6.780 Notes

Lecturer:

Andrew Liu

Spring 2024

Last updated on Saturday 10th February, 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. 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.