

CMPUT 654: Mathematical Foundations of Modern AI Systems

Fall 2026

Lecture 3: Probabilistic models, log loss, and compression

September 8, 2026

Lecturer: Csaba Szepesvári

Scribe: Student name

3.1 Exercises

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

1. **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 $h : \mathcal{X} \rightarrow \{0, 1\}$, 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 $h : \mathcal{X} \rightarrow \{0, 1\}$. The minimum value

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

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

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

- (c) Let $q : \mathcal{X} \rightarrow (0, 1)$ 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

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

- (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.

2. **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 $h : \mathcal{X} \rightarrow \{0, 1\}$, define

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

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

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

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

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

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

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

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

- (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 :

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

Compute $\hat{R}_D(h_D)$ 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 $\mathbb{I}\{h(X_i) \neq Y_i\}$.