CS 578 Statistical Machine Learning (Fall 2026) Problem Set #1
Name: _____ **Due: September 6**

Problem 1. Bayes predictor.

For a measurable predictor $f : \mathcal{X} \to \mathcal{Y}$, define its population risk by

$$\mathcal{R}(f) := \mathbb{E}\big[\ell(Y, f(X))\big].$$

A Bayes predictor $f_\star : \mathcal{X} \to \mathcal{Y}$ satisfies, for P_X-almost every $x' \in \mathcal{X}$,

$$f_\star(x') \in \arg\min_{z \in \mathcal{Y}} \mathbb{E}\big[\ell(Y, z) \mid X = x'\big].$$

When the minimizer is nonunique, any measurable choice of a minimizer is called a Bayes predictor.

1. Consider binary classification with $\mathcal{Y} = \{-1, 1\}$ and the loss function

$$\ell(y, z) = \begin{cases} 0 & \text{if } y = z, \\ c_{\mathrm{FP}} > 0 & \text{if } y = -1, z = 1, \\ c_{\mathrm{FN}} > 0 & \text{if } y = 1, z = -1. \end{cases}$$

   Compute a Bayes predictor at $x = x'$ as a function of $\mathbb{E}[Y \mid X = x']$.
   Hint: Let $\mu(x') = \mathbb{E}[Y \mid X = x']$ and compare the conditional risks of predicting 1 and -1.

2. Show that every Bayes predictor minimizes the population risk; that is, $\mathcal{R}(f_\star) \le \mathcal{R}(g)$ for every measurable $g : \mathcal{X} \to \mathcal{Y}$.

3. *(Bonus)* Suppose $\mathcal{Y} = \mathbb{R}$, $\mathbb{E}|Y| < \infty$, and the loss is $\ell(y, z) = |y - z|$. Characterize all Bayes predictors in terms of the conditional distribution of $Y \mid X = x'$, and give one explicit choice using the conditional cumulative distribution function (CDF).

Solution.

1. Fix $x' \in \mathcal{X}$ and write

$$\mu := \mu(x') = \mathbb{E}[Y \mid X = x'].$$

   Since $Y \in \{-1, 1\}$,

$$\Pr(Y = 1 \mid X = x') = \frac{1 + \mu}{2}, \qquad \Pr(Y = -1 \mid X = x') = \frac{1 - \mu}{2}.$$

   The conditional risk of predicting 1 is therefore

$$\begin{aligned} L_{x'}(1) &:= \mathbb{E}[\ell(Y, 1) \mid X = x'] \\ &= c_{\mathrm{FP}} \Pr(Y = -1 \mid X = x') = \frac{c_{\mathrm{FP}}}{2}(1 - \mu), \end{aligned}$$

whereas the conditional risk of predicting -1 is

$$\begin{aligned} L_{x'}(-1) &:= \mathbb{E}[\ell(Y,-1) \mid X = x'] \\ &= c_{\mathrm{FN}} \Pr(Y = 1 \mid X = x') = \frac{c_{\mathrm{FN}}}{2}(1+\mu). \end{aligned}$$

We should predict 1 precisely when $L_{x'}(1) \le L_{x'}(-1)$. This inequality is equivalent to

$$\begin{aligned} & c_{\mathrm{FP}}(1-\mu) \le c_{\mathrm{FN}}(1+\mu) \\ \iff \qquad & \mu \ge \frac{c_{\mathrm{FP}} - c_{\mathrm{FN}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}}. \end{aligned}$$

Thus, breaking ties in favor of 1, one Bayes predictor is

$$f_\star(x') = \begin{cases} 1, & \mathbb{E}[Y \mid X = x'] \ge \dfrac{c_{\mathrm{FP}} - c_{\mathrm{FN}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}}, \\ -1, & \mathbb{E}[Y \mid X = x'] < \dfrac{c_{\mathrm{FP}} - c_{\mathrm{FN}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}}. \end{cases}$$

At equality, the two conditional risks are equal, so either label is Bayes optimal. Equivalently, if

$$\eta(x') := \Pr(Y = 1 \mid X = x'),$$

then the threshold rule is

$$f_\star(x') = 1 \quad \iff \quad \eta(x') \ge \frac{c_{\mathrm{FP}}}{c_{\mathrm{FP}} + c_{\mathrm{FN}}},$$

again with an arbitrary choice at equality.

2. For each $x \in \mathcal{X}$ and $z \in \mathcal{Y}$, define the conditional risk

$$L_x(z) := \mathbb{E}[\ell(Y,z) \mid X = x].$$

By the defining property of a Bayes predictor,

$$L_x(f_\star(x)) \le L_x(g(x))$$

for P_X-almost every x. Applying the law of total expectation gives

$$\begin{aligned} \mathcal{R}(f_\star) &= \mathbb{E}\big[\ell(Y, f_\star(X))\big] \\ &= \mathbb{E}\left[\mathbb{E}\big[\ell(Y, f_\star(X)) \mid X\big]\right] \\ &= \mathbb{E}\big[L_X(f_\star(X))\big] \\ &\le \mathbb{E}\big[L_X(g(X))\big] \\ &= \mathbb{E}\left[\mathbb{E}\big[\ell(Y, g(X)) \mid X\big]\right] \\ &= \mathcal{R}(g). \end{aligned}$$

Hence every Bayes predictor minimizes the population risk.

