Towards a continuous evaluation of calibration

Nicolas POSOCCO, Antoine BONNEFOY
{nicolas.posocco, antoine.bonnefoy}@euranova.eu

Contributions

We propose a **continuous version** of the **reliability plot**, which allows the introduction of a **more robust Expected Calibration Error** (ECE) estimator. We define the notion of local calibration error (LCE) in that context, and propose a **new calibration method**.

For simplicity, the binary calibration case (calibration with respect to the positive class in a 2-class setting) is the only one shown below. Extentions to confidence and class-wise settings are covered following the extension for the discrete approaches [2].

We tackle the calibration of a binary classifier, that uses a dataset ($X_i \in X$ for $i \in [1, N]$ and associated $Y_i \in \{0, 1\}$) to learn a classification function building upon a decision score $s : X \to [0, 1]$. Legacy methods use a binning scheme, let thus B_m be the set of indices of samples whose scores for the positive class fall into the intervall $m = (\frac{m-1}{M}, \frac{m}{M}]$. Let, for any sample X_i , $b(X_i)$ be the set of all samples that fall into the same bin as X_i. Finaly, let $f_{s(X)}$ represent the probability density function of the scores given by the model for the positive class (estimated via KDEs in our implementations) and K be a convolution kernel.

Reliability diagram : from discrete to continuous

We rewrite the legacy formula for the ECE to **isolate the contribution of each sample**. **Local kernel methods** are then resorted to introduce the **Local Calibration Error (LCE)**, which indicates **how scores should be altered** to be properly calibrated :

$$\forall\ s \in [0,1], LCE(s) = \big[K * \sum_{i \in [1,N]} \frac{1}{N}(\mathbb{1}_{Y_i=1} - s(X_i))\ \delta_{s(X_i)}\big](s)$$

This further enables the production of a **continuous reliability plot**, that we call **reliability curve** (RC), as well as **a new estimator for the Expected Calibration Error** :

$$ECE \triangleq \mathbb{E}\big[|P(Y=1 \mid s) - s|\big]$$

$$\sum_{i \in [1,N]} \frac{1}{N} \Big| \frac{1}{N_{b(X_i)}} \sum_{j \in b(X_i)} (\mathbb{1}_{Y_j=1} - s(X_j)) \Big|$$

