Part I
Calibration

1 Class 1: Calibrated Predictions

We will start with a very simple model of forecasting a binary outcome. We will call this the *batch setting*.

- There are labeled examples denoted (x, y) where $x \in \mathcal{X}$ is a feature vector and $y \in \{0, 1\}$.
- There is a distribution $\mathcal{D}$ over pairs and (X, Y) denotes the corresponding random variables.
- There is a predictor $f : \mathcal{X} \to [0, 1]$ that maps features to a forecast of the outcome.

A useful toy example is predicting whether or not it will rain on a given day. Here x would consist of a set of features describing the weather, like the date, temperature, humidity, cloud cover, wind speed, etc. And the outcome y is 1 if it rained and 0 otherwise. A predictor f takes the features and produces a forecast $f(x) = p$.

We can also consider a *sequential* prediction problem. Here, instead of learning a fixed model, we receive feature vectors one at a time and we have to make predictions as we go. On each round, we receive some features x_t, make a prediction p_t, and then see the outcome y_t. We can think of this as a game between the forecaster and nature.

For each day $t = 1, \ldots, T$**:**
1. Nature reveals the features x_t.
2. The forecaster makes a prediction $p_t \in [0, 1]$.
3. Nature reveals the outcome $y_t \in \{0, 1\}$.

This setting is more challenging than the batch setting in two ways: each data point is revealed only after the last one, and, more importantly, the outcomes can be chosen adversarially rather than being drawn from some distribution. We will focus mostly on batch predictions to start because the mathematics and notation are simpler.

What are good predictions? What does it really mean to predict the probability of an event that will only occur once? It will either rain today or it wont and today will never happen again. The sequential setting allows us to talk about predictions without assuming any probabilistic model for the outcomes, which is nice! But it also means that we will inherently be confronting the fact that there is no such thing as the ground truth probability of rain today. In the batch setting, we do posit that there is a probabilistic model for the data, so it makes mathematical sense to talk about the true probability of rain conditioned on the feature vector we observe. However, even in this setting we will allow these probabilities to be arbitrary, the chance of rain for each specific x can be totally unrelated to the probability for any other x. So even though the batch setting is a convenient model, we are often faced with the reality that we will see no more than one outcome corresponding to a given x. Thus, having a probabilistic model *for the data* is a useful mathematical convention, but we are still thinking of settings where we cannot really know the true probability of rain given the feature vector. Classical machine learning is based on the idea that there is some probabilitic model *for the outcomes* that constraints what the ground truth forecast f can be, and makes it possible to

learn f from data. That's great, machine learning works really well! But it makes it challenging to talk about uncertainty and trustworthyness unlessyou really believe in that probabilistic model.

Now that we've cast a lot of doubt on the idea that there is a ground truth forecaster and that probabilities of events make sense, we can state our key question for today.

> *When should we treat the forecast $f(x) = p$ as if it were really the probability that it will rain today?*

How is this different from accuracy? Going back to our core question, it sounds like we're asking the usual question about measuring a model's accuracy. A common notion of accuracy of a real-valued prediction is the squared error

$$\mathrm{MSE}_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{\mathcal{D}}[(f(X) - Y)^2].$$

Note that the unique predictor f^* that minimizes MSE is the one that outputs the true probability it rains given the features x,

$$f^*(x) = \mathop{\mathbb{P}}_{\mathcal{D}}[Y = 1 \mid X = x]$$

So one might think that MSE is a good way to measure how trustworthy a forecast is. The problem is that MSE doesn't distinguish between inherent uncertainty of the outcome and inaccuracy of the model, as in the following example:

Example 1.1. Consider two cities and two different weather forecasts.

Las Vegas. Here it rains on about 10% of days, and consider the uninformative predictor $f_{\mathrm{LV}}(x) = 0$, which always predicts that it will not rain. Its MSE is

$$\mathrm{MSE}_{\mathcal{D}_{\mathrm{LV}}}(f_{\mathrm{LV}}) = 0.9(0-0)^2 + 0.1(0-1)^2 = 0.1.$$

Portland. Here it rains on about 50% of days. However, half of the days are cloudy and half are sunny, with

$$\mathbb{P}[Y = 1 \mid X = \text{cloudy}] = 0.75, \qquad \mathbb{P}[Y = 1 \mid X = \text{sunny}] = 0.25.$$

Consider the predictor that reports these conditional probabilities:

$$f_{\mathrm{PO}}(\text{cloudy}) = 0.75, \qquad f_{\mathrm{PO}}(\text{sunny}) = 0.25.$$

Its MSE is 0.1875 under either type of weather condition, so

$$\begin{aligned}\mathrm{MSE}_{\mathcal{D}_{\mathrm{PO}}}(f_{\mathrm{PO}}) &= 0.5(0.75(0.75-1)^2 + 0.25(0.75-0)^2) + 0.5(0.25(0.25-1)^2 + 0.75(0.25-0)^2) \\ &= 0.1875.\end{aligned}$$

The Portland predictor has higher MSE even though it reports the true conditional probabilities and conveys useful information about cloudy and sunny days. The Las Vegas predictor gives no useful information but has low MSE just because rain is rare.

