USN □□□□□□□□□□

P E S INSTITUTE OF TECHNOLOGY CS 363

(AN AUTONOMOUS INSTITUTE UNDER VTU, BELGAUM)

Seventh Semester End Examination(SEE) B.E Degree August 2011
(Session June-August 2011)

CS 363 DATA MINING

Time: 3 Hrs **Max. Marks : 100**

Note: All the Questions Are Compulsory

1 a Explain whether each of the following is a data mining task. — 4

(i) Dividing the customers according to their gender.
(ii) Predicting the outcome of tossing a pair of dice.
(iii) Predicting the future stock price of a company using historical records.
(iv) Monitoring the heart beat of a patient for abnormalities

1 b Describe **data characterization** and **data discrimination**. How can you apply these concepts to the data on car sales in India given in the table below? — 4

Car Manufacturer	July 2011 Sales Number of units	July 2010 Sales Number of units	Change %
.Maruti	66.504	90,134	-26
Hyundai	25,462	28,811	-11
Tata	17, 192	28,865	-38
Mahindra	17,312	12,725	35
Toyota	13,192	6,834	99
GM	9,508	7,125	33
VW	6,529	2597	151
Ford	7,504	8,729	-14
Skoda	2,412	1,222	97
Honda	4,725	4,685	1

What additional information will be needed to do further analysis to discover interesting inferences or knowledge? Please explain.

1 c. A market research firm collected data on consumers' car buying behavior and a data mining system was deployed to analyze the data and following association rules were found. — 4

age(X, "40-60") ^ annual income(X,"2 Lakhs…5 Lakhs") => buys(X,"Maruti") (support=>30%, confidence=70%)
age(X,"30-40) ^ annual income(X,"5 Lakhs..15Lakhs") => buys(X,"Huyndai") (support=20%, confidence=50%)

Explain the above rules in your own words (plain English).

1 d Describe data classification models and its uses. Name any two forms using which a data classification model can be represented. Explain with an example. — 4

1 e What is Data pre-processing? Why is it important? What steps does it include? Briefly describe each step with an example. — 4

2 a Describe the architecture of a typical data ware housing/data mining system with a neat diagram. — 4

Define Data Warehouse, Multidimensional Data Model, Data Cube and Data Mart. How does Data Warehouse differ from an operational Database.?

2 b A data warehouse can be modeled either by a star schema or a snowflake schema. Briefly describe the similarities and differences between these two models and analyze their advantages or disadvantages with regard to one another. Give your opinion of which might be empirically more useful and explain the reasons behind your answer. 4

2 c Suppose that a data warehouse consists of the three dimensions *time*, *doctor* and *patient* and the two measures *count and charge*, where *charge* is the fee that a doctor charges a patient for a visit. 4

(a) Enumerate three classes of schemas that are popularly used for modeling data warehouses.

(b) Draw a schema diagram for the above data warehouse using one of the schema classes listed in (a).

(c) Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in a particular year, say 2008?).

2 d What is descriptive data mining? How is it different than predictive data mining? What is data generalization? Is it descriptive or predictive? 2

2 e Often the aggregate measure value of many cells in a large data Cuboid is zero, resulting in a huge, yet sparse, multi-dimensional matrix. 4

(a) Design an implementation method that can elegantly overcome this sparse matrix problem. Note that you need to explain your data structures in detail and discuss the space needed, as well as how to retrieve the data from your structures.

(b) Modify the design in (a) to handle incremental data updates. Give the reasoning behind your new design.

3 a. A database has 5 transactions. Let min_sup=60% and min_conf=80%. Find all frequent item sets using Apriori algorithm. 4

TID	Date	Items bought
T100	05/05/2011	(M,O,N,K,E,Y)
T200	05/05/2011	(D,O,N,K,E,Y)
T300	19/05/2011	(M,A,K,E)
T400	22/06/2011	(M, U, C,K, Y)
T500	01/07/2011	(C,O,O,K,I,E)

3 b What are the techniques used to improve the efficiency of Apriori algorithm? Give brief description of each of these techniques. 6

3 c For the data given in question 3 a, find all frequent item sets using FP-growth algorithm. Compare the efficiency of FP-Growth Algorithm with Apriori algorithm. 5

3 d Using a neat diagram and an example, describe concept hierarchy and multi-level mining. In which situations, multi-level mining is particularly useful. Differentiate multi-level mining with uniform minimum support, reduced minimum support and item or group-based minimum support. 5

4 a Briefly outline the major steps in a decision tree classification. What are the stopping conditions for the decision tree? Give an example and draw a neat diagram of the decision tree corresponding to that. 6

4 b Given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules or (b) pruning the decision tree and then converting the pruned tree to rules. What disadvantages does (b) have over (a)? 2

4 c Why naïve Bayesian classification is called "*naïve*"? Briefly outline the major steps of Bayesian classification. 5

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day 1	Sunny	Hot	High	Weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 3	Overcast	Hot	High	Weak	Yes
Day 4	Rain	Mild	High	Weak	Yes
Day 5	Rain	Cool	Normal	Weak	Yes
Day 6	Rain	Cool	Normal	Strong	No
Day 7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day 9	Sunny	Cool	Normal	Weak	Yes
Day 10	Rain	Mild	Normal	Weak	Yes

	The table above provides training date for Bayesian classification system. Using the training data, predict the value of "Play Tennis", when Outlook="Sunny", Temperature="Cool", Humidity="High", and Wind="Strong".	5
4 e	Compare the advantages and disadvantages of eager classification(e.g. decision tree, Bayesian, neural network) versus lazy classification (e.g. , k-nearest neighbor, case based reasoning).	2
5 a	Briefly describe the following approaches to clustering: Partitioning Methods, Hierarchical Methods, Density Based Methods, Model Based Methods and Constraint based Methods.	10
5 b	Present conditions under which Density based clustering is superior to hierarchical and partition methods of clustering. Give some sample data sets to support your argument.	2
5 c	Why is outlier mining important? Explain in brief different types of outlier detection.	4
5 d	Where do you think data mining applications can be useful in real life? Give examples and state intended benefits.	4