Week7-notes

Monday, 1 March 2021 12:15

Textbook:

4.7)

Non-parametric model

$Y_i = \underline{f(x_{i1}, \dots, x_{ip})} + \varepsilon_i$

$E[\varepsilon_i] = 0$, $Var(\varepsilon_i) = \sigma^2$

(2) ε_i are iid, $Cov(Y_i, Y_i)$

$Var(Y_i) = Var(\varepsilon_i) = \sigma^2$, for $i = 1, \dots, n$.

Linear smother: $\underline{\hat{Y}} = S\underline{Y}$

$n \times n$

Smother matrix

Predicted values.

$$\underline{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

$$S = \begin{bmatrix} \boxed{s_{11} \cdots s_{1n}} \\ \vdots \quad \ddots \quad \vdots \\ s_{n1} \cdots s_{nn} \end{bmatrix} = \begin{bmatrix} \boxed{\underline{s}_1} \\ \vdots \\ \underline{s}_n \end{bmatrix}$$

Then we can write

(1) $\hat{Y}_i = \underline{s}_i \underline{Y}$.

We want to show that

$$\sum_{i=1}^{n} Cov(\hat{Y}_i, Y_i) = tr(S)\sigma^2.$$

Let's look at $Cov(\hat{Y}_i, Y_i)$

$$Cov(\hat{Y}_i, Y_i) \overset{(1)}{=} Cov(\underline{s}_i \underline{Y}, Y_i)$$

$= cov(s_{i1}Y_1 + \dots + s_{ii}Y_i + \dots s_{in}Y_n, Y_i)$

Rules of covariance

$= cov(s_{i1}Y_1, Y_i) + \dots + cov(s_{ii}Y_i, Y_i)$

$+ \dots + Cov(s_{in}Y_n, Y_i)$

$= s_{i1} cov(Y_1, Y_i) + \dots + s_{ii} Cov(Y_i, Y_i)$

$+ \dots + s_{in} \underbrace{Cov(Y_n, Y_i)}_{= 0 \text{ due to independence}}$

$\overset{(2)}{=} s_{ii} Cov(Y_i, Y_i)$

$= s_{ii} Var(Y_i)$

$= s_{ii} \sigma^2$

We can now look at the sum

$$\sum_{i=1}^{n} Cov(\hat{Y}_i, Y_i) = \sum_{i=1}^{n} s_{ii}\sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} s_{ii}$$

$$= \sigma^2 \cdot tr(S).$$

$$tr(X) = tr\left(\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}\right) = \sum_{i=1}^{n} x_{ii}$$

sum of diagonal elements

$\boxed{\hat{Y} = HY},$ $H = \underline{X}\underbrace{(X^T X)^{-1}}_{\text{symmetric}}\underline{X}^T$

KNN - We look at $k = 2$.

so $\hat{Y}_2$ will be based on the values of 2 closest neighbors. So in our example, that is Y_1 and Y_3.

$\hat{Y}_2$ = avg of Y_1 and Y_3

$= \frac{1}{2}(Y_1 + Y_3) = \frac{1}{2}Y_1 + 0Y_2 + \frac{1}{2}Y_3$

So if $\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$

$$\hat{Y}_2 = \underbrace{\begin{array}{ccc} y_1 & y_2 & y_3 \\ \hline \frac{1}{2} & 0 & \frac{1}{2} \end{array}}_{\underline{s}_2} \underline{Y} = \underline{s}_2 \underline{Y}$$

original respons

So $s_{21} = 1/2$, that means that Y_1 contributes half of the value of $\hat{Y}_2$.

