SAND2022-16321

SANDIA REPORT

Printed

Sandia National Laboratories

Preliminary Results for Using Uncertainty and Out-of-distribution Detection to Identify Unreliable Predictions

Justin E. Doak and Michael C. Darling

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185
Livermore, California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from

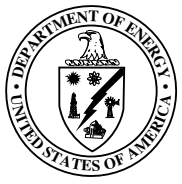
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods



ABSTRACT

As machine learning (ML) models are deployed into an ever-diversifying set of application spaces, ranging from self-driving cars to cybersecurity to climate modeling, the need to carefully evaluate model credibility becomes increasingly important. Uncertainty quantification (UQ) provides important information about the ability of a learned model to make sound predictions, often with respect to individual test cases. However, most UQ methods for ML are themselves data-driven and therefore susceptible to the same knowledge gaps as the models themselves. Specifically, UQ helps to identify points near decision boundaries where the models fit the data poorly, yet predictions can score as *certain* for points that are under-represented by the training data and thus out-of-distribution (OOD). One method for evaluating the quality of both ML models and their associated uncertainty estimates is out-of-distribution detection (OODD). We combine OODD with UQ to provide insights into the reliability of the individual predictions made by an ML model.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PRELIMINARY RESULTS FOR USING UNCERTAINTY AND OUT-OF-DISTRIBUTION DETECTION TO IDENTIFY UNRELIABLE PREDICTIONS

ML models are integrated into almost every imaginable application space with a few examples being cybersecurity, image analysis, self-driving cars, and recommender systems. Many of these applications are of high consequence, such as those pertaining to national security, and thus evaluating the quality of ML predictions is of utmost importance. Typically, ML models are evaluated using measures of performance averaged over samples in a fixed test set. Examples include the use of cross-validation and hold-out to estimate metrics such as classification error, F-score, precision, and recall. However, for critical decisions, we need to know whether a model's prediction is reliable for individual instances.

Many ML models will provide a probability estimate that indicates the confidence of a model in its prediction. Yet, these estimates may or may not be credible depending on how well the model was trained in the part of the feature space where a specific point resides. Therefore, there have been efforts to quantify the uncertainty of a model's estimates by considering sources of uncertainty in an ML pipeline [1].

Uncertainty estimates can evaluate the quality of a model's probability estimate. However, if an unseen sample lies in an area of the feature space where the training data is underrepresented, the model's probability estimate is often incorrect and the associated UQ score overly optimistic. OODD methods can identify these underrepresented areas.

Figure 0-1 shows three overlapping data sets, all three are sampled from Gaussian distributions. In this example, an ML model would have difficulty discriminating between the points that lie in the overlap regions. A model could also potentially be overly optimistic in ares with limited points if the sparsity is due to under sampling and not the behavior of the generating distribution. In real-world problems, such regions may exist though be hard to detect in high-dimensional spaces—even when large quantities of training data are available. For example, Moosavi-Dezfooli et al. [2] and Szegedy et al. [3] show that Deep learning models that often make high-confidence predictions on points that are significantly removed from the training distribution.

Figure 0-2 illustrates how UQ and OODD measurements can provide useful, complementary information. The points are classified using a logistic regression model trained only on data generated from distributions 1 and 2. We choose not to place samples from distribution 3 in the training set to represent a data class unknown to the model.

The OOD points are determined by a Local Outlier Factor (LOF) model [4] and the uncertainty scores are calculated using minimum prediction deviation (MPD) [5]. The data samples are shaded by their uncertainty scores with darker shades representing high levels of uncertainty. The OOD points are marked with Xs.

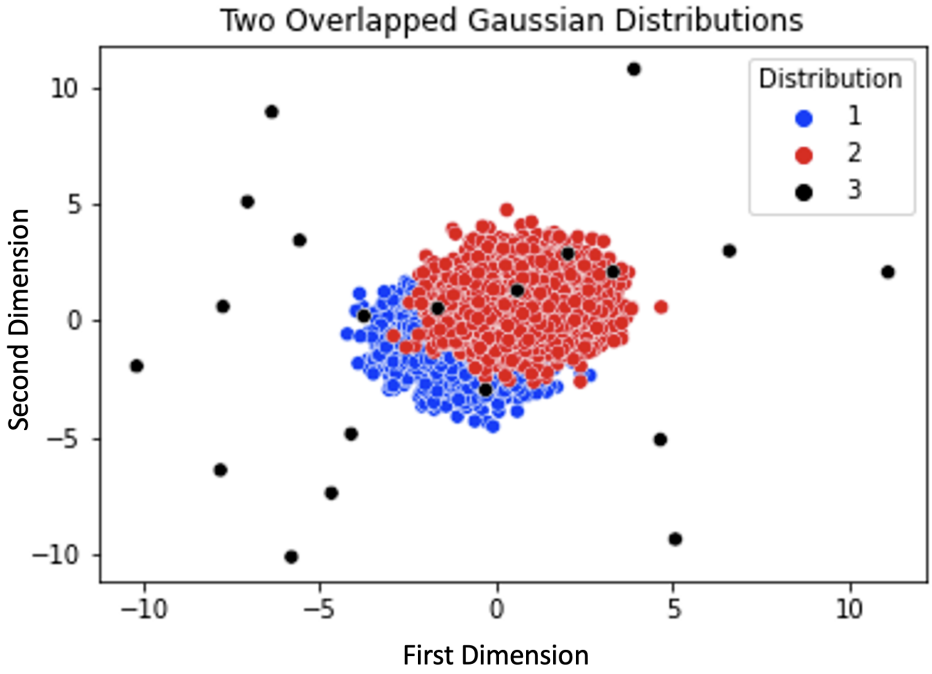


Figure 0-1. Three Data sets drawn from Gaussian Distributions.

