Register Number: ……

M.E. DEGREE EXAMINATIONS: NOV/DEC 2010

Third Semester

COMPUTER SCIENCE & ENGINEERING

CSE587: Data Mining

Time: Three Hours **Maximum Marks: 100**

Answer ALL Questions:-

PART A (10 x 2 = 20 Marks)

1. What is Data Mining? List out any two major challenges that motivated the development of Data Mining.
2. Define Precision.
3. List out any four methods used to evaluate the performance of a classifier.
4. Distinguish between rule based ordering and class based ordering.
5. What is meant by Closed Frequent Itemset?
6. Define confidence.
7. What is meant by cluster analysis? List out the various types of clusters.
8. List out the strengths and weaknesses of K-means clustering.
9. What are SOMs?
10. What is DENCLUE? What are its strengths?

PART B (5 X 16 = 80 Marks)

11. a) (i) Explain the major categories of data mining tasks. (8)

(ii) Given x = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0) (8)

y = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)

Calculate the SMC, Jaccard Coefficient and Cosine Similarity.

OR

b) What is the need for data preprocessing? Explain the various data preprocessing strategies in detail.

12. a) (i) Explain the various techniques used for visualizing higher dimensional data. (6)

(ii) With an algorithm explain the working of Nearest Neighbor Classifier. (10)

OR

b) (i) What is a data cube? How is it created? Explain the operations performed on data cubes. (10)

(ii) Write and explain the Decision Tree Induction algorithm. (6)

13. a) (i)Explain how frequent itemsets are generated in the Apriori algorithm. (8)

(ii) What are the different methods used for applying association analysis to continuous data? Explain. (8)

OR

b) (i) Explain the FP-Growth algorithm for frequent itemset generation with a suitable example. (10)

(ii) Write a short note on compact representation of frequent itemsets. (6)

14. a) (i) Explain in detail the Agglomerative hierarchical clustering algorithm. (10)

(ii) Perform Single link and Complete link clustering using the Euclidean distance matrix given .Show your results by drawing a dendrogram. (6)

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
P1	0.00	0.24	0.22	0.37	0.34	0.23
P2	0.24	0.00	0.15	0.20	0.14	0.25
P3	0.22	0.15	0.00	0.15	0.28	0.11
P4	0.37	0.20	0.15	0.00	0.29	0.22
P5	0.34	0.14	0.28	0.29	0.00	0.39
P6	0.23	0.25	0.11	0.22	0.39	0.00

OR

b) (i) Explain in detail the K-means clustering technique. (12)

(ii) Write a brief note on cluster validation. (4)

15. a) (i) Explain the various characteristics of data that can affect the cluster analysis. (8)

(ii) Discuss about the various anomaly detection techniques. (8)

OR

b) Explain the following:

(i) CLIQUE (8)

(ii) Grid Based Clustering (8)

****