DSSV 2023 Data Science, Statistics & Visualisation
ECDA 2023 European Conference on Data Analysis

Reliable uncertainty estimation via proper scores

Florian Buettner[1,2]

[1]Goethe University Frankfurt, Departments of Informatics and Medicine, Frankfurt, Germany
[2]German Cancer Consortium (DKTK) and German Cancer Research Center Heidelberg, Germany

With model trustworthiness being crucial for sensitive real-world applications, practitioners are putting more and more focus on improving the uncertainty awareness of deep neural networks. This raises the need to quantify and improve the quality of predictive uncertainty, ideally via a dedicated metric. An uncertainty-aware model should give probabilistic predictions which represent the true likelihood of events depending on the very prediction. To quantify the extend to which this condition is violated, calibration errors have been introduced and post-hoc recalibration methods are commonly used to improve them. However, estimators of calibration errors are usually biased and inconsistent. In practice, this means that common calibration estimators are highly sensitive w.r.t. the test set size. In recent work [1], we demonstrate that for commonly used estimators, the estimated improvement of recalibration methods is heavily biased and becomes monotonically worse with fewer test data. We introduce the framework of **proper calibration errors**, which gives important guarantees and relates every calibration error to a proper score. We can reliably estimate the improvement of an injective recalibration method w.r.t. a proper calibration error via its related proper score.

The most common way to measure predictive uncertainty is via the predicted confidence. While this tends to work well for in-domain samples, these estimates are unreliable under domain drift and restricted to classification. A core principle behind the success of modern machine learning approaches are loss functions (usually derived from a proper score), which are used to optimize and compare the goodness-of-fit of predictive models. Proper scores are a common occurrence as loss functions for probabilistic modelling since their defining criterion is to assign the best value to the target distribution as prediction. As alternative to confidence scores, proper scores can be used directly for estimating the uncertainty of a prediction as a composite measure: Typical loss functions, such as the Brier score or the negative log-likelihood, capture not only predictive power (in the sense of accuracy) but also predictive uncertainty. However, for such loss functions, it is not clear how we can decompose them such that a specific component capturing predictive uncertainty alone arises.
In recent work [2], we discover the Bregman Information as a natural replacement of model variance via a bias-variance decomposition for strictly proper scores. The Bregman Information generalizes the variance of a random variable via a closed-form definition based on a generating function. Via Bregman Information, we give novel formulations for decompositions of exponential families and the classification log-likelihood in the logit space. We show how ensemble predictions marginalize out a specific source of uncertainty and propose a general way to give confidence regions for predictions. Finally, we showcase experiments on how typical classifiers differ in their Bregman Information and demonstrate that the Bregman Information can be a more meaningful measure of out-of-domain uncertainty compared to the confidence score.

Taken together, in this talk, I will present recent work on quantifying predictive uncertainty and its quality from two different angles: First, I will introduce proper calibration errors as a summary metric to quantify the quality of a model's confidence scores [1]. Next, I will demonstrate how the a new general bias-variance decomposition leads to the Bregman Information as a meaningful measure of out-of-domain uncertainty [2].

References

[1] S. G. Gruber and F. Buettner, "Better uncertainty calibration via proper scores for classification and beyond," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8618–8632, 2022.
[2] S. G. Gruber and F. Buettner, "Uncertainty estimates of predictions via a general bias-variance decomposition," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* (F. Ruiz, J. Dy, and J.-W. van de Meent, eds.), vol. 206 of *Proceedings of Machine Learning Research*, pp. 11331–11354, PMLR, 25–27 Apr 2023.