C

APJ Abdul Kalam Technological University

Ernakulam II Cluster

First Semester M.Tech Degree Examination December 2017

05CS 6005– DATA MINING AND WAREHOUSING

Time: 3 hrs. Max. Marks: 60

I.

a) Explain normalization. Use suitable normalisation technique to normalize each of the given marks to the range (0,1). (6 Marks)

Mark
6
9
14
20

b) Analyse the Data Preprocessing in detail. (6 Marks)

II.

a) Build a decision tree classification model using the given training set. (8 Marks)

Sl. no	**age**	**income**	**student**	**Credit_rating**	**Class:buys_computer**
1	youth	high	no	Fair	no
2	youth	high	no	Excellent	no
3	middle_aged	high	no	Fair	yes
4	senior	medium	no	Fair	yes
5	senior	low	yes	Fair	yes

6	senior	low	yes	Excellent	no
7	middle_aged	low	yes	Excellent	yes
8	youth	medium	no	Fair	no
9	youth	low	yes	Fair	yes
10	senior	medium	yes	Fair	yes
11	youth	medium	yes	Excellent	yes
12	Middle_aged	medium	no	Excellent	yes
13	middle_aged	high	yes	Fair	yes
14	senior	medium	no	Excellent	no

b) Why is naïve Bayesian classification called "naïve"? Briefly outline the major ideas of naïve Bayesian classification. (4 Marks)

III.

a) Explain the 3 tier data warehouse architecture with neat diagram. (6 Marks)

b) Discuss the major steps in CLIQUE. How does CLIQUE use the Apriori principle to arrive at best clustering? (6 Marks)

c) Describe the working of PAM(partitioning around medoids) algorithm. (6 Marks)

OR

IV.

a) Give the K means clustering algorithm. Suppose that the data mining task is to cluster the following eight points (with (x; y) representing location) into three clusters.

A1(2; 10); A2(2; 5); A3(8; 4); B1(5; 8); B2(7; 5);

B3(6; 4); C1(1; 2); C2(4; 9):

The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show:

i. The three cluster centers after the first round of execution and

ii. The final three clusters (10 Marks)

b) Explain the working of hierarchical clustering methods. (8 Marks)

V.

a) Explain outliers and its types. Discuss the different applications of outlier analysis. (7 Marks)

b) Explain supervised, semi-supervised and unsupervised methods for outlier detection. (6 Marks)

c) Describe outlier detection using a histogram. (5 Marks)

OR

VI.

a) Explain parametric statistical method for outlier detection. (12 Marks)

b) Discuss CELL, the grid based method for distance based outlier detection. (6 Marks)