Tools and Techniques for Machine Learning
Homework 5

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. It's preferred that you write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or Jupyter), though if you need to you may scan handwritten work. For submission, you can also export your Jupyter notebook and merge that PDF with your PDF for the written solutions into one file. **Don't forget to complete the Jupyter notebook as well, for the programming part of this assignment**.

1 Calibration error

We the defined **calibration error** of $f : \mathcal{X} \to [0,1]$ is as

$$\mathrm{CE}(f) = \left(\mathbb{E}\left[(f(X) - \mathbb{P}[Y = 1 \mid f(X)])^2\right]\right)^{1/2}$$

and we defined the **integrated squared error as**

$$\mathrm{ISE}(f) = \left(\mathbb{E}\left[(f(X) - \mathbb{P}[Y = 1 \mid X])^2\right]\right)^{1/2}.$$

We claimed that without any knowledge or assumption about $\mathbb{P}[Y = 1 \mid X]$, such as it being in some smooth class of functions, it can be impossible to get a good estimate of $\mathrm{ISE}(f)$. The essence of the issue is that we need to have data from all possible values of X to estimate $\mathbb{P}[Y = 1 \mid X]$, and this is a problem if X is a continuous variable. In fact, we'll have the same problem estimating $\mathrm{CE}(f)$ if $f(X)$ take continuous values (or takes uncountably many different values).

1. Show that if $f : \mathcal{X} \to [0,1]$ is an injective function (i.e. $x \neq x' \implies f(x) \neq f(x')$ for any $x, x' \in \mathcal{X}$), then $\mathrm{CE}(f) = \mathrm{ISE}(f)$. [Hint: Let $g(x) = \mathbb{P}[Y = 1 \mid f(X) = f(x)]$and let $h(x) := \mathbb{P}[Y = 1 \mid X = x]$ and show that $g(x) = h(x)$ for all $x \in \mathcal{X}$.] [Discussion: When f is injective, we can say that $f(x)$ maintains all the information in x. The implication of this question is that if we want to be able to estimate $\mathrm{CE}(f)$, we're going to need to make some assumptions about f. The assumption that is typically made is that $f(x)$ takes on only finitely many different values. If this isn't the case, we can approximate f by "binning", as discussed in lecture, then estimate the CE for the binned f.]

2 Plug-in estimator of calibration error

Suppose we want to estimate the calibration error of $f : \mathcal{X} \to [0,1]$, but $f(x)$ takes too many different values to estimate $\mathbb{P}[Y = 1 \mid f(X) = f(x)]$ for each. We decide to approximate f with a "binned version" $f_{\mathcal{B}}(x)$ that takes only finitely many values, as follows: Let $\mathcal{B}$ be a partition of $[0, 1]$ into disjoint sets (i.e. "bins") $I_1, \ldots, I_B$, and define

$$f_{\mathcal{B}}(x) = \mathbb{E}[f(X) \mid f(X) \in I_b] \qquad \text{where } f(x) \in I_b.$$

Without knowledge of the marginal distribution of X, we can't compute $f_{\mathcal{B}}(x)$. However, we'll assume we have a labeled sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, and we'll use the natural estimate

$$\hat{f}_{\mathcal{B}}(x) := \text{mean}\left\{f(X_i) \mid f(X_i) \text{ and } f(x) \text{ are in the same bin}\right\}.$$

We'll now define the "plug-in estimator" for the [squared] CE of $f_{\mathcal{B}}$ as

$$\widehat{\text{CE}^2}(f_{\mathcal{B}}) := \sum_{b=1}^{B} \hat{p}_b(\hat{f}_{\mathcal{B}}(x_b) - \hat{\mu}_b)^2,$$

where x_b is any value for which $f(x_b) \in I_b$, $\hat{\mu}_b = \text{mean}\left\{Y_i \mid f(X_i) \in I_b\right\}$ and $\hat{p}_b = n_b/n$, where n_b is the number of X_i's for which $f(X_i) \in I_b$.

1. Assuming the partition $\mathcal{B}$ is determined independently of the sample, give an expression for $\text{CE}^2_\infty(f_{\mathcal{B}})$, the limit of $\widehat{\text{CE}}(f_{\mathcal{B}})$ as $n \to \infty$. [Hint: You'll want to define $\phi_b := \mathbb{P}\left(f(X) \in I_b\right)$, the probability that a prediction is in bin b. You can also use the expression $\mathbb{E}\left[Y \mid f(X) \in I_b\right]$ in your answer.] (Just provide the expression – you don't have to prove it. But if you feel like being rigorous, the proof is a straightforward application of the weak law of large numbers, Slutsky's theorem, and the continuous mapping theorem.)

2. Show that $\text{CE}^2_\infty(f_{\mathcal{B}}) = [\text{CE}(f_{\mathcal{B}})]^2$. Combined with our previous problem, this will imply that $\widehat{\text{CE}}(f_{\mathcal{B}})$ is a consistent estimator of the calibration error of $f_{\mathcal{B}}$.

3. [Optional – no credit]Above, we assumed that the binning was determined independently of our sample. Now suppose we use the same sample to determine the bins $\mathcal{B}$ as we use to compute $\widehat{\text{CE}}(f_{\mathcal{B}})$. Show that if the $f(X_i)$ are distinct for all $i = 1, \ldots, n$ and if we use quantile binning with n bins, then $\widehat{\text{CE}}(f_{\mathcal{B}})$ becomes the empirical Brier score of f (i.e. mean squared error). [This is bad if our goal is to estimate $\text{CE}(f)$, because the Brier score is quite a different thing from calibration error, as we've discussed.]