A different starting point: average bias. This example shows that MSE is conflating model errors with inherent uncertainty, and is failing to capture why our perfect model of Portland weather is "better" than our silly model of Las Vegas weather. Let's start by thinking of a different way to say what is wrong with the Las Vegas model.

Definition 1.2. For a distribution $\mathcal{D}$ and a predictor f, the *average bias* is

$$\beta_{\mathcal{D}} := \mathop{\mathbb{E}}_{\mathcal{D}}[f(X) - Y]$$

If the average bias is 0 the predictor is *unbiased on average*. Note that average bias can be positive or negative, so when we informally say a predictor has low average bias we really mean that $|\beta_{\mathcal{D}}|$ is small.

The optimal predictor f^* is unbiased on average. But the nearly uninformed predictor

$$f(x) := \mathop{\mathbb{E}}_{\mathcal{D}}[Y]$$

that just outputs the average outcome is also unbiased on average. So clearly this is a very weak property that can't distinguish good and bad predictors. However, it does help us identify why our predictor for Las Vegas is bad despite having low MSE. Note that

$$\begin{aligned}
\beta_{\mathcal{D}_{\text{LV}}}(f_{\text{LV}}) &= \mathop{\mathbb{E}}_{\mathcal{D}_{\text{LV}}}[f_{\text{LV}}(X)] - \mathop{\mathbb{E}}_{\mathcal{D}_{\text{LV}}}[Y] = 0 - 0.1 = -0.1, \\
\beta_{\mathcal{D}_{\text{PO}}}(f_{\text{PO}}) &= \mathop{\mathbb{E}}_{\mathcal{D}_{\text{PO}}}[f_{\text{PO}}(X)] - \mathop{\mathbb{E}}_{\mathcal{D}_{\text{PO}}}[Y] = 0.5(0.75) + 0.5(0.25) - 0.5 = 0.
\end{aligned}$$

So the Las Vegas predictor is biased while the Portland predictor is unbiased on average.

What is wrong the fact that the Las Vegas predictor is biased? Well, it might lead to bad decisions. Suppose you have a very light umbrella that is easy to carry around, and you absolutely hate getting rained on. You are also an expected utility maximizer. It makes sense to take an umbrella if there is even a 5% chance of rain. Now, if you interpret the Las Vegas predictor as if there really were a 0% chance of rain, you would never take an umbrella, and would be unhappy! On the other hand, if you had a similarly uninformed, but unbiased prediction that always outputs $f(x) = 0.1$, then you would take your umbrella every day and would be happier.

> *If you have a biased predictor, then you should not treat its predictions as if they were real probabilities.*

We will revisit this point a lot more when we talk about decision theory.

Fortunately, there is an easy way to make the Las Vegas predictor unbiased on average, just replace every output of 0 with an output of 0.1! More generally, given a distribution $\mathcal{D}$ and a predictor f, we can define the debiased predictor to be

$$f_{\beta \to 0}(x) := f(x) - \beta_{\mathcal{D}}$$

Intuitively, reducing bias is good. The debiased predictor for Las Vegas still reports that there is a 10% chance of rain every day, which isn't very informative, but it's clearly better than saying 0% every day. But being unbiased on average is a weak property, so we should ask what are we giving

up by unbiasing? Well, if we still care about MSE, then the answer is nothing. In fact, the debiased predictor also has lower MSE. Indeed, since $\beta_{\mathcal{D}} = \mathbb{E}_{\mathcal{D}}[f(X) - Y]$,

$$\begin{aligned}
\mathrm{MSE}_{\mathcal{D}}(f_{\beta\to 0}) &= \mathop{\mathbb{E}}_{\mathcal{D}}\left[(f(X) - Y - \beta_{\mathcal{D}})^2\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}}\left[(f(X) - Y)^2\right] - 2\beta_{\mathcal{D}} \mathop{\mathbb{E}}_{\mathcal{D}}[f(X) - Y] + \beta_{\mathcal{D}}^2 \\
&= \mathrm{MSE}_{\mathcal{D}}(f) - 2\beta_{\mathcal{D}}^2 + \beta_{\mathcal{D}}^2 \\
&= \mathrm{MSE}_{\mathcal{D}}(f) - \beta_{\mathcal{D}}^2.
\end{aligned}$$

Thus debiasing reduces the MSE by exactly $\beta_{\mathcal{D}}^2$.

