

Question-01: The decision-tree model was built using X_1, X_2, X_3 and X_4 to predict the response of Y . The following table contains training data from a database. Let Y be the class label attribute. Obtain the Gain Ratio of for the attribute X_4 - the attribute selection measure in classification.

X_1	X_2	X_3	X_4	Y
N	F	W	M	N
G	F	W	L	Z
N	D	M	M	Z
N	F	S	M	Z
G	F	S	L	Z
G	D	W	M	Z
N	F	W	H	N
N	D	W	H	N
N	F	M	H	Z
N	D	S	M	N
G	D	S	L	N
G	D	M	L	Z
G	F	M	H	Z
G	F	S	M	Z

Solution: To find the Gain Ratio, firstly, we have to construct the following table -

X_4		N		Z	Total
		High(H)	Low(L)		
	High(H)	2		2	4
	Low(L)	1		3	4
	medium(m)	2		4	6

$$\text{Info}(D) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right)$$

$$= 0.94028 \text{ bits}$$

$$\text{Info}_{X_4}(D) = \frac{4}{14} \left(-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} \right) +$$

$$\frac{4}{14} \left(-\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} \right) +$$

$$\frac{6}{14} \left(-\frac{2}{6} \log_2 \frac{2}{6} - \frac{4}{6} \log_2 \frac{4}{6} \right)$$

$$= 0.911063$$

$$\text{Gain}(X_4) = \text{Info}(D) - \text{Info}_{X_4}(D)$$

$$= 0.94028 - 0.911063$$

$$= 0.029217 \text{ bits}$$

$$\text{Split Info}_{X_4}(D) = -\frac{4}{14} \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \log_2\left(\frac{6}{14}\right) -$$

$$\frac{4}{14} \log_2\left(\frac{4}{14}\right) = 1.557$$

$$\text{Gain Ratio}(X_4) = \frac{0.029}{1.557} = 0.019$$

Question-02: The predicted probability and the true label are given in the following table. For each threshold, classify any instance with a predicted probability greater than or equal to the threshold as positive (1) and those below the threshold as negative (0). Construct the ROC curve and calculate the Area under the ROC curve (AUC).

ID	Predicted Probability	True Label
1	0.899	1
2	0.827	0
3	0.772	1
4	0.734	1
5	0.660	0
6	0.546	1
7	0.412	0
8	0.360	0

Solution: To construct the ROC curve, firstly we have to construct the table with TPR and FPR at different thresholds.

Threshold	TP	FP	TN	FN	TPR	FPR
0.899	1	0	4	3	0.25	0.00
0.827	1	1	3	3	0.25	0.25
0.772	2	1	3	2	0.50	0.25
0.734	3	1	3	1	0.75	0.25
0.660	3	2	2	1	0.75	0.50
0.546	4	2	2	0	1.00	0.50
0.412	4	3	1	0	1.00	0.75
0.360	4	4	0	0	1.00	1.00

Here, TPR (True Positive Rate) or sensitivity is calculated as -

$$TPR = \frac{TP}{TP + FN}$$

and FPR (False Positive Rate) is calculated as -

$$FPR = \frac{FP}{FP + TN}$$

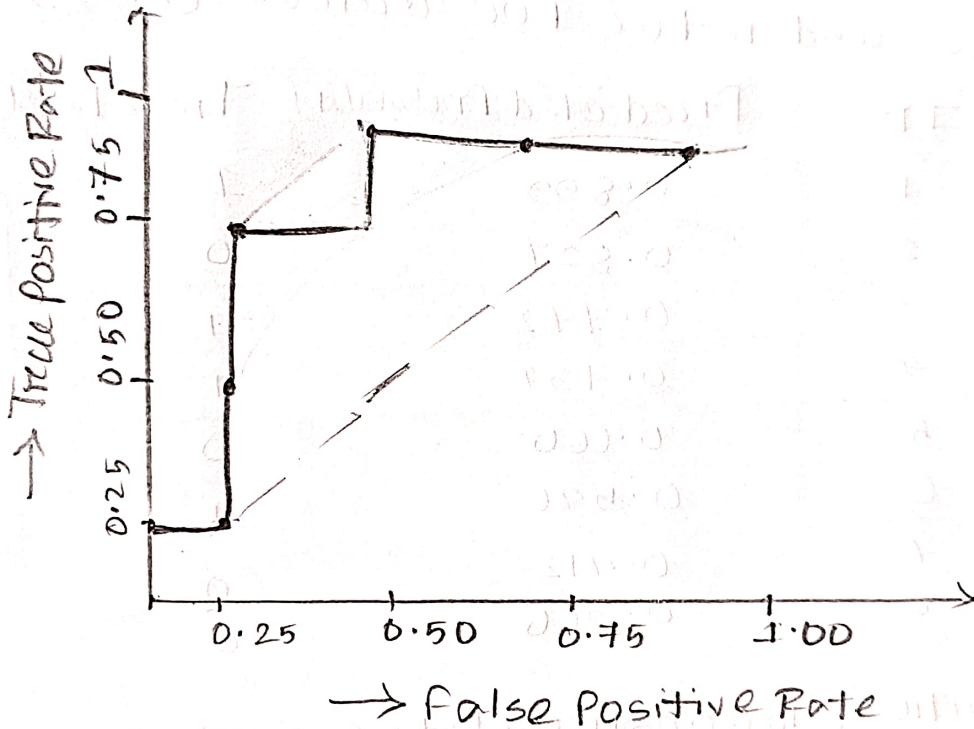


Figure: Receiver Operating Characteristic (ROC) curve

Now, we want to calculate the Area under curve,

$$\begin{aligned}
 AUC &= \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \times \frac{(TPR_{i+1} + TPR_i)}{2} \\
 &= 0.0625 + 0 + 0 + 0.1875 + 0 + 0.25 + 0.25 \\
 &= 0.75
 \end{aligned}$$

Question: 03: The final partition matrix for different values of K is obtained for the fuzzy C -means clustering. Find the optimal number of clusters to cluster the items A, B, C, D, E, F and G. Comment on your results.

Solution:

Item	K = 2		Row's Sum Total
A	0.045	0.055	0.89605
B	0.938	0.062	0.883688
C	0.191	0.809	0.690962
D	0.163	0.837	0.727138
E	0.507	0.403	0.518818
F	0.648	0.352	0.543808
G	0.214	0.786	0.663592
			4.924056

Fuzzy Partition Coefficient will be -

$$FPC = \frac{4.924056}{7} = 0.7034 \quad \left[\text{Here, number of data points} = 7 \right]$$

Item	K = 3			Row Square Total
A	0.026	0.967	0.007	0.935814
B	0.033	0.958	0.009	0.918934
C	0.094	0.004	0.002	0.988056
D	0.063	0.021	0.015	0.928035
E	0.340	0.464	0.196	0.369312
F	0.333	0.504	0.162	0.391149
G	0.004	0.002	0.005	0.000045
Total				5.521345

$$FPC = \frac{5.521345}{7} = 0.789$$

Item	K = 4				Row Square Total
A	0.000	1.000	0.000	0.000	1.000
B	0.000	1.000	0.000	0.000	1.000
C	0.990	0.003	0.002	0.005	0.980138
D	0.991	0.003	0.002	0.004	0.982128
E	0.001	0.002	0.001	0.006	0.002022
F	0.002	0.002	0.001	0.005	0.000034
G	0.000	0.000	1.000	0.000	1.000
Total					6.044304

$$FPC = \frac{6.044304}{7} = 0.8633$$