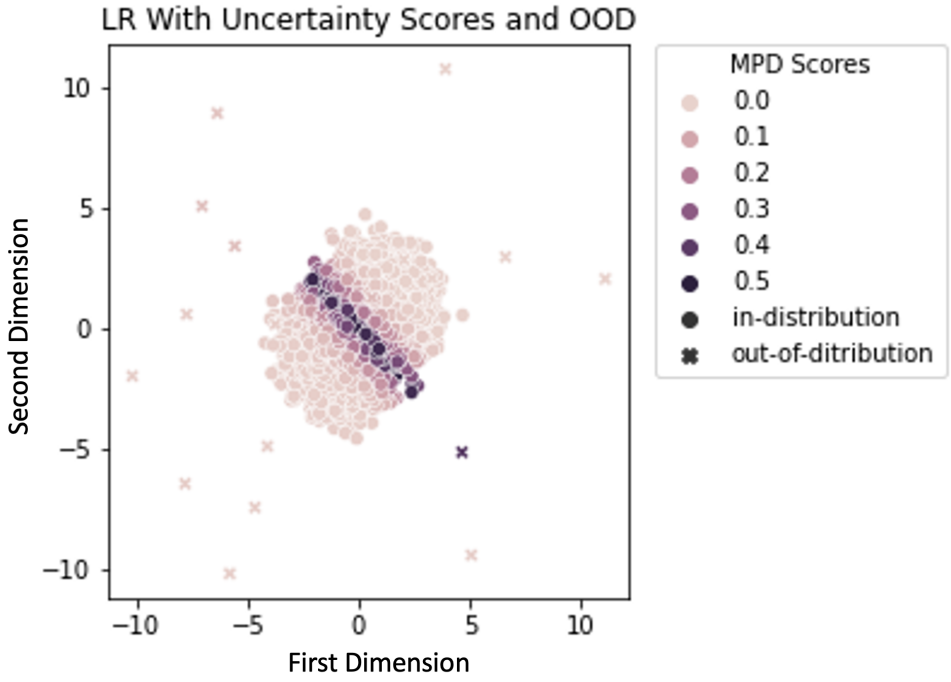


Figure 0-2. OOD points and uncertainty scores for a linear regression (LR) model classifying the raw data seen in Figure 1.

	Certain	Uncertain
In-distribution	1	2
Out-of-distribution	3	4

Table 0-1. Categories of created when applying MPD and OOD to ML predictions.

As expected, there are higher levels of uncertainty for points near the model's decision boundary. However, the MPD scores are low (certain) in the sparsely populated areas of the space. Depending on where these points lie, this level of certainty may be correct. However, the points that fall in the sparsely populated overlap regions, present high levels of certainty. Such samples would present higher levels of uncertainty if we merely sampled more data in their local regions. Thus, while uncertainty can be highly informative about the reliability of the model's prediction, the MPD scores can themselves be unreliable in low density regions and provide misleading information.

An OODD method, such as LOF, provides a means for detecting such spaces. While such issues are obvious in these simple 2-dimensional plots, such low density regions could exist in small pockets in higher dimensional space—even if the data is well sampled overall. The out-of-distribution samples might also belong to an unanticipated, unknown class for which the model is untrained

We can summarize the data shown in Table 0-1. Category 1 are points that are in-distribution (ID) and far from a decision boundary; these are both certain and ID. Category 2 are the points that are ID, but close to a decision boundary and thus the predictions are uncertain. Category 3, denoted by peach Xs in Figure 0-2, are OOD, but since they are also far from the decision boundary have low uncertainty scores. Category 4 points, also marked by peach Xs, are OOD and close to a decision boundary making them uncertain. We can improve ML outomes by identifying the points about which the model is uncertain (2 and 4) *and* that are not representative of the training data (3 and 4).

REFERENCES

[1] David J. Stracuzzi, Michael C. Darling, Maximillian G. Chen, and Matthew G. Peterson. Data-driven uncertainty quantification for multisensor analytics. In *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR IX*, volume 10635, page 106350N. International Society for Optics and Photonics, 2018.

[2] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[4] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[5] Michael C Darling. *Using Uncertainty To Interpret Supervised Machine Learning Predictions*. PhD thesis, University of New Mexico, 2019.

DISTRIBUTION

Email—Internal

Name	Org.	Sandia Email Address
CA Technical Library	8551	cateclib@sandia.gov

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.