CS590D Spring 2006 (midterm) exam (partial) solutions, March 9, 2006

Prof. Chris Clifton

Time will be tight. If you spend more than the recommended time on any question, **go on to the next one**. If you can't answer it in the recommended time, you are either going in to too much detail or the question is material you don't know well. You can skip one or two parts and still demonstrate what I believe to be an A-level understanding of the material.

Note: It is okay to use abbreviations in your answers, as long as the abbreviations are unambiguous and reasonably obvious.

1 Association Rules (25 minutes, 6 points)

You have run the a-priori algorithm to find association rules in a grocery store transaction database. It takes an unexpectedly long time to complete. On completion, the following is one (of many) rules:

<milk, butter, cheese, bread, flour, sugar, salt, chocolate, apples> ⇒ vanilla

1.1 Execution time (10 minutes, 3 points)

Based on seeing the above rule, you should be able to make a good guess as to why the algorithm took a long time. Explain why.

The key problem is with the a-priori principle itself: All subsets of a frequent itemset are frequent. Given an itemset with 10 items, the number of possible subsets is very large. Each becomes a candidates, and is checked. While the number of passes through the data is small (10), the cost of each pass can become high.

Scoring: 1 for understanding a-priori, 1 for idea, 1 for good reason.

1.2 FP-growth (10 minutes, 3 points)

Would the FP-growth algorithm likely do better, or would it have problems as well? Explain.

Scoring: 1 for understanding FP-growth, 1 for idea, 1 for good reason.

2 Classification (5 minutes, 3 points)

Would Naïve Bayes be effective for developing a classifier for the following data set? Why or why not?

Temp.	*Season*	*Electricity use*
Below Average	Winter	High
Above Average	Winter	Low
Below Average	Summer	Low
Above Average	Summer	High

(*Electricity use* is the class to be predicted.)

Naïve Bayes would not be good for this problem, as it assumes that the impact of attributes on the class are independent. It is clear that it is the combination of low temperatures and winter or high temperatures and summer that lead to the high class.

Scoring: One for showing you know how classification works, one for understanding Naïve Bayes/correct answer, one for explanation.

3 Clustering (30 minutes, 8 points)

For this question, you will perform k-means clustering of the following data set.

Pressure	*Temperature*
145	9
152	11
57	21
54	20
147	32
144	33
82	31

3.1 Distance (2 minutes, 1 point)

State the distance function you will be using. Feel free to choose one that is easy to calculate, even if it is not the most appropriate for the problem.

I would use Manhattan distance: $|p_1 - p_2| + |t_1 - t_2|$*, as it is easy to compute.*

Scoring: 1 for showing you know what a distance function is.

3.2 k-means Clustering (20 minutes, 5 points)

Demonstrate k-means clustering of the data, with $k = 3$. You are allowed to make some approximations and take short-cuts, as long as you demonstrate that you know the steps of the algorithm and what it is likely to produce. Use as much space as necessary.

Scoring: 1 for k initial, 1 for assignment, 1 for calculating means, 1 for repeat, 1 for termination.

3.3 Preprocessing (5 minutes, 2 points)

Name a preprocessing step that would be particularly appropriate for this dataset, and would likely give different results. Explain why.

Normalization. Since the variability in Pressure is much greater than Temperature, it will dominate the impact. Scaling to a 0-1 range would equalize the impact of the two attributes.

Scoring: 1 for step, 1 for reason why.