

# Final Exam 2006-12-15

## DATA MINING - 1DL105, 1DL111

Date ..... Friday, Dec 15, 2006  
Time ..... 15:00-20:00  
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.
- This examination contains **40** points 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**. To get VG, you must score at least **30**. 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 paper with notes that you have brought with you, but **no other material**.

1. **Classification:**

8 pts

- (a) In text of not more than two pages, present the main ideas (perhaps in the form of pseudocode) of the K-Nearest Neighbor (KNN) technique for classification. **(3pts)**
- (b) What is the complexity of the KNN algorithm as a function of the number of elements in the training set ( $q$ ), and the number of elements ( $n$ ) to be classified? **(1pt)**
- (c) Discuss issues that are important to consider when employing a Decision Tree-based classification algorithm. **(2pts)**
- (d) What are the main advantages and disadvantages of Decision Tree classification algorithms? **(2pts)**

2. **Evaluation measures in Rule-based classification:**

8 pts

- (a) In rule-based-classification, consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules,

R1 : A  $\rightarrow$  + (covers 4 positive and 1 negative examples),  
R2 : B  $\rightarrow$  + (covers 30 positive and 10 negative examples),  
R3 : C  $\rightarrow$  + (covers 100 positive and 90 negative examples),

determine which is the best and worst candidate rule according to:

- i. Rule accuracy. **(3pts)**
  - ii. FOIL's information gain:  $p_1 \times (\log_2(p_1/(p_1 + n_1)) - \log_2(p_0/(p_0 + n_0)))$ . **(3pts)**
- (b) Explain why rule accuracy and FOIL's information gain rank the rules differently. **(2pts)**

3. **Association patterns evaluation and sequential patterns:**

8 pts

- (a) In the context of association analysis, classify the following objective interestingness measures as either symmetric or asymmetric measures and state if they are invariant under inversion, row-column scaling and null-addition operations. **(2pts)**
  - i. support
  - ii. confidence
- (b) To discover sequential patterns, the Generalized sequential pattern (GSP) algorithm can be applied that include the following steps:
  - step 1: Make the first pass over the sequence database to yield all the 1-element frequent sequences
  - step 2: Repeat until no new frequent sequences are found
    - Candidate generation
    - Candidate pruning
    - Support Counting
    - Candidate Elimination

Assume that we we have the following frequent 3-sequences:

- < {a} {b} {c} >
- < {a} {b e} >
- < {a} {e} {c} >
- < {b} {c} {d} >
- < {b e} {c} >
- < {c} {d} {e} >
- < {e} {c d} >

You should perform the candidate generation and pruning steps by:

- i. Generating valid 4-sequence candidates by merging appropriate 3-sequences. Also explain how you generated these sequences. **(3pts)**
- ii. Pruning the candidate set of 4-sequences ending up with a set of pruned candidate 4-sequences. Also explain how this pruning step was carried out. **(3pts)**

**4. Clustering:**

8 pts

In the figure below we see the sitting arrangement of students A..O in a lecture hall. The lecturer clusters the students based on where they are sitting in the classroom using the DBSCAN algorithm. He is using the manhattan distance measure, an EPS of 2.1 and MinPts of 4 (*Note that the MinPts value includes the point itself*). The lecturer finds that there are two clusters.

1				A					
2		B		C					D
3				E	F		G		
4				H	I				
5									
6	J								
7		K	L		M	N			
8									
9			O						
	1	2	3	4	5	6	7	8	9

- (a) Which students are core points? **(1 p)**
- (b) Which students are border points? **(1 p)**
- (c) Which students are directly density reachable from student I? **(1 p)**
- (d) Which students are directly density reachable from student M? **(1 p)**
- (e) Student P enters the room. He feels connected to both clusters of students. Where should he sit if he wants to belong to both clusters, but he does not want the clusters to merge into one? **(2 pts)**
- (f) Where should Student P sit if he wants there to be only one cluster? **(2 pts)**

**5. Association Rules:**

8 pts

- (a) Perform the Apriori algorithm for the given transaction database, using the minimum support 0.3 and the minimum confidence 0.77. Note that you have to show how the algorithm is performed, it is not enough to simply state the end result. **(4 pts)**

Row	Transaction
1	{a b}
2	{b c}
3	{a b c}
4	{d e f}
5	{a b c}
6	{d f}
7	{c d e f}
8	{a b c d e}

- (b) Assuming we have a rule  $I_1 \rightarrow I_2$ . In not more than a few sentences describe how to *interpret* the situation when the rule has: **(2 pts)**
- Low support and high confidence.
  - High support and low confidence.
- (c) Assume that the rule  $\{1\ 2\} \rightarrow \{3\ 4\}$  is in the final set of rules, and the rule  $\{3\ 4\} \rightarrow \{1\ 2\}$  is *not* in the final set. For each of the following rules, state if the rule definitely appears in the final set, if there is a possibility that it appears in the final set, or if it definitely does not appear in the final set of rules. **(2 pts)**
- i.  $\{1\ 2\ 3\} \rightarrow \{4\}$
  - ii.  $\{1\} \rightarrow \{2\ 3\ 4\}$
  - iii.  $\{2\ 3\ 4\} \rightarrow \{1\}$
  - iv.  $\{3\} \rightarrow \{1\ 2\ 4\}$

Good luck and a Merry, Merry Christmas!

/ Kjell