Set - D

DA End Sem

Sayam Kumar
S20180010158

Q-1

X = Average ambient temp (x) in F

(a) Y = Average monthly power consumption

For simple LR, $Y = \beta X + \alpha$

$\bar{x} = 73.2$

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = 70.4$

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\beta = \frac{(82-73.2)(76-70.4) + (73-73.2)(83-70.4) + \dots}{(82-73.2)^2 + (73-73.2)^2 + \dots}$$

$\beta = 0.7651$

$\alpha = \bar{y} - \beta \bar{x}$

$= 70.4 - (0.7651)(73.2)$

$= 14.39$

So, equation is

$Y = 0.7651 X + 14.39$

$y = 0.7651 X + 14.39$

(4) The validity of the model canbe done

SSE = Residual sum of squared error

$$= \sum_{i=1}^{n} (\text{actual output} - \text{predicted output})^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = 1714.62$$

SST = Total corrected sum of squares

$$= \sum_{i=1}^{n} (\text{acutal output} - \text{average of output})$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 = 4275.6$$

$$R^2 = 1 - \frac{SSE}{SST} = 0.599$$

The significance level of our regression analysis is (0.599).

Q-2 (a) for the given data in question, spearman correlation is applicable

The sample data is of ordinal type

And for ordinal data, the spearman correlation analysis is applicable.

(h) Calculate coefficient of determination and interpret the results.

Sample #	Rank x	Rank y	d	d^2
1	1	4.5	-3.5	12.25
2	2	2	0	0
3	9	2	7	49
4	8	2	6	36
5	7	6.5	0.5	0.25
6	6	8.5	-2.5	6.25
7	5	8.5	-3.5	12.25
8	4	6.5	-2.5	16.25
9	3	4.5	-1.5	2.25

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 134.5}{9 \times 80} = 0.5018$$

Coefficient of determination $= r_s^2$

$= 0.2518$

$$t = r\sqrt{\frac{n-1}{1-r^2}} = 1.63$$

Yes, Around 60% are same

Q.3

(a) X	3	2	0	5	0	0	0	2	0	0
Y	1	0	0	0	0	0	0	1	0	2

metric used => cosine similarity

$$\text{similarity} = \frac{A \cdot B}{\|A\| * \|B\|}$$

$$= \frac{5}{\sqrt{6}\sqrt{42}} = 0.3149$$

(b) Test set size = 100

Predicts 80 test tuples correctly

i) observed accuracy = 0.80 = 80/100 = 0.8

ii) standard error rate = $\sqrt{\frac{\varepsilon(1-\varepsilon)}{N}} = \sqrt{\frac{(0.8)(0.2)}{100}}$

$= 0.04$

iii) True accuracy

$\alpha = 95\%$, $T_\alpha = 1.96$

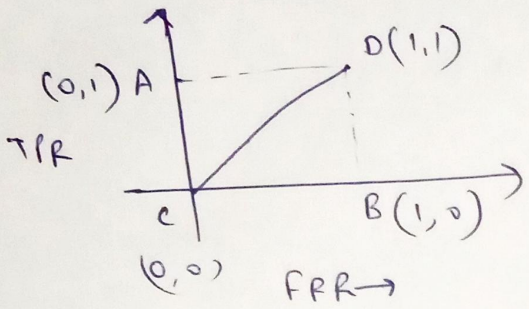
$$\tilde{\varepsilon} = \varepsilon \pm T_\alpha * \sqrt{\varepsilon(1-\varepsilon)/N}$$

$$= 0.8 \pm (1.96 \times 0.04)$$

$= 0.7216$

(c) Next Page 2

(c) Roc plot



A → FPR = 0 and TPR = 1 ⇒ Ideal classifier

B → FPR = 1 and TPR = 0 ⇒ Worst classifier

C ⇒ FPR = 0 and TPR = 0 ⇒ Ultra conservative classifier
(predict everything to be -ve)

D = FPR = 1 and TPR = 1 ⇒ Ultra liberal classifier
(predict everything to be of +ve class)

Any other classifier that lies on diagonal ⇒ random classifier
eg → (0.5, 0.5)
(FPR, TPR)

(d) Consider Confusion matrix TP = 80, FP = 15
FN = 25, TN = 70

i) Precision = $\frac{TP}{TP+FP} = \frac{80}{80+15} = \frac{80}{95} = 0.8421$

ii) Recall
$$= \frac{TP}{TP+FN} = \frac{80}{105} = 0.7619$$

iii) Sensitivity $= \frac{TP}{TP+FN} = \frac{80}{80+25} = 0.7619$

Recall and Sensitivity are same.

Q.4 ~~After 3 steps~~, After step 1
the centroids are -

$$\text{centroid} = \begin{bmatrix} 5.882, & 3.528 \\ 5.290, & 3.3 \\ 5.350, & 3.34 \end{bmatrix}$$

Step 2

$$\text{centroid} = \begin{bmatrix} 5.94, & 3.5 \\ 5.29, & 3.3 \\ 5.4, & 3.3 \end{bmatrix}$$

Step 3

$$\begin{bmatrix} 5.995, & 3.6 \\ 5.3, & 3.35 \\ 5.5, & 3.51 \end{bmatrix}$$

---

(4)) When there are outliers

⇒ Anamolies