T-61.5060: Algorithmic Methods of Data Mining

Course Instructor: Panagiotis Papapetrou

Final Exam

December 15, 2010

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) 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 divides the transactions of a database D into n non-overlapping partitions. 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 K-means and K-median. 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 (Classification) **[10 points]**

(a) Suppose you are given a dataset D of N records and M attributes. Also, you are given ten classification models (classifiers) M1, M2, ..., M10. A model can, for example, be a naïve Bayes classifier. Describe briefly what process you would follow in order to choose the classifier that works best for your data. For this process you should include **all** of the following tools: k-fold cross-validation, confusion table, and ROC curve. [5 points]

T-61.5060: Algorithmic Methods of Data Mining

Course Instructor: Panagiotis Papapetrou

Final Exam

December 15, 2010

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) 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 divides the transactions of a database D into n non-overlapping partitions. 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 K-means and K-median. 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 (Classification) **[10 points]**

(a) Suppose you are given a dataset D of N records and M attributes. Also, you are given ten classification models (classifiers) M1, M2, ..., M10. A model can, for example, be a naïve Bayes classifier. Describe briefly what process you would follow in order to choose the classifier that works best for your data. For this process you should include **all** of the following tools: k-fold cross-validation, confusion table, and ROC curve. [5 points]