Combining Aleatoric and Epistemic Uncertainties for Robust Healthcare Decision-Making

Zhiyin Lin, Sam Saarinen,[1] Michael L. Littman
Brown University
Providence, Rhode Island
[1]contact: sam_saarinen@brown.edu

Abstract

Automating and improving healthcare decision-making using machine learning is limited by the inability for practitioners to trust decisions handed to them by black-box models. Other work has explored making model decisions more interpretable, but this may reduce model accuracy, and often requires expertise to determine cases where the model is unreliable (such as when data are encountered that aren't represented in the training set). This paper presents an approach to unify two types of uncertainty in the context of regression problems, giving the novel model the ability to provide accurate per-instance confidence regions, without compromising on model accuracy. These techniques are evaluated in terms of likelihood of the true data under the confidence regions and ability to distinguish out-of-distribution test points. This paper also shows that the techniques presented robustly distinguish two types of uncertainty: uncertainty due to inherent variability (aleatoric risk) and uncertainty due to a lack of experience (epistemic uncertainty).

Introduction

As noted by previous work in human-machine collaboration in medicine (Jorritsma, Cnossen, and van Ooijen, 2015), establishing *appropriate* trust between healthcare practitioners and machine-learning models is critical to enabling automation of medical processes by models with human-level or super-human performance, thus improving outcomes and lowering costs. In cases of under-trust, the healthcare practitioners may unduly reject the prediction provided by the machine-learning model. In cases of over-trust, medical practitioners may default to an automated prediction that turns out to be imperfect. In both cases, a more appropriate level of trust may be established by accurately identifying and communicating a case-specific confidence in the model's prediction (Jorritsma, Cnossen, and van Ooijen, 2015). This paper considers regression problems and provides a set of techniques for augmenting arbitrary models such that, in addition to outputting a mean prediction, the model also outputs a variance. These outputs together can be interpreted as providing a Gaussian confidence region over where the true value is.

Evaluation Metrics

When a model outputs only a mean prediction for a regression, it is common to express its loss as the mean squared-error of the predictions with respect to a test set. How can a model that outputs means *and* variances be evaluated? Here, we take inspiration from the Bayesian modeling literature, where the goal is often to maximize the likelihood of a test set $T : X \to Y$ under a model M. Since the model's outputs are conditioned only on the current input, the outputs are independent of each other given these inputs and our goal becomes finding a maximum likelihood model $\operatorname{argmax}_M \prod_{i=1}^{n} M(x_i)[y_i]$. Such a model maximizes the product of the probability densities of each test output under the distribution given by the model for the corresponding test input. For notational and computational reasons, it is common to take the logarithm of the probability, so the objective becomes $\operatorname{argmax}_M \sum_{i=1}^{n} \ln M(x_i)[y_i]$. All of the regression models considered here are designed to output Gaussian distributions, so the intuitive nature of this performance metric will be further illustrated on this special case. An elementary application of calculus shows that, independent of the variance σ_i^2 given by the model, the model's score on a single test point is maximized when $\mu_i = y_i$. Similarly, independent of the mean prediction μ_i, the model's score on a single test point is maximized when $\sigma_i^2 = (\mu_i - y_i)^2$.

Uncertainty Modeling

Invariably, in the real world, there are discrepancies between expectation and result. Some deviations are unavoidable, due to apparent random chance. We will call the source of these mismatches **aleatoric risk** (from the Latin for dice-player: an aleator).[1] Other discrepencies are due to simple ignorance, and might be eliminated with additional experience or data. We will call these influences the **epistemic uncertainty** (from the Greek for knowledge: epistēmē). Interpretationally, these quantities roughly correspond to variance in the data and uncertainty in the mean, respectively. By naïvely assuming that these two sources of error are independent, we may add the variances together to get a measure of total uncertainty.

[1]Note that some documents favor the alternate form of "aleatory" for better alignment with the Latin root word. We favor "aleatoric" for better parallelism with "epistemic".

There are other forms of uncertainty that are outside the scope of this paper. For example, while epistemic uncertainty is able to deal with nonstationarity in the input distribution, it is not able to capture nonstationarity in the output distribution—changes to the output labels for previously seen inputs. None of these models is designed to deal with uncertainty due to the presence of other agents (whose behavior may be history-dependent and thus a source of nonstationarity over outputs). We note that these sources of temporal dynamism are distinct from simply dealing with temporal or sequential data.

