Homework 1: Prediction, generalization, and evaluation

Due Friday, September 18, 2026, at 11:59 p.m. This assignment is worth 100 points and 10% of the course grade. Late homework is not accepted; the two-lowest-grade replacement described in the syllabus covers ordinary illness, overload, and missed work. Complete only the assigned parts; optional extensions are ungraded and need not be submitted. Submit one ZIP archive named hw01-ccid.zip, containing a single top-level directory named hw01-ccid with your PDF, LaTeX source, and any supporting files. Replace ccid with your University of Alberta CCID.

General instructions. Justify every answer, including yes/no answers, even when the question does not explicitly ask for a justification.

Name: YOUR NAME Student ID: YOUR STUDENT ID

Assigned work: Complete parts (a) and (c). Part (b) is an optional, ungraded extension.
Points: (a): 8, (c): 12; total 20.

1. **Bounded scores and attention concentration.** Let $T \geq 2$ be an integer and let $a, b \in \mathbb{R}$ satisfy $a \leq b$. Write $R = b - a$ for the score range. For a score vector $s \in [a, b]^T$, define the probability vector $\alpha(s)$ by
$$\alpha_j(s) = \frac{e^{s_j}}{\sum_{r=1}^{T} e^{s_r}}, \qquad j \in [T].$$

   (a) Compute $\sup_{s \in [a,b]^T} \max_{j \in [T]} \alpha_j(s)$ and give a score vector attaining this value.

   (b) *Optional extension.* Writing $H(\alpha) = -\sum_{j=1}^{T} \alpha_j \log \alpha_j$ for the *entropy*, show that for every $s \in [a, b]^T$,
$$\log T - R \leq H(\alpha(s)) \leq \log T.$$
   Deduce that, for fixed a, b, $H(\alpha(s))/\log T \to 1$ uniformly over $s \in [a, b]^T$ as $T \to \infty$. This concerns entropy relative to its maximum; it does not imply $\sum_{j=1}^{T} |\alpha_j(s) - 1/T| \to 0$.

   (c) For a fixed $c \in (0, 1)$, determine the smallest $R \geq 0$ for which some $s \in [a, a + R]^T$ satisfies $\max_{j \in [T]} \alpha_j(s) \geq c$. Give such a score vector at this minimum range. Describe how this minimum range grows with T for fixed c, and deduce whether a score range independent of T can maintain $\max_j \alpha_j(s) \geq c$ as $T \to \infty$.

Solution

Write your solution here.

Assigned work: Complete all parts.
Points: (a): 14, (b): 6; total 20.

2. **Which boxes should the data cover?** Let $N = 2n$, with $n \geq 2$. The boxes form two known groups of n boxes each. Labels are constant within each group, but the two groups may have different labels. The test index is uniform over all N boxes.

   (a) Compute the minimax risk for each of two training samples: two distinct boxes from the first group, and one box from each group. Give an optimal learner in each case and prove that its worst-case risk cannot be improved. Allow randomized learners.

   (b) Suppose the training sample reveals every box in the first group. Compute the minimax risk and compare it with the risks from observing just one box from that group and from observing one box from each group.

Solution

Write your solution here.

Assigned work: Complete parts (a), (b), and (c). Part (d) is an optional, ungraded extension.
Points: (a): 8, (b): 10, (c): 12; total 30.

3. **Classification from probabilistic prediction.** Let (X, Y) have an arbitrary distribution on $\mathcal{X} \times \{0, 1\}$, where $\mathcal{X}$ is finite. For every x with $\mathbb{P}(X = x) > 0$, define

$$\eta(x) = \mathbb{P}(Y = 1 \mid X = x).$$

At the remaining inputs, $\eta(x)$ may be defined arbitrarily. For a decision $a \in \{0, 1\}$ and an x with $\mathbb{P}(X = x) > 0$, define the conditional risk

$$r_x(a) = \mathbb{P}(a \neq Y \mid X = x).$$

For a classifier $h : \mathcal{X} \to \{0, 1\}$, define its risk by

$$R_{01}(h) = \mathbb{P}(h(X) \neq Y).$$

A classifier is *Bayes-optimal* if it minimizes R_{01} over all classifiers $h : \mathcal{X} \to \{0, 1\}$. The minimum value

$$R^*_{01} = \min_{h:\mathcal{X}\to\{0,1\}} R_{01}(h)$$

is the *Bayes risk*.

