References

1. Pattern Classification
   Duda, Hart, Stork

2. Probabilistic Theory of Pattern Recognition
   Devroye, Gyorfi, Lugosi

3. Pattern recognition and Machine Learning
   Chris Bishop

Recap

$\mathcal{X} \to$ Instance space

$\mathcal{Y} \to$ Label space

Binary Classification

$\mathcal{Y} = \{-1, 1\}$

Classifier

$h : \mathcal{X} \to \{-1, 1\}$ $\quad \mathcal{X} \subseteq \mathbb{R}^d$

Prior

$p_1 = P(Y=1)$, $p_2 = P(Y=-1)$

Class conditional Distribution

$P(X=x \mid Y=1)$, $P(X=x \mid Y=-1)$

[Densities / p.m.f]

$$\eta(x) = P(Y=1 \mid X=x)$$

$$1-\eta(x) = P(Y=-1 \mid X=x)$$

$$P(Y=1 \mid X=x) = \frac{P(Y=1)\,P(X=x \mid Y=1)}{P(X=x)}$$

Bayes Classifier

$$h(x) = \begin{cases} 1 & P(Y=1 \mid X=x) > P(Y=-1 \mid X=x) \\ -1 & P(Y=-1 \mid X=x) \geq P(Y=1 \mid X=x) \end{cases}$$

$$h(x) = \begin{cases} 1 & \eta(x) > 1-\eta(x) \\ -1 & 1-\eta(x) \geq \eta(x) \end{cases}$$

$$\boxed{h^*(x) = \mathrm{sign}(2\eta(x) - 1)} \rightarrow \text{BC}$$

$$\mathrm{sign}(z) = \begin{cases} 1 & z > 0 \\ -1 & z \le 0 \end{cases}$$

Specific case

$x \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$

$C \in \mathbb{R}^{d\times d}$, Symmetric positive semi-definite

$$N(x \mid \mu, C)$$

$$= \frac{1}{(\sqrt{2\pi})^d |C|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)}$$

$|C| \equiv \det(C)$

$$P(X = x \mid Y = 1) = N(x \mid \mu_1, C_1)$$

$$P(X = x \mid Y = -1) = N(x \mid \mu_2, C_2)$$

If $C_1 = C_2 = C$

$$h^*(x) = \text{sign}(w^T x + b)$$

$$w = C^{-1}(\mu_1 - \mu_2)$$

$$b = \log \frac{P_1}{P_2} - \frac{1}{2}\left(\mu_1^T C^{-1} \mu_1 - \mu_2^T C^{-1} \mu_2\right)$$

How good is the Bayes Classifier (BC)

Thm2.1 DG2

$$P(h(X) \neq Y) - P(h^*(X) \neq Y) \geq 0$$

$$P(h(X) \neq Y) = E_X E_{Y|X}(h(X) \neq Y))$$

$$E_{Y|\cancel{X}=x} \mathbb{1}\{h(x) \neq Y\}$$

$$= \eta(x) \mathbb{1}\{h(x) \neq 1\} + (1-\eta(x))\mathbb{1}\{h(x) \neq -1\}$$

$$= \begin{cases} (1-\eta(x)) & h(x) = 1 \\ \eta(x) & h(x) = -1 \end{cases}$$

If $Y = -1$, but $h(X) = 1$, then there is a mistake.

If $Y = 1$, but $h(X) = -1$ then again there is a mistake

If $\eta(x) > (1 - \eta(x))$ predict $h(x) = 1$

The minimum over h (all possible classifiers) is attained at choosing $h(x)$ such that

$$E_{Y|X=x} \mathbb{1}\{h(X) = Y\} = \min(\eta(x), 1 - \eta(x))$$

$$h^*(x) = \begin{cases} 1 & \eta(x) > 1 - \eta(x) \\ -1 & 1 - \eta(x) > \eta(x) \end{cases}$$

Bayes decision theory

We have $x \in \mathbb{R}^d$ with label $y \in \{-1, 1\}$

We predict $\hat{y} \in \{1, 1\}$

$$\ell(\hat{y}, y) : \mathcal{Y} \times \mathcal{Y} \Rightarrow \mathbb{R}_+$$

Expected loss

$$\boxed{R(h) = \mathbb{E}_{X,Y}\, \ell(h(X), Y)}$$

$$\boxed{\min_h R(h)}$$

$$\underset{h}{\text{minimize}} \; R(h) = R(\tilde{h})$$

$$\tilde{h}(x) = \min_{h} E_{Y|X=x} \, \ell(h(x), Y)$$

$R(\tilde{h})$ is called the Bayes error-rate.

For a given instance it chooses the label which yields minimum loss.

$$\underset{h}{\text{minimize}} \; R(h)$$

$$\tilde{h}(x) = \min_{h} E_{Y|X=x} \, \ell(h(x), Y)$$

2 class problem

$$\min_{h} E_{Y|X=x} \, \ell(h(X), Y) \quad Y \in \{-1, 1\}$$

$$E_{Y|X=x} \, \ell(1, Y) = \ell(1,1) P(Y=1 | X=x) + \ell(1,-1) P(Y=-1 | X=x)$$

$$= \ell(1,1)\, \eta(x) + \ell(1,-1)\,(1-\eta(x))$$

$$E_{Y|X=x} \, \ell(-1, Y) = \ell(-1,1)\, \eta(x) + \ell(-1,-1)\,(1-\eta(x))$$

Expected Loss

Choose h such that

$E_{Y|X=x}\, \ell(h(x), Y)$ should be minimum

Choose class $\boxed{\tilde{h}(x) = 1}$

if $E_{Y|X}\, \ell(1, Y) < E_{Y|X}\, \ell(-1, Y)$

Choose class $\boxed{\tilde{h}(x) = -1}$

if $E_{Y|X=x}\, \ell(-1, Y) < E_{Y|X}\, \ell(1, Y)$

$\tilde{h}(x) = 1$

$$\therefore\ l(1,1)\eta(x) + l(1,-1)(1-\eta(x)) < l(-1,1)\eta(x) + l(-1,-1)(1-\eta(x))$$

$$(l(1,1) - l(-1,1))\eta(x) < (l(-1,-1) - l(1,-1))(1-\eta(x))$$

$\tilde{h}(x) = -1$

$$l(-1,1)\eta(x) + l(-1,-1)(1-\eta(x)) < l(1,1)\eta(x) + l(1,-1)(1-\eta(x))$$

$$\Big(\ell(-1,-1) - \ell(1,-1)\Big)\big(1-\eta(x)\big)$$

$$< \Big(\ell(1,1) - \ell(-1,1)\Big)\eta(x)$$

Properties of ℓ

1. non-negative
2. should penalize mistakes more

Choose $\ell(-1, 1) = \ell(1, -1) = 1$

$\ell(1, 1) = \ell(-1, -1) = 0$

loss should penalize mistakes more

$\tilde{h}(x) = 1$ if

$(0-1)\eta(x) < (0-1)(1-\eta(x))$

$\Rightarrow \eta(x) > 1-\eta(x)$

$\tilde{h}(x) = -1$ if

$1-\eta(x) > \eta(x)$.

Then $\tilde{h}(x) = h^*(x)$

$R(\tilde{h}) < R(h)$

$R(\tilde{h})$ is called the B error rate