WEE2 - Lecture-5

Statistical decision Theory

$X \in \mathbb{R}^p$, $Y \in \mathbb{R}$ $\quad Pr(X,Y) \leftarrow$ Assumption - Datas are drawn from a distribution.

input $\quad$ output (regression) $\quad$ ↳ Joint distribution $\rightarrow \{(x_1, y_1), \ldots, (x_n, y_n)\}$

Goal: $f(x): \mathbb{R}^p \rightarrow \mathbb{R}$

$$\hat{Y} = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots + \beta_p x_p$$

$$x = (x_1, x_2, x_3, \ldots x_p)^T$$

$$= \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

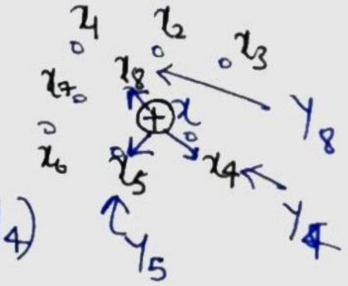
set $\Rightarrow x_0 = 1$

$$= \sum_{j=0}^{p} x_j \beta_j \quad \text{[Linear Regression]}$$

K- Nearest Neighbour:

$$\hat{Y}(x) = \frac{1}{K} \sum_{x_i \in N(x)} y_i$$

$K=3$



$$\hat{Y}(x) = \frac{1}{3} \sum (Y_5 + Y_8 + Y_4)$$

Loss f^n: $L(Y, f(x))$

Squared error: $(Y - f(x))^2$

Expected prediction Error $(EPE(f)) = E\{(Y - f(x))^2\}$

$$= \int (Y - f(x))^2 \cdot pr(dx, dy)$$

$Pr(x,y) = P_r(Y|x) P_r(x)$

$$= E_x E_{Y/x}\left([Y - f(x)]^2 | x \right)$$

minimise over diff values of c. (single point)

Take that value of c to calculate $f(x)$

$$\boxed{f(x) = \underset{c}{\arg\min}\, E_x E_{Y/x}\left([Y - c]^2 | x = x_0 \right)}$$

(Conditioning on a point)

The function $f(x)$ that minimises mean-squared error is given as

$f(x) = E[Y|X=x]$ (Conditional Expectation or regression function)

$\boxed{f(x) = \text{Avg}(y_i \mid x_i = x)}$ → ~~we~~ suppose for $x_i = 3$, $y_1 = 2$, $y_1' = 3$, $y_1'' = 7$, $y_1''' = 9$

$$f(x) = \frac{y_1 + y_1' + y_1'' + y_1'''}{4} = \left(\frac{2+3+7+9}{4}\right)$$

1. Problem: If we have only one sample of x. Its difficult to ~~get~~ get avg.
2. unable to predict on unseen data.

→ $\boxed{\text{NOT WORKING}}$

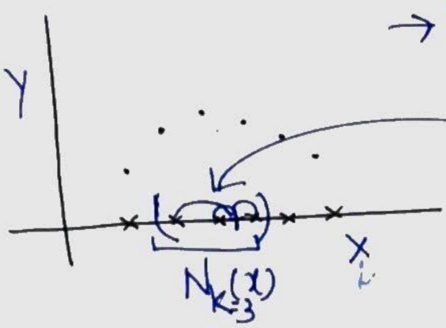
Conclusion:

Instead of conditioning on single data point, we will condition on a region (many data points)

$$\hat{f}(x) = \text{Avg}(y_i \mid x_i \in N_k(x))$$

$N_k(x)$ → Neighborhood of x

Assumption: output of the function will be constant on that region.



→ output will be same for this region

As $K \uparrow$, estimate more stable

As $N, K \to \infty$, $K/N \to 0$, $\hat{f}(x) \to E[Y|x = x]$

In case of linear regression,

$$f(x) = x^T \beta \qquad X = \begin{bmatrix} x_{11}, x_{12}, \cdots & x_{1p} \\ x_{21}, x_{22}, \cdots & x_{2p} \\ & \\ x_{n1} \;\; x_{n2} \;\cdots & x_{np} \end{bmatrix}_{n \times p}$$

$$(Y - X\beta)^2 = EPE(\hat{f})$$

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T Y}$$

$$(Y - X\beta)^T (Y - X\beta) = Y^T (Y - X\beta) - (X\beta)^T (Y - X\beta)$$

$$= Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X \beta$$

$$E(\cancel{f(x)}) = Y^T Y - 2Y^T X\beta + \beta^T X^T X \beta$$

$$\frac{\partial E}{\partial \beta} = 0 - 2Y^T X + 2X^T X \beta = 0$$

$$\Rightarrow X^T X \beta = X^T Y$$

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T Y}$$

Classification

$X \in \mathbb{R}^p$, $G \in \mathcal{G}$ $\qquad Pr(X, G)$ $\qquad \{(x_1, G_1), (x_2, G_2), (x_3, G_3), \ldots, (x_n, G_n)\}$

$f(x): \mathbb{R}^p \rightarrow \mathcal{G}$

LOSS = is a $(K \times K)$ matrix where $\boxed{\text{(zero on diagonal)}}$ $\quad K = \text{Card}(\mathcal{G})$

$\boxed{L(k, l) \text{ cost of classifying } k \text{ as } l.}$

0 - 1 loss function, $K = 3$

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}_{3 \times 3}$$

$$EPE(\hat{f}) = E\left[L(G,\hat{f})\right]$$

(Discrete distribution)

$$= E_X E_{G|X} \left\{ L[G,\hat{f}] \,|\, x \right\}$$

$$= E_X \sum_{k=1}^{K} L[k,\hat{f}(x)]\, pr(k|x)$$

original class

$$\hat{f}(x) = \underset{g}{\arg\min} \sum_{k=1}^{K} L(k,g).\, pr(k|X=x)$$

↳ predicted class

0-1 class	3 classes

$pr(1|x) = 0.6$

$pr(2|x) = 0.2$

$pr(3|x) = 0.2$

let's $\underline{\underline{g=2}}$

$$L = \begin{bmatrix} 0 & \boxed{1} & 1 \\ 1 & 0 & \phi \\ \phi & \phi & 0 \end{bmatrix}$$

$$\hat{f}(x) = \sum_{k=1}^{K} L(k,g)\, pr(k|\cancel{x}=x)$$

for $g=2$, $L(1,2)pr(1/x) + L(2,2)pr(2/x) + L(3,2)pr(3/x)$

$= 0.8$

$g=1$, $L(1,1)pr(1/x) + L(2,1)pr(2/x) + L(3,1)\, pr(3/x)$

Baiyes optimal classifier

$$\hat{f}(x) = \underset{g}{\arg\max}\; pr(g|x)$$

K-NN: (pick "K" nearest neighbour and take majority)

Week-2 lecture-6

Binary classifier

$(x_1, 0)$

$(x_2, 1)$

$(x_3, 0)$

$(x_1, 1)$

$(x_5, 0)$

$$f(x) \begin{cases} \geq 0.5 & \text{class-1} \\ < 0.5 & \text{class 0} \end{cases}$$

Bias- variance Week-2 lecture7