Data mining and machine learning (Assignment-2)

(3) (i) $(x, y) = [(5, 40)\ (7, 120)\ (12, 180)\ (16, 210)]$

Computing X^TX and $(X^TX)^{-1}$

step 1: Construct the input matrix X

for the linear regression the input matrix X will be like

$$X = \begin{bmatrix} 1 & 5 \\ 1 & 7 \\ 1 & 12 \\ 1 & 16 \end{bmatrix}$$

step 2: Construct the output vector y

The output vector y is constructed from the y values of data points

$$y = \begin{bmatrix} 40 \\ 120 \\ 180 \\ 210 \end{bmatrix}$$

step 3: Compute X^TX

Now we compute the transpose of x and multiply by x

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 7 & 12 & 16 \end{bmatrix}$$

Calculate X^TX

$$X^TX = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 7 & 12 & 16 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ 1 & 7 \\ 1 & 12 \\ 1 & 16 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} 4 & 40 \\ 40 & 5^2+7^2+12^2+16^2 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} 4 & 40 \\ 40 & 474 \end{bmatrix}$$

<u>step 4</u>: Compute $(X^TX)^{-1}$

If $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ then $A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

Then $A = \begin{bmatrix} 4 & 40 \\ 40 & 474 \end{bmatrix}$

Calculating the determinant ad-bc

$\det(A) = (4)(474) - (40)(40) = 1896 - 1600 = 296$

Now applying the inverse formula:

$$(X^TX^{-1}) = \frac{1}{296} \begin{bmatrix} 474 & -40 \\ -40 & 4 \end{bmatrix} = \begin{bmatrix} 474/296 & -40/296 \\ -40/296 & 4/296 \end{bmatrix}$$

Calculating the fractions:

$$(X^T X)^{-1} = \begin{bmatrix} 1.5973 & -0.1351 \\ -0.1351 & 0.0135 \end{bmatrix}$$

Therefore the

matrix $X^T X = \begin{bmatrix} 4 & 40 \\ 40 & 474 \end{bmatrix}$

Inverse $(X^T X)^{-1} \approx \begin{bmatrix} 1.5973 & -0.1351 \\ -0.1351 & 0.0135 \end{bmatrix}$

(3) (ii) Computing $\hat{\omega}_0$ and $\hat{\omega}_1$

$$\begin{bmatrix} \hat{\omega}_0 \\ \hat{\omega}_1 \end{bmatrix} = (X^T X)^{-1} X^T y$$

step 1: Constructing the output vector y from the data points $(x, y) = [(5, 40)\ (7, 120)\ (12, 180)\ (16, 210)$ the output vector y is

$$y = \begin{bmatrix} 40 \\ 120 \\ 180 \\ 210 \end{bmatrix}$$

step 2 : Compute $X^T y$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 7 & 12 & 16 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 7 & 12 & 16 \end{bmatrix} \begin{bmatrix} 40 \\ 120 \\ 180 \\ 210 \end{bmatrix}$$

Calculating the entries

$$X^T y = \begin{bmatrix} 550 \\ 6560 \end{bmatrix}$$

step 3 : Compute $\begin{bmatrix} \hat{\omega}_0 \\ \hat{\omega}_1 \end{bmatrix}$

$$\begin{bmatrix} \hat{\omega}_0 \\ \hat{\omega}_1 \end{bmatrix} = (X^T X)^{-1} X^T y$$

$$\begin{bmatrix} \hat{\omega}_0 \\ \hat{\omega}_1 \end{bmatrix} = \begin{bmatrix} 1.5973 & -0.1351 \\ -0.1351 & 0.0135 \end{bmatrix}$$

now multiply $(X^T X)^{-1}$ by $X^T y$:

$$\begin{bmatrix} \hat{\omega}_0 \\ \hat{\omega}_1 \end{bmatrix} = \begin{bmatrix} 1.5973 & -0.1351 \\ -0.1351 & 0.0135 \end{bmatrix} \begin{bmatrix} 550 \\ 6560 \end{bmatrix}$$

