T-61.5060: Algorithmic Methods of Data Mining

Course Instructor: Panagiotis Papapetrou

Final Exam

March 10, 2011

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** in this exam in order to have the remaining course points (assignments, project, and quizzes) count towards your final grade.

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

(a) Assume a transaction database *D* where the possible items that can occur in *D* are A, B, or C. Suppose that we have mined all frequent closed itemsets in *D* with the minimum support count threshold *min_sup = 3*. These itemsets are {A, B, C} with support count = 3, {A} with support count = 5, and {A, B} with support count = 4. Using only this information, infer the remaining frequent itemsets in *D* and their support values. [5 points]

(b) A partitioning version of Apriori divides the transactions of a database *D* into *n* non-overlapping partitions not necessarily of equal size. By the pigeon-hole principle it holds that any itemset that is frequent in *D* must be frequent in at least one partition in *D*. Prove this using the definition of the support measure. [5 points]

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

(a) What is the difference between K-means and K-medoids? [5 points]

(b) As we discussed in class it is possible that for the same data, K-means and Agglomerative clustering (using, e.g., single-linkage) may produce different clusters. Discuss a scenario where such result may be possible. [5 points]

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

John was given a dataset *D* of N records and M attributes. Also, he was given three classification models (classifiers) M1, M2, and M3. A model can, for example, be a naïve Bayes classifier. He was then asked to decide which classifier is best for that dataset. To solve this problem, John took the whole dataset *D* and trained each of the three models. After the models were trained, John tested their classification accuracy on the same dataset *D* and concluded that all three models are close to perfect achieving almost 100% accuracy. At the same time, Mary followed a different methodology in finding the best classifier and concluded that M2 is the best for this dataset.

(a) Comment on John's computation. Is his result reasonable? Did he do anything wrong? [5 points]

(b) Describe what could be the methodology that Mary used to solve this problem. [5 points]

[10 points]

Question 4 (Ranking)

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. [4 points]

(b) How should the above linking structure change if the spammer is additionally allowed to remove 1 page from the existing graph? [4 points]

(c) Would such an approach be possible when the hubs and authorities algorithm is used? Why? [2 points]

[10 points]

Question 5 (Time Series)

(a) In class we saw that Dynamic Time Warping (DTW) does not satisfy the metric property. Specifically, it violates the triangle inequality. Provide an intuitive explanation on what causes this violation. [5 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). 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 within a factor of *r* of the query length, i.e., |Q|-r ≤ |Y| ≤ |Q|+r, where r is the parameter of the Sakoe-Chiba band. 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 time complexity of this approach in terms of *n*, *m*, *k*, and r? [2 points]