But where do we get the distribution? There is something a little fishy about this discussion, since the debiased predictor depends on the distribution. If we knew the distribution, we would just use the optimal predictor and we wouldn't be discussing how to measure error. But intuitively, knowing the bias is a lot easier than learning the optimal predictor. I just looked up the fraction of days it rains in Las Vegas on wikipedia, even though I don't know anything about weather modeling.

We can capture our knowledge of the distribution using the idea of a *sample*. Suppose we are given n independent, identically distributed samples $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn from $\mathcal{D}$. We can use these samples to estimate the bias

$$\hat{\beta}_S = \frac{1}{n}\sum_{i=1}^{n} f(x_i) - y_i$$

Of course, for a finite number of samples n, $\hat{\beta}_S$ wont be exactly equal to the true bias $\beta_{\mathcal{D}}$, but we can quantify how far apart they will be.

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}[\hat{\beta}_S] = \beta_{\mathcal{D}}, \qquad \mathop{\mathrm{Var}}_{S\sim\mathcal{D}^n}(\hat{\beta}_S) \le \frac{1}{n}.$$

This means that the gap between the true bias and the bias we estimate on the sample will typically be about $|\hat{\beta}_S - \beta_{\mathcal{D}}| \approx n^{-1/2}$. Since this quantity goes to 0 rapidly as the sample size grows, we think of debiasing the predictor as an easy thing to do.

I

Average bias for sequential predictions. In the sequential setting we can still measure the bias of the forecaster's predictions in hindsight at the end of the sequence.

Definition 1.3. For a sequence of predictions $p_1, \ldots, p_T$ and outcomes $y_1, \ldots, y_T$, the *average bias* is

$$\beta := \frac{1}{T}\sum_{t=1}^{T}(p_t - y_t).$$

If the average bias is 0, the forecaster is *unbiased on average*. Note that average bias can be positive or negative, so when we informally say a forecaster has low average bias we really mean that $|\beta|$ is small.

In the sequential setting there is no exact analogue of the constant predictor that outputs the average outcome, but there is still an uninformed forecaster that will achieve very low bias.

Specifically, consider the forecaster that always outputs the last day's outcome

$$p_1 = 0 \qquad \text{and} \qquad p_t = y_{t-1} \text{ for } t = 2, \ldots, T.$$

where the first prediction p_1 can be any value in $[0, 1]$. The errors telescope, so for every sequence of outcomes,

$$\begin{aligned} |\beta| &= \left| \frac{1}{T} \sum_{t=1}^{T} (p_t - y_t) \right| \\ &= \left| \frac{1}{T}(p_1 - y_1) + \frac{1}{T} \sum_{t=2}^{T} (p_t - y_t) \right| \\ &= \left| -y_1 + \frac{1}{T} \sum_{t=2}^{T} (y_{t-1} - y_t) \right| \\ &= \frac{1}{T}|y_T| \leq \frac{1}{T}. \end{aligned}$$

Thus this forecaster always has average bias at most $1/T$ in absolute value.

Defining calibration. The sequential forecaster that always predicts the last outcome makes a good example for understanding why we need a stronger notion of trustworthiness and what it could be. Suppose that we live in a town where it rains on odd days but not on even days. The last outcome forecaster will predict rain on even days but not on odd days.

$$p_t = \begin{cases} 1 & \text{if } t \text{ is even} \\ 0 & \text{if } t \text{ is odd} \end{cases} \qquad \text{and} \qquad y_t = \begin{cases} 1 & \text{if } t \text{ is odd} \\ 0 & \text{if } t \text{ is even} \end{cases}$$

This predictor is unbiased on average, but clearly we cannot trust the predictions. If we look at the even days, the preditor is telling us rain is guaranteed, but in fact there will never be rain, and the reverse for the odd days. In other words, despite having low bias, the predictions don't behave like real probabilities.

The definition of calibration is designed to rule out this type of behavior. Intuitively, it says that each value of p that the forecaster outputs, should behave as if it were the true probability of rain on those days. Formally, we can define calibration in the batch and sequential settings as follows.

Definition 1.4. A predictor f is *calibrated* with respect to a distribution $\mathcal{D}$ if, for every prediction p in the range of f with $\mathbb{P}_{\mathcal{D}}[f(X) = p] > 0$,

$$\mathop{\mathbb{E}}_{\mathcal{D}}[Y \mid f(X) = p] = p.$$

Similarly, for a sequence of predictions $p_1, \ldots, p_T$ and outcomes $y_1, \ldots, y_T$, let

$$S_p := \{t \in \{1, \ldots, T\} : p_t = p\}.$$

The sequence is *calibrated* if, for every prediction p with $S_p \neq \emptyset$,

$$\frac{1}{|S_p|} \sum_{t \in S_p} y_t = p.$$

In the batch setting, the uninformed predictor that outputs the average outcome is still calibrated! So calibration on its own is still a very weak property. However, like with average bias, we can still argue that, given a predictor f, we always do better by recalibrating it.

