6.780 Notes

Lecturer:

Andrew Liu

Spring 2024

Last updated on Sunday 10th March, 2024.

Contents

1 February 6, 2024

2 February 8, 2024

The setup for today's lecture is a model family $\mathcal{H} \in \{H_0, H_1, \ldots, H_{M-1}\}$. In the classification problem, we can think of $\mathcal{H}$ as a set of class labels, and we want to determine the correct label given some test data.

2.1 Bayesian Binary Hypothesis Testing

In this case, $M = 2$, so there are only two hypotheses. Our model has two major components. The first is some a priori information

$$\begin{aligned} P_0 &= \mathbb{P}[H = H_0] \\ P_1 &= \mathbb{P}[H = H_1] = 1 - P_0. \end{aligned}$$

We also have the observation model, which is given by likelihood functions

$$\begin{aligned} H_0 &: p_{Y|H}(\cdot|H_0) \\ H_1 &: p_{Y|H}(\cdot|H_1). \end{aligned}$$

Our goal is to create a **decision rule**, a.k.a. a **classifier**, which maps every $y \in \mathcal{Y}$ to some hypothesis $H_i \in \mathcal{H} = \{H_0, H_1\}$. This is somewhat confusing with the standard terminology of a hypothesis class being the set of possible solutions to a model, but we accept it for now.

Definition 2.1 (Cost)
In its most general form, we let

$$C(H_j, H_i) \triangleq C_{ij}$$

denote the cost of predicting H_j when the correct class is H_i.

Using cost to drive the notion of "best", our best possible decision rule takes

the form

$$\hat{H}(\cdot) = \arg\min_f \mathbb{E}_{Y,H}[C(H, f(Y))].$$

The expected cost on the RHS is called **Bayes risk**, which we denote as $\varphi(f)$ for any decision rule f.

We can explicitly calculate this quantity:

$$\begin{aligned}\varphi(f) &= \mathbb{E}_{Y,H}[C(H, f(Y))] \\ &= \mathbb{E}_Y[\mathbb{E}_{H|Y}[C(H, f(Y))|Y = y]] \\ &= \int p_Y(y)\mathbb{E}[C(H, f(Y))|Y = y]\mathrm{d}y.\end{aligned}$$

Notice that we have control over the expected risk for each point, so to minimize $\varphi(f)$, we only have to solve a solution for individual points. For a fixed $y* \in \mathcal{Y}$, there are two possibilities; if $f(y^*) = H_0$, then

$$\mathbb{E}[C(H, f(y^*))|y = y^*] = C_{00}\mathbb{P}[H = H_0|y = y^*] + C_{01}\mathbb{P}[H = H_1|y = y^*],$$

otherwise

$$\mathbb{E}[C(H, f(y^*))|y = y^*] = C_{10}\mathbb{P}[H = H_0|y = y^*] + C_{11}\mathbb{P}[H = H_1|y = y^*].$$

This already technically gives us the optimal decision rule; for any given input y, we can explicitly compute both values, and return the hypothesis that gives the lesser of the two values. We can also express this in a simpler form. Since

$$\mathbb{P}[H = H_i|Y = y] = \frac{p_{Y|H}(y|H_i)p_H(H_i)}{p_Y(y)},$$

we can substitute into the above expressions:

$$C_{00}p_{Y|H}(y|H_0)P_0 + C_{01}p_{Y|H}(y|H_1)P_1 \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} C_{10}p_{Y|H}(y|H_0)P_0 + C_{11}p_{Y|H}(y|H_1)P_1$$

We can rewrite this expression in terms of the ratios

$$L(y) \triangleq \frac{p_{Y|H}(y|H_1)}{p_{Y|H}(y|H_0)} \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \triangleq \eta.$$

We call $L(y)$ the **likelihood ratio**.

Theorem 2.2 (Likelihood Ratio Test)
Given a priori probabilities P_0, P_1, data y, observation models $p_{Y|H}(\cdot|H_0), p_{Y|H}(\cdot|H_1)$, and costs $C_{00}, C_{01}, C_{10}, C_{11}$, the Bayesian decision rule form

$$L(y) \triangleq \frac{p_{Y|H}(y|H_1)}{p_{Y|H}(y|H_0)} \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \triangleq \eta,$$

meaning that the decision is $\hat{H}(y) = H_1$ when $L(y) > \eta$, $\hat{H}(y) = H_0$ when $L(y) < \eta$, and it is indifferent when $L(y) = \eta$.

Note that the optimal rule is simple and deterministic. Prof. Wornell makes a point about $L(y)$ being a scalar that we can always calculate. This is the heart of classification models; in larger neural nets, like ImageNet, ultimately what the large network of weights allows us to do is to express the intractable probabilities and compute a scalar value.