3. Fix $x' \in \mathcal{X}$ and let

$$F_{x'}(t) := \Pr(Y \le t \mid X = x')$$

be the conditional CDF. Also write

$$F_{x'}(t^-) := \Pr(Y < t \mid X = x')$$

for its left limit. A number $m \in \mathbb{R}$ is a conditional median of Y given $X = x'$ if

$$F_{x'}(m) \ge \frac{1}{2} \qquad \text{and} \qquad F_{x'}(m^-) \le \frac{1}{2}.$$

Equivalently,

$$\Pr(Y \le m \mid X = x') \ge \frac{1}{2} \quad \text{and} \quad \Pr(Y \ge m \mid X = x') \ge \frac{1}{2}.$$

We claim that the minimizers of

$$L_{x'}(z) := \mathbb{E}[|Y - z| \mid X = x']$$

are exactly these conditional medians.

To see this, for any $a < b$, we have

$$\begin{aligned} L_{x'}(b) - L_{x'}(a) &= \mathbb{E}\big[|Y - b| - |Y - a| \mid X = x'\big] \\ &= \int_a^b \big(2F_{x'}(t) - 1\big)\, dt. \end{aligned}$$

Now suppose that m is a conditional median. If $z > m$, then $F_{x'}(t) \ge F_{x'}(m) \ge 1/2$ for every $t \in [m, z]$, and hence

$$L_{x'}(z) - L_{x'}(m) = \int_m^z \big(2F_{x'}(t) - 1\big)\, dt \ge 0.$$

If $z < m$, then $F_{x'}(t) \le F_{x'}(m^-) \le 1/2$ for every $t < m$, and therefore

$$L_{x'}(m) - L_{x'}(z) = \int_z^m \big(2F_{x'}(t) - 1\big)\, dt \le 0.$$

Thus $L_{x'}(m) \le L_{x'}(z)$ for every $z \in \mathbb{R}$.

Conversely, if $F_{x'}(z) < 1/2$, then by right-continuity of the CDF, moving a sufficiently small distance to the right strictly decreases the conditional risk. If $F_{x'}(z^-) > 1/2$, moving a sufficiently small distance to the left strictly decreases it. Consequently, a minimizer must satisfy

$$F_{x'}(z) \ge \frac{1}{2} \qquad \text{and} \qquad F_{x'}(z^-) \le \frac{1}{2}.$$

Hence the Bayes predictors under absolute loss are exactly the measurable selections of conditional medians.

One explicit choice is the lower conditional median

$$f_\star(x') := F_{x'}^{-1}\left(\frac{1}{2}\right) := \inf\left\{t \in \mathbb{R} : F_{x'}(t) \ge \frac{1}{2}\right\}.$$

Problem 2. Empirical risk minimization and PAC learning.

Let $\mathcal{X}$ be a discrete (finite or infinite) domain, $\mathcal{Y} = \{0, 1\}$, and use the 0–1 loss. The training sample consists of m i.i.d. examples drawn from a distribution *realizable* by $\mathcal{H} = \{h_z : z \in \mathcal{X}\} \cup \{h_-\}$, where, for each $z \in \mathcal{X}$, h_z is the indicator function of z:

$$h_z(x) = \begin{cases} 1 & \text{if } x = z, \\ 0 & \text{otherwise.} \end{cases}$$

The hypothesis h_- is the all-zero function: $h_-(x) = 0$ for every $x \in \mathcal{X}$.

1. Describe an algorithm that implements the ERM rule for learning $\mathcal{H}$.
2. Show that $\mathcal{H}$ is PAC learnable. Provide an upper bound on the sample complexity.

Solution. Let

$$S = ((X_1, Y_1), \ldots, (X_m, Y_m))$$

be the training sample.

1. Consider the following learning algorithm:

   - If the sample contains an example (X_j, Y_j) with $Y_j = 1$, output h_{X_j}.
   - If the sample contains no positive example, output h_-.

   Under the realizability assumption, all positively labeled examples, if any, have the same input value. Indeed, the only hypotheses in $\mathcal{H}$ that assign label 1 do so at a single point. Thus the first case is well defined.

   This algorithm is an ERM. If a positive example $(z, 1)$ occurs, realizability implies that the target hypothesis is h_z, so h_z correctly labels every example in the sample and has empirical risk zero. If no positive example occurs, h_- correctly labels every sample point and also has empirical risk zero. Since empirical risk is nonnegative, zero is the minimum possible empirical risk.

2. Let $h^* \in \mathcal{H}$ be the target hypothesis, and let $\widehat{h}$ be the output of the ERM algorithm.

   If $h^* = h_-$, then every sample label is zero, the algorithm outputs h_-, and

   $$\mathcal{R}(\widehat{h}) = 0.$$

   It remains to consider the case $h^* = h_z$ for some $z \in \mathcal{X}$. Set

   $$q := \Pr(X = z).$$

   If at least one copy of z appears in the training sample, then it is labeled 1, and the algorithm outputs $h_z = h^*$. In this event the true error is zero.

