Midterm Take-Home Exam Questions

Due date: Friday March 2nd 2007 at 11:59pm.

Your answers should be handed in printable format. No hand written answers would be accepted. Keep your answers concise and to the point. Do not exceed one typed page on average per question. Figures, if needed, can be added in appendix. **Remember, this is individual work.**

Question 1: (10 points)
A telecommunication company is designing a data warehouse containing information about their customers and their telephone calls. The data warehouse consists of four dimensions: *customer*, *time*, *location* and *destination* representing respectively the customer information, the date when the call is recorded, the location from which the call was initiated, and the location to which the call was destined. The warehouse also contains three measures: a *count* for the number of calls, an *amount* for the total dollars spent, and *time* the total numbers of minutes. Draw a star schema diagram for this data warehouse.

Question 2: (20 points)
We discussed in class some clustering algorithms. There are also many papers about other clustering algorithms or related to the problem of clustering. Compare the different methods CLARANS, Rock and DBSCAN and then present the major differences in a table.

Question 3: (20 points)
We discussed in class different methods to learn a classification model. In a table highlight the advantages and disadvantages of Neural networks, Naïve Bayesian Classifiers, k-Nearest Neighbor Classifiers and Associative Classifiers. Add to your table Support Vector Machines.

Question 4: (20 points)
Suppose I have sequences of symbols belonging to 2 different classes and I want to build a classifier using lazy learning with k-nearest neighbors, I would need a distance measure. The sequences are of variable lengths. What kind or distance functions would you propose? What are the advantages and disadvantages. Notice that distance can be applied to sequences directly or to features extracted from the sequences.

Question 5: (30 points)
The FP-growth idea is presented in the paper: Mining Frequent Patterns without Candidate Generation , J. Han, J. Pei, and Y. Yin, ACM-SIGMOD 2000, Dallas, May 2000. FP-growth, like the apriori algorithm, considers the presence of individual items in transactions. If items reoccur in a same transaction, they are not taken into account, but the flag for that item in the transaction is set to 1. We call these binary transactions. However, there are cases where we want to consider the reoccurrence of items. Discuss a method, and possibly present an algorithm, that enhances the FP-tree idea with recurring items. For an example on mining itemsets with reoccurring items, see the paper: Mining Recurrent Items in Multimedia with Progressive Resolution Refinement, Osmar R. Zaïane, Jiawei Han, Hua Zhu Conf. on Data Engineering (ICDE'2000), pp. 461-470, San Diego, CA, February, 2000 http://www.cs.ualberta.ca/~zaiane/postscript/icde00.pdf