2.2 0-1 Loss

In the case of "0-1 loss", i.e., $C_{00} = C_{11} = 0$, $C_{01} = C_{10} = 1$, in which case our test simplifies to

$$p_{H|Y}(H_1|y) \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} p_{H|Y}(H_0|y).$$

This is the **maximum a posteriori** (MAP) decision rule.

If we additionally assume that $P_0 = P_1$, i.e., that our prior belief is indifferent, then our test further simplifies to

$$p_{Y|H}(y|H_1) \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} p_{Y|H}(y|H_0).$$

This is the **maximum likelihood** (MLE) decision rule. In either case, the expected rate of error is given by

$$\varphi(\hat{H}) = \mathbb{P}[\hat{H}(Y) = H_0, H = H_1] + \mathbb{P}[\hat{H}(Y) = H_1, H = H_0].$$

Example 2.3 (Communicating a Bit)
We have a signal y, randomly distributed with variance σ^2, and with two possible sources s_0, s_1.

The likelihood ratio test gives

$$\ln L(y) = \ln\left(\frac{e^{-(y-s_1)^2/(2\sigma^2)}}{e^{-(y-s_0)^2/(2\sigma^2)}}\right) = \frac{1}{2\sigma^2}((y-s_0)^2 - (y-s_1)^2).$$

Assuming $0-1$ loss, $\ln L(y) = 0$, so the decision boundary is

$$y \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \frac{s_0+s_1}{2}.$$

We could compute the expected rate of error as follows:

$$\begin{aligned}
\varphi(\hat{H}) &= \frac{1}{2}\left(\mathbb{P}[\hat{H}(Y) = H_0 | H = H_1] + \mathbb{P}[\hat{H}(Y) = H_1 | H = H_0]\right) \\
&= \frac{1}{2}\left(\mathbb{P}\left[y < \frac{s_0+s_1}{2}\middle| H = H_1\right] + \mathbb{P}\left[y \geq \frac{s_0+s_1}{2}\middle| H = H_0\right]\right) \\
&= \frac{1}{2}\left(\mathbb{P}\left[\frac{y-s_1}{\sigma} < \frac{s_0-s_1}{2\sigma}\middle| H = H_1\right] + \mathbb{P}\left[\frac{y-s_0}{\sigma} \geq \frac{s_1-s_0}{2\sigma}\middle| H = H_0\right]\right) \\
&= Q\left(\frac{s_1-s_0}{2\sigma}\right).
\end{aligned}$$

The quantity $(s_1 - s_0)/\sigma$ is a measure of signal-to-noise; the larger the SNR, the more uncertain we are about our prediction, i.e., the higher our expected rate of error.

3 February 13, 2023

Bad weather day today. Prof. Wornell tells us about how this is the first time he's given a lecture over zoom since the pandemic. He was hoping that he would never have had to give a zoom lecture again, but here we are.

3.1 Review: Bayesian Hypothesis Testing

Overarching goal: optimize the expected cost of choosing one hypothesis over another. This can be accomplished with the Likelihood Ratio Test:

$$L(y) \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \eta$$

Some issues with this solution: we require some notion of costs, as well as priors $P_i = P_H(H_i)$. It can be difficult to assign probabilities to abstract concepts, like $\mathbb{P}$[patient has disease].

3.2 Non-Bayesian Hypothesis Testing

The "folk theorem" that we will be proving today is that all optimum decision rules takes the form of a Likelihood Ratio Test:

$$\frac{p_{Y|H}(y|H_1)}{p_{Y|H}(y_{H_0})} \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \eta,$$

for some η. This is largely true, but not always.

3.3 General Performance Measures

Let $\hat{H}(\cdot)$ be any rule. An equivalent characterization of this rule is some partition of the observation space Y:

$$y_0 = \{y \in Y : \hat{H}(y) = H_0\}$$
$$y_1 = y \backslash y_0$$

Then $P_D = \mathbb{P}[\hat{H} = H_1 | H = H_1] = \int_{y_1} p_{Y|H}(y|H_1)\mathrm{d}y$ is called the **detection probability**, and $P_F = \mathbb{P}[\hat{H} = H_1 | H = H_0] = \int_{y_1} p_{Y|H}(y|H_0)\mathrm{d}y$ is called the **"false alarm" probability**. Some related terminology: $P_M = 1 - P_D$ is called the "miss" probability.

- Statistics terminology: $P_E^1 = P_F$, $P_E^2 = P_M$, "probability of error of each kind". The probability of error of the first kind is called the "size" of the test, while the probability of error the second kind is called the "power" of the test.
- Medical terminology: P_F is the false positive rate, while P_M is the false negative rate.
- Learning / pattern classification: $P_R = P_D$ is the "recall" or "sensitivity". The "precision" is defined as $P_P = \mathbb{P}[H = H_1 | \hat{H} = H_1] = 1/(1 + P_F/P_D \cdot P_0/P_1)$.

