DS 102 Data, Inference, and Decisions **Spring 2020**

Lecture 3: Neyman-Pearson and Likelihood Ratio Tests

Lecturer: Moritz Hardt

3.1 Review of Decision Theory

In Lecture 2, we introduced statistical decision theory, and connected it to both Bayesian and frequentist perspectives. This discussion introduced several new terms, which we review in this section.

We denote by $\mathbb{P}$ the joint distribution over data X and θ, a parameter of interest. For example, θ can be the state of reality that generated the data, or a label corresponding to the data X.

It is often useful to study the corresponding marginal distributions. $\mathbb{P}(X)$ denotes the marginal over the data X. Taking a Bayesian perspective, $\mathbb{P}(\theta)$ denotes the marginal over the parameter θ. Also known as the prior, $\mathbb{P}(\theta)$ represents our understanding of the distribution of θ before observing any data.

The conditional distributions related to $\mathbb{P}$ also play an important role, as we will see in today's discussion of the Likelihood Ratio Test. The likelihood, $\mathbb{P}(X \mid \theta)$, is the distribution of the data conditional on (a particular value of) the parameter θ. For example, in the setting of a hypothesis test, $\mathbb{P}(X \mid \theta_0)$ represents the distribution of the data X under the null hypothesis. The posterior, $\mathbb{P}(\theta \mid X)$, is the distribution of θ given the observed data X. The posterior is a Bayesian concept, and captures the information that X gives us about the distribution of θ.

We also define $\delta(X)$ to be our decision function, and $\ell(\theta, \delta(X))$ as our loss function. Using these component, we defined a few different notions of risk:

- Frequentist risk is the expected loss, where the expectation is taken according to the likelihood:
$$r(\theta) = \mathbb{E}_{X \sim \mathbb{P}(X|\theta)}[\ell(\theta, \delta(X))].$$
In this frequentist setting, we assume that there is one true (but unknown) θ. We do not have a prior over θ, instead we work exclusively with the likelihood.
- The Bayesian posterior risk is the expected loss when we average over the posterior,
$$\rho(X) = \mathbb{E}_{\theta \sim \mathbb{P}(\theta|X)}[\ell(\theta, \delta(X))],$$
and it measures how well $\delta(X)$ captures θ given the observed data.
- Averaging either $r(\theta)$ over θ or $\rho(X)$ over X gives the same number, known as the Bayes risk:
$$R(\delta) = \mathbb{E}_{\theta \sim \mathbb{P}(\theta)}[r(\theta)] = \mathbb{E}_{X \sim \mathbb{P}(X)}[\rho(X)] = \mathbb{E}_{(X,\theta) \sim \mathbb{P}}[\ell(\theta, \delta(X))].$$
Once the decision rule $\delta(\cdot)$ is fixed, the Bayes risk is a real number.

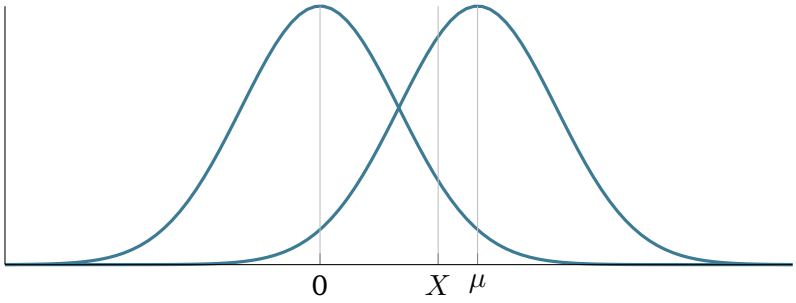
3.2 Binary Testing

Today, we will restrict our focus to the binary setting: $\theta \in \{\theta_0, \theta_1\}$.

Example 3.1. We may want to distinguish two Gaussians with the same variance,

$$\mathbb{P}(X|\theta_0) = N(0, \sigma^2) \text{ and,}$$

$$\mathbb{P}(X|\theta_1) = N(\mu, \sigma^2) \text{ where } \mu > 0.$$



We would like to tell these two distributions apart given a sample X. The decision we make will ultimately depend on the data we observe and how we determine the cost of making false positives.

3.3 Neyman-Pearson Formulation

What do optimal decision rules look like in this binary setting?

