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 \( F^{\text{cal}} \) be a forecaster making predictions in the set \( \{ \frac{i}{N} \mid i = 0, ..., N \} \), where \( 1/N \) is called the resolution of \( F^{\text{cal}} \). We define calibra-
tion error as 
\[ C_T = \sum_{t=0}^{N} \left( \rho_T(i/N) - \frac{i}{N} \right)^2 \left( \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(p_T^F = \frac{i}{N}) \right), \]
where 
\[ \rho_T(p) = \frac{\sum_{i=1}^{T} y_i 1_{y_i = p}}{\sum_{i=1}^{T} 1_{y_i = p}} \]
is the frequency at which event \( y = 1 \) occurred over the times when we predicted \( p \). We say that \( F_{\text{cal}} \) is \( \epsilon \)-calibrated algorithm with resolution \( 1/N \) if \( \lim_{T \to \infty} C_T \leq \epsilon \). We say \( F_{\text{cal}} \) is well-calibrated if it is \( \epsilon \)-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 \( \theta \) 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 Recalibration procedure](image)

**Algorithm 1** Recalibration procedure

| Require: Number of buckets \( M \), online calibration algorithm \( F_{\text{cal}} \), uncalibrated \( p_t^F \).
| 1: Let \( I = \{0, \frac{1}{M}, \frac{2}{M}, ..., \frac{M-1}{M}, 1\} \) be a set of intervals that partition \([0, 1]\).
| 2: Let \( F = \{F_{\text{cal}}| i = 0, ..., M-1\} \) be a set of \( M \) independent instances of \( F_{\text{cal}} \).
| 3: for \( t = 1, ..., T \):
| 4: Let \( [\frac{i}{M}, \frac{i+1}{M}] \) be the interval containing \( p_t^F \).
| 5: Let \( p_t \) be the forecast of \( F_{\text{cal}} \). Output \( p_t \).
| 6: Observe \( y_t \) and pass it to \( F_{\text{cal}} \) in \( F \).
| 7: end for

**Lemma 1.** Suppose that \( M \geq N \) and \( F_{\text{cal}} \) is \( \epsilon \)-calibrated. Then Algorithm 1 is \( \epsilon \)-calibrated and the recalibrated forecasts \( p_t \) are not worse than the \( p_t^F \) in the sense that the \( \ell_2 \) 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 \( F_{\text{cal}} \) as a black box; its convergence depends on that of \( F_{\text{cal}} \). 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 \( \frac{1}{\sqrt{M}} \approx \frac{1}{N} \approx \frac{1}{\sqrt{T}} \), our method has \( O(\frac{1}{\sqrt{T}}) \) relative to \( F_{\text{cal}} \). Using the best online calibration algorithms [10], we obtain \( C_T(\text{Alg 1}) \leq \frac{1}{\sqrt{T}} + \epsilon \).

**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 \( \ell_1 \)-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 \( m_t = \max_{s \in [s_t]} |s_t| \). 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 \( \ell_2 \) loss of the \( p_t \) approaches to within 0.01 of the \( p_t^F \) loss as \( T \) increases.
References