Clinical Applicability

The goal of this work is to adapt models that produce output of the form "I predict the value μ" to models of the form "I predict the value $\mu \pm 2\sigma$". When the uncertainty is very high, the medical professional may confidently overrule the prediction provided by the model. But, if the model's uncertainty is low, the medical professional ought to consider the track record of the model or whether there is relevant information the model is not privy to.

Where techniques distinguish between aleatoric risk and epistemic uncertainty (as in this paper), this provides additional information for clinical practice. For example, the dataset used in the Experiments section involves regressing to a clinically relevant value 1 year in the future, which has a range of several hundred. On different points, the uncertainties (and their clinical ramifications) differ. For one point, the model predicts the value 175 ± 91, with 64% of the uncertainty being epistemic. This is a point where the mean estimate could be wrong by a large margin, mostly because the model simply does not have enough data on similar patients. In this case, the clinician may defer to their own experience and insights or seek more definitive testing and data collection. In contrast, the predicted value for another point is 213 ± 121, with 98% of the uncertainty being aleatoric. For this patient, there have been many similar patients in the data, so the mean value is quite accurate, but the outcome for the patient is still highly uncertain. In this case, it is best that the clinician use the mean estimate provided by the model and communicate with the patient what additional factors (such as behavioral choices) may affect their outcome. On a third point, the model is more precise, predicting 113 ± 48, with 77% of the uncertainty being epistemic. The model is already much more confident than average, and additional data on similar patients might further improve the quality of the prediction.

Contributions

The core contributions of this paper are threefold: first, a novel combination of fitted random priors, regression to variance, and uncertainty calibration using isotonic regression; second, experimental validation of the usefulness of this technique in accurately assessing uncertainty; and third, experimental evidence that the technique correctly distinguishes between aleatoric risk and epistemic uncertainty.

Method

All problems and techniques discussed below are in the context of regression problems.

Fitting Random Priors

There are several techniques that can be used for estimating epistemic uncertainty, but the one with possibly the best theoretical justification to date is the technique of fitting random priors, analyzed by Ciosek et al. (2020). Although the analysis is quite sophisticated, the algorithm is straightforward: simulate the core learning problem (consisting of data from the true function and a core model to fit that data) using a handful of synthetic learning problems. See fig. 1.

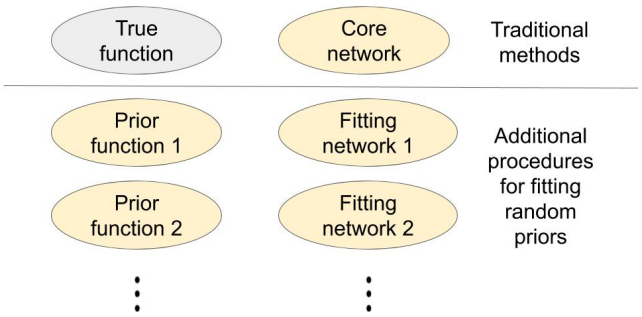

Figure 1: Fitting random priors creates simulated learning problems using randomly sampled prior functions to generate artificial datasets from the same inputs. If the distribution of prior functions is well-chosen, the performance of the fitting networks in matching the prior functions on some test input should be correlated with the performance of the core network on that same test input.

For each simulated problem, draw a "prior function" from a distribution over functions (for example, randomly initialized neural networks) that (hopefully) assigns non-negligible probability density to something closely approximating the true function (that is, the prior functions have comparable complexity and characteristics to the true function). Generate synthetic datasets by feeding the *true inputs* to the synthetic prior functions. Then, train a "fitting function", which has the same structure as the core model, for each synthetic dataset.

At inference time, a test input can be fed to both the prior functions and their respective fitting functions, and these can provide an estimate of the error of the core model with respect to the true output. Since all of the models were trained on the same inputs, the models will be tightly constrained by the data in areas near training inputs, and relatively unconstrained elsewhere. In fact, Ciosek et al. (2020) prove that these variance estimates are conservative with respect to a Bayesian estimate of error.

Regression to Variance

The idea of regressing to the distance between a model's predictions and the true values in a dataset has been around

for a while, whether in gradient boosting or heteroskedastic regression (Fan and Yao, 1998). Here, it is a useful technique for modeling aleatoric risk. If the dataset is separated into training, validation, and test sets, we perform regression to variance on the validation set once the core model has been trained on the training set. More specifically, we measure the squared error between the predictions and true values and subtract out the epistemic uncertainty estimated by the fitted random priors. Note that in practice this may result in negative regression targets. In our experiments, we naïvely clipped these values at zero. This change may bias our uncertainty estimates to be overlarge, but that is preferable in the application domain to the possibility of providing variance estimates that are too small, zero, or even negative.

