Uppsala University
Department of Information Technology
Kjell Orsborn

Final Exam 2007-08-16
DATA MINING - 1DL105, 1DL111

```
Date ..................................... Thursday, Aug 16, 2007
Time ..................................................... 9:00-14:00
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. **Data mining:** 8 pts

Discuss (shortly) whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their profitability.
Answer: No. This is an accounting calculation, followed by the application of a threshold. However, predicting the profitability of a new customer would be data mining.

(b) Computing the total sales of a company.
Answer: No. Again, this is simple accounting.

(c) Sorting a student database based on student identification numbers.
Answer: No. Again, this is a simple database query.

(d) Predicting the outcomes of tossing a (fair) pair of dice.
Answer: No. Since the die is fair, this is a probability calculation. If the die were not fair, and we needed to estimate the probabilities of each outcome from the data, then this is more like the problems considered by data mining. However, in this specific case, solutions to this problem were developed by mathematicians a long time ago, and thus, we wouldnÕt consider it to be data mining.

(e) Predicting the future stock price of a company using historical records.
Answer: Yes. We would attempt to create a model that can predict the continuous value of the stock price. This is an example of the area of data mining known as predictive modelling. We could use regression for this modelling, although researchers in many fields have developed a wide variety of techniques for predicting time series.

(f) Monitoring the heart rate of a patient for abnormalities.
Answer: Yes. We would build a model of the normal behavior of heart rate and raise an alarm when an unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection. This could also be considered as a classification problem if we had examples of both normal and abnormal heart behavior.

(g) Monitoring seismic waves for earthquake activities.
Answer: Yes. In this case, we would build a model of different types of seismic wave behavior associated with earthquake activities and raise an alarm when one of these different types of seismic activity was observed. This is an example of the area of data mining known as classification.

(h) Extracting the frequencies of a sound wave.
Answer: No. This is signal processing.

2. **Classification:** 10 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? (2pts)
Answer: O(qn) with q elements in the training set and n element to classify, i.e. comparing each element to be classified with each element in the training set.

(c) Discuss issues that are important to consider when employing a Decision Tree based classification algorithm. (3pts)
Answer: choosing splitting attributes, ordering of splitting attributes, splits, tree structure, stopping criteria, training data and pruning.

(d) What are the main advantages and disadvantages of Decision Tree classification algorithms? (2pts)

Answer: easy to use and efficient, rules generated that are easy to interpret and understand, scale well, trees can be constructed for data with many attributes

Do not handle continous data well, division of domain space into rectangular region not suitable for all classification problems, handling missing data difficult, overfitting may occur, correlations among attributes are ignored

3. **Clustering:** 8 pts

(a) Suppose you want to cluster the eight points shown below using k-means.

	A_1	A_2
x_1	2	10
x_2	2	5
x_3	8	4
x_4	5	8
x_5	7	5
x_6	6	4
x_7	1	2
x_8	4	9

Assume that $k = 3$ and that initially the points are assigned to clusters as follows: $C_1 = \{x_1, x_2, x_3\}$, $C_2 = \{x_4, x_5, x_6\}$, $C_3 = \{x_7, x_8\}$. Apply the k-means algorithm until convergence (i.e., until the clusters do not change), using the Manhattan distance. (Hint: the Manhattan distance of point x_1 from the centroids of C_1, C_2, and C_3 in the initial clustering assignment is $5\frac{2}{3}$, $8\frac{1}{3}$, and 5, respectively.) Make sure you clearly identify the final clustering and show your steps. (4pts)

(b) Consider the set of points given in Figure 1. Assume that $eps = \sqrt{2}$ and $minpts = 3$ (including the center point). Using Euclidian Distance find all the density-based clusters in the figure using the DBSCAN algorithm. List the final clusters (with the points in lexicographic order, i.e., from A to J) and outliers. (4pts)

Answer: Clusters are C,F,G and D,E,I, Outliers A, B, H, J (A,B is a cluster with minpts = 2)

4. **Association rules:** 6 pts

Consider the market basket transactions shown in Table 1. Assume that $minsup > 0$.

(a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)? (1pt)

Answer: There are six items in the data set. Therefore the total number of rules is 602.

(b) What is the maximum size of frequent itemsets that can be extracted? (1pt)

Answer: Because the longest transaction contains 4 items, the maximum size of frequent itemset is 4.

(c) What is the maximum number of size-3 itemsets that can be derived from this data set. (An expression to calculate this number is also a valid answer to this sub-question.) (2pts)

Answer: $\binom{6}{3} = 20$.

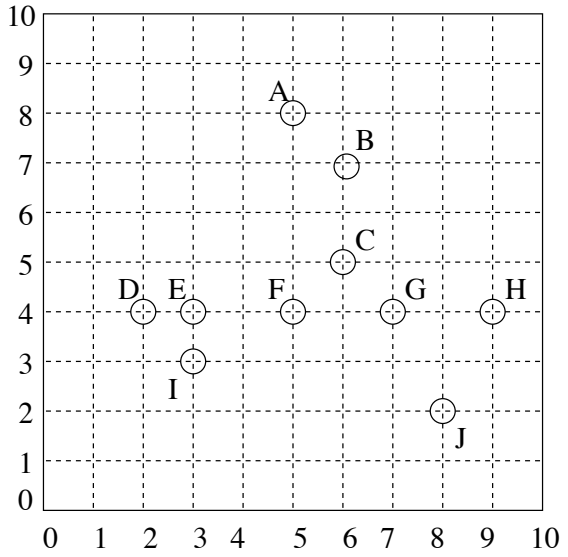

Figure 1: Points for question 3b.

Table 1: Market basket transactions for Question 4.

Transaction ID	Items Bought
1	{ Milk, Beer, Diapers }
2	{ Bread, Butter, Milk }
3	{ Milk, Diapers, Cookies }
4	{ Bread, Butter, Cookies }
5	{ Beer, Cookies, Diapers }
6	{ Milk, Diapers, Bread, Butter }
7	{ Bread, Butter, Diapers }
8	{ Beer, Diapers }
9	{ Milk, Diapers, Bread, Butter }
10	{ Beer, Cookies }

(d) Find an itemset (of size 