Bayesian decision theory (Chp 2)

$$\begin{cases} H_0: & Y \sim & P_0\left(y \middle| H_0\right) \\ H_1: & Y \sim & P_1\left(y \middle| H_1\right) \end{cases}$$

1) Underlying Theory
2) Learning Algorithm
3) Interpretation

Bayesian decision rule $\delta\left(y\right)$

Observation set: $\Gamma = \Gamma_1 \cup \Gamma_0 \quad \left(\Gamma_0 = \Gamma_1^c\right)$

such that

$$\delta\left(\mathrm{y}\right) = \begin{cases} 1, & \text{if} \quad y \in \Gamma_1 \\ 0, & \text{if} \quad y \in \Gamma_0 \end{cases}$$

How to determine Γ_1 or $\delta\left(y\right)$ in an optimum way?

Bayesian decision theory (Chp 2)

Class conditional risks (prediction or classification error rate)

$$\begin{cases} R_0(\delta) & = P_0(\Gamma_1) \quad \text{false positive rate} \\ R_1(\delta) & = P_1(\Gamma_0) \quad \text{false negative rate} \end{cases}$$

Class prior probability (the probability of H_0/H_1 occurance unconditioned on y):

$$\begin{cases} P(H_0) = \pi_0, & \pi_0 + \pi_1 = 1. \\ P(H_1) = \pi_1, & \end{cases}$$

Average or Bayesian risks :

$$\begin{aligned} r(\delta) &= \pi_0 R_0(\delta) + \pi_1 R_1(\delta) \\ &= \pi_0 P_0(\Gamma_1) + \pi_1 \left\{ 1 - \int_{\Gamma_1} p_1(y)\,dy \right\} \\ &= \pi_1 + \int_{\Gamma_1} \left(\pi_0 p_0(y) - \pi_1 p_1(y) \right) dy \end{aligned}$$

What would happen if decisions are only based on priori(s)? While cannot know, the 2nd term involves conditional probability of the 'observed'.

Bayesian decision theory (Chp 2)

Optimality: minimize $r(\delta)$, that is,

$$\Gamma_1 = \left\{ y \in \Gamma \,\middle|\, \pi_0 p_0(y) - \pi_1 p_1(y) \le 0 \right\}.$$

Optimum decision rule:

$$\delta(y) = \pi_1 p_1(y) \ge \pi_0 p_0(y) \text{ or } \frac{\pi_1 p_1(y)}{\pi_0 p_0(y)} \ge 1 \Rightarrow \Gamma_1$$

Likelihood ratio text:

$$\delta(y): \ \frac{P(y, H_1)}{P(y, H_0)} \begin{cases} \ge 1 \Rightarrow \Gamma_1 \Rightarrow H_1 \\ < 1 \Rightarrow \Gamma_0 \Rightarrow H_0 \end{cases}$$

Bayesian decision theory (Chp 2)

Bayesian decision rule:

Apply the Bayes law, we have

$$\frac{\pi_1 p_1(y)}{\pi_0 p_0(y)} = \frac{\dfrac{P(y, H_1)}{p(y)}}{\dfrac{P(y, H_0)}{p(y)}} = \frac{P(H_1|y)}{P(H_0|y)} \begin{array}{c} \geq 1 \\ < 1 \end{array}$$

That is:

$$P(H_1|y) \geq P(H_0|y) \Rightarrow \Gamma_1 \Rightarrow H_1$$

$$P(H_1|y) < P(H_0|y) \Rightarrow \Gamma_0 \Rightarrow H_0$$

Bayesian decision theory (Chp 2)

The extension of Bayesian decision rule to multiclass hypothesis testing M-ary classification is called "maximum a posterior probability".

MAP decision rule:

$$H_j = \underset{j=0,1,\ldots,M-1}{\operatorname{argmax}} P\left(H_j \middle| y\right)$$

Homework #1:

Self-reading chapters 1 and 2

Chapter 2, problems 2 and 6.

Statistical learning from “data”

Data $\left(x_i, y_i\right)$

x ~ observation (e.g., blood pressure)

y ~ desired output (class label) (e.g., normal versus abnormal)

when π_j, P_i are unknown, so instead, they must be estimated or learned from the available "training samples", i.e., data.

An generic example of "learning from data $\left(x_i, y_i\right)$"

$$x \Rightarrow \boxed{f\left(x,\theta\right)} \Rightarrow y = f\left(x,\theta\right)$$

where θ is the model parameter set.

Statistical learning from "data"

For a binary classifier, we have

$$y = f\left(x,\theta\right) = \begin{cases} a & x \in C_1 \\ b & x \in C_2 \end{cases}$$

Suppose we have total N training sample points and we adopt "error-correction learning" strategy using mean-squared error (MSE) criterion, we have

$$\varepsilon = \frac{1}{N}\left[\sum_{x \in C_1}\left[f\left(x,\theta\right) - a\right]^2 + \sum_{x \in C_2}\left[f\left(x,\theta\right) - b\right]^2\right]$$

Statistical learning from "data"

When N is sufficiently large, we then have

$$\varepsilon \approx \int_{\Gamma} \left[f\left(x,\theta\right) - a \right]^2 p\left(x, C_1\right) dx + \int_{\Gamma} \left[f\left(x,\theta\right) - b \right]^2 p\left(x, C_2\right) dx$$

where $p\left(x, C_j\right)$ is the joint probablity density function of x and C_j.

Furthermore, we have

$$\varepsilon \approx \int f^2\left(x,\theta\right)\left[p\left(x,C_1\right) + p\left(x,C_2\right) \right] dx$$

$$-2\int f\left(x,\theta\right)\left[ap\left(x,C_1\right) + bp\left(x,C_2\right) \right] dx$$

$$+a^2 \int p\left(x,C_1\right) dx + b^2 \int p\left(x,C_2\right) dx$$

Recall the "total probablity law", we have

$$p\left(x\right) = p\left(x,C_1\right) + p\left(x,C_2\right)$$

Statistical learning from "data"

By the Bayes law, we can define

$$d(x) = \frac{ap(x, C_1) + bp(x, C_2)}{p(x, C_1) + p(x, C_2)} = aP(C_1|x) + bP(C_2|x)$$

where $P(C_j|x)$ is the posterior probablity of class C_j given the observation x.

We can re-express MSE

$$\varepsilon \approx \int \left[f(x,\theta) - d(x) \right]^2 p(x)\, dx +$$

$$+a^2 P(C_1) + b^2 P(C_2) - \int d^2(x)\, p(x)\, dx$$

Minimize $\varepsilon \Rightarrow$ MMSE

$$\theta = \underset{\theta}{\operatorname{argmin}} \int \left[f(x,\theta) - d(x) \right]^2 p(x)\, dx$$

Statistical learning from "data"

What can we see from here?

"Adjusting the parameters of $f(x,\theta)$ to minimize ε is equivalent to minimizing the MSE between the model outpur $f(x,\theta)$ and $d(x)$."

Let $a=1, b=0$, i.e., following what we did before

$$y = f(x,\theta) = \begin{cases} 1 & x \in C_1 \\ 0 & x \in C_2 \end{cases}$$

that leads to

$$d(x) = P(C_1|x)$$ consistent with Bayes decision theory!

Remark: $f(x,\theta) \Rightarrow d(x) = P(C_1|x)$ is what the prediction model is learning to approximate in a weighted MSE sense, where the weight is $p(x)$.