Uppsala University
Department of Information Technology
Kjell Orsborn

Final Exam 2007-12-14
DATA MINING - 1DL105, 1DL025

```
Date ........................................... Friday, Aug 12, 2007
Time ....................................................... 8:00-13:00
Teacher on duty ....... Kjell Orsborn, phone 471 11 54 or 070 425 06 91
```

Instructions:

Read through the complete exam and note any unclear directives before you start solving the questions.

The following guidelines hold:

- Write readably and clearly! Answers that cannot be read can obviously not result in any points and unclear formulations can be misunderstood.
- Assumptions outside of what is stated in the question must be explained. Any assumptions made should not alter the given question.
- Write your answer on only one side of the paper and use a new paper for each new question to simplify the correction process and to avoid possible misunderstandings. Please write your name on each page you hand in. When you are finished, please staple these pages together in an order that corresponds to the order of the questions.
- NOTE! This examination contains **40** points for 1DL105 (6hp) and **48** points for 1DL025 (7,5hp) in total and their distribution between sub-questions is clearly identifiable. Note that you will get **credit only for answers that are correct**. To pass, you must score at least **22** and **28** respectively. The examiner reserves the right to lower these numbers.
- You are allowed to use dictionaries to and from English, a calculator, and the one A4 single-sided paper with your hand-written notes that you have brought with you, but **no other material**.

1. **Data in data mining:** 8 pts

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity (Example: Age in years. Answer: Discrete, quantitative, ratio).

(a) Brightness as measured by a light meter.
Answer: Continuous, quantitative, ratio

(b) Angles as measured in degrees between 0 ° and 360 °.
Answer: Continuous, quantitative, ratio

(c) Bronze, Silver, and Gold medals as awarded at the Olympics.
Answer: Discrete, qualitative, ordinal

(d) Number of patients in a hospital.
Answer: Discrete, quantitative, ratio

(e) Ability to pass light in terms of the following values: opaque, translucent, transparent.
Answer: Discrete, qualitative, ordinal

(f) Military rank. Answer: Discrete, qualitative, ordinal

(g) Density of a substance in grams per cubic centimeter.
Answer: Discrete, quantitative, ratio

(h) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

Answer: Discrete, qualitative, nominal

2. **Classification:** 8 pts

After a data mining course the results of the exam was recorded along with some data about the students. The results can be found in the table below. (GPA is the Grade Point Average.)

ID	Phone number	Language	Passed all assignments	GPA	Passed exam
1	555 - 3452	Java	No	3.1	Yes
2	555 - 6294	Java	No	2.0	No
3	555 - 9385	C++	Yes	3.5	Yes
4	555 - 9387	Python	Yes	2.5	Yes
5	555 - 9284	Java	Yes	3.9	No
6	555 - 0293	C++	No	2.9	No
7	555 - 9237	Java	No	1.9	No
8	555 - 3737	Python	Yes	3.2	Yes

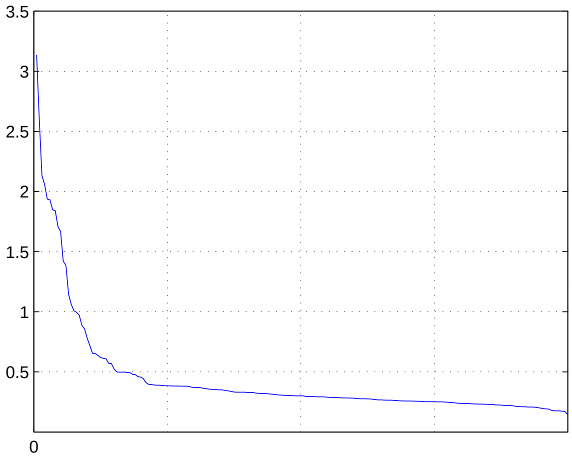
(a) In no more than one page of text, describe the design of a K-Nearest Neighbor classifier to predict if a student will fail or pass the exam. (6pts)

(b) Use your K-NN classifier to predict whether the following student (who overslept and missed the original exam) will pass the re-exam. (2pts)

ID	Phone number	Language	Passed all assignments	GPA	Passed exam
9	555 - 6295	C++	Yes	3.0	?

