Hall Ticket Number:

Code No. : 41115

VASAVI COLLEGE OF ENGINEERING *(Autonomous)*, HYDERABAD

B.E. (C.S.E.) IV Year I-Semester Main Examinations, December-2017

Data Mining

Time: **3** hours Max. Marks: **70**

*Note: Answer **ALL** questions in **Part-A** and any **FIVE** from **Part-B***

Part-A (10 ×2 = 20 Marks)

1. Identify the type of attribute
   i) Hair color ii) grade iii) military rank iv) outcome of a medical test
2. Compute the Cosine similarity for the following document vectors
   D1= (2,2,0,0,0,3,0,0,1,0)
   D2= (0,1,0,0,0,1,0,2,0,0)
3. In a supermarket dataset the customer income attribute values are missing. How do you fill these missing values of this attribute?
4. List OLAP operations.
5. What is closed frequent itemset? Give an example
6. Find the relation between the attributes A and B by using the following contingency table

	A	A^-	
B	800	100	900
B^-	150	50	200
	950	150	1100

7. What is Entropy?
8. List methods to measure classifier accuracy.
9. Differentiate Supervised and Unsupervised Learning.
10. Illustrate one example where outliers are useful that regular data.

Part-B (5 × 10=50 Marks)
(All bits carry equal marks)

11. a) Describe the steps involved in data mining when viewed as a process of knowledge discovery.
    b) Consider the following matrix of TV viewing. Values indicate number of times a person watched the given show. Calculate Euclidean distance between Praful and the other viewers and order them from most similar to least similar

	CID	BigBoss	vahrehvah	Jabardasth	dumdam	Mahabharat
Praful	0	4	4	1	0	0
Vineet	1	2	2	1	0	3
Raja	0	0	1	1	1	0
Lahiri	5	6	1	3	5	0

12. a) Explain techniques to perform Data Reduction.
    b) Describe AOI algorithm for Class Characterization.

Contd... 2

13. a) A database has five transactions. Let **min_sup**=60% and **min_conf**=80%. Find all frequent itemsets using Apriori.

TID	Items brought
T100	{M,O,N,K,E,Y}
T200	{D,O,N,K,E,Y}
T300	{M,A,K,E}
T400	{M,U,C,K,Y}
T500	{C,O,O,K,I,E}

b) Illustrate how mining is done in multidimensional association rules.

14. a) Describe the major steps of naive bayesian classification. Find the classification label for X given as X= {A=T, B=F}

A	B	C	TARGET
T	T	T	YES
T	T	F	NO
T	F	T	YES
F	T	T	YES
F	T	F	NO
F	F	F	YES

b) Exemplify k nearest neighbor classifier.

15. a) We have a set of one dimensional points {6, 12, 18, 24, 30, 42, 48}. For two initial centroids {18, 45}, create two clusters by K-Means; then calculate SSE for the clustering result.

b) Explain the disadvantages of Partitioning based clustering algorithms. How DBSCAN algorithm is used to address those issues while forming clusters?

16. a) Summarize major issues in Data mining

b) Demonstrate different data normalization methods.
Suppose that the minimum and maximum values for the attributes income are 12,000 and 98,000 respectively. Map income to the range [0, 0, 1, 0]. By min-max normalization how a value of 73,600 for income is transformed?

17. Answer any ***two*** of the following:
    a) Explain mining multilevel associations.
    b) Describe the steps in Decision tree induction algorithm
    c) What is outlier? Give a brief explanation about types of outlier.

§§§§§