T-61.5060: Algorithmic Methods of Data Mining

Course Instructor: Panagiotis Papapetrou

Final Exam

January 12, 2012

Instructions:

You have **three (3)** hours to complete this exam. You are allowed to use one **two-sided cheat-sheet** (A4 page, both sides hand-written), which you have to submit together with the exam paper. No additional material can be used. The total score that can be obtained is **50 points**. As described in the course requirements, you need to score **at least 25/50 points** to pass this exam.

Question 1 (Frequent itemsets and association rules) **[10 points]**

(a) The Apriori algorithm uses prior knowledge of subset support. Given a frequent itemset y and a subset x of y, prove that the confidence of rule "$x \Rightarrow (l - x)$" cannot be more that the confidence of rule "$y \Rightarrow (l - y)$", where l is a superset of both x and y. [5 points]

(b) A partitioning version of Apriori groups the transactions of a database D into n non-overlapping partitions. A partition may contain at least one transaction and each transaction should be contained in one partition. Prove that any itemset that is frequent in D must be frequent in at least one partition in D. [5 points]

Question 2 (Clustering) **[10 points]**

(a) Two common clustering problems are the K-means and the K-median problems. Explain what is the main difference between the two problems, name an algorithm that solves each problem and identify at least one advantage and at least one disadvantage of each algorithm. [5 points]

(b) One technique for speeding up clustering algorithms is to sample uniformly from the available data points. For example, consider an input consisting of n data points and a clustering algorithm that requires time $O(n^2)$. Then, the running time of the algorithm in a sample consisting of $\sqrt{n}$ data points would require time $O(n)$. Then, clusters can be extracted for the whole dataset simply by assigning the remaining points (those that are not in the sample) to their closest cluster (e.g., the cluster with the closest representative). For each of the following types of data discuss briefly if (1) sampling may cause problems for this approach and (2) what these problems are. Assume that the sampling technique randomly chooses points from the total set of n points. Finally assume, that the number of clusters k required for clustering is much smaller than n:

1. Data with clusters of different sizes (that includes very small clusters and very big clusters).
2. Data with outliers, i.e., very atypical points.
3. Data with a small number of noise points. [5 points]

Question 3 (Ranking) **[10 points]**

Consider N web-pages interlinked among each other. PageRank is computed for these web-pages. A spammer adds his/her web page rolexx.com to this set of N web-pages and wants rolexx.com to be ranked as high as

possible. The spammer is allowed to add a maximum of k web-pages including rolexx.com. That creates a total of N+k pages in the web-page graph.

(a) Given that k is fixed, how should the pages be linked in order to maximize the PageRank value of rolexx.com? Assume that the spammer can create new links between any pair of pages in the whole graph (including the new pages that have been added), but cannot remove existing links. [5 points]

(b) How should the above linking structure change if the spammer is additionally allowed to remove at most p pages from the existing graph? [5 points]

Question 4 (Social networks) **[10 points]**

(a) Consider the influence maximization problem in a social network. For each the following cases, which of the two common models (linear threshold or independent cascade) would you use and why? Note that you do not have to explain how you are going to use the model.

1. Find the best 100 people to vaccinate in order to minimize the spread of a deadly virus.
2. Find in polynomial time the tweeter user whose message will be propagated to as many users as possible.

[5 points]

(b) Bob claims to have hacked Facebook and has managed to obtain the complete social graph information of Facebook. Mary, who works for Google, becomes quite interested in this and offers Bob 1 million euros for his graph. However, before making the deal she demands to have a look at the data. Bob refuses to give her access to the graph itself, but instead he sends her a list that contains, for each node in the graph, the number of incoming and outgoing links. After making a simple plot of this list, Mary immediately becomes suspicious and rejects the offer. Why did Mary change her mind when she saw the list that Bob sent her? [5 points]

Question 5 (Time Series) **[10 points]**

(a) Dynamic time warping (DTW) violates the triangle inequality. What is the main reason that causes this violation? Show an example to support your argument. [3 points]

(b) Consider a large 1-dimensional time series X of n points and a query sequence Q of m points with $|Q|<<|X|$ (the size of Q is much smaller than the size of X). The problem you want to solve is the following: you want to find the best subsequence Y of X that matches Q. Also, you are given the additional information that the matching subsequence you are looking for is of length equal to the query length, i.e., $|Y| = |Q|$. Using the *LB_Keogh* lower-bound and the DTW distance measure (employing the Sakoe-Chiba band with parameter r):

1. Describe a sliding window approach to solve this problem. [3 points]
2. What is the best and worst case runtime complexity of this approach in terms of n, m, and r? [2 points]
3. Based on the above complexity it should be clear that your algorithm performs much worse than SPRING. In what setting would both methods (yours and SPRING) have similar worst-case time complexity? [2 points]