Therefore $\hat{\omega}_0 = -6.541$

$\hat{\omega}_1 = 14.591$

thus the linear regression model can be given as: $\hat{y} = -6.541 + 14.591x$

(6) (i) the F-value for the given ANOVA problem will be given as:

$$F\text{-static } (F) = \frac{MSB}{MSW}$$

where MSB is the mean square b/w ~~groups~~

MSW is the mean square within groups

→ Calculating the group means

X_1 mean $(\bar{X}_1)$:

$$\bar{X}_1 = \frac{8+8+10+7+10}{5} = \frac{43}{5} = 8.6$$

X_2 mean $(\bar{X}_2)$:

$$\bar{X}_2 = \frac{5+6+6+4+8}{5} = \frac{29}{5} = 5.8$$

X_3 mean $(\bar{X}_3)$:

$$\bar{X}_3 = \frac{7+6+7+7+9}{5} = \frac{36}{5} = 7.2$$

→ Calculating overall mean:

$$\bar{X} = \frac{8+8+10+7+10+5+6+6+4+8+7+6+7+7+9}{15}$$

$$= \frac{93}{15} = 6.2$$

→ Calculating SSB

$$SSB = n_1(\bar{X}_1 - \bar{X}_2)^2 + n_2(\bar{X}_2 - \bar{X})^2 + n_3(\bar{X}_3 - \bar{X})^2$$

Assuming $n_1 = n_2 = n_3 = 5$

$$SSB = 5 \times (8.6 - 6.2)^2 + 5 \times (5.8 - 6.2)^2 + 5 \times (7.2 - 6.2)^2$$

$$= 34.6$$

→ Calculating SSW:

X_1 group: $(8-8.6)^2 + (8-8.6)^2 + (10-8.6)^2 + (7-8.6)^2 + (10-8.6)^2$

$$= 7.2$$

X_2 group: $(5-5.8)^2 + (6-5.8)^2 + (6-5.8)^2 + (4-5.8)^2 + (8-5.8)^2$

$= 8.8$

X_3 group: $(7-7.2)^2 + (6-7.2)^2 + (7-7.2)^2 + (7-7.2)^2 + (9-7.2)^2$

$= 4.8$

Therefore $SSW = 7.2 + 8.8 + 4.8 = 20.8$

→ Calculate mean square:

$MSB = \frac{SSB}{k-1} = \frac{34.6}{2} = 17.3$ (where $k=3$ is the no. of groups)

$MSW = \frac{SSW}{N-k} = \frac{20.8}{12} \approx 1.73$ (where $N=15$ is the total no. of observation)

→ Calculate the F-value

$$F = \frac{MSB}{MSW} = \frac{17.3}{1.73} = 10$$

(6) (ii) As the critical value is 3.89 and the F-value is 10 we reject the null hypothesis H_0

If the calculated f-value is less than or equal to the critical f-value, fail to reject the null hypothesis H_0.

(8) (i) Given sample points $X = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix}$

Calculating the each dimension

$$\mu_1 = \frac{4+2+5+1}{4} = \frac{12}{4} = 3$$

$$\mu_2 = \frac{1+3+4+0}{4} = \frac{8}{4} = 2$$

Center the data

$$X_{centered} = X - [\mu_1 \ \mu_2] = \begin{bmatrix} 4-3 & 1-2 \\ 2-3 & 3-2 \\ 5-3 & 4-2 \\ 1-3 & 0-2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix}$$

Covariance of the matrix

$$S = \frac{1}{n-1} \left(X^T_{centered} \, X_{centered} \right)$$

$$X_{centered}^{T} X_{centered} = \begin{bmatrix} 1 & -1 & 2 & -2 \\ -1 & 1 & 2 & -2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix}$$

$$= \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

Covariance matrix $S = \frac{1}{4-1} \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$

$$= \frac{1}{3} \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix} = \begin{bmatrix} 10/3 & 2 \\ 2 & 10/3 \end{bmatrix}$$

$$= \frac{1}{3}$$

∴ the final covariance matrix S is :

$$S = \begin{bmatrix} 10/3 & 2 \\ 2 & 10/3 \end{bmatrix}$$

(8) (ii) Computing the eigen values λ_1 and λ_2

$\det(S - \lambda I) = 0$ where I = identity matrix

$$S - \lambda I = \begin{bmatrix} 10/3 - \lambda & 2 \\ 2 & 10/3 - \lambda \end{bmatrix}$$

Calculating the determinant:

$$\det(S - \lambda I) = \left(\frac{10}{3} - \lambda\right)^2 - 4 = 0$$

$$\left(\frac{10}{3} - \lambda\right)^2 = 4$$

$$\frac{10}{3} - \lambda = \pm 2$$

solving for λ

1. $\frac{10}{3} - \lambda = 2$

$$\lambda_1 = \frac{10}{3} - 2 = \frac{10}{3} - \frac{6}{3} = \frac{4}{3}$$

2. $\frac{10}{3} - \lambda = -2$

$$\lambda_2 = \frac{10}{3} + 2 = \frac{10}{3} + \frac{6}{3} = \frac{16}{3}$$

(8) (iii) $(S - \lambda I)V = 0$

$$\lambda_1 = \frac{4}{3}$$

$$S - \frac{4}{3}I = \begin{bmatrix} \frac{10}{3} - \frac{4}{3} & 2 \\ 2 & \frac{10}{3} - \frac{4}{3} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