In general, our the optimal decision rule makes the true positive rate as high as possible, while making very few false positives. The Neyman-Pearson formulation, introduced in Lecture 1, aims to find a decision rule δ that maximizes the true positive rate (TPR) subject to a constraint on the false positive rate (FPR):

$$\begin{aligned} \max_{\delta} \quad & \text{TPR}(\delta) \\ \text{s.t.} \quad & \text{FPR}(\delta) \leq \alpha. \end{aligned}$$

We can rewrite the TPR(δ) and FPR(δ) terms in this optimization problem in a few different ways. In our old notation, TPR(δ) = Pr($D = 1 \mid R = \theta_1$), and in our decision theoretic notation,

$$\text{TPR}(\delta) = \int \delta(X)\mathbb{P}(X \mid \theta_1)\, dX = \mathbb{E}_{X\sim\mathbb{P}(X|\theta_1)}[\delta(X)].$$

Although we usually think of $\delta(X)$ as being either 0 or 1 (where 1 represents guessing that θ_1 is the state of reality and 0 represents accepting that θ_0 as reality), this expression is valid whenever $\delta(X) \in [0, 1]$ if we interpret $\delta(X)$ as the probability of choosing 1.

We have similar formulae for the false positive rate:

$$\begin{aligned}\text{FPR}(\delta) &= \Pr(D = 1 \mid R = \theta_0) \\ &= \int \delta(X)\mathbb{P}(X \mid \theta_0)\, dX \\ &= \mathbb{E}_{X\sim\mathbb{P}(X|\theta_0)}[\delta(X)].\end{aligned}$$

In our decision theoretic notation, the optimization problem corresponding to the Neyman-Pearson formulation is

$$\begin{aligned}\max_{\delta} \quad & \int \delta(X)\mathbb{P}(X \mid \theta_1)\, dX \\ \text{s.t.} \quad & \int \delta(X)\mathbb{P}(X \mid \theta_0)\, dX \leq \alpha.\end{aligned} \tag{$\star$}$$

3.4 Likelihood Ratio Tests

As we will see in the next section, the problem ($\star$) is intimately related to a type of hypothesis test known as the **Likelihood Ratio Test** (LRT). Define the likelihood ratio by $L(X) = \frac{\mathbb{P}(X \mid \theta_1)}{\mathbb{P}(X \mid \theta_0)}$. The LRT with threshold $\lambda \geq 0$ is the decision rule

$$\delta_\lambda(X) = \begin{cases} 0 & L(X) \leq \lambda \\ 1 & L(X) > \lambda \end{cases}$$

Intuitively, δ_λ guesses 1 if the data X looks more than λ times more likely under $\mathbb{P}(X \mid \theta_1)$ than under $\mathbb{P}(X \mid \theta_0)$, and guesses 0 otherwise. We can write the the LRT more succinctly as

$$\delta_\lambda(X) = \mathbb{1}\{L(X) > \lambda\}.$$

3.5 Neyman-Pearson Lemma

It turns out that optimal tests according to ($\star$) look like likelihood ratio tests. This result is known as the **Neyman-Pearson Lemma**.

Lemma 3.2 (Neyman-Pearson)**.** *The optimal solution to the optimization problem* ($\star$) *is the likelihood ratio test δ_λ, where the choice of λ depends on the chosen significance level α.*

Proof of Neyman-Pearson. We begin by applying the Lagrange multiplier method to the constrained optimization problem ($\star$), which gives

$$\max_{\delta} \quad \min_{\lambda \geq 0} \quad \int \delta(X)\mathbb{P}(X \mid \theta_1)\, dX - \lambda \left(\int \delta(X)\mathbb{P}(X \mid \theta_0)\, dX - \alpha \right).$$

We can relax this further by exchanging the order of the optimization problems and invoking weak duality:

$$\begin{aligned}\max_{\delta} \quad \min_{\lambda \geq 0} \quad & \int \delta(X)\mathbb{P}(X \mid \theta_1)\, dX - \lambda \left(\int \delta(X)\mathbb{P}(X \mid \theta_0)\, dX - \alpha \right) \\ & \leq \min_{\lambda \geq 0} \quad \max_{\delta} \quad \int \delta(X)\mathbb{P}(X \mid \theta_1)\, dX - \lambda \left(\int \delta(X)\mathbb{P}(X \mid \theta_0)\, dX - \alpha \right) \\ & = \min_{\lambda \geq 0} \quad \max_{\delta} \quad \int \delta(X) \left(\mathbb{P}(X \mid \theta_1) - \lambda \mathbb{P}(X \mid \theta_0) \right) dX + \lambda\alpha.\end{aligned}$$

Consider the inner optimization problem

$$\max_{\delta} \quad \int \delta(X) \left(\mathbb{P}(X \mid \theta_1) - \lambda \mathbb{P}(X \mid \theta_0) \right) dX \tag{$\star\star$}$$

The problem ($\star\star$) is solved by the LRT δ_λ. To see why, note that the objective is maximized by choosing $\delta(X) = 1$ for all X where $\mathbb{P}(X \mid \theta_1) - \lambda \mathbb{P}(X \mid \theta_0) > 0$, and $\delta(X) = 0$ wherever $\mathbb{P}(X \mid \theta_1) - \lambda \mathbb{P}(X \mid \theta_0)$ is non-positive.

