EE559

HOMEWORK-1

1. $K=3$

$\therefore \hat{y}_{KNN} = \frac{y_1+y_2+y_3}{3}$

for 150 cm,

$\because K=3$, we choose 3-nearest neighbours.

distance metric = Manhattan $\Sigma|x_i-y_i|$

Person	Height	Weight	Distance
1	171	80	21 ✓
2	168	78	18 ✓
3	191	100	41
4	182	80	32
5	150	65	0 ✓
6	178	83	28

$\therefore$ weight for 150cm height

$= \frac{80+78+65}{3} = \frac{223}{3} \simeq 74.33$ kg.

$\Rightarrow$ for Height = 155cm, 165cm

Person	Height	Weight	for Distance (155)	for Distance (165)	for 190.
1	171	80	16 ✓	6 ✓	19
2	168.	78	13 ✓	3 ✓	22
3	191	100	36	26	1 ✓
4	182	80	27	17	8 ✓
5	150	65	5 ✓	15	40
6	178	83	23	13 ✓	12 ✓

∴ weight for 155cm,

$$= \hat{y}_{KNN} = \frac{65+78+80}{3} = 74.33\text{kg}$$

weight for 165 cm,

$$\hat{y}_{KNN} = \frac{78+80+83}{3} = 80.33\text{kg}$$

for, 190 cm,

$$\hat{y}_{KN} = \frac{100+80+83}{3} \approx 87.67\text{kg}$$

Answer:

for height(cm)	Weight $\hat{y}_{KNN}$ (kg)
150	74.33
155	74.33
165	80.33
190	87.67

2. using

$$\hat{y}_{KNN} = \frac{w_1 y_1 + w_2 y_2 + w_3 y_3}{w_1 + w_2 + w_3}$$

where, $w_i = \frac{1}{d_i}$

d_i = distance metric.

Person	Height (cm)	Weight (kg)	1/distance (150) wi	wi for 155	wi for 165	wi for 190
	171	80	$1/21 = 0.046$	0.0625	0.1667	0.0526
	168	78	$1/18 = 0.055$	0.0769	0.33	0.0455
	191	100	0.0244	0.0278	0.038	$1/1 = 1$
	182	80	0.0313	0.0370	0.0588	0.125
	150	65	$1/0$ (use actual weight) (60)	0.2	0.066	0.025
	178	83	$1/28 = 0.035$	0.0435	0.076	0.0833

for 150 cm,

$\hat{y}_{KNN} = 65\ kg$

$\because$ for 150cm, weight $= 1/0 =$ dominant over all others.

for 155,

$$\hat{y}_{KNN} = \frac{(0.2)(65) + (0.0769)(78) + (0.0625)(80)}{0.2 + 0.0769 + 0.0625}$$

$$= 70.69\ kg$$

for 165 cm,

$$\hat{y}_{KNN} = \frac{(0.333)(78) + (0.1667)(80) + (0.076)(83)}{0.333 + 0.166 + 0.076}$$

$$= 77.99\ kg.$$

for 190 cm,

$$\hat{y}_{KN} = \frac{1(100) + (0.125)(80) + (0.0833)(83)}{1 + 0.125 + 0.083}$$

$$\hat{y}_{KNN} = 96.03\ kg$$

Answer:

Height (cm)	Weight (kg) $\hat{y}_{KNN}$
150	65 $\rightarrow$ $\because$ dominant weight.
155	70.69
165	77.99
190	96.03

3. $J(x) = x^T Q x + d^T x + c$

where,

$Q = Q^T \in R^{n\times n}$ $\Rightarrow Q_{n\times n}$

$x, d \in R^n$ $\Rightarrow x^T_{1\times n}$ $\Rightarrow x = n\times 1$

$d^T_{1\times n}$

$c \in R$

To show, $\nabla_x J(x) = 2Qx + d$

$H = \frac{\partial^2 J}{\partial x \partial x^T} = 2Q$

sol:

$\nabla_x J(x)$

for $x^T Q x$:

$x^T Q x$ can be decomposed as,

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & & \\ q_{n1} & \cdots & q_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$\cancel{= x_1 q_{11} x_1 + x_2 q_{22} x_2 + \cdots + x_i q_{ij} x_j}$

~~∴ differentiating w.r.t x.~~

Consider $Q_{2\times 2}$

$\therefore x^T Q x$

$$= \underset{1\times 2}{\begin{bmatrix} x_1 & x_2 \end{bmatrix}} \underset{2\times 2}{\begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{bmatrix} q_{11}x_1 + q_{21}x_2 & q_{12}x_1 + q_{22}x_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= q_{11}x_1^2 + q_{21}x_1x_2 + q_{12}x_1x_2 + q_{22}x_2^2 \quad = f(x)$$

differentiating w.r.t x

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix}$$

$$\frac{\partial f(x)}{\partial x_1} = 2x_1 q_{11} + x_2 q_{12} + x_2 q_{21}$$

$$\frac{\partial f(x)}{\partial x_2} = q_{12} x_1 + x_1 q_{21} + 2 q_{22} x_2$$

$$\Rightarrow \frac{\partial f(x)}{\partial x} = \underbrace{\begin{bmatrix} 2q_{11} & q_{12}+q_{21} \\ q_{12}+q_{21} & q_{22} \end{bmatrix}}_{2\times2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \left(\underbrace{\begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}}_{Q} + \underbrace{\begin{bmatrix} q_{11} & q_{21} \\ q_{12} & q_{22} \end{bmatrix}}_{Q^T} \right) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$\because Q = Q^T$

we get $\frac{\partial f(x)}{\partial x} = 2Qx$ --(i) { generalising where, $Q \in R^{n\times n}$, $x \in R^n$ }

now, consider $d^T x$

$d^T = [d_1 \ d_2 \ \cdots \ d_n]$

$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$

$\because d^T x = d_1 x_1 + d_2 x_2 + \cdots + d_n x_n = f(x)$

d. w.r.t x,

$$\frac{\partial f(x)}{\partial x_i} = d_i \Rightarrow \frac{\partial f(x)}{\partial x} = \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} = d \text{ ------ (ii)}$$

$\therefore \boxed{\nabla_x J(x) = 2Qx + d}$ from (i) & (ii)

Hessian:

$$\frac{\partial f(x)}{\partial x \partial x^T} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1 \partial x_1} & \frac{\partial f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2 \partial x_1} & \frac{\partial f(x)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \\ \frac{\partial f(x)}{\partial x_n x_1} & \frac{\partial f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

$$\therefore \frac{\nabla J(x)}{\partial x \partial x^T} = ?$$

from (i)

$$\frac{\partial f(x)}{\partial x \partial x^T} = \begin{bmatrix} 2q_{11} & q_{12}+q_{21} \\ q_{12}+q_{21} & 2q_{22} \end{bmatrix} = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} + \begin{bmatrix} q_{11} & q_{21} \\ q_{12} & q_{22} \end{bmatrix}$$

$$= Q + Q^T$$

$$= 2Q$$

from (ii)

$$\frac{\partial f(x)}{\partial x \partial x^T} = \frac{\partial^2 f(x)}{\partial x \partial x^T} = \frac{\partial d}{\partial x} = 0$$

$$\boxed{\therefore \frac{\nabla J(x)}{\partial x \partial x^T} = 2Q}$$ = Hessian.