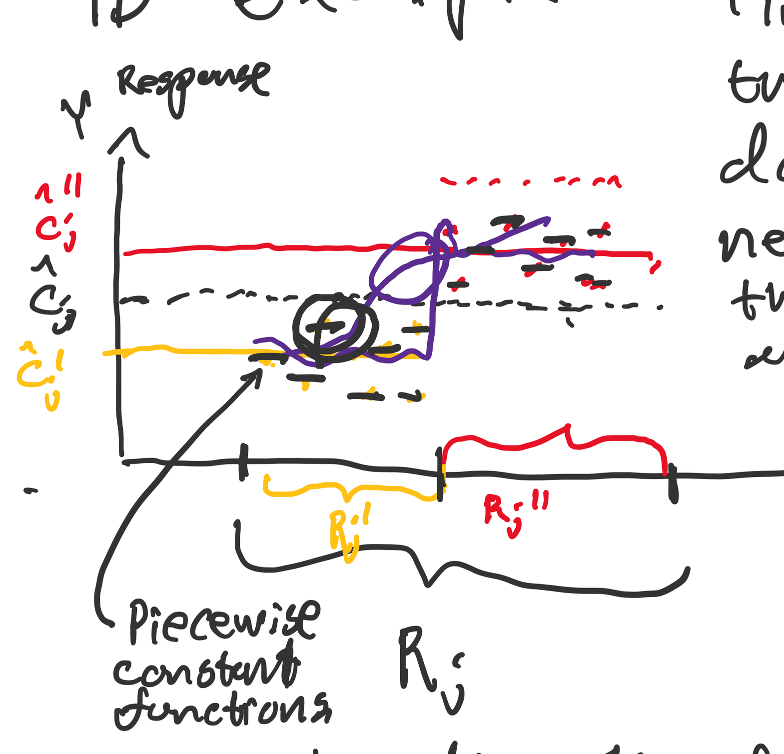
4.8)

Tree growth algorithm on page 99-103.

Show that $\underbrace{D_j - D_j^* > 0}$.

(Indicates that traing RSS always decreas in each new iteration of the tree growth algorithm.

1D-example



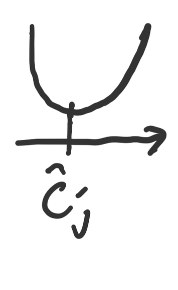
So in iteration k of the algorithm we split R_j into two subregions R_j' and R_j''.

$D_j = \sum_{i \in R_j} (y_i - \hat{c}_j)^2$, where $\hat{c}_j$ is the average respons in region R_j.

This means that any other value for $\hat{c}_j$ will yield a larger D_j.

And

$$D_j^* = \sum_{i \in R_j'} (y_i - \hat{c}_j')^2 + \sum_{i \in R_j''} (y_i - \hat{c}_j'')^2$$



$$< \sum_{i \in R_j'} (y_i - \hat{c}_j)^2 + \sum_{i \in R_j''} (y_i - \hat{c}_j)$$

$$= \sum_{i \in R_j} (y_i - \hat{c}_j)^2$$

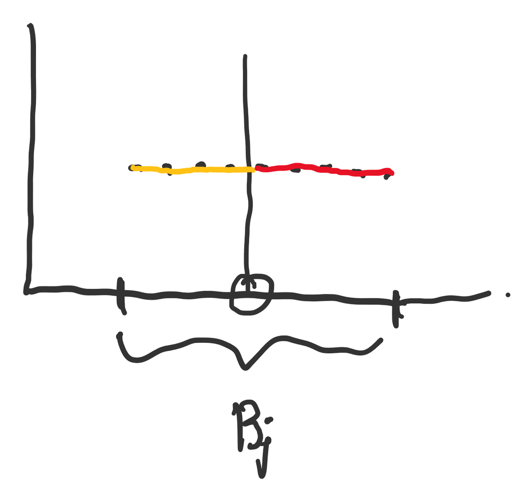
$$= D_j$$

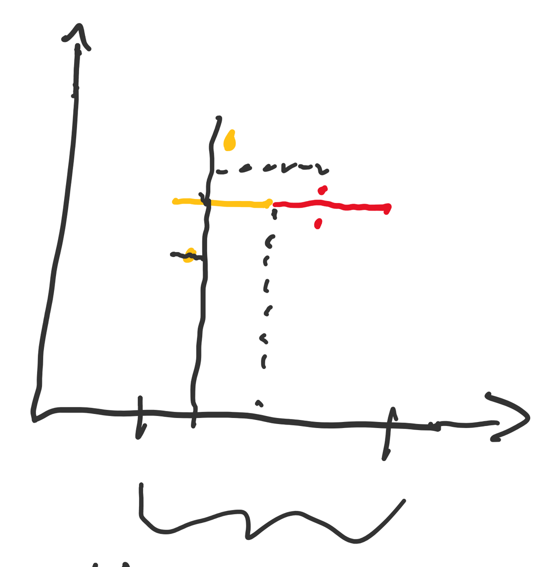
We have that $D_j^* < D_j$, this means that $0 < D_j - D_j^*$.

Degenerate cases.

1. If there is only 1 obs. in R_j.
2. If all the responses in R_j have the same respons value.

Then $\hat{c}_j = \hat{c}_j' = \hat{c}_j''$.





(Maybe enough with the means of R_j' and R_j'' being the same).