T-61.5060 ALGORITHMIC METHODS OF DATA MINING

EXAMINATION

T1, 21 December 2007 at 9–12.

To pass the course you must also pass the exercise project. Results of this examination are valid for one year after the examination date.

This examination has six problems. To get full points you must complete all problems. You can answer in Finnish, Swedish or English. Please write clearly and leave a wide (left or right) margin. You can use a calculator.

The results will be posted to the notice board and also emailed to an address of form `12345X@students.hut.fi`, where `12345X` is your student number, on 21 January 2008, at latest.

You can keep this paper.

1. Describe informally the count-min data structure and the types of problems it can be used for.

2. Recall the notation for sets of variables: $ABC = \{A, B, C\}$ etc. Frequent sets: (a) Suppose ABC, BCE, ABE, ABD, BCD, ACE, ADE, BDE are frequent. Which sets of size 4 can be frequent? (b) Consider a 0-1 dataset over the attributes A, B, and C. Suppose $f(A) = 0.5, f(B) = 0.4, f(C) = 0.6, f(AB) = 0.35$. What can be said about the frequencies of AC, BC, and ABC?

3. Borders: (a) Define the concepts of negative and positive border for frequent sets. (b) Consider data consisting of strings, and the class of substring patterns. I.e., $D = \{w_1, \ldots, w_n\}$, where each w_i is a string in some alphabet Σ. A pattern p is also a string in the same alphabet, and p occurs in w_i if w_i can be written as xpy for some x and y. A string pattern p is frequent if p occurs in sufficiently many strings w_i. Assume $\Sigma = \{a, b, c, d\}$ and that the positive border of the frequent strings is $\{abcd\}$. What is the negative border?

4. Define the concept of a covering problem and describe the greedy approximation algorithm for this problem. What can be said about the quality of the approximations produced by the greedy algorithm?

5. Define the clustering aggregation problem and describe one algorithm for it.

6. Describe Kleinberg's impossibility result for clustering. No proof is needed.