In general, P_D and P_F are conflicting objectives. We seek large P_D and small P_F. The bayesian approach to this "multi-objective" optimization is to choose the rule that satisfies:

$$\min_{\hat{H}(\cdot)}(\alpha P_F - \beta P_D).$$

3.4 Operating Characteristic of the LRT

Other tradeoffs are possible. Consider the family of ratio tests:

$$\{\hat{H}(\cdot) = \text{LRT}, \text{for some } \eta\}.$$

Example 3.1

Consider two hypotheses

$$H_0 : y \sim \mathcal{N}(0, \sigma^2)$$
$$H_1 : y \sim \mathcal{N}(m, \sigma^2)$$

The LRT is given by

$$y \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \frac{m}{2} + \frac{\sigma^2 \ln \eta}{m} \triangleq \gamma.$$

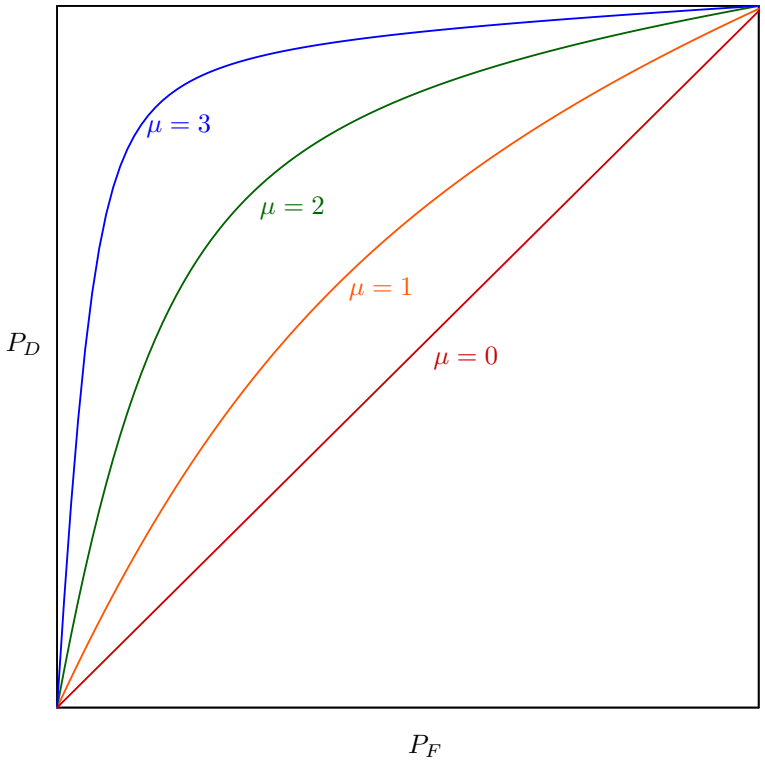
The false positive rate

$$P_F = \int_\gamma^\infty p_{Y|H}(y|H_0)\mathrm{d}y = Q\left(\frac{\gamma}{\sigma}\right),$$

and the detection probability

$$P_D = \int_\gamma^\infty p_{Y|H}(y|H_1)\mathrm{d}y = Q\left(\frac{\gamma - m}{\sigma}\right).$$

Graphing a plot of (P_F, P_D) over different values of $\mu = m/\sigma$ gives



This curve is called the **OC-LRT**: Operating Characteristic of the Likelihood Ratio Test.

$$\text{OC-LRT} : \{(P_F, P_D) : \hat{H}(\cdot) \text{ is LRT for some } \eta\}.$$

Claim 3.2
OC-LRT is monotonic and non-decreasing.

Proof. If $\eta_2 > \eta_1$, then $P_D(\eta_2) \leq P_D(\eta_1)$ and $P_F(\eta_2) \leq P_F(\eta_1)$. □

A few other key properties here:

- All of the frontiers are concave-down, i.e., we should always be "better" at maximizing P_D than minimizing P_F
- As $\mu \to \infty$, we approach the optimal curve, which is when P_D is 1 everywhere and P_F is 0 everywhere. This makes sense intuitively, since μ large separates the two hypotheses.

- Similarly, when $\mu = 0$, the two hypotheses are indistinguishable, so the frontier is as good as random guessing.
- When $P_F = P_D = 1$ when $\gamma = 0$ and $P_F = P_D = 0$ when $\gamma \to \infty$. Therefore, as γ increases to infinity, we travel from top-right to bottom-left along each of the curves (this will always be the case for any OC-LRT, for this reason).