To prove that the optimal solution to the original problem ($\star$) is a LRT, it remains to show that the inequality we introduced by invoking weak duality is actually tight for the LRT δ_λ, provided that we choose λ carefully so as to give us false positive rate α. To do so, first note that

$$\text{FPR}(\delta_\lambda) = \int \delta_\lambda(X)\mathbb{P}(X \mid \theta_0)\, dX = \int_{L(X) > \lambda} \mathbb{P}(X \mid \theta_0)\, dX$$

is a continuous function of λ. Therefore, we can always choose a λ_* such that $\text{FPR}(\delta_{\lambda_*}) = \alpha$. But then,

$$\begin{aligned}\text{TPR}(\delta_{\lambda_*}) & \leq \max_{\delta} \quad \text{TPR}(\delta) \quad \text{s.t.} \quad \text{FPR}(\delta) \leq \alpha \\ & \leq \min_{\lambda \geq 0} \quad \max_{\delta} \quad \int \delta(X) \left(\mathbb{P}(X \mid \theta_1) - \lambda \mathbb{P}(X \mid \theta_0) \right) dX + \lambda\alpha \\ & \leq \max_{\delta} \quad \int \delta(X) \left(\mathbb{P}(X \mid \theta_1) - \lambda_* \mathbb{P}(X \mid \theta_0) \right) dX + \lambda_*\alpha \\ & = \int \delta_{\lambda_*}(X) \left(\mathbb{P}(X \mid \theta_1) - \lambda_* \mathbb{P}(X \mid \theta_0) \right) dX + \lambda_*\alpha \\ & = \int \delta_{\lambda_*}(X)\mathbb{P}(X \mid \theta_1) - \lambda_* \int \delta_{\lambda_*}(X)\mathbb{P}(X \mid \theta_0)\, dX + \lambda_*\alpha \\ & = \text{TPR}(\delta_{\lambda_*}) - \lambda_* \text{FPR}(\delta_{\lambda_*}) + \lambda_*\alpha \\ & = \text{TPR}(\delta_{\lambda_*}).\end{aligned}$$

Since the first and last quantities are clearly equal, all the inequalities must actually be equalities. We have therefore shown that the LRT δ_{λ_*} solves the Neyman-Pearson problem ($\star$).

□

Returning to **Example 3.1**, the likelihood ratio in this case has the form

$$L(X) = \frac{e^{\frac{-X^2}{2\sigma^2}}}{e^{\frac{-(X-\mu)^2}{2\sigma^2}}} = e^{\frac{(X-\mu)^2}{2\sigma^2} - \frac{X^2}{2\sigma^2}}.$$

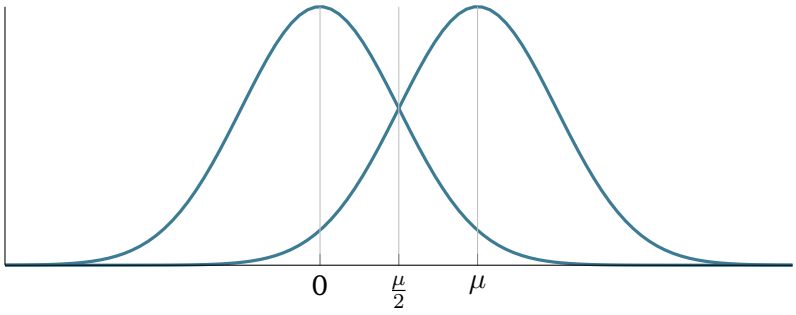
We can simplify the likelihood ratio condition $L(X) \leq \lambda$ by taking logarithms on both sides, which gives

$$\log L(X) \leq \log(\lambda) \Leftrightarrow X \leq \frac{\mu}{2} + \frac{\log(\lambda)\sigma^2}{\mu}$$

by expanding the square and rearranging terms.

We can therefore write the LRT for this problem as

$$\delta_\lambda(X) = \begin{cases} 0 & X \leq \frac{\mu}{2} + \frac{\log(\lambda)\sigma^2}{\mu} \\ 1 & X > \frac{\mu}{2} + \frac{\log(\lambda)\sigma^2}{\mu} \end{cases}$$



The condition $X \leq \frac{\mu}{2}$ is an intuitive threshold to choose between mean 0 and mean μ, but it isn't clear how the decision rule based on this threshold will fare in terms of true and false positive rates. The likelihood ratio condition $X \leq \frac{\mu}{2} + \frac{\log(\lambda)\sigma^2}{\mu}$ contains an additive term which adjusts the intuitive threshold $\frac{\mu}{2}$ to achieve the desired false positive rate.

In this example, the false positive rate corresponds to a Gaussian tail:

$$\text{FPR}(\delta_\lambda) = \int_{\frac{\mu}{2} + \frac{\log(\lambda)\sigma^2}{\mu}}^{\infty} \mathbb{P}(X \mid \theta_0)\, dX$$