4. Prediction $\hat{y}$ for a test input x' is given by

$$\hat{y} = x'_{1\times p}\,\hat{\beta}_{p\times 1}$$

where, $\hat{\beta}$ = estimated regression co-efficients, calculated by normal equation,

$$RSS = \|y - X\beta\|^2 \quad \text{where, } X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}_{n\times p+1}$$

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

setting gradient = 0, we get $\hat{\beta}$ (min).

$$\frac{\partial}{\partial\beta}(y - X\beta)^T(y - X\beta) = 0$$

$$= \frac{\partial}{\partial\beta}\left(y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta\right) = 0$$

$$= \frac{\partial}{\partial\beta}\left(-2y^TX\beta + \beta^TX^TX\beta\right) = 0 \quad \text{—(i)}$$

from (8) we have, if $J(x) = x^TQx + d^Tx + c$ we have, $\nabla_x J(x) = 2Qx + d$. $\{Q$ is symm$\}$

∴ (i) becomes, (also taking derivative).

$$2X^TX\beta - 2X^Ty = 0$$

$$\Rightarrow \hat{\beta} = (X^TX)^{-1}X^Ty$$

$$\therefore \boxed{\hat{y} = x'_{1\times p}(X^TX)^{-1}X^Ty}$$

$(X^TX)^{-1}X^T$

Viewing as special case of KNN

Linear regression uses training data (K=n) where weights are obtained from ~~solution~~ solving normal equation. In KNN, weights are determined by distance of test data to each train data.

∴ linear regression can be seen as a special case where weights are based on normal equation & not distance and also as KNN with k=n.

5. To show:

for $y \in \mathbb{R}^n$

$\hat{y} = X(X^TX)^{-1}X^Ty \in R(X)$ { Range Space of $X \Rightarrow$ linear combination of X

where, $X \in \mathbb{R}^{n \times (p+1)}$

We have,

$$\hat{y} = X(X^TX)^{-1}X^Ty$$

From the definition/identifying in vector v belongs to column space of some vector X, we need to have

$$v = Xc$$

where c = some scalar matrix (α). $\mathbb{R}^n$

here,

if $X \in \mathbb{R}^{n \times (p \times 1)}$

we need to show if any vector, say

$$v = c_1x_1 + c_2x_2 + \cdots + c_{p+1}x_{p+1}$$

for some scalars $c_1, c_2 \cdots c_{p+1}$.

$$\hat{y} = X(X^TX)^{-1}X^Ty$$

if $X = X_{n \times (p+1)}$

$X^TX = X_{(p+1) \times n} X_{n \times (p+1)}$

$\Rightarrow (X^TX)^{-1}_{(p+1) \times (p+1)}$

$\Rightarrow (X^TX^{-1})X^T = (X^TX)^{-1}_{p+1 \times p+1} X^T_{p+1 \times n} \Rightarrow [(X^TX^{-1})X^T]_{p+1 \times n}$

$\boxed{\therefore \hat{y} = Xc}$ where $c = (X^TX)^{-1}X^T_{p+1} y_{n \times 1} \in \mathbb{R}^{p+1}$

where, $c \in \mathbb{R}^{p+1}$ = vector of co-efficients

$\therefore \hat{y}$ is linear combination of cols of X

6. In Linear Regression, if $\hat{\beta}$ min $RSS(\beta)$ then, $(y-\hat{y}) \perp C(X)$

Proof: $\because$ If $(y-\hat{y}) \perp C(X)$ then $(y-\hat{y})^T X = 0$

Let $y-\hat{y} = e$

$\therefore$ from orthogonality we need find if $e^T Xc = 0$

Proof:

we have $\hat{\beta} = (X^T X)^{-1} X^T y$

$\hat{y} = X\hat{\beta}$; $e = y-\hat{y}$

$\therefore e^T Xc = (y-\hat{y})^T Xc$

$e^T Xc = (y - X(X^T X)^{-1} X^T y)^T Xc$

$e^T Xc = y^T Xc - y^T X \underbrace{(X^T X)^{-1} X^T X}_{I}$

$= y^T Xc - y^T Xc$

$\Rightarrow e^T Xc = 0$

$\Rightarrow$ for any vector $c \in R^{p+1}$, residual vector e is orthogonal to Column Space of X.