SIGEcom Exchanges Annotated Reading List: Multiclass Calibration

RABANUS DERR
University of Tübingen, Tübingen AI Center
and
JESSIE FINOCCHIARO
Boston College

ML model evaluation often takes one of two main approaches: *risk minimization*, associated with "high accuracy" or *calibration*, meaning that predictions are "trustworthy" and can be interpreted from a probabilistic lens. There is an extensive line of work which has studied the relationship between risk minimization and calibration, mostly focusing on the binary outcome setting. Even in the binary setting, there are a variety of proposed calibration metrics which non-trivially interact. In the multiclass label setting, the choices to be made are even more complex and particularly there are different semantics for different notions. Here, we briefly present an annotated reading list reviewing some of the proposed definitions and their relationships.

Introduction

The classical understanding of calibration is that a prediction is calibrated if among the days on which the probability of rain was forecasted is p, the average number of rainy days is p, e.g., [DeGroot and Fienberg 1983]. This is formalized by saying a predictor $f : \mathcal{X} \to [0, 1]$ is *calibrated* if

$$\Pr[Y = 1 | f(X) = p] \approx p \qquad \forall p \in \mathbf{im}(f) \ . \tag{1}$$

In the binary setting, Equation (1) satisfies two nice desiderata of trustworthiness:

a. *self-referential*: predicting p means the true probability of the positive label is p, and
b. it *precisely estimates the loss* incurred by a decision maker using the prediction.

The first of the two desiderata is relatively self-explanatory. The second requires more context in terms of a decision maker. We understand a decision maker as an agent equipped with a loss function mapping from actions and outcomes to scores. The decision maker can *precisely estimate the loss*, if the given prediction allows the decision-maker to precisely compute the expected loss for a taken action in comparison to the actual incurred loss for the same taken action. It is a matter of some computations to show that Equation (1) fulfills this requirement, if the decision maker orients its action only based on the prediction, e.g., [Zhao et al. 2021].

Vanilla calibration (Equation (1)) as provided above gives trustworthiness desiderata when the prediction task is binary, e.g., rain or no rain. If the set of considered possible outcomes grows, e.g., rain, sun, cloudy, i.e., $\mathcal{Y}$ is non-binary but finite, then

Authors' addresses: rabanus.derr@uni-tuebingen.de, finocch@bc.edu

the formal definitions of calibration, and their achieved trustworthiness desiderata need reconsideration.

A naïve extension of Equation (1) following [Kull et al. 2019] to,

$$\Pr[Y = y | f(X) = p] \approx p_y \qquad \forall y \in \mathcal{Y}, \forall p \in \mathbf{im}(f) \ , \tag{2}$$

where $f : \mathcal{X} \to \Delta(\mathcal{Y})$ and p_y denotes the y-component of the probability vector p, is problematic. The sample complexity of calibration grows in the number of conditional probabilities considered (cf., the "for all"-quantifier over $\mathbf{im}(f)$). Hence, Equation (2) called *full distribution calibration* results in an exponential blowup in sample complexity. Scholarship essentially suggests two ways around the problem.

One solution is to consider only a relevant subset of conditions usually defined by the downstream decision makers. That is, calibration is aimed to be achieved only around decision boundaries of the decision makers. This requirement can be further weakened by focusing only on the loss estimates instead of the action recommendations made through the predictions; this notion is called *decision calibration*.

One alternative proposal is that calibration is generalized and achieved for relevant summary statistics, such as the mean or class-wise distributions. The variety of definitions for multiclass calibration have been mainly proposed with a focus on computationally constructing predictors with small statistical complexity. This final semantic generalization of calibration is called *property calibration*.

Interestingly, recent work shows that the semantic notions of decision calibration and property calibration have a strict separation in the multiclass setting. In the following reading list, we try to reflect the approaches to multiclass calibration. The list is not meant to be exhaustive, but rather should demonstrate differing notions.

Reading List

Distribution calibration

(1) Kull et al. [2019] propose a "natively multiclass calibration" method. In doing so, they offer a clean introduction of a natural extension of Equation (1) to multiple classes Equation (2), albeit with exponential computational and sample complexity. They further relate the full calibration extension to multiple classes to other suggestions made in literature which demand for class-wise calibration, respectively best-class calibration.

(2) Gopalan et al. [2024] discuss the fragile relationship between definitions of calibration in multiclass settings and the need to balance (a) sample complexity, (b) computational complexity, and (c) robustness of calibration notions. To this end, they propose the metric of *smooth projected calibration error* for multiclass settings and analyze the sample and computational complexities of attaining a predictor with low calibration error in this sense. Their work focuses on distributional predictors which might be used for binary subset selection problems as the downstream decision.

Decision calibration

(3) Zhao et al. [2021] propose a definition of calibration for multiclass settings that is motivated by the usefulness of predictions for downstream decision-

making. Motivated by *loss outcome indistinguishability*, they say a predictor is *decision calibrated* if a loss-minimizing decision-maker attains near-optimal loss by trusting a model's probabilistic predictions. Importantly, they require the action space of the decision-maker to be polynomially bounded in the number of classes, which stands in contrast to distribution calibration, where action spaces are not considered, and distance is simply measured from a predicted to observed distribution.

