Calibrated Confidence Measures With Guarantees

Volodymyr Kuleshov
Department of Computer Science
Stanford University
Stanford, CA 94305

Stefano Ermon
Department of Computer Science
Stanford University
Stanford, CA 94305

Abstract

In medical and healthcare applications of machine learning, assessing the confidence of predictions is often as important as obtaining high accuracy. Here, we present an algorithm for constructing *calibrated* confidence estimates, which are probabilities that match empirical event frequencies in the long run. Unlike previous approaches, our confidence scores are generated in an online fashion and are guaranteed to be calibrated on any input (even in adversarial settings). We analyze the convergence rate of our method and validate it empirically on the task of predicting diabetes from genomic data.

In medical and healthcare applications of machine learning, assessing the confidence of predictions is often as important as obtaining high accuracy. A natural way of expressing uncertainty is via *calibrated* probability estimates, which match the true empirical frequencies of the occurrences of an event. For example, if we predict 60% chance of disease for 100 patients, then we expect that about 60 patients will eventually become sick.

The importance of calibrated confidence estimates is well-understood in medical diagnosis [1], language understanding [2], speech recognition [3], meteorology [4], econometrics [5], and psychology [6]. In machine learning and statistics, there exist many popular techniques that *recalibrate* the scores used for classification by existing models, including SVMs, neural networks, and decision trees in a *batch* setting. For binary classification problems, Platt scaling [7] and isotonic regression [8] are popular heuristics that typically produce good results in practice [8].

Here, we introduce recalibration methods in the *online* and *adversarial* settings, which are more suitable to healthcare problems like medical diagnosis. Previous approaches in this setting [9, 10] were motivated by game theory and are not applicable to standard prediction tasks: for example, they do not consider covariates x_t (e.g. genomic data affecting outcome y_t); moreover, they provide no guarantees on the forecast *sharpness* [4]). For example, predicting 0.5 on a sequence 01010.. formed by alternating 0s and 1s is considered a valid calibrated forecaster.

Here, we introduce a new problem called *online recalibration* that is suitable for machine learning applications, allowing for covariates x_t and encouraging sharp predictions. We propose algorithms for this problem and analyze their convergence properties theoretically as well as empirically on the task of predicting diabetes from genomic data.

Online Learning and Calibration We place our work in the classical framework of online optimization [11, 12]. At each of a series of time steps $t = 1, ..., T$, Nature chooses $(x_t, y_t) \in \mathcal{X} \times \{0, 1\}$ and reveals features $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ to the forecaster F. F makes a prediction $p_t^F = \sigma(w_{t-1} \cdot x_t) \in [0, 1]$ where $w_{t-1} \in S$ is a parameter vector, $S \subseteq \mathbb{R}^d$ is a convex set, and $\sigma : \mathbb{R} \to \mathcal{P}$ is a *transfer function*. Nature then reveals outcome y_t. F incurs a loss of $\ell(p_t^F, y_t)$, where $\ell : [0, 1] \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function that is convex in p_t^F for all y_t. Finally, the forecaster sets a new w_t.

Our goal in this setting is to produce calibrated predictions p_t^F. Let F^{cal} be a forecaster making predictions in the set $\{\frac{i}{N} \mid i = 0, ..., N\}$, where $1/N$ is called the *resolution* of F^{cal}. We define calibra-

tion error as $C_T = \sum_{i=0}^{N} \left(\rho_T(i/N) - \frac{i}{N}\right)^2 \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\{p_t^F = \frac{i}{N}\}}\right)$, where $\rho_T(p) = \frac{\sum_{t=1}^{T} y_t \mathbb{I}_{p_t^F = p}}{\sum_{t=1}^{T} \mathbb{I}_{p_t^F = p}}$ is the frequency at which event $y = 1$ occurred over the times when we predicted p. We say that F^{cal} is ϵ-calibrated algorithm with resolution $1/N$ if $\limsup_{T\to\infty} C_T \leq \epsilon$. We say F^{cal} is well-calibrated if it is ϵ-calibrated for all $\epsilon > 0$ [12].

It is well-known that there exist algorithms with resolution $1/N$ that produce calibrated estimates p_t of y_t from $y_1, ..., y_{t-1}$ [12, 9]. We will call these *online calibration algorithms*. Unfortunately, they do not admit covariates x_t and do not encourage forecast sharpness.

Online Recalibration To address these shortcomings, we propose a new problem called *online recalibration*, in which an algorithm A produces calibrated predictions $p_t = A_\theta(p_t^F) \in [0, 1]$, where θ are parameters that depend on previous observations and forecasts p_s, y_s, x_s for $s < t$. We seek to produce p_t that are calibrated and that have similar accuracy to the original p_t^F.

