6.S951 Modern Mathematical Statistics Fall 2024

Lecture 6 — September 24, 2024

Prof. Stephen Bates

Scribe: Santiago Torres

1 Outline

So far:

1. Statistical decision theory (notions of optimality).
2. Least-squares in the Gaussian Linear Model.
3. High probability results for prediction tasks.

Agenda:

1. Hypothesis testing basics.
2. Optimal simple-simple tests (Neyman-Pearson).
3. Total variation distance.

Recap - Taxonomy of tasks:

1. **Estimation:** Find a property of the underlying distribution - *Action space:* $\mathbb{R}^d$.
2. **Prediction:** Find a function that can predict well over unseen data coming from the underlying distribution - *Action space:* $\mathcal{H}$, some space of functions.
3. **Hypothesis testing:** Distinguish over two possible sets of distributions. - *Action space:* $\{0, 1\}$, binary choice.

2 Hypothesis testing basics

Setup:

- *Space of distributions (parametrized by $\theta \in \Theta$):* $\Theta = \underbrace{\Theta_0}_{\text{Null}} \cup \underbrace{\Theta_1}_{\text{Alternative}}$, s.t. $\Theta_0 \cap \Theta_1 = \emptyset$.
- *Action Space:* $\mathcal{A} = \{0, 1\}$, where 0 corresponds to choosing the null, and 1 to choosing the alternative.

- *Loss function:* For some numbers $c_{\text{FP}}, c_{\text{FN}} > 0$:

$$L(a,\theta) = \begin{cases} 0 & \text{if } a = 0 \text{ and } \theta \in \Theta_0, \\ c_{\text{FP}} & \text{if } a = 1 \text{ and } \theta \in \Theta_0, \\ 0 & \text{if } a = 1 \text{ and } \theta \in \Theta_1, \\ c_{\text{FN}} & \text{if } a = 0 \text{ and } \theta \in \Theta_1. \end{cases}$$

Why study testing?

1) *Fundamental theoretical problem:* How well can we distinguish from different distributions possibly generating the data?

2) *Applied importance:* Is there evidence of a given structure beyond noise?

3) *Provides general insights for binary statistical decision-making.*

3 Optimal simple vs. simple hypothesis tests

Setup:

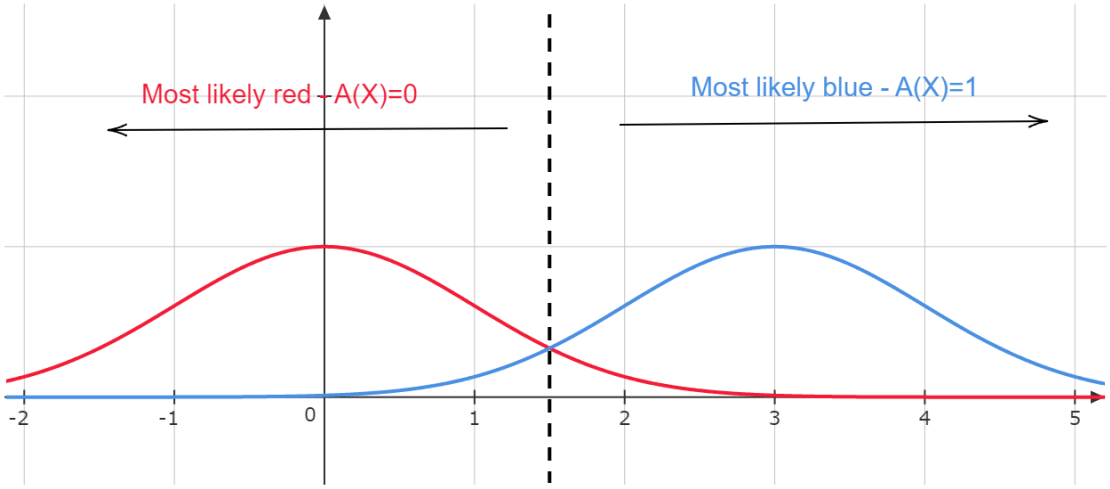
- We want to distinguish between two distributions, namely $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.
- We observe data $X \sim P_\theta$.
- We want to design optimal procedures $A : \mathcal{X} \to \{0, 1\}$ to select from these distributions.

