and blue for other); Row 2 shows each pixel shaded by its MPD value with lighter pixels signifying greater levels of uncertainty; Row 3 is a combination of the first two: each pixel is assigned a color for its most likely class and then shaded by its MPD value; Row 4 is the same as 3 except white pixels are above an MPD threshold of 0.4.

The models’ accuracy on each scene is (a) 0.92, (b) 0.954, (c) 0.761, (d) 0.810, (e) 0.917.

To understand the level uncertainty, we set a threshold on each pixel’s MPD value at 0.4 and take the ratio of the number of pixels above the threshold over the total number in the image: (a) 0.012, (b) 0.001, (c) 0.154, (d) 0.017, (e) 0.025.

For all of the scenes, the classifiers are fairly accurate and certain in their assessment of the non-tree class and their primary confusion is between trees and roofs. This is probably since the buildings and trees are similar heights and none of the other objects are nearly as tall.

The boathouse in column (c) is has the highest number of pixels above the threshold. The brightness of the roof Figure 6.2 column (c) shows relatively high MPD values for these pixels, which is unsurprising, since this is the only scene where the
6.2. MPD ANALYSIS

Figure 6.2: Results of pixel classification: (1) each pixel’s most likely prediction (red for tree and blue for other); (2) pixel shaded by its MPD value (the pixel becomes brighter when more uncertain); (3) is a combination of rows 1 and 2; (4) the same as row 3 except pixels that are above an MPD threshold of 0.4 are white.

boathouse has a red roof (all the others have gray colored roofs).

Figure 6.3 shows a stratified plot for the most a least accurate tests: scenes b and c respectively. The plots are partitioned into two strata to measure the models’ accuracy on the pixels above and below our specified MPD threshold of 0.4. For scene b, there are only 446 pixels above the threshold with an accuracy of 0.772, which results in an accuracy of 0.965 for the remaining pixels. For scene c, there are
6.2. **MPD ANALYSIS**

![Figure 6.3](image)

Figure 6.3: Stratified plots for the most and least accurate test runs: the scenes from Figure 6.1 columns (b) and (c)

2330 pixels above the threshold which have an accuracy of 0.448, with the remaining samples having an accuracy of 0.856.

The results in this chapter show, as with the URL application, MPD allows us to define uncertain examples and use that information to improve our analyses.
Chapter 7

Conclusion

This dissertation demonstrates an approach to uncertainty analysis for the supervised two-class learning problem and derives a new measure for interpreting the uncertainty of a supervised model’s predictions. We show how measurements of uncertainty allow us to assess the credibility of predictions in the absence of a sample’s true label. The goal of this work is not to build the best classifier, rather, we explore how uncertainty can help to understand the performance properties of any classifier.

7.1 Contributions

We summarize the contributions of this dissertation.

1. A novel measure of uncertainty for classification systems that quantifies three notions:

   (a) Uncertainty is higher in areas of the input space where class distributions are more evenly distributed: class assignments for samples from such re-
7.1. CONTRIBUTIONS

- Regions are ambiguous compared to areas with greater separation between class distributions.

(b) Uncertainty is higher in areas of the input space where the learning model is less stable: regions where there is a high variability in the predictions made by classifiers trained on same-size sampled data sets drawn from the same distribution.

(c) The significance of uncertainty is dictated by the relative sparseness of the data. In sparse regions, we infer from the input data that we are unlikely to see future samples. This weighting factor highlights future samples that are outliers with respect to the input distribution.

In applications where outliers are inconsequential to the question of interest, this measure could be used to drive down uncertainty. For applications such as threat detection, we define an alternate measure that could be used as a kind of anomaly detection.

2. An algorithm for estimating uncertainty in supervised learning predictions based on bootstrap sampling.

3. Method for post-learning decision making based on uncertainty. Using the new uncertainty measure we can determine which predictions can be accepted based on application-specific tolerance.

4. A comparison of the uncertainty measure to the traditional measures of uncertainty such as covariate-dependent confidence intervals such that:

(a) the new method makes less restrictive assumptions

(b) the new method works with any classification method that produces probabilistic class estimates whereas covariates-dependent confidence intervals are not available for many of these classification methods.
7.2. DISCUSSION

(c) comparisons are made on multiple data sets, where both confidence intervals and the new uncertainty measure are used to identify uncertain regions of the input space.

5. The new uncertainty measure can be used to identify salient characteristics of a learning method. For example, the novel uncertainty measure shows that decision tree learning methods are unstable in low density regions of the input space, even when the learning method employs a regularization scheme to prevent this.

7.2 Discussion

In the context of machine learning, uncertainty quantification provides an objective measure of the sufficiency of the available data and the selected modeling approach for answering a question of interest. Traditional evaluation methods, such as precision-recall, ROC curves (Powers, 2011) and cross-validation (Kohavi et al., 1995), provide a global measure of a classifier’s ability to discriminate among examples of different classes based on a set of known test data. As estimates of future performance, they assume that new examples are drawn from the same distribution as the test examples.

