Uppsala University
Department of Information Technology
Kjell Orsborn

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 subquestions 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 $\longrightarrow$ + (covers 4 positive and 1 negative examples),
   R2 : B $\longrightarrow$ + (covers 30 positive and 10 negative examples),
   R3 : C $\longrightarrow$ + (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 arrangment 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

Answers to the Exam in Data Mining 2006-12-15

1. **Classification**

   (a) See any textbook on the subject.

   (b) $O(qn)$

   (c)
   - How to split nodes (binary split, multiway split)
   - How to evaluate how good splits are (GINI-meassure, entropy)
   - Stopping conditions.

   (d)
   - **Pros:**
     - Fast classification, O(depth of tree).
     - Easy to interpret.
     - Inexpensive to construct.
     - ...
   - **Cons:**
     - Difficult to construct the optimal decision tree.
     - Works poorly for some data since we are splitting on one characteristic at the time, which leads to rectangular classification borders.
     - ...

---

2. **Evaluation meassures in rule-based classification**

   (a) i. The accuracies of the rules are 80% (for R1), 75% (for R2), and 52.6% (for R3), respectively. Therefore R1 is the best candidate and R3 is the worst candidate according to rule accuracy.

   ii. Assume the initial rule is $\emptyset \rightarrow +$. This rule covers $p_0 = 100$ positive examples and $n_0 = 400$ negative examples.
   - R1 covers $p_1 = 4$ positive examples and $n_1 = 1$ negative example. Therefore, the FOIL's information gain for this rule is

$$4 * \left(log_2 \left(\frac{4}{5} \right) - log_2 \left(\frac{100}{500} \right) \right) = 8$$

- R2 covers $p_1 = 30$ positive examples and $n_1 = 10$ negative example. Therefore, the FOIL's information gain for this rule is

$$30 * \left(log_2 \left(\frac{30}{40} \right) - log_2 \left(\frac{100}{500} \right) \right) = 57.2$$

- R3 covers $p_1 = 100$ positive examples and $n_1 = 90$ negative examples. Therefore, the FOIL's information gain for this rule is

$$100 * \left(log_2 \left(\frac{100}{190} \right) - log_2 \left(\frac{100}{500} \right) \right) = 139.6$$

R3 is the best candidate and R1 is the worst candidate.

(b) Rule accuracy is only concerned with the accuracy of the rule when it is applied. FOIL's information gain also takes into account how often the rule can be applied, and how much better it is than the default rule.

---

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

(a) Support:

- Symmetric
- Not invariant under inversion
- Not invariant under row-column scaling
- Not invariant under null-addition

Confidence:

- Asymmetric
- Not invariant under inversion
- Not invariant under row-column scaling
- Invariant under null-addition

(b) i. <{a} {b} {c} {d}>
<{a} {b e} {c}>
<{a} {e} {c d}>
<{b} {c} {d} {e}>
<{b e} {c d}>

ii. <{a} {b e} {c}>

---

4. **Clustering**

(a) Core Points: C, E, F, H, I, L

(b) Border Points: A, B, G, K, O, M

(c) Directly Density Reachable from I: E, F, H

(d) Directly Density Reachable from M: None (M is not a core point)

(e) (Row, Column) = (5, 3) P then becomes a border point for both clusters.

(f) (6, 4), (5, 5), (6,5). P becomes a core point, reachable from both clusters.

---

5. **Association rules**

(a) MinSuppCount = ceil$(0.3 * 8) = 3$
MinConf = 0.77

C1	**L1**	**C2**	**L2**	**C3**	**C'3**	**L3**
{a}	{a}	{a b}	{a b}	{a b c}	{a b c}	{a b c}
{b}	{b}	{a c}	{a c}	{d e f}		
{c}	{c}	{a d}	{b c}			
{d}	{d}	{a e}	{d e}			
{e}	{e}	{a f}	{d f}			
{f}	{f}	{b c}				
		{b d}				
		{b e}				
		{b f}				
		{c d}				
		{c e}				
		{c f}				
		{d e}				
		{d f}				
		{e f}				

Final rules
{a c}→{b}
{a}→{b}
{b}→{a}
{b}→{c}
{c}→{b}
{e}→{d}
{f}→{d}

(b) **Low support and high confidence** $I_1 \cup I_2$ is seldom bought, but when I_1 is bought we know that there is a high probability that I_2 is also bought. The high confidence tells us that I_1 is relatively uncommon, and if I_2 is also uncommon we have a strong rule, however seldom applicable.
Example: {Expensive beer} → {plastic bag} is a pretty uninteresting rule, but {Ipod} → {Special Ipod headphones} is more interesting.

High support and low confidence $I_1 \cup I_2$ is relatively often bought together, but since I_1 is bought even more often, we cannot say for sure that somebody

interested in I_1 will also be interested in I_2. These rules are only interesting if the other rules also have a low confidence.
Example: {plastic bag} $\rightarrow$ {Cheap beer}.

(c) Given facts and relations between support of subsets:

$$\frac{S_{\{1234\}}}{S_{\{12\}}} \geq C_{\texttt{min}} \quad S_{\{123\}} \leq S_{\{12\}} \leq S_{\{1\}}$$
$$\frac{S_{\{1234\}}}{S_{\{34\}}} < C_{\texttt{min}} \quad S_{\{234\}} \leq S_{\{34\}} \leq S_{\{3\}}$$

i. Will definitely appear in the final set
ii. Might appear in the final set
iii. Might appear in the final set
iv. Will definitely not appear in the final set