Uncertainty Calibration

The regression to variance for modeling aleatoric risk may be noisy and inaccurate. Therefore, we make use of a robust technique for improving the accuracy and generalizability of these values from Kuleshov, Fenner, and Ermon (2018). This technique uses isotonic regression (the hypothesis class is piece-wise monotonic functions) to model the one-dimensional map from the predicted aleatoric variances to the true targets. In the case that the regression to variance was highly accurate, this step does not harm the precision of the model, and in the case that the values were inaccurate, calibration produces accurate values at the expense of sharpness.

To illustrate the versatility of the calibration procedure, we calibrate a poor approximation of case-specific uncertainty (1 divided by the distance to the nearest training point) to accurate values for a 1-Nearest Neighbor model on the UCI Wisconsin Breast Cancer (Diagnostic) Dataset (Street, Wolberg, and Mangasarian, 1993). See fig. 2.

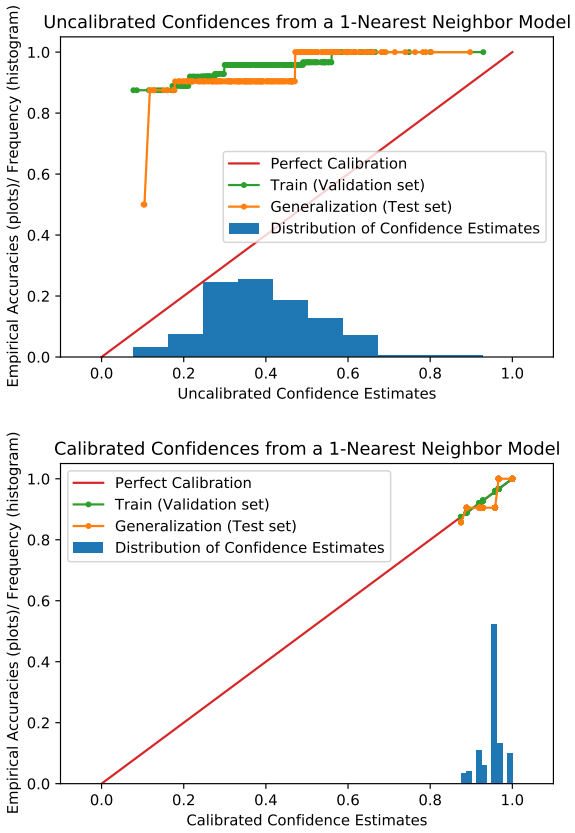

Figure 2: Data distribution and accuracy predictions pre- and post-calibration using isotonic regression. The calibration on the validation set (green) generalizes well to the test set (orange).

Experiments

We empirically test three claims. First, that combining epistemic uncertainty and aleatoric risk models gives us better performance in terms of the log-likelihood of the true labels under the output distributions. Second, we test that the combined technique is sensitive to perturbations of the data in an out-of-distribution experiment. Finally, we test whether aleatoric risk and epistemic uncertainty are usefully distinguished by our technique.

Dataset

For the experiments that follow, we make use of the Diabetes Progression Dataset from Efron et al. (2004), which can be viewed as a regression problem with 442 total datapoints and 10 features. This dataset was selected due to its ready availability and greater complexity relative to other commonly used medical datasets.

Ablation Study on Log-Likelihood Performance

In fig. 3, we compare the scores of 3 different models: the model combining epistemic uncertainty and aleatoric risk estimates; epistemic uncertainty *only* using fitted random priors; and aleatoric risk estimates *only* by regression to variance and uncertainty calibration using value clipping. Note that all models used the same mean predictions and differ only in the uncertainty estimates they provide. Core model, prior functions, fitting functions, and regression to variance all used a shallow 3-layer-deep fully-connected neural-network architecture with layers 100 nodes wide; the prior functions were sampled using the Kaiming normal weight initialization (He et al., 2015) with whitened inputs and rescaled output weights to match the natural range of the training data.