3. **Clustering:** 8 pts

   (a) The pictures shows a 5-dist graph for a data set.



When we run dbscan on the data with $MinPts = 5$ and $\epsilon = 0.5$

i. Which is the largest region in the 5-dist graph that contains only core points? (1pt)

ii. Which is the largest region in the 5-dist graph that contains only noise points? (2pts)

iii. Does the regions above cover all points? If not, what points can appear in the remaining region? (1pt)

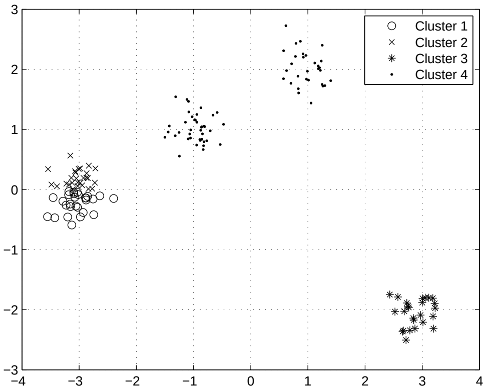
(b) In the figures below two bad clusterings based on K-means is shown. What is the main reason for the bad results, and what can be done to address the problems.

i. In figure (a) (2pts)

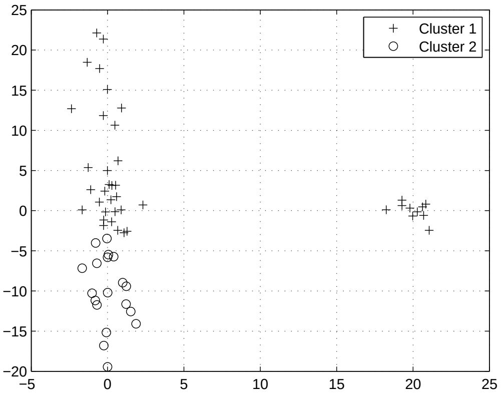
ii. In figure (b) (2pts)

4. **Association rules:** 6 pts

   (a) After mining a transaction database for large itemsets (i.e., itemsets with enough support), there is only one large itemset of size 8. Let N be the total number of large itemsets (including the one of size 8). What is the minimal value of N? (2pts)



(a)



(b)

(b) The $F_{k-1} \times F_{k-1}$ method is a method for generating apriori candidates.

i. Describe the method. (2pts)

ii. With a short motivation, does this method eliminate the need for candidate pruning? (1pt)

(c) For each of the following, say if the statement is always true, always false or sometimes true and sometimes false. S is the support and C is the confidence. No motivation is needed. **(0.5 p for correct answer, -0.5 points for wrong answer and 0 points for no answer. You can never get negative result on the whole question.)** (3pts)

i. $S(a \rightarrow b) \geq S(a \rightarrow b, c)$

ii. $C(a \rightarrow b) \geq C(a \rightarrow b, c)$

iii. $S(a \rightarrow b) = S(b \rightarrow a)$

iv. $C(a \rightarrow b) = C(b \rightarrow a)$

v. $C(a \rightarrow b, c) > C(a, b \rightarrow c)$

vi. $Min(C(a \rightarrow b), C(b \rightarrow c)) > C(a \rightarrow c)$

5. Sequence mining: 8 pts

Consider the following customer sequence dataset:

Sequence ID	Sequence
1	< {1 5}{2}{3}{4} >
2	< {1}{3}{4}{3 5} >
3	< {1}{2}{3}{4} >
4	< {1}{3}{5} >
5	< {4}{5} >