(4) Fröhlich and Williamson [2024] study the evaluation of imprecise forecasts. That means that forecasts are not single probability distributions, but sets of probability distributions. The paper focuses on loss functions and calibration as evaluation metrics. Even though imprecise forecasting is a rather exotic topic, they are a perfect ground to study the meaning of evaluations. In particular, the authors argue that a distinction between the goal of trustworthy uncertainty estimates and the goal of recommending favorable actions is required for imprecise forecasts, but not for precise ones.

(5) Noarov et al. [2025] study the computation of sequential predictions which fulfill a polynomial number of unbiasedness conditions. In particular, the authors can guarantee sample efficient predictions which are calibrated in a multiclass setting. This is achieved by putting focus on decision relevant conditions, i.e., the unbiasedness conditions can be defined through action policies by the decision maker.

Property calibration

(6) Jung et al. [2021] provide computational methods to predict such that "moment multicalibration" is met. Multicalibration is the extension of calibration as in Equation (1) to simultaneously hold on a set of subgroups $\mathcal{G} \subseteq 2^{\mathcal{X}}$. Moment calibration refers to the understanding of Equation (1) as a moment matching task. That is, the expected value of the output should be equal to the prediction, conditioned on the prediction. The authors extend this moment matching idea beyond the first order moment to higher order moments, including the variance.

(7) Noarov and Roth [2023] generalize the notion of moment multicalibration [Jung et al. 2021] into Γ-multicalibration for continuous, real-valued properties.[1] In this work, the authors propose a definition of Γ-multicalibration, which intuitively suggests that, conditioned on a model f predicting a property value r (e.g., predicting *the mean* is r), the property value should be approximately r. They characterize the set of "calibratable" properties Γ and present batch and online algorithms to Γ-multicalibrate a given predictor $f : \mathcal{X} \to \mathbb{R}$ for a set of labels $\mathcal{Y} \subseteq \mathbb{R}$ for continuous, real-valued Γ.

(8) Gneiting and Resin [2023] take a statistical perspective on forecast evaluation and model diagnostics. The aim of the paper is to develop calibration for real-valued forecasts. In particular, the authors suggest the notion of T-calibration using the concept of identifiability for properties[2]. Their definition is, up to

[1]Properties are functions $\Gamma\colon \Delta_{\mathcal{Y}} \to \mathcal{R}$ mapping distributions over labels to descriptive statistics, such as the mean $\Gamma(p) = \mathbb{E}_{Y\sim p}[Y]$, or mode $\Gamma(p) = \arg\max_y p_y$.

[2]An identifiable property is a property $\Gamma\colon \Delta_{\mathcal{Y}} \to \mathcal{R}$ such that there exists a function $\nu\colon \mathcal{Y}\times\mathcal{R} \to \mathcal{R}$ with $\Gamma(p) = \gamma \Leftrightarrow \mathbb{E}_{Y\sim p}[\nu(Y,\gamma)] = 0$. For instance, the mean or the median.

minor details, equivalent to Γ-calibration from [Noarov and Roth 2023]. The developments within this paper and [Noarov and Roth 2023] have, even though strongly related, happened independently.

Relationships between semantic notions

(9) Derr et al. [2025] examine the works above (among others), and proposes the semantic clusters of distribution calibration, property calibration, and decision calibration to characterize the differences and relationships between the semantic notions. In the binary setting, Derr et al. [2025] shows the semantic notions are equivalent, but establishes that decision calibration and property calibration are strictly separate in multiclass settings.

REFERENCES

DeGroot, M. H. and Fienberg, S. E. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician) 32,* 1-2, 12–22.

Derr, R., Finocchiaro, J., and Williamson, R. C. 2025. Three types of calibration with properties and their semantic and formal relationships.

Fröhlich, C. and Williamson, R. C. 2024. Scoring rules and calibration for imprecise probabilities.

Gneiting, T. and Resin, J. 2023. Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics 17,* 2, 3226 – 3286.

Gopalan, P., Hu, L., and Rothblum, G. N. 2024. On computationally efficient multi-class calibration. In *Proceedings of Thirty Seventh Conference on Learning Theory*, S. Agrawal and A. Roth, Eds. Proceedings of Machine Learning Research, vol. 247. PMLR, 1983–2026.

Jung, C., Lee, C., Pai, M., Roth, A., and Vohra, R. 2021. Moment multicalibration for uncertainty estimation. In *Proceedings of Thirty Fourth Conference on Learning Theory*, M. Belkin and S. Kpotufe, Eds. Proceedings of Machine Learning Research, vol. 134. PMLR, 2634–2678.

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Vol. 32. Curran Associates, Inc.

Noarov, G., Ramalingam, R., Roth, A., and Xie, S. 2025. High-dimensional prediction for sequential decision making. In *Forty-second International Conference on Machine Learning*.

Noarov, G. and Roth, A. 2023. The statistical scope of multicalibration. In *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds. Proceedings of Machine Learning Research, vol. 202. PMLR, 26283–26310.

Zhao, S., Kim, M., Sahoo, R., Ma, T., and Ermon, S. 2021. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Vol. 34. Curran Associates, Inc., 22313–22324.