Setting up the equation

$$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$2x_1 + 2x_2 = 0 \implies x_1 = -x_2$$

$$v_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

for $\lambda_2 = \frac{16}{3}$

$$S - \frac{16}{3} I = \begin{bmatrix} \frac{10}{3} - \frac{16}{3} & 2 \\ 2 & \frac{10}{3} - \frac{16}{3} \end{bmatrix}$$

$$= \begin{bmatrix} -6/3 & 2 \\ 2 & -6/3 \end{bmatrix} = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}$$

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$-2x_1 + 2x_2 = 0$$

$$x_1 = x_2$$

$$v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

(8) (iv) $\lambda_2 = \frac{16}{3}$ $\quad v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

The line represents $y = x$

The sample points are:

(4, 1)

(2, 3)

(5, 4)

(1, 0)

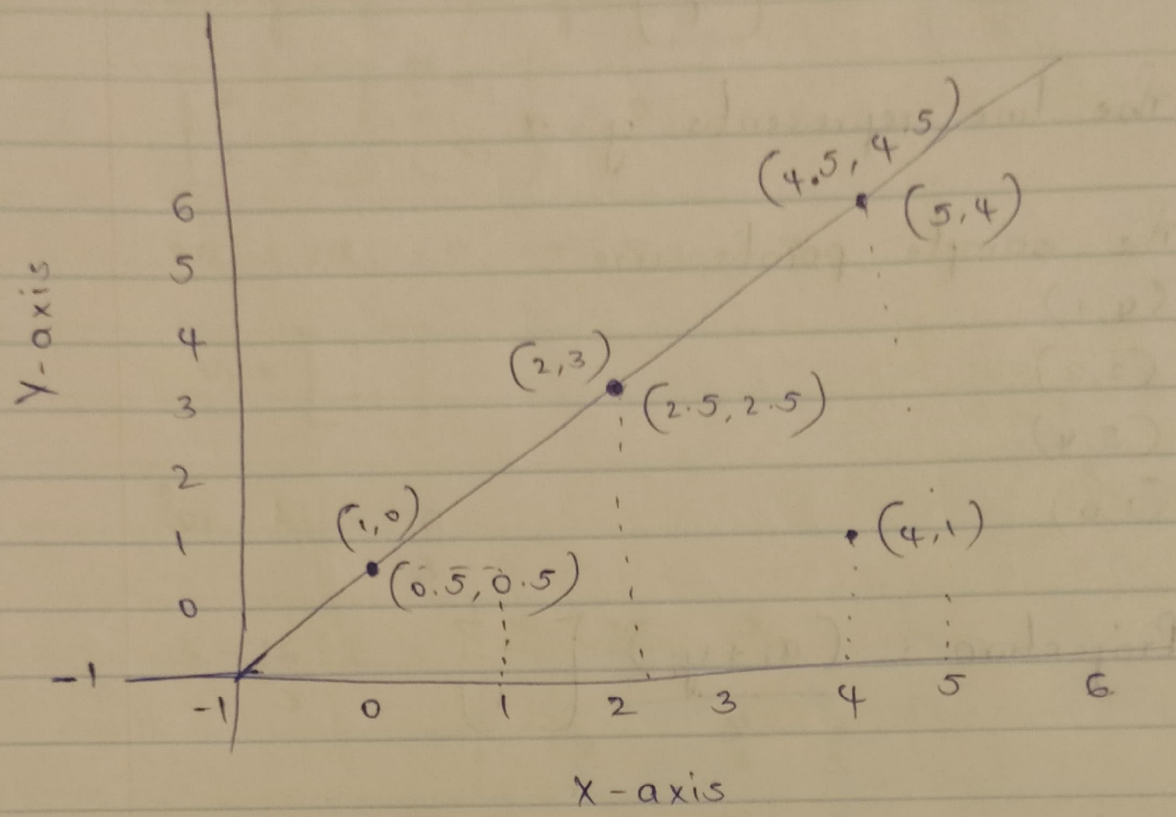
$$\text{Projection} = \frac{(x_i + y_i)}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$p(4,1) : \frac{4+1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{5}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = (2.5, 2.5)$$

$$p(2,3) : \frac{2+3}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{5}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = (2.5, 2.5)$$

$$p(5,4) : \frac{5+4}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{9}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = (4.5, 4.5)$$

$$p(1,0) : \frac{1+0}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = (0.5, 0.5)$$