3.5 Neyman-Pearson Criterion

To avoid the problem of costs and priors, a common alternate criteria choose a rule subject is the **Neyman-Pearson Criterion**:

$$\max_{\hat{H}(\cdot)} P_D \text{ s.t. } P_F \leq \alpha.$$

In words, choose the hypothesis with largest detection power given a fixed upper bound on the false alarm size.

Theorem 3.3 (Neyman-Pearson Lemma, Specialized)
For deterministic $\hat{H}(\cdot)$, a solution to the Neyman-Pearson Criterion is an LRT when the LRT is continuous. In other words,

$$\hat{H}(y) = H_{\mathbb{1}_{L(y)\geq\eta}},$$

where η is the smallest threshold s.t.

$$P_F = \mathbb{P}(L(y) \geq \eta | H = H_0) \leq \alpha.$$

This statement of NP is considered 'specialized' because we are not employing randomization in our hypotheses. We will see and prove the full version next lecture.

Proof. We can prove this with lagrange multipliers. Fix $P_F = \alpha' \leq \alpha$. Then, we want

to optimize

$$\begin{aligned} min_{\hat{H}(\cdot)}\varphi(\hat{H}) &= (1-P_D)+\lambda(P_F-\alpha') \\ &= \int_{y_0} p_{Y|H}(y|H_1)\mathrm{d}y + \lambda\left(\int_{y_1} p_{Y|H}(y|H_0)\mathrm{d}y - \alpha'\right) \\ &= \lambda(1-\alpha') + \int_{y_0}\Big(p_{Y|H}(y|H_1) - \lambda p_{Y|H}(y|H_0)\Big)\mathrm{d}y. \end{aligned}$$

The min φ occurs when we assign y to y_0 whenever the the integrand is ≤ 0, as this minimizes the cost. For the same reason, we want α' to be as large as possible. Therefore,

$$\frac{p_{Y|H}(y|H_1)}{p_{Y|H}(y|H_0)} \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \lambda,$$

where α' is chosen to be the largest threshold achievable by LRT. □

We also present an alternate proof below.

Proof. We compare the LRT decision region with an arbitrary decision region, and show that we cannot do better than the LRT decision region. Let $\mathcal{Y}_1^\eta$ be the set of points for which $\hat{H}^\eta(y) = H_1$, and $\mathcal{Y}_1$ be the set of points for which $\hat{H}(y) = H_1$ for an arbitrary decision rule. We have

$$(\mathbb{1}(L(y)\geq\eta) - \mathbb{1}(y\in\mathcal{Y}_1))(L(y)-\eta)\geq 0,$$

which can be verified with casework. Then,

$$\begin{aligned} &\int_{\mathcal{Y}} (\mathbb{1}(L(y)\geq\eta) - \mathbb{1}(y\in\mathcal{Y}_1))(L(y)-\eta)p_{Y|H}(y|H_0)\mathrm{d}y \\ &= \int_{\mathcal{Y}} (\mathbb{1}(L(y)\geq\eta) - \mathbb{1}(y\in\mathcal{Y}_1))(p_{Y|H}(y|H_1) - \eta p_{Y|H}(y|H_0))\mathrm{d}y \geq 0. \end{aligned}$$

Some expansion and rearranging gives

$$\int_{\mathcal{Y}_1^\eta} p_{Y|H}(y|H_1)\mathrm{d}y - \int_{\mathcal{Y}_1} p_{Y|H}(y|H_1)\mathrm{d}y \geq \eta\left(\int_{\mathcal{Y}_1^\eta} p_{Y|H}(y|H_0)\mathrm{d}y - \int_{\mathcal{Y}_1} p_{Y|H}(y|H_0)\mathrm{d}y\right),$$

which (term-by-term) is just

$$P_D^\eta - P_D \geq \eta(P_F^\eta - P_F).$$

When P_F^η is fixed at α, the constraints of the problem force us to pick $P_F \leq P_F^\eta$, so $P_D^\eta - P_D \geq 0$. But this implies that any hypothesis that is not the LRT can only have worse detection power, hence $\hat{H}_1^\eta$ is a valid solution. □

3.6 Hypothesis Testing in the LRT framework

The typical LRT criteria:

$$L(y) \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} \eta.$$

When we apply a monotonic function to both sides, the criteria remains the same. In particular, when we apply P_F, we get

$$p_*(y) \overset{\Delta}{=} P_F(L(y)) \underset{\hat{H}=H_1}{\overset{\hat{H}=H_0}{\lessgtr}} P_F(\eta) = \alpha.$$

Note that the direction of the hypothesis changed, since P_F is monotonically decreasing in η. The RHS is the significance level of the test, while p_* is a function that maps each data point to a p-value. Roughly speaking, if p_* is large, this means L was small, so the data was not very significant. If p_* is small, this means L was large, so the data was significant. The threshold of "significant" is determined by our threshold α.