Out-of-Distribution Experiment

While performance on the test set of a dataset is compelling, there are significant concerns around robustness to distributional shift or to data that is not well-represented in the training set. This kind of robustness can be evaluated by using an out-of-distribution test; performance on the test data is compared to performance on test data from another dataset. In the medical domain, it is difficult to find datasets with identical input and output structure and different statistical input distributions, so we synthesize an out-of-distribution test set by perturbing the values of our test set in each fea-

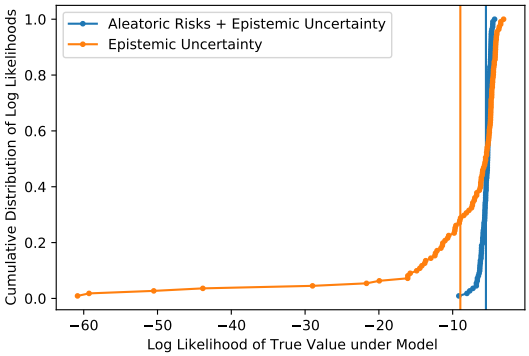

Figure 3: The Cumulative Distribution of Scores (Log Likelihood of True Value under Model) for Combined Method (Blue) and Epistemic-Only Method (Orange) on the Diabetes Dataset. Mean performance scores are represented as vertical bars. (Further to the right is better.) Aleatoric-Only data could not be plotted due to some 0-uncertainty estimates producing infinitely negative performance values.

ture in either direction by up to 20% of the feature's value range (for an average 10% perturbation). As seen in fig. 4, the distribution of uncertainties over the test set is shifted for the out-of-distribution points, resulting in a greater average uncertainty even for a relatively mild perturbation to the data.

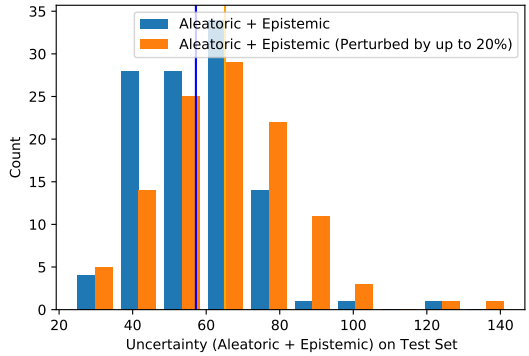

Figure 4: Out-of-Distribution Test: The Distribution of Combined Uncertainty of Original Samples and Perturbed ($\leq 20\%$) Samples on Test Set of the Diabetes Dataset

Distinguishing Aleatoric Risk and Epistemic Uncertainty

It is conceivable that a combination of uncertainty modeling techniques could benefit from a kind of heterogeneous ensemble effect, even if the techniques were not measuring semantically different sources of error. To test this idea, we measure the average estimated aleatoric risk and epistemic uncertainty predicted by our combined model as we increase the number of training points available to it. The results can be seen in fig. 5.

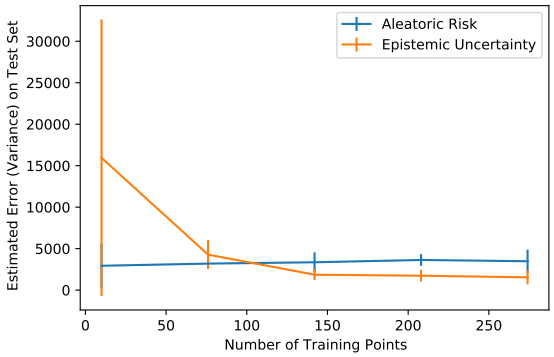

Figure 5: Number of Training Points versus Estimated Error (Variance) on Test Set of the Diabetes Dataset. Confidence intervals represent the distribution across 5 independent trials with different initializations and randomly selected training points.

As the number of training points increases, the aleatoric risk remains almost constant while the epistemic uncertainty decreases greatly, demonstrating that our method robustly distinguishes between uncertainty due to statistical randomness and uncertainty due to limited knowledge.

Conclusions and Future Work

This paper has presented a promising direction in increasing the clinical relevance of machine-learning models (black-box or otherwise) by providing accurate case-by-case uncertainty estimates that distinguish between two forms of uncertainty. Future work should consider and compare to other methods for modeling uncertainty, such as ensemble methods, and may also consider how to adapt these techniques to classification problems.

References

Ciosek, K.; Fortuin, V.; Tomioka, R.; Hofmann, K.; and Turner, R. 2020. Conservative uncertainty estimation by fitting prior networks. In *International Conference on Learning Representations*.

Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R.; et al. 2004. Least angle regression. *The Annals of statistics* 32(2):407–499.

Fan, J., and Yao, Q. 1998. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 85(3):645–660.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Jorritsma, W.; Cnossen, F.; and van Ooijen, P. M. 2015. Improving the radiologist–cad interaction: designing for appropriate trust. *Clinical radiology* 70(2):115–122.

Kuleshov, V.; Fenner, N.; and Ermon, S. 2018. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*.

Street, W. N.; Wolberg, W. H.; and Mangasarian, O. L. 1993. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, 861–870. International Society for Optics and Photonics.