If no copy of z appears, then the sample has no positive example and the algorithm outputs h_-. The hypotheses h_- and h_z differ only at z, so in this event

$$\mathcal{R}(\widehat{h}) = \Pr\big(h_-(X) \neq h_z(X)\big) = \Pr(X = z) = q.$$

Therefore, for any $0 < \epsilon < 1$,

$$\begin{aligned}\Pr\big(\mathcal{R}(\widehat{h}) > \epsilon\big) &= \begin{cases} 0, & q \leq \epsilon, \\ (1-q)^m, & q > \epsilon, \end{cases} \\ &\leq (1-\epsilon)^m \\ &\leq e^{-m\epsilon}.\end{aligned}$$

Consequently, it is sufficient to take

$$m \geq \frac{1}{\epsilon} \log \frac{1}{\delta}$$

in order to guarantee

$$\Pr\big(\mathcal{R}(\widehat{h}) > \epsilon\big) \leq \delta.$$

Thus $\mathcal{H}$ is PAC learnable, with the sample-complexity upper bound

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{1}{\epsilon} \log \frac{1}{\delta} \right\rceil.$$

A slightly sharper sufficient bound, obtained before using $(1-\epsilon)^m \leq e^{-m\epsilon}$, is

$$m \geq \frac{\log(1/\delta)}{-\log(1-\epsilon)}.$$

Notice that the bound does not depend on the cardinality of $\mathcal{X}$; in particular, it remains valid when $\mathcal{X}$ is infinite.

Problem 3. Probability and concentration.

Suppose that we have m coins. Each coin has the same probability p of landing on heads. For each coin, perform n tosses, and assume that all mn tosses are mutually independent. The probability of obtaining exactly k heads in n tosses of any fixed coin is

$$\binom{n}{k} p^k (1-p)^{n-k}, \qquad k \in \{0, 1, \ldots, n\}.$$

Define

$$\widehat{p}_i := \frac{\text{number of times coin } i \text{ lands on heads}}{n}.$$

1. Compute the exact probability that at least one coin has $\widehat{p}_i = 0$.
2. For $\epsilon > 0$, apply Hoeffding's inequality and a union bound to prove an upper bound on

$$\Pr\left[\max_{1 \le i \le m} |\widehat{p}_i - p| > \epsilon\right].$$

Solution.

1. For each $i \in \{1, \ldots, m\}$, let

$$A_i := \{\widehat{p}_i = 0\}$$

be the event that coin i produces no heads in its n tosses. All n tosses of coin i must be tails, so

$$\Pr(A_i) = (1-p)^n.$$

Because the tosses of different coins are independent, the events $A_1, \ldots, A_m$ are independent. Hence

$$\begin{aligned} \Pr\left(\bigcap_{i=1}^{m} A_i^c\right) &= \prod_{i=1}^{m} \Pr(A_i^c) \\ &= \left(1 - (1-p)^n\right)^m. \end{aligned}$$

Taking the complementary event gives

$$\boxed{\Pr\left(\exists i \in \{1, \ldots, m\} : \widehat{p}_i = 0\right) = 1 - \left(1 - (1-p)^n\right)^m.}$$

2. Let Z_{ij} be the indicator that the jth toss of coin i is heads. Then

$$Z_{ij} \in \{0, 1\}, \qquad \mathbb{E}[Z_{ij}] = p, \qquad \widehat{p}_i = \frac{1}{n} \sum_{j=1}^{n} Z_{ij}.$$

For each fixed coin i, Hoeffding's inequality yields

$$\Pr(\widehat{p}_i - p > \epsilon) \le e^{-2n\epsilon^2},$$
$$\Pr(p - \widehat{p}_i > \epsilon) \le e^{-2n\epsilon^2}.$$

Therefore,

$$\Pr\big(|\widehat{p}_i - p| > \epsilon\big) \le 2e^{-2n\epsilon^2}.$$

Applying the union bound over the m coins gives

$$\begin{aligned}\Pr\left[\max_{1\le i\le m} |\widehat{p}_i - p| > \epsilon\right] &= \Pr\left(\bigcup_{i=1}^{m}\{|\widehat{p}_i - p| > \epsilon\}\right)\\ &\le \sum_{i=1}^{m} \Pr\big(|\widehat{p}_i - p| > \epsilon\big)\\ &\le 2me^{-2n\epsilon^2}.\end{aligned}$$

Thus

$$\boxed{\Pr\left[\max_{1\le i\le m} |\widehat{p}_i - p| > \epsilon\right] \le \min\left\{1,\, 2me^{-2n\epsilon^2}\right\}.}$$

Equivalently, for any $0 < \delta < 1$, if

$$n \ge \frac{1}{2\epsilon^2}\log\frac{2m}{\delta},$$

then, with probability at least $1 - \delta$,

$$\max_{1\le i\le m} |\widehat{p}_i - p| \le \epsilon.$$