Point Prediction Uncertainty

At ING-bank we tend to evaluate machine learning models based on group-level metrics such as precision-recall, F1 or AUCROC, evaluated on an entire dataset. However, in many business settings knowing the prediction uncertainty on individual data points is critical, as a model should refrain from predicting when there is high uncertainty. You will work on an innovative method to find the prediction uncertainty on individual data points, and try to apply this to some models used at ING.

Problem statement

Let $S = \{(X_1, y_1), ..., (X_n, y_n)\}$ be a set of n i.i.d. (input, target) pairs that follow an unknown distribution $\mathcal{D}$. Where $y \in \{0, 1\} \subset \mathbb{Z}$ and $X \subset \Phi \in \mathbb{R}^d; \Phi \sim F_\Phi$.

Our goal is to learn a function $h \in \mathcal{H}; h : \mathbb{R}^d \to \{0, 1\}$ that maximises (minimises) the target metric e: $\max_{x,y} \mathbb{E}[e(h(x), y)]$. A properly calibrated model predicts $P(y_i = C_j | X_i)$, the probability that observation y_i belongs to class C_j.

Yet a model can predict a score of 0.5 for an observation with very low uncertainty, meaning the model is very sure the probability of either class is equal. Existing model confidence scores are based on evaluating a hold-out set over (many) permutations of the classifier. Evaluating the variability of the predicted class for these points over the permutations gives a measure of uncertainty. Various flavours of model confidence scores exist(Mandelbaum and Weinshall 2017; Lakshminarayanan, Pritzel, and Blundell 2017; Gruber and Buettner 2023).

However, all these methods fail in one regard: the further an observation lies from the decision boundary the higher the model confidence tends to be. This confidence may not be warranted if the model has seen a few or even no observations for a particular region.

Many machine learning models are essentially curve fitters which can exhibit erratic behaviour for out-of-domain observations. This problem quickly becomes worse in high dimensional feature spaces which are often heterogeneous in density and can result in highly non-linear decision boundaries. What is needed is a unified measure of uncertainty that not only incorporates the model's uncertainty but also the statistical uncertainty inherent to having only a limited sample.

To summarise, let $\mathcal{X} \sim F_\Phi \mid \mathcal{X} \cap X = \emptyset$ be a set of inputs for which we want to estimate the corresponding target. For a given observation $x \in \mathcal{X}$ there will be the uncertainty for the result of the classifier, which we can divided it into mainly 2 parts, the out of distribution classifier uncertianty and rest:

- σ_{OODC}: scince we trained the classifier on the training set, so the training set is all the information we know about the data distribution, while the situation that the new data(test set) from the distribution looks like an

outlier for the training set, so it's like the probability that the feature of the new data point is not captured by the training set.

$$\begin{aligned}\sigma_{\text{OODC}} &= \mathbf{P}\left(x \nsim \widehat{F}_X \mid X, x \in \mathcal{X}\right) \\ &= 2|0.5 - \widehat{F}_X(x)|\end{aligned}$$

For this method to work one would need to have a way to estimate the CDF (F_X) of the distribution from which X was drawn. If one can estimate the PDF but not necessarily the CDF it can be approximated using for example:

$$\widehat{F}_X(x) \approx \text{percentile}(\hat{f}_X(X))$$

When we have a learning algorithm, in real life cases, we are going to train the model on a training set, while the extant of the training set capturing the feature of the distribution has variance, and the training environments, including software parameter initialization, which leads to the variance of the point predictions by models. This is the uncertainty of the classifier generating process itself.

Given a learning algorithm and given a set of training sets and training environments, it can generate a hypothiesis class $\mathcal{H}$, where

$$\sigma_{\text{classifier}} = var(h(x) \mid h \in \mathcal{H})$$

References

Gruber, Sebastian, and Florian Buettner. 2023. "Uncertainty Estimates of Predictions via a General Bias-Variance Decomposition." In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, edited by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, 206:11331–54. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v206/gruber23a.html.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." *Advances in Neural Information Processing Systems* 30.

Mandelbaum, Amit, and Daphna Weinshall. 2017. "Distance-Based Confidence Score for Neural Network Classifiers." *arXiv Preprint arXiv:1709.09844*.