(a) Apply the GSP algorithm to the dataset in the table using minimum support s = 33% to determine all large sequences. (6pts)

Answer:

1-itemsets	sup count
< {1} >	4
< {2} >	2
< {3} >	4
< {4} >	4
< {5} >	4

2-itemsets	sup count
~~<{1}, {1}>~~	~~0~~
<{1}, {2}>	2
<{1}, {3}>	4
<{1}, {4}>	3
<{1}, {5}>	2
~~<{2}, {1}>~~	~~0~~
~~<{2}, {2}>~~	~~0~~
<{2}, {3}>	2
<{2}, {4}>	2
~~<{2}, {5}>~~	~~0~~
~~<{3}, {1}>~~	~~0~~
~~<{3}, {2}>~~	~~0~~
~~<{3}, {3}>~~	~~1~~
<{3}, {4}>	3
<{3}, {5}>	2
~~<{4}, {1}>~~	~~0~~
~~<{4}, {2}>~~	~~0~~
~~<{4}, {3}>~~	~~1~~
~~<{4}, {4}>~~	~~0~~
<{4}, {5}>	2
~~<{5}, {1}>~~	~~0~~
~~<{5}, {2}>~~	~~1~~
~~<{5}, {3}>~~	~~1~~
~~<{5}, {4}>~~	~~1~~
~~<{5}, {5}>~~	~~0~~

~~<{1, 2}>~~	~~0~~
~~<{1, 3}>~~	~~0~~
~~<{1, 4}>~~	~~0~~
~~<{1, 5}>~~	~~1~~
~~<{2, 3}>~~	~~0~~
~~<{2, 4}>~~	~~0~~
~~<{2, 5}>~~	~~0~~
~~<{3, 4}>~~	~~0~~
~~<{3, 5}>~~	~~1~~
~~<{4, 5}>~~	~~0~~

3-itemsets	sup count
<{1}, {2}, {3}>	2
<{1}, {2}, {4}>	2
<{1}, {3}, {4}>	3
<{1}, {3}, {5}>	2
~~<{1}, {4}, {5}>~~	~~1~~
<{2}, {3}, {4}>	2
~~<{2}, {3}, {5}>~~	(pruned)
~~<{2}, {4}, {5}>~~	(pruned)
~~<{3}, {4}, {5}>~~	~~1~~

4-itemsets	sup count
<{1}, {2}, {3}, {4}>	2

(b) Identify the maximal sequence patterns. (2pts)

Answer:

```
1-itemsets sup count
< {1}, {2}, {3}, {4} >, < {1}, {3}, {5} >, < {4}, {5} >
```

NOTE! You SHOULD only solve the following problem 6 if you fulfil the following conditions:

You are a student registered to course 1DL025 - 7.5hp (but not to 1DL105 - 6hp) and you have NOT handed in AND passed the extra assignment on assignment 3, regarding direct hashing.

If you have handed in and passed the extra assignment on assignment 3 you will automatically get 8 points for this exam problem below and SHOULD NOT hand in any solution to problem 6.

---

6. **Direct hashing and pruning:** 8 pts

For the transaction database below, perform the following steps of the apriori algorithm with direct hashing and transaction pruning.

1. Count the itemsets of size 1.

2. Prune unsupported itemsets.

3. Perform transaction pruning.

4. Perform direct hashing for itemsets of size 2.

Minimum support is 0.5. Use the hash function $\left(\sum_i x_i i^2\right) \mathit{mod}\ 4$, where x_i is one element of an itemset and $x_{i+1} > x_i$.

Which candidate itemsets can be pruned based on the hash table? Measured in the total count in the hash table, how much work do you save by using the pruned transaction table, compared to the non-pruned version?

Transactions
1 3 6
1 2 3 6
1 4 5
2 3 6
1 3 5
2 3 4 5 6
1 2
1 2 4 6

Good Luck and a Merry, Merry Christmas!

/ Kjell