classification

Y categorical (not numeric)

ground truth: for given x,

$$\left.\begin{aligned} &P_1(x_0) = P(Y=1 \mid X=x_0) \\ &\quad\vdots \\ &P_k(x_0) = P(Y=k \mid X=x_0) \end{aligned}\right\}$$

randomness
some x_0 can give different values of Y
irreducible error

want $\hat{f}(x_0) \longrightarrow$ returns the category of Y most likely at x_0

measure error

misclassification rate: $\frac{1}{n}\sum_{i=1}^{n} 1\{Y_i \neq \hat{Y}_i\}$

between 0 and 1

accuracy = 1 − misclassification rate

Confusion matrix ($k=2$)

pred \ true	1	0
1	a	b
0	c	d

$a =$ # samples where $\hat{Y}_i = 1$ and $Y_i = 1$

$b =$ . . . . $\hat{Y}_i = 1$ $\quad Y_i = 0$

$c =$ $\quad\hat{Y}_i = 0$ $\quad Y_i = 1$

$d =$ $\quad\hat{Y}_i = 0$ $\quad Y_i = 0$

$n = a+b+c+d$

misclassification rate $= \frac{b+c}{n}$

accuracy $= \frac{a+d}{n}$

before, used CV to estimate MSPE

now use CV to estimate misclassification rate

$k=2$

Compute $P(\hat{Y}=1 \mid X)$, $P(\hat{Y}=0 \mid X)$

predict $\hat{Y}=1$ if $P(\hat{Y}=1 \mid X) > 0.5 \leftarrow$ threshold

Why not consider different values of threshold?
might want to consider different threshold

↳ can get smaller misclassification error with different threshold

linear regression gives $\hat{Y}$ any number $\in \mathbb{R}$

want to get $P(\hat{Y}=1)$ instead

probability p $\;0 \le p \le 1$

$$x \in \mathbb{R} \longrightarrow \frac{e^x}{1+e^x} \in (0,1)$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \longrightarrow \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x)} \in (0,1)$$

$P(\hat{Y}=1 \mid X)$

$$\Longrightarrow \ln\left(\frac{P(\hat{Y}=1 \mid X)}{1-P(\hat{Y}=1 \mid X)}\right) = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$P(\hat{Y}=1 \mid X) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 X)} \longrightarrow 1 \text{ as } x \to \infty$$

$\hat{\beta}_1 > 0$ then X increasing means $P(\hat{Y}=1 \mid X)$ increases also

$\hat{\beta}_1 < 0$ then x increasing means $P(\hat{Y}=1 \mid X)$ decreases