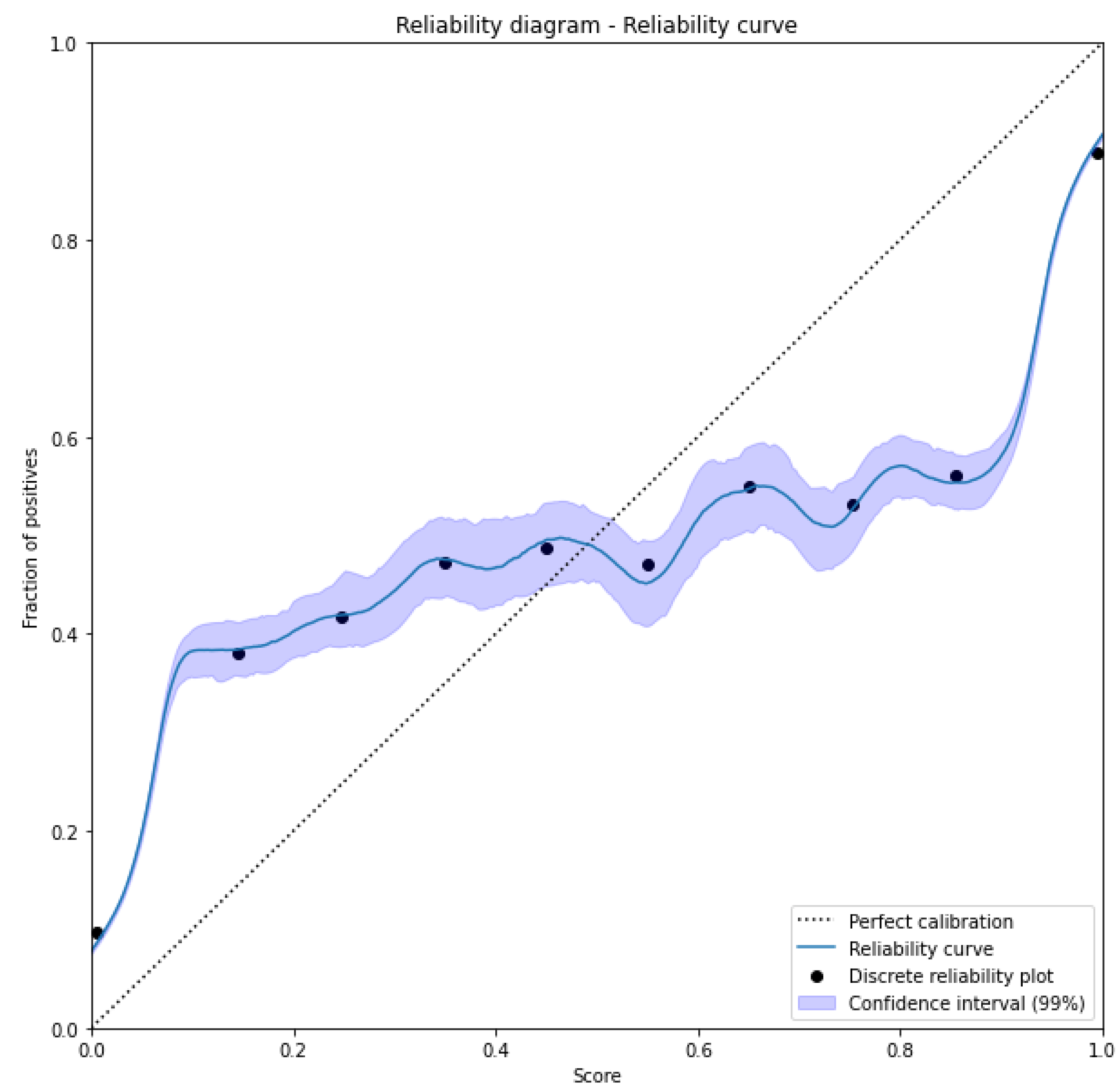
Legacy binning-based estimator [3]

$$\int_{[0,1]} \big|\ LCE(s) \big|\ f_{s(X)}(s)\ ds$$

Kernel-inspired estimator

$$\forall\ s \in [0,1], RC(s) = LCE(s) + s$$

Quantiles trajectories (median and percentiles of interest) computed via bootstrapping of the calibration evaluation set, allow a **robust estimate** of the calibration trajectory, as well as **confidence intervals** surrounding the local estimated calibration error.



Comparison of the discrete reliability diagram (10 bins) with the introduced reliability curve for a Naïve Bayes model trained on a generated dataset, with a bootstrapped confidence interval

Introduction of a new calibration method

The local calibration error **can be used as a way to calibrate scores**. In order to do so, we apply the same procedure **as the one used in the naïve histogram binning calibration** method [1], yet we use the LCE function instead of the binned error of the histogram binning.

$$\forall\ s \in [0,1], cal(s) = s - LCE(s)$$

Using local calibration for prediction uncertainty evaluation

Local ECEs can be given to the user of the classifier on top of the class probabilities, giving him **a relevant grasp of the error on the estimates of these probabilities**.

	Class	Posterior probability estimates	Uncertainty on Posterior probability estimates
Classifier	✓		
Calibrated classifier	✓	✓	
Calibrated classifier + LCE	✓	✓	✓

Future works

These approaches can be generalized to a **multidimensional** score space, and thus can be used to compute a **multiclass-ECE estimator**, with more flexibility than currently available options, thus helping the relieve of the computation constraint, which forces us to use the weaker notion of class-wise-ECE.

Empirical evaluation of the proposed calibration method is still in progress.

References

0. Jochen Bröcker and Leonard Smith. 2007. Increasing the reliability of reliability diagrams. *Weather and Forecasting - WEATHER FORECAST*, 22, 06.
1. Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. - ICML, pp. 609–616.
2. Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling : Obtaining well-calibrated multiclass probabilities with dirichlet calibration.
3. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. On calibration of modern neural networks.