(a) Compute $r_x(0)$ and $r_x(1)$. Show that

$$h^*(x) = \mathbb{I}\{\eta(x) \geq 1/2\}$$

is Bayes-optimal and that the Bayes risk is

$$R^*_{01} = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}].$$

(b) Prove the exact excess-risk identity

$$R_{01}(h) - R^*_{01} = \mathbb{E}\left[|2\eta(X) - 1|\mathbb{I}\{h(X) \neq h^*(X)\}\right].$$

(c) Let $q : \mathcal{X} \to (0, 1)$ be a probabilistic predictor, and define

$$L_{\log}(q) = \mathbb{E}[-Y \log q(X) - (1 - Y) \log(1 - q(X))].$$

Let $L^*_{\log} = L_{\log}(\eta)$, with the usual conventions at $\eta(x) \in \{0, 1\}$, and let $h_q(x) = \mathbb{I}\{q(x) \geq 1/2\}$. Prove

$$R_{01}(h_q) - R^*_{01} \leq \sqrt{2\big(L_{\log}(q) - L^*_{\log}\big)}.$$

(d) *Optional extension.* Construct a family of probabilistic predictors whose induced classifiers h_q are Bayes-optimal but whose excess log loss tends to infinity.

Hint for (c): You may use Pinsker's inequality for Bernoulli distributions:

$$|p - q| \leq \sqrt{\frac{1}{2}\left(p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}\right)}.$$

What this exercise is meant to teach. A probabilistic predictor and a classifier answer different questions. The excess classification risk weights a wrong decision by twice the distance of $\eta(x)$ from the decision boundary $1/2$. Small excess log loss controls excess classification risk, but Bayes-optimal classification does not imply accurate probabilities or small log loss.

Solution

Write your solution here.

Assigned work: Complete all parts.
Points: (a): 8, (b): 8, (c): 6, (d): 8; total 30.

4. **Evaluation on a fresh test set.** Let $\mathcal{X}$ be finite. Let $D = ((X_i, Y_i))_{i=1}^m$ be an i.i.d. test sample from a distribution on $\mathcal{X} \times \{0,1\}$, and let (X, Y) be an independent draw from the same distribution. For a classifier $h : \mathcal{X} \to \{0,1\}$, define

$$R(h) = \mathbb{P}(h(X) \neq Y), \qquad \widehat{R}_D(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\{h(X_i) \neq Y_i\}.$$

(a) Show that every classifier $h : \mathcal{X} \to \{0,1\}$ satisfies, for every $\epsilon > 0$,

$$\mathbb{P}\big(|\widehat{R}_D(h) - R(h)| > \epsilon\big) \leq 2e^{-2m\epsilon^2}.$$

(b) Fix a finite set $\mathcal{H} = \{h_1, \ldots, h_K\}$ of classifiers. Use the union bound to prove

$$\mathbb{P}\left(\max_{h \in \mathcal{H}} |\widehat{R}_D(h) - R(h)| > \epsilon\right) \leq 2Ke^{-2m\epsilon^2}.$$

Give a sufficient sample size for the displayed maximum to be at most ϵ with probability at least $1 - \delta$.

(c) Let $f : (\mathcal{X} \times \{0,1\})^m \to \mathcal{H}$ and let $\widehat{h} = f(D)$. Show that

$$\mathbb{P}\left(\left|\widehat{R}_D(\widehat{h}) - R(\widehat{h})\right| > \epsilon\right) \leq 2Ke^{-2m\epsilon^2}.$$

(d) The previous part uses a simultaneous guarantee over all classifiers that the selection rule may return. Show what can happen without such a guarantee: a classifier defined using the test sample can have zero test error but large population risk. In particular, consider the following: Let $\mathcal{X} = [N]$, let X be uniform on $[N]$, and let $Y = 1$ almost surely. Define the following function of the sample D:

$$h_D(x) = \mathbb{I}\{x \in \{X_1, \ldots, X_m\}\}.$$

Compute $\widehat{R}_D(h_D)$ and $R(h_D)$. Show that $N \geq 2m$ implies $R(h_D) \geq 1/2$, and explain why this does not contradict part (a) or part (b).

Hint for (a): Apply Hoeffding's inequality to the independent random variables $\mathbb{I}\{h(X_i) \neq Y_i\}$.

Solution

Write your solution here.