Similar to bias, given any predictor f, we can make it calibrated by separately debiasing the examples that receive each prediction. For every p in the range of f, define the bias among examples receiving prediction p by

$$\beta_p := \mathop{\mathbb{E}}_{\mathcal{D}}[f(X) - Y \mid f(X) = p] = p - \mathop{\mathbb{E}}_{\mathcal{D}}[Y \mid f(X) = p].$$

Now define the recalibrated predictor

$$f_{\text{cal}}(x) := f(x) - \beta_{f(x)} = \mathop{\mathbb{E}}_{\mathcal{D}}[Y \mid f(X) = f(x)].$$

Thus, whenever the original predictor outputs p, the recalibrated predictor outputs the average outcome among examples on which the original prediction was p. Each such group is unbiased by construction. If several original predictions are mapped to the same new prediction q, their union is also unbiased at q, so

$$\mathop{\mathbb{E}}_{\mathcal{D}}[Y \mid f_{\text{cal}}(X) = q] = q.$$

Therefore f_{cal} is calibrated. As with bias, we can also show that this procedure reduces the MSE.

Unlike average bias, we only defined what it means for a predictor to be calibrated. The first definition you would consider is called *expected calibration error*.

Definition 1.5. For a distribution $\mathcal{D}$ and a predictor f, the *expected calibration error (ECE)* is

$$\text{ECE}_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{P=f(X)}\left[\left|\mathop{\mathbb{E}}_{\mathcal{D}}[Y \mid f(X) = P] - P\right|\right].$$

For a sequence of predictions $p_1, \ldots, p_T$ and outcomes $y_1, \ldots, y_T$, let $\mathcal{P} := \{p_t : t = 1, \ldots, T\}$ be the set of distinct predictions. The expected calibration error is

$$\text{ECE} := \sum_{p \in \mathcal{P}} \frac{|S_p|}{T}\left|\frac{1}{|S_p|}\sum_{t \in S_p} y_t - p\right| = \frac{1}{T}\sum_{p \in \mathcal{P}}\left|\sum_{t \in S_p}(p_t - y_t)\right|.$$

In either setting, the predictor is calibrated exactly when its ECE is 0.

Calibration error and calibration in the sequential setting. Unlike the batch setting, it's not at all clear how to produce a calibrated predictor in the sequential setting. In fact, it almost seems too good to be true! However, a beautiful result of Foster and Vohra [FV98] shows that it is possible to produce a calibrated predictor in the sequential setting, even when the outcomes are chosen adversarially. The proof is constructive, and gives a simple algorithm for producing calibrated predictions.

Theorem 1.6 ([FV98]). *There is a randomized sequential forecasting strategy such that for every adversarial sequence of outcomes $y_1, \ldots, y_T \in \{0, 1\}$,*

$$\mathbb{E}[\mathrm{ECE}] \leq O(T^{-1/3}).$$

The guarantee holds when nature chooses each outcome based on the previous outcomes and predictions $p_1, y_1, \ldots, p_{t-1}, y_{t-1}$.

If you're interested in seeing how this theorem is proven, I highly recommend the note by Sergiu Hart [Har22]. Not only does it give a short proof of the thoerem, but it introduces both online learning models and the minimax theorem, which are probably two of the most useful things I know about.

Interestingly, this theorem was interpreted for a long time as a *negative result* showing that calibration can't be a meaningful guarantee, because it seems like someone who knows nothing about weather should not be able to produce a "good" forecast, and therefore calibration must not be ensuring "good" forecasts. However, nowadays we have good justification for calibration as a meanignful guarantee, and we tend to think of this theorem as a *positive result* showing that there are remarkable algorithms for learning to make good sequenital predictions from the past history.

Acknowledgments and AI Use

The overall structure of this lecture is inspired by Aaron Roth's lecture notes [Rot26]. The content of the notes is more or less AI-free, but I used AI throughout to help write the LaTeX code for examples, definitions, and calculations.

References

[FV98] Dean P. Foster and Rakesh V. Vohra. "Asymptotic Calibration." *Biometrika* 85.2 (1998). URL: https://doi.org/10.1093/biomet/85.2.379.

[Har22] Sergiu Hart. "Calibrated Forecasts: The Minimax Proof." arXiv:2209.05863. 2022. URL: https://arxiv.org/abs/2209.05863.

[Rot26] Aaron Roth. "Uncertain: Modern Topics in Uncertainty Quantification." Incomplete draft textbook. 2026. URL: https://www.cis.upenn.edu/~aaroth/uncertain.html.