Algorithm 1 is an example of an online recalibration algorithm. This method partitions the sequence $(y_t)_{t=1}^T$ into M disjoint sets $S_j = \{y_t \mid p_t^F \in [\frac{j}{M}, \frac{j+1}{M})\}$ for $j = 0, 1, ..., M-1$ and trains an online calibration algorithm F_j^{cal} on each set S_j. After observing p_t^F, Algorithm 1 outputs the prediction of the F_j^{cal} associated with the interval $[\frac{j}{M}, \frac{j+1}{M}) \ni p_t^F$. One can show that with sufficiently high resolution, F_j^{cal} is guaranteed to do as well as constantly predicting $\frac{j}{M}$. By construction, whenever F_j^{cal} is called, $p_t^F \approx \frac{j}{M}$ and so we are also guaranteed to do about as well as p_t^F. Since this is true for any j, we will do at least as well as p_t^F on the entire input sequence. Formally, we can show that

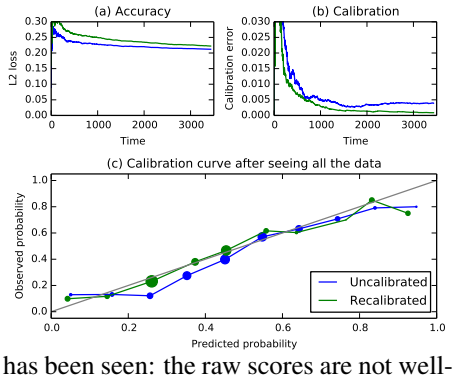
Algorithm 1 Recalibration procedure

Require: Number of buckets M, online calibration algorithm F^{cal}, uncalibrated $(p_t^F)_{t=1}^T$.

Let $\mathcal{I} = \{[0, \frac{1}{M}), [\frac{1}{M}, \frac{2}{M}), ..., [\frac{M-1}{M}, 1]\}$ be a set of intervals that partition $[0, 1]$.
Let $\mathcal{F} = \{F_j^{\text{cal}} \mid i = 0, ..., M-1\}$ be a set of M independent instances of F^{cal}.
for $t = 1, ..., T$: **do**
  Let $[\frac{j}{M}, \frac{j+1}{M})$ be the interval containing p_t^F.
  Let p_t be the forecast of F_j^{cal}. Output p_t.
  Observe y_t and pass it to $F_j^{\text{cal}} \in \mathcal{F}$.
end for

Lemma 1. *Suppose that $M \geq N$ and F^{cal} is ϵ-calibrated. Then Algorithm 1 is ϵ-calibrated and the recalibrated forecasts p_t are not worse than the p_t^F in the sense that the ℓ_2 norm regret $\lim_{T\to\infty}\left(\frac{1}{T}\sum_{t=1}^{T}(y_t - p_t)^2 - \frac{1}{T}\sum_{t=1}^{T}(y_t - p_t^F)^2\right) < 4/N$, and can be made arbitrarily small.*

Note that Algorithm 1 will not alter the original system F, and if F possesses favorable convergence properties (e.g. the ability to exploit sparsity or speed up its convergence via adaptive learning rates like Adagrad), our calibrated p_t will inherit all of these properties. Also, Algorithm 1 can use any online calibration procedure F^{cal} as a black box; its convergence depends on that of F^{cal}. In particular, if $C_T(F^{\text{cal}}) \leq R_T + \epsilon$, where $R_T \to 0$, we can show that $C_T(\text{Alg } 1) \leq \frac{R_T}{\sqrt{M}} + \epsilon$. Since typically $\frac{1}{M} \approx \frac{1}{N} \approx \frac{1}{\epsilon}$, our method has $O(\frac{1}{\sqrt{\epsilon}})$ relative to F^{cal}. Using the best online calibration algorithms [10], we obtain $C_T(\text{Alg } 1) \leq \frac{1}{\epsilon\sqrt{T}} + \epsilon$.



Experiments Finally, we used our method to predict the occurrence of type 1 diabetes in $T = 3,443$ patients from genotype data at 447,221 SNPs [13]. We recalibrated scores from an online ℓ_1-regularized linear support vector machine (SVM) that observes the patients one at a time; uncalibrated p_t^F were computed by normalizing the raw SVM score s_t within the range of all the previously seen scores, i.e. $p_t^F = (s_t + m_t)/2m_t$, where $m_t = \max_{1\leq r\leq t} |s_r|$. The adjacent figure (part c) compares the calibration plots [8] of p_t^F and p_t obtained from Algorithm 1 after all the data has been seen: the raw scores are not well-calibrated outside of the interval $[0.4, 0.6]$; recalibration makes them almost perfectly calibrated. Part (b) quantifies this difference and shows that it holds over the entire learning process. Finally, part (a) shows that the ℓ_2 loss of the p_t approaches to within 0.01 of the p_t^F loss as T increases.

References

[1] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *JAMIA*, 19(2):263–274, 2012.

[2] Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *CoRR*, abs/1508.05154, 2015.

[3] Dong Yu, Jinyu Li, and Li Deng. Calibration of confidence measures in speech recognition. *Trans. Audio, Speech and Lang. Proc.*, 19(8):2461–2473, November 2011.

[4] J. Brocker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009.

[5] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[6] S. Lichtenstein, B. Fischhoff, and L. D. Phillips. *Calibration of Probabilities: The State of the Art to 1980*. Cambridge University Press, 1982.

[7] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[8] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 625–632, New York, NY, USA, 2005. ACM.

[9] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration, 1998.

[10] Jacob Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In Sham M. Kakade and Ulrike von Luxburg, editors, *COLT 2011 - The 24th Annual Conference on Learning Theory, June 9-11, 2011, Budapest, Hungary*, volume 19 of *JMLR Proceedings*, pages 27–46. JMLR.org, 2011.

[11] Shai Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. Phd thesis, Hebrew University, 2007.

[12] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[13] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, June 2007.