

Extra 6

Let $f: \mathcal{X} \rightarrow \{0,1\}$ be a classifier

and $l(y, f(x))$ a loss function.

Suppose

$$l(y, f(x)) = \begin{cases} 0 & \text{if } f(x) = y \\ c_0 & \text{if } f(x) \neq y \text{ \& } y=0 \\ c_1 & \text{if } f(x) \neq y \text{ \& } y=1 \end{cases}$$

$$= c_0 I(y=0 \cap f(x)=1) + c_1 I(y=1 \cap f(x)=0)$$

where $I(A) = 1$ if A is true, 0 otherwise ("indicator function")

↳ Different cost of misclassifying the negatives ($y=0$) and the positives ($y=1$)

① The expected prediction error (EPE) is

$$E[l(y, f(x))] = E_x \left\{ E_{y|x} [l(y, f(x)) | x] \right\}$$

where

$$\begin{aligned} E_{y|x} [l(y, f(x)) | x] &= P(Y=0|x) l(0, f(x)) + P(Y=1|x) l(1, f(x)) \\ &= \underbrace{P(Y=0|x) c_0 I(f(x)=1)}_{\equiv Q_0(x)} + \underbrace{P(Y=1|x) c_1 I(f(x)=0)}_{\equiv Q_1(x)} \end{aligned}$$

Hence,

$$\begin{aligned} E[l(y, f(x))] &= E_x \left[I(f(x)=1) Q_0(x) + I(f(x)=0) Q_1(x) \right] \\ &= \int_{\mathcal{X}} I(f(x)=1) Q_0(x) \overset{p(x)}{dx} + \int_{\mathcal{X}} I(f(x)=0) Q_1(x) p(x) dx \quad (*) \end{aligned}$$

as we were asked to show.

②

Note that when $f(x) \in \{0,1\}$, then

$$I(f(x)=1) = 1 - I(f(x)=0)$$

The EPE (*) thus can be rewritten as

$$E[l(y, f(x))] = \int_{\mathcal{X}} (1 - I(f(x)=0)) Q_0(x) p(x) dx + \int_{\mathcal{X}} I(f(x)=0) p(x) dx$$

$$= \underbrace{\int_{\mathcal{X}} Q_0(x) p(x) dx}_{\text{some constant that does not depend on } f(x)} + \int_{\mathcal{X}} I(f(x)=0) \{Q_1(x) - Q_0(x)\} p(x) dx$$

some constant that does not depend on $f(x)$

The optimal predictor should minimize the EPE.

From the last expression, we see that

- When $Q_1(x) > Q_0(x)$ the EPE increases if $f(x) = 0$, and we should set $f(x) = 1$
- When $Q_1(x) < Q_0(x)$ the EPE decreases if $f(x) = 0$, and we should set $f(x) = 0$

The optimal prediction rule is thus

$$f(x) = \begin{cases} 1 & \text{if } Q_1(x) > Q_0(x) \\ 0 & \text{otherwise} \end{cases}$$

Note that

$$Q_1(x) > Q_0(x) \quad (\Leftrightarrow) \quad P(Y=1|x) > \frac{C_0}{C_1} P(Y=0|x)$$

which gives us the result we are asked for.

This rule is reasonable, as we predict $Y=1$ ($Y=0$) if the probability is very high (low) or the relative cost of making a mistake is low (high).

Example:

Suppose it is twice as costly to misclassify a true negative, $C_0 = 2C_1$. To predict $Y=1$, we would then require that the probability of being correct, $P(Y=1|x)$, is at least twice of the probability that we are wrong, $P(Y=1) > 2 P(Y=0|x)$.