(10) (i) To compute the entropy H (Target) of the target variable we have to follow the required steps

step 1: Identify unique classes
the target class can be given as:
T (True)
F (False)

step 2: Count the instances
Count the occurences of each class

Count of T : 2 (from rows 1 and 5)
Count of F : 3 (from rows 2, 3 and 4)

step 3 : Calculate probabilities
Total number of instances = 5

Probability of T :

$$P(T) = \frac{\text{Count of } T}{\text{Total}} = \frac{2}{5} = 0.4$$

Probability of F :

$$P(F) = \frac{\text{Count of } F}{\text{Total}} = \frac{3}{5} = 0.6$$

step 4 : Using the entropy formula

Entropy H :

$$H(\text{Target}) = -\Sigma(P(x) \cdot \log_2(P(x))$$

where x represents each class

Calculating H(Target) :

$$H(\text{Target}) = -(P(T) \cdot \log_2(P(T)) + P(F) \cdot \log_2(P(F)))$$

Substituting the probabilities

$H(\text{Target}) = -(0.4 \cdot \log_2(0.4) + 0.6 \cdot \log_2(0.6))$

step 5: Calculate logarithms

$\log_2(0.4) \approx -1.3219$

$\log_2(0.6) \approx -0.73697$

Substituting these values back to entropy values:

$H(\text{Target}) = -(0.4 \cdot (-1.32193) + 0.6 \cdot (-0.73697))$

$H(\text{Target}) = -(0.528772 + -0.442182)$

$H(\text{Target}) = 0.528772 + 0.442182 \approx 0.970954$

$\therefore H(\text{Target}) \approx 0.971$

(10) (iii) step 1: Calculate $H(\text{Target} | A)$

Split the data by attribute A

the attribute A has two values: Y and N

for A = Y

Instances: (Y, N, N, T)

(Y, N, Y, F)

(Y, Y, Y, F)

(Y, Y, N, T)

Target counts :

T : 2

F : 2

Total instances for A = Y : 4

Probabilities :

$$P(T|A=Y) = \frac{2}{4} = 0.5$$

$$P(F|A=Y) = \frac{2}{4} = 0.5$$

Entropy for A = Y

$$H(\text{Target}|A=N) = -(P(T|A=N) \cdot \log_2(P(T|A=N)) + P(F|A=N) \cdot \log_2(P(F|A=N))$$

$$H(\text{Target}|A=N) = -(0.5 \cdot \log_2(0.5) + 0.5 \log_2(0.5))$$

$$= 1$$

<u>step 2</u>: Calculate $H(\text{Target}|A)$

$$H(\text{Target}|A) = P(A=Y) \cdot H(\text{Target}|A=Y) + P(A=N) \cdot H(\text{Target}|A=N)$$

Total instances = 5

$$P(A=Y) = \frac{4}{5} = 0.8$$

$P(A=N) = \frac{2}{5} = 0.4$

Calculating $H(\text{Target}|A)$:

$0.8 \times 1 + 0.4 \times 1 = 1.2$

step 3: Calculate information gain $IG(A)$

$IG(A) = H(\text{Target}) - H(\text{Target}|A)$

$H(\text{Target}) \approx 0.971$

$H(\text{Target}|A) \approx 1.2$

Calculating $IG(A) = 0.971 - 1.2 = -0.229$

step 4: Calculate gain ratio

$IV(A) = -(P(A=Y) \cdot \log_2(P(A=Y)) + P(A=N) \cdot \log_2(P(A=N)))$

$P(A=Y) = \frac{4}{5} = 0.8$

$P(A=N) = \frac{1}{5} = 0.2$

Calculating $IV(A)$

$IV(A) = -(0.8 \cdot \log_2(0.8) + 0.2 \cdot \log_2(0.2))$

$\log_2(0.8) \approx -0.32193$

$\log_2(0.2) \approx -2.3219$

calculating $IV(A) = -(0.8 \cdot (-0.32193) + 0.2 \cdot (-2.32193))$

$IV(A) = 0.2575 + 0.4643 = 0.72193$

finally the gain ratio (A):

$$\frac{IG(A)}{IV(A)} = \frac{-0.229}{0.72193} \approx -0.317$$

$\therefore$ H (Target | A) ≈ 1.2

IG (A) ≈ -0.229

Gain ratio (A) ≈ -0.317