Increasing Trust By Quantifying Uncertainty

David J. Stracuzzi and Michael C. Darling
Sandia National Laboratories
Albuquerque, NM 87123
djstrac@sandia.gov, mcdarli@sandia.gov

In the context of machine learning, uncertainty quantification provides an objective measure of sufficiency of the available data and the selected modeling approach for answering a question of interest. When a machine learning model generates a prediction, a decision maker needs to determine how much to trust its output. Currently, such determinations depend heavily on expert opinion, using a mix of domain and modeling expertise, and accuracy-based validation metrics, such as precision-recall and ROC curves. While these methods evaluate a model's performance relative to a fixed set of validation data, they do not tell us the model's certainty with respect to a particular prediction on an unseen, perhaps critical, data point.

Models that produce labels probabilistically provide the user a deeper understanding than a binary class assignment. A probability value indicates goodness-of-fit for candidate labels, so high probability labels fit the associated data well given the model. An uncertainty analysis increases understanding a step further by providing a measure of model credibility. Stated simply, a high uncertainty (low credibility) model indicates that alternate valid interpretations of the data exist and that the model cannot distinguish among them. For example, a distribution over the probability estimated for a given label indicates the degree of confusion over the most likely label. Thus, 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 high point-estimate probability.

In many machine learning problems, a combination of heuristics, domain knowledge and guesswork determine the amount of data required to construct a reliable model. If we take, for example, the problem of classifying malicious websites using their associated URLs, determining the distribution of URL characteristics is practically impossible. 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 a single URL — let alone the URLs found on the internet as a whole. Currently, there are over 1.2 billion unique domains each with a number of associated URLs (for many prominent websites there can be thousands, millions, or even billions of URLs for a single domain, e.g., Facebook).

Determining which predictions are reliable with respect to a model's decision surface and which are uncertain is key to increasing trust (Ribeiro et al., 2016). 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 the model classifies each sample. In many real-world problems, no matter how much training data we collect, we can always find a population of samples on which the model will perform greatly or poorly. Even if we collect enough data to construct a training set representative of the entire distribution, there will always be some subset of outliers, however small, that the model will classify poorly.

To illustrate how an uncertainty analysis can increase trust in a model's output, we expand on the example of classifying malicious URLs in Figure 1 using data from Darling et al. (2015). The left panel shows probabilistic distributions produced by an ensemble of classifiers tasked with

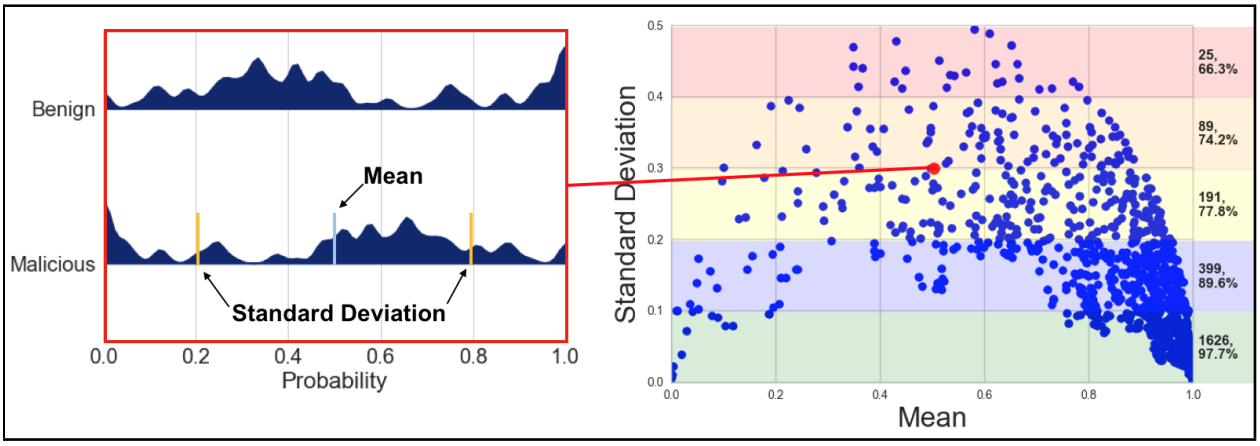

Figure 1: The left panel shows the ensemble probability distributions of predicting the specified URL. The right panel plots the means and standard deviations of the predictions of each sample. The color bars annotate the number of samples within each standard deviation range and the average accuracy with which they were classified.

predicting whether the given URL is malicious or benign. The distributions are constructed by sampling the data many times with replacement (bootstrap sampling) and training a classifier for each set of sampled data. These distributions show that small perturbations in the training data can lead to high variability in predictions. The right panel plots the means and standard deviations of the prediction distributions for each URL in the test data. The plots are overlaid with bars of various colors which annotate the average accuracy of the classifiers' predictions on the samples which fall within a range of the standard deviation.

For example, the green bar contains 1626 samples which have a standard deviation between 0 and 0.1 and have been classified with an accuracy of 97.7%; the red bar contains 25 samples which have a standard deviation between 0.4 and 0.5 and have been classified with an accuracy of 66.3%. In other words, the higher the standard deviation, the less likely the samples have been accurately classified. Therefore, we can have a greater degree of trust in the predictions which produce distributions with lower standard deviations and higher means.

Although uncertainty analysis uncovers information not provided by current methods, using it to increase 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. Similarly, we showed that the standard deviation of prediction probability distributions correlate with accuracy, but other measures such as highest density intervals may prove more informative. In general, application-specific analytic goals dictate the value of a given uncertainty measure. 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.

The unifying theme of research into uncertainty quantification for machine learning is to advance the data-to-decisions process by improving the user's ability to evaluate and trust a model's results. Recent advances have made both data and the patterns they contain more accessible. However, simply finding patterns of interest is not sufficient to support high consequence decision making. We must also assess confidence in results, which includes consideration of uncertainty.

Acknowledgments

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

Darling, M., Heileman, G., Gressel, G., Ashok, A., and Poornachandran, P. (2015). A lexical approach for classifying malicious urls. In *High Performance Computing & Simulation (HPCS), 2015 International Conference on*, pages 195–202. IEEE.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. *arXiv preprint arXiv:1602.04938*.