Example:

Suppose $P_{\theta_0} = \mathcal{N}(0, 1)$ and $P_{\theta_1} = \mathcal{N}(3, 1)$, and we observe a single data point X. The rule

$$A(X) = \begin{cases} 1 & \text{if } X > 3/2, \\ 0 & \text{otherwise.} \end{cases}$$

seems like a sensible procedure to determine the most probable underlying distribution (see figure below).



Naturally, any hypothesis testing procedure will not always be correct. Therefore, we usually distinguish between the types of mistakes, and the probabilities thereof, that we can incur by implementing a particular decision rule:

- *Type I error - False Positive:* $P_{\theta_0}(A(X) = 1)$
- *Type II error - False Negative:* $P_{\theta_1}(A(X) = 0)$

Since hypothesis testing falls within the general statistical decision theory framework studied in class, we can use the notions of optimality we have introduced in previous lectures. In particular, we can study Bayes optimal hypothesis testing rules:

Let Q be a prior over Θ. Since Θ consists only of two points, let $Q(\theta_0) = \pi_0 \in (0, 1)$ and $Q(\theta_1) = \pi_1 = 1 - \pi_0$. Then the Bayes risk of an arbitrary procedure $A : \mathcal{X} \to \{0, 1\}$ under Q is

$$\begin{aligned} R_B(A(\cdot), Q) &= \mathbb{E}_{\theta \sim Q}[\mathbb{E}_{X \sim P_\theta}[L(A(X), \theta) \mid \theta]] \\ &= \pi_0 P_{\theta_0}(A(X) = 1) c_{\text{FP}} + \pi_1 P_{\theta_1}(A(X) = 0) c_{\text{FN}} \end{aligned}$$

Note: Observe that the Bayes risk is a linear combination of the probabilities of committing the various errors.

In particular, if $\pi_0 = \pi_1 = 1/2$, and $c_{\text{FN}} = c_{\text{FP}} = 1$, this reduces to

$$R_B(A(\cdot), Q) = \frac{1}{2}(P_{\theta_0}(A(X) = 1) + P_{\theta_1}(A(X) = 0))$$

Theorem 1 (Bayes-optimal test)**.** *Suppose P_{θ_0} has density f_0 and P_{θ_1} has density f_1. Then the Bayes-optimal test is*

$$A(x) = \begin{cases} 1 & \text{if} \quad \dfrac{f_1(x)}{f_0(x)} > \dfrac{\pi_0 c_{FP}}{\pi_1 c_{FN}}, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. To find the Bayes-optimal procedure we calculate the posterior distribution and choose the action that minimizes the expected loss. By the Bayes rule, we have that

$$\mathbb{P}(\theta = \theta_0 | X = x) = \frac{\mathbb{P}(x|\theta = \theta_0)\mathbb{P}(\theta = \theta_0)}{\mathbb{P}(x|\theta = \theta_0)\mathbb{P}(\theta = \theta_0) + \mathbb{P}(x|\theta = \theta_1)\mathbb{P}(\theta = \theta_1)} = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

Likewise,

$$\mathbb{P}(\theta = \theta_1 | X = x) = \frac{\mathbb{P}(x|\theta = \theta_1)\mathbb{P}(\theta = \theta_1)}{\mathbb{P}(x|\theta = \theta_0)\mathbb{P}(\theta = \theta_0) + \mathbb{P}(x|\theta = \theta_1)\mathbb{P}(\theta = \theta_1)} = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