In many cases, it may be impossible, or at least expensive, to verify the similarity of new test examples relative to the data the model was trained on. Furthermore, if there are any changes in the distribution, for example due to some shift in sensor properties, conditions or targets, then the performance estimates will likely be overly optimistic. Uncertainty analysis provides a means to identify such cases.

Assessing the uncertainty of a classifier’s predictions provides information not provided by traditional evaluation methods. If models that produce probabilistic estimates provide users more information than a binary class assignment, an uncer-
7.2. DISCUSSION

tainty analysis further increases the information available by providing a means to measure credibility in a model’s prediction.

A prediction presenting high uncertainty (low model credibility) indicates that alternate valid interpretations of the data exist and the degree to which the model can distinguish between them. For example, a model that assigns a label with high uncertainty, such as a wide distribution over a label’s probability, should be viewed with skepticism even if the label has a relatively high point-estimate probability. Uncertainty measures, like performance metrics, are also biased by the specifics of the training data. Yet, uncertainty still provides knowledge not otherwise available.

In applications where every classification is potentially critical, cyber security for example, it is important to have the ability to assess the credibility of a model’s output. One of the examples we use is a classifier charged with protecting a network by identifying malicious URLs. Regardless of the model’s overall performance, it will occasionally misclassify some examples, risking infection and infiltration of the network.

Obtaining and characterizing a model’s uncertainty with respect to each sample provides a means to differentiate between the predictions that are more or less reliable. This knowledge can be used to send the samples presenting higher prediction uncertainty for an alternative, perhaps more computationally intense, analysis—or human verification.

In many complex machine learning problems, determining the amount of data required to construct a reliable model is a problem solved with a combination of heuristics, domain knowledge, and guess work. For the problem of classifying malicious URLs, it is practically impossible to accurately model the space of all URLs. The accuracy-based validation metrics tell us how well we model the data on hand, but tell us nothing of the efficacy of our classifier on the URLs found on the internet.
7.3. FUTURE WORK

as a whole. According to Internet Live Stats (2019), there are, currently, over 1.7 billion websites.

Even if we could hypothetically collect enough URLs to construct a training set representative of the entire distribution of URLs, there will always be some subset of samples that the model will classify poorly and these are exactly the set of URLs that hackers will attempt to acquire. In this domain, no matter how much training data we collect, we can always find a population of samples on which the model will perform poorly. Regardless of the reason for samples being outliers with respect to a learning model’s hypothesis, uncertainty quantification allows us to measure the credibility with which a model classifies each sample.

7.3 Future Work

Obtaining a model’s uncertainty when predicting the label of a sample enables us to infer a measure of confidence. This understanding can lead to considerations of trade-offs between the accuracy and variability in results produced by learning algorithms.

Uncertainty information can also be used to make decisions such as when a human needs to be alerted, when an alternative model should be used, or when additional parameter tuning or data may be necessary.

In this context we outline the following research goals:

1. Multiple-Classes: The obvious approach here would be to employ our current methods with a one-vs-rest scheme. However, it may be necessary to reform our approach for the multi-class classifier.

2. Uncertainty in the machine learning pipeline: In Chapter 2, we discuss how
7.3. FUTURE WORK

uncertainty can originate from any area of the machine learning pipeline. In this work we are primarily concerned with the uncertainty stemming from data and subsequent model induction. In future experiments we could keep the data fixed and see how other areas of the pipeline affect the uncertainty in a model’s outcomes—we have the ability to examine uncertainty in a model’s output with the methods introduced here.

3. Learning methods: The experiments discussed in this dissertation use decision tree and logistic regression algorithms. Since our approach is applicable to any learning method, we wish to do a comprehensive study of uncertainty analyses using common learning strategies.

4. Model Selection: we wish to develop methods for comparing uncertainty between classifiers and model validation.

5. Feature Selection: Which features contribute to uncertainty? We can use uncertainty as a criterion to analyze the addition or subtraction of features or data sources.

6. Label Validation: For some problems, ground truth can be hard to define. For pixel labeling of overhead imagery, when the same analyst labels the same image twice, the results slightly differ. Uncertainty could be used a means of validating a labeling method.

7. Boosting: We would like to study the effects of boosting using uncertainty as a weighting criterion instead of accuracy. Current approaches overcorrect on outliers. The hypothesis is that by boosting over uncertainty, we would clarify the decision boundaries, akin to margin maximization, though not obviously equivalent.

An uncertainty analysis applied to supervised learning increases the information available to analysts and automated systems. Although uncertainty analysis uncovers
7.3. FUTURE WORK

information not provided by other machine learning methods, using it to increase user trust requires research into methods, metrics, and visualizations. For example, the definition and calculation of uncertainty depends on both application context and the modeling algorithm.

We show that the standard deviation and confidence intervals of the probability distributions has some correlation with accuracy, but minimum prediction uncertainty proved more informative. In general, application-specific analytic goals dictate how informative a given uncertainty measure will be. As a result, human-machine interaction and uncertainty visualizations also require extensive research and development to convey the implication of an analysis and engender trust.
References


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


