Data Mining and Web Algorithms
15B22CI621

Tutorial 8

Q1: Consider the following data set with Shape, Color, Size and Edible2 (Class) attributes:

Color	Size	Shape	Edible?
Yellow	Small	Round	+
Yellow	Small	Round	-
Green	Small	Irregular	+
Green	Large	Irregular	-
Yellow	Large	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Green	Small	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Small	Irregular	+
Yellow	Large	Irregular	+

(a) What is Entropy of this collection of training data?
(b) What is Gain Ratio of an attribute (color)?

Q2. Differentiate between predictive and descriptive analysis. Clustering analysis falls under which category. Justify your answer.

Q 3: Consider the data consisting mixed attributes: Tax (Numeric), Marital Status (nominal) and Designation (ordinal)

ID	Tax Income	Marital Status	Designation	diabetes	Blood pressure
1	125000	Single	Dean	y	y
2	10000	Married	Professor	n	y
3	70000	Single	Associate Professor	n	n
4	120000	Married	Assistant Professor	y	n
5	95000	Married	Associate Professor	n	y
6	60000	Married	Assistant Professor	n	y

a) Apply z-score scaling on Tax income using mean absolute deviation and calculate the similarity between (3, 1).
b) Find the dissimilarity between (1,2) using only Tax Income,
c) Find the dissimilarity matrix using only Marital Status
d) Find the dissimilarity matrix using only Designation.
e) Find the dissimilarity between (1, 2) using diabetes and blood pressure.

Q4. Build the term-document matrix from given set of documents and calculate cosine similarity between document (1,3) and (3,4).

- **Doc 1 :** new Covid cases top forecasts
- **Doc 2:** **C**ovid cases rise in March
- **Doc 3:** increase in Covid cases in February
- **Doc 4:** March new Covid cases rise

Q5. Apply K-means clustering/K-Mediods using Manhattan distance on the following data using k =2. Choose initial centroids (1, 1), (3,4) and (5, 7) and do it for 2 iterations.

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Q6. Apply Agglomerative hierarchical clustering (Complete link as well as Single Link).

	A	B	C	D	E
A	0	1	2	7	5
B		0	3	8	6
C			0	5	9
D				0	4
E					0