Next, we note that

- *Expected loss of choosing* $a(x) = 0$*:*

$$\frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} c_{\text{FN}}$$

- *Expected loss of choosing* $a(x) = 1$*:*

$$\frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} c_{\text{FP}}$$

whence it is optimal to choose $a(x) = 1$ whenever

$$\frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} c_{\text{FP}} < \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} c_{\text{FN}} \longleftrightarrow \frac{f_1(x)}{f_0(x)} > \frac{\pi_0 c_{\text{FP}}}{\pi_1 c_{\text{FN}}}$$

□

Note: The function $\frac{f_1(x)}{f_0(x)}$ is known as the *likelihood ratio.*

Example (continued):

Suppose $P_{\theta_0} = \mathcal{N}(0, 1)$ and $P_{\theta_1} = \mathcal{N}(3, 1)$. Thus, the likelihood ratio takes the form:

$$\frac{f_1(x)}{f_0(x)} = \frac{(2\pi)^{-1/2} \exp\left(-\frac{1}{2}(x-3)^2\right)}{(2\pi)^{-1/2} \exp\left(-\frac{1}{2}x^2\right)} = \exp\left(-\frac{1}{2}(9 - 6x)\right)$$

Since the likelihood ratio is increasing in x, it suggests that the Bayes optimal decision procedure follows a thresholding rule:

$$A(x) = \begin{cases} 1 & \text{if } x > \tau, \\ 0 & \text{otherwise.} \end{cases}$$

In the case $\pi_0 = \pi_1 = 1/2$, and $c_{\text{FN}} = c_{\text{FP}} = 1$, then $\tau = 3/2$.

Alternative Hypothesis Testing Setup: Often, hypothesis testing occurs within a different framework than the one just stated. In many cases, we would like to design a procedure that guarantees that the type I error is less than some level $\alpha \in (0, 1)$. In these settings, we usually fix Θ_0, the null, as a no effect or status quo world, while Θ_1, the alternative, contains some interest effect or structure we want to establish. Thus, the idea is that the burden of proof is on us to establish if it is really the case that $\theta \in \Theta_1$.

Definition (Neyman-Pearson Optimality): An statistical test $A^*(\cdot)$ satisfies the Neyman-Pearson Optimality if it is a solution to the problem

$$\sup_{A} \underbrace{P_{\theta_1}(A(X) = 1))}_{\text{Power} = 1 - \text{Prob. of Type II error}} \quad \text{s.t.} \quad \underbrace{P_{\theta_0}(A(X) = 1)}_{\text{Prob. of Type I error}} \leq \alpha \tag{1}$$

Theorem 2 (The Neyman-Pearson Lemma)**.** *Suppose P_{θ_0} has density f_0 and P_{θ_1} has density f_1. Then a solution $A^{NP} : \mathcal{X} \to \{0, 1\}$ to (1) is*

$$A^{NP}(x) = \begin{cases} 1 & \textit{if} \quad \dfrac{f_1(x)}{f_0(x)} > \lambda, \\ 0 & \textit{otherwise.} \end{cases}$$

where λ satisfies

$$P_{\theta_0}\left(\frac{f_1(x)}{f_0(x)} > \lambda\right) = \alpha$$

Note: The Neyman-Pearson solution resembles the Bayes-optimal test for some combination of $\pi_1, \pi_0, c_{\text{FP}}, c_{\text{NP}}$, which are implicitly determined by the choice of f_1, f_0, and α.

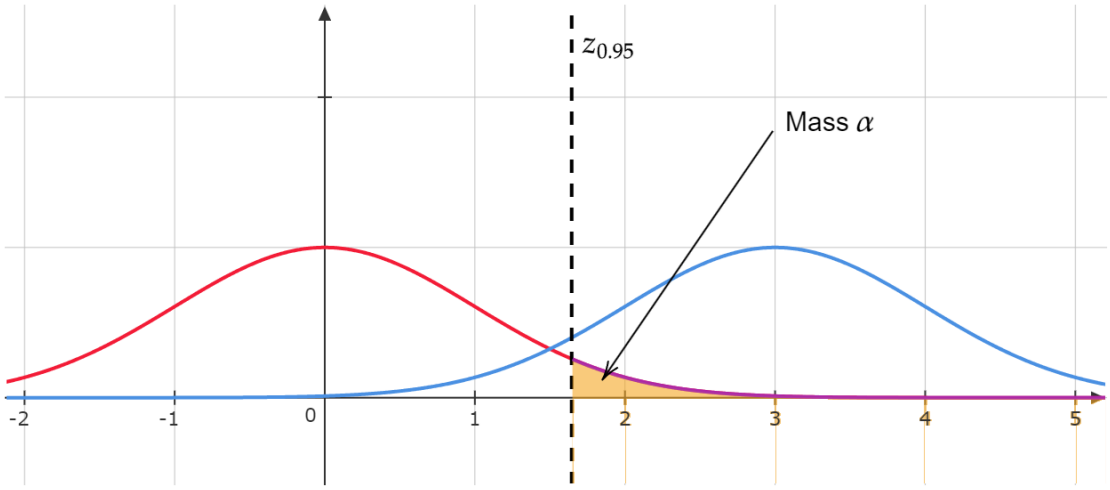
Example (continued):

Suppose $P_{\theta_0} = \mathcal{N}(0, 1)$ and $P_{\theta_1} = \mathcal{N}(3, 1)$, and let $\alpha = 0.05$. The Neyman-Pearson optimal procedure is given by

$$A^{\text{NP}}(x) = \begin{cases} 1 & \text{if} \quad x > \lambda = \exp\left(3z_{0.95} - \dfrac{9}{2}\right), \\ 0 & \text{otherwise.} \end{cases}$$

where $z_{0.95}$ is such that the excess mass to the right of this point in the normal standard pdf is exactly $\alpha = 0.95$.[1]

[1] This λ is the 95th percentile of the random variable $\frac{f_1(X)}{f_0(X)}$ when $X \sim \mathcal{N}(0, 1)$.



4 Total Variation Distance

Definition: The total variation distance between two distributions P_{θ_0} and P_{θ_1} over some common space $\mathcal{X}$ is

$$||P_{\theta_0} - P_{\theta_1}||_{\text{TV}} = \sup_{B \subseteq \mathcal{X}} |P_{\theta_1}(X \in B) - P_{\theta_0}(X \in B)| \tag{2}$$

Lemma 3. *If P_{θ_0} has density f_0 and P_{θ_1} has density f_1, the set B^{Opt} that optimizes (2) is*

$$B^{Opt} = \{x : f_1(x) > f_0(x)\}$$

Proof. Consider some alternative set $B \subseteq X$, with $B \neq B^{\text{Opt}}$. Without loss of generality, assume that $P_{\theta_1}(B) > P_{\theta_0}(B)$ (else take the complements). Then,

$$\begin{aligned} P_{\theta_1}(B^{\text{Opt}}) - P_{\theta_0}(B^{\text{Opt}}) - (P_{\theta_1}(B) - P_{\theta_0}(B)) &= \int_{B^{\text{Opt}}} f_1(x)dx - \int_{B^{\text{Opt}}} f_0(x)dx - \left(\int_B f_1(x)dx - \int_B f_0(x)dx\right) \\ &= \underbrace{\int_{B^{\text{Opt}} \setminus B} (f_1(x) - f_0(x))dx}_{>0} - \underbrace{\int_{B \setminus B^{\text{Opt}}} (f_1(x) - f_0(x))dx}_{<0} \\ &> 0 \end{aligned}$$

□

Theorem 4 (Testing hardness bound by the TV distance)**.** *To be completed in next lecture*

$$\inf_{A:\mathcal{X} \to \{0,1\}} P_{\theta_0}(A(X) = 1) + P_{\theta_1}(A(X) = 0) = 1 - ||P_{\theta_0} - P_{\theta_1}||_{TV}$$