classification

Y categorical (not numeric)

ground truth: for given x,

$P_1(x_0) = P(Y=1 \mid X=x_0)$

$\vdots$

$P_k(x_0) = P(Y=k \mid X=x_0)$

} randomness
some x_0 can give different values of Y
irreducible error

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

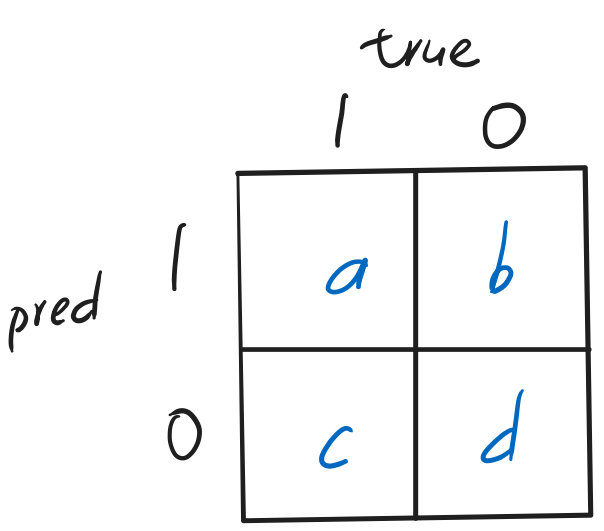
measure error

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)



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

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

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

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

$n = a+b+c+d$

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 $P(\hat{Y}=1|x)$, $P(\hat{Y}=0|x)$

predict $\hat{Y}=1$ if $P(\hat{Y}=1|x) > 0.5$ ← threshold

Why not consider different values of threshold?

might want to consider different threshold.

↳ can get smaller misclassification error with different threshold