6.S951 Modern Mathematical Statistics Fall 2024

Lecture 5 — September 19, 2024

Prof. Stephen Bates

Scribe: Andreas Petrou-Zeniou

1 Outline

Agenda:

1. Statistical decision theory for prediction
2. PAC for finite function classes

Last time:

1. Minimax in Gaussian linear model
2. Least-squares with model misspecification

Next two weeks:

1. Optimal hypothesis testing
2. Information-theoretic minimax lower bounds

2 Statistical Decision Theory for Prediction

Data: $(Z_i, Y_i) \in \mathcal{Z} \times \mathcal{Y}$, $i = 1, \ldots, n$, i.i.d. from $P_\theta \in \mathcal{P}$, where $\mathcal{P}$ is the set of all distributions over $\mathcal{Z} \times \mathcal{Y}$.

Intuitive Goal: We want to find $h : \mathcal{Z} \to \mathcal{Y}$ that is good at predicting Y_i from Z_i. Our action space is therefore $\mathcal{A} = \{h; h \in \mathcal{H}\}$ where $\mathcal{H}$ is a class of functions $h : \mathcal{Z} \to \mathcal{Y}$.

Definition: We define the single point loss $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. We similarly define the loss:

$$L(h, P_\theta) = \mathbb{E}_{(Z_0,Y_0)\sim P_\theta} \left(l(h(Z_0), Y_0) \right)$$

We can think of this as the average performance in a holdout set, or the average forecast performance.

Examples: We may have 0-1 loss for classification problems, defined by $l(\hat{y}, y) = \mathbb{I}(\hat{y} \neq y)$. We may also have mean squared error $l(\hat{y}, y) = (\hat{y} - y)^2$.

Definition: We define a prediction procedure as a function A from our sample to our class of functions $\mathcal{H}$, that is $A : (\mathcal{Z} \times \mathcal{Y})^n \to \mathcal{H}$.

Definition: For a P_θ and a function class $\mathcal{H}$, the optimal loss is:

$$L^* = \inf_{h \in \mathcal{H}} L(h, P_\theta)$$

Formal Goal: Find A such that $L(A(X), P_\theta)$ is close to L^*

Definition: A procedure is (ε-δ)-PAC (probably approximately correct) if:

$$\sup_{P_\theta \in \mathcal{P}} P\left(L(A(X), P_\theta) > L^* + \varepsilon\right) < \delta$$

We notice that the probability above depends crucially on $\mathcal{P}$, $\mathcal{H}$, and n. Visually, we have:

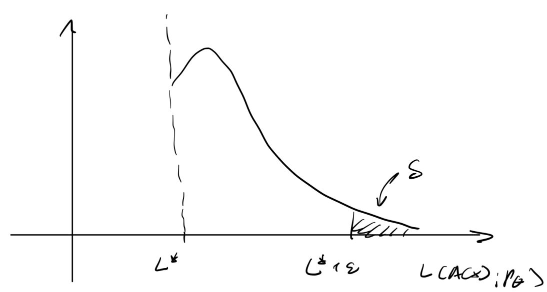

Figure 1: (ε-δ)-PAC visually

Definition: We say that a procedure A minimizes empirical risk if:

$$A(X) = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l(h(Z_i), Y_i) =: \arg\min_{h \in \mathcal{H}} \bar{L}(h, P_\theta)$$

Example: Let $\mathcal{H}$ be the class of linear functions from $\mathbb{R}^d$ to $\mathbb{R}$. Let $l(\hat{y}, y)$ be the squared error loss. Here, the $A(X)$ that minimizes empirical risk is simply least-squares.

3 PAC for finite function classes

Let us consider a finite function class $\mathcal{H} = \{h_1, h_2, \ldots, h_K\}$, and loss $l(\hat{y}, y) \in [0, 1]$. Our goal is to provide finite-sample rates on ε and δ for empirical risk minimization, which is (ε-δ)-PAC.

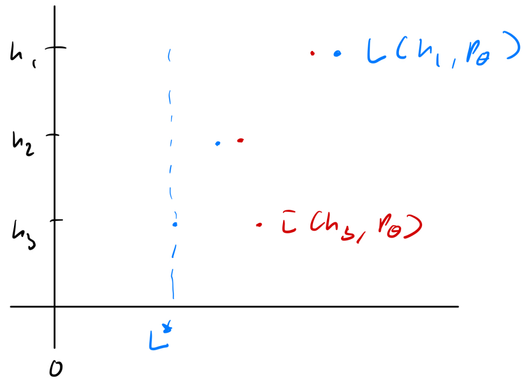

Figure 2: Minimizing empirical risk with finite $\mathcal{H}$

In the above figure, h_2 minimizes empirical loss. Now, let us provide the following proposition:

Proposition (Hoeffding's Inequality): Let W_i be i.i.d. supported on $[0, 1]$. We have:

$$P\left(\left|\frac{1}{n}\sum W_i - \mathbb{E}(W_i)\right| > \varepsilon\right) \leq 2\exp(-2\varepsilon^2/n)$$

Proof: See here. ■

Theorem (PAC for finite $\mathcal{H}$): For empirical loss minimizing A and any $\varepsilon > 0$, we have:

$$P\left(L(A(X), P_\theta) > L^* + \varepsilon\right) \leq 2K\exp(-n\varepsilon^2/2)$$

Proof: We first notice that A selects h with $L(h, P_\theta) \geq L^* + \varepsilon$ only if:

$$\max_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} l(h(Z_i), Y_i) - L(h, P_\theta)\right| > \varepsilon/2$$

Thus, first applying a union bound and Hoeffding's inequality, we get:

$$\begin{aligned}
&P\left(L(A(X), P_\theta) > L^* + \varepsilon\right) \\
&\leq P\left(\max_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} l(h(Z_i), Y_i) - L(h, P_\theta)\right| > \varepsilon/2\right) \\
&\leq \sum_{k=1}^{K} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} l(h_k(Z_i), Y_i) - L(h_k, P_\theta)\right| > \varepsilon/2\right) \\
&\leq 2K\exp(-n\varepsilon^2/2)
\end{aligned}$$

We can take the supremum over P_θ in $\mathcal{P}$ and the result follows. ■

Interpreting the Finite-Sample Bound: Finally, we can fix δ (say at 1%) to find the corresponding ε, getting:

$$\varepsilon = \sqrt{\frac{2\log(2K) + 2\log(1/\delta)}{n}}$$

We can notice that we are converging at rate $1/\sqrt{n}$. Moreover, our precision is decreasing in the size of the function class K.

Parting Comments: These results can be extended to infinite function classes $\mathcal{H}$. Moreover, we can produce hardness bounds (as we did for minimax estimation) by imposing a prior over Θ.