Name - Sradha Kedia
Exam Roll no. - 20234757053

Ans 5 Clustering feature (LS, SS, N)

LS = linear sum of N points

$$\sum_{i=1}^{N} X_i$$

N is total no. of data points

SS = square sum of N points

$$\sum_{i=1}^{N} X_i^2$$

$$\text{Diameter} = \left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} (\vec{x_i} - \vec{x_j})^2}{N(N-1)} \right)^{1/2} \quad \text{(given)}$$

$$= \left[\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} [\vec{x_i}^2 + \vec{x_j}^2 - 2\vec{x_i}\vec{x_j}]}{N(N-1)} \right]^{1/2}$$

$$= \left[\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \vec{x_i}^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} \vec{x_j}^2 - 2\sum_{i=1}^{N}\sum_{j=1}^{N} \vec{x_i}\vec{x_j}}{N(N-1)} \right]^{1/2}$$

$$= \left[\frac{\sum_{i=1}^{N} \vec{x_i}^2 + \sum_{j=1}^{N} \vec{x_j}^2 - 2\sum_{i=1}^{N} \vec{x_i} \sum_{j=1}^{N} \vec{x_j}}{N(N-1)} \right]^{1/2} = \left[\frac{SS + SS - 2\,LS\,LS}{N(N-1)} \right]^{1/2}$$

$$\left[\frac{2\,SS - 2(LS)^2}{N(N-1)}\right]^{1/2}$$

$$\therefore \text{Diameter} = \left[\frac{2\,(SS - (LS)^2)}{N(N-1)}\right]^{1/2}$$

Ans 2. $S = \{1, 3, 6, 10, 20, 100\}$

i) $K = 2$

initial seed values = 1 and 100

<u>iteration 1</u>: Choose minimum distance points in 1 cluster by checking Euclidean distance b/w points.

x	distance from 1	distance from 100
1	0	99
3	2	97
6	5	94
10	9	90
20	19	80
100	99	0

$\therefore K_1 = \{1, 3, 6, 10, 20\}$, $K_2 = \{100\}$

<u>iteration 2</u>: find mean of K_1 and K_2

$$m_1 = \frac{1+3+6+10+20}{5} = \frac{40}{5} = 8$$

$$m_2 = \frac{100}{1} = 100$$

x	distance from 8	distance from 100
1	7	99
3	5	97
6	2	94
10	2	90
20	12	80
100	92	0

$K_1 = \{1, 3, 6, 10, 20\}$, $K_2 = \{100\}$

(ii) K-medoid algorithm.

K=2, Initial medoids be 1 and 100

given data = {1, 3, 6, 10, 20, 100}

Soln. case 1:

x	distance from 1	distance from 100
3	2	97
6	5	94
10	9	90
20	19	80

So, Taking minimum from distance 1 and 100

cluster : $K_1 = \{1, 3, 6, 10, 20\}$, $K_2 = \{100\}$

Now, According to given scenario, mediod new value will be 3.

∴

x	distance from 3	distance from 100
1	2	99
6	3	94
10	7	90
20	17	80

Again, min. distance in 1 cluster

$\therefore K_1 = \{1, 3, 6, 10, 20\}$, $K_2 = \{100\}$

now, let's calculate cost for case 1 $= 2 + 5 + 9 + 19$

$= 35$

cost for case 2 $= 2 + 3 + 7 + 17 = 29$

cost in case 2 < cost in case 1 so, 3 can replace 1 in this case.

Ans 4: False. Overfitting is an inherent characteristics of decision tree and its occurrence depends on training data set also. In Overfitting, the learning system tightly fits the given training data so much that it would be inaccurate in predicting the outcomes of the untrained data. It is almost impossible before you test the data. It helps to address the inherent characteristics of overfitting which is inability to generalize data sets. In decision trees, over fitting occurs when the the tree is designed so as to perfectly fit all samples in the training dataset. But it is wrong to say that it occurence does not depend on training data set.

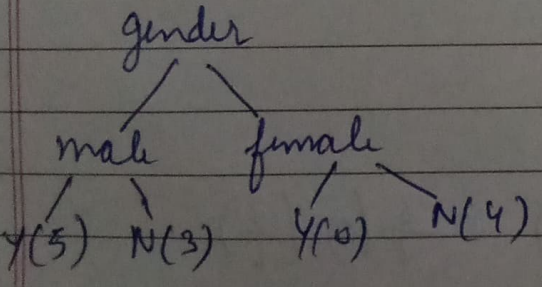
Ans 3. Moving representative points towards the cluster centroid helps in overcoming the effects of outliers. In the proposed approach, moving only a fixed distance towards the centroid would be less effective against outliers, since the distance between the outliers and centroid may be much larger than the distance b/w the other points and the centroid.

Ans 1. Total yes = 5, No = 7 (given)

$$E(S) = -P_1 \log_2(P_1) - P_2 \log_2(P_2)$$

$$= -\frac{5}{12} \log_2\left(\frac{5}{12}\right) - \frac{7}{12} \log_2\left(\frac{7}{12}\right)$$

$$= 0.979$$



$$E(\text{gender})_{\text{male}} = [5+, 3-]$$

$$= -\frac{5}{8} \log_2\left(\frac{5}{8}\right) - \frac{3}{8} \log\left(\frac{3}{8}\right)$$

$$= 0.95442$$

$$E(\text{gender female}) = [0+, 4-]$$

$$= -\frac{4}{4} \log_2\left(\frac{4}{4}\right) = 0$$

$$\text{gender gain} = E(S) - \frac{4}{12}(0.954) - \frac{8}{12}(0)$$

$$= 0.343$$

Date
Page

E(lecture I MCSC 201) = [3+, 3-]

$= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{3}\log_2\left(\frac{3}{6}\right)$

$= -\frac{1}{2}(-1) - \frac{1}{2}(-1) = 1$

E(lecture MSC 202) = [2+, 4-]

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

$= 0.918.$

lecture I gain $= E(S) - \frac{6}{12}(1) - \frac{6}{12}(0.918)$

$= 0.979 - 0.5 - \frac{1}{2}(0.918).$

$= 0.02$

E (lecture II MCSC 301)

$= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$

$= 1$

E (lecture II MCSC 302)

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

$= 0.918$

Date
Page

gain elective $= 0.979 - \frac{1}{2}(1) - \frac{1}{2}(0.918)$

$= 0.979 - 0.5 - \frac{1}{2}(0.918)$

$= 0.02$

$\therefore$ Best feature = best $(0.343, 0.02, 0.02)$

Gender

$= 0.343$, is best feature.

(ii) In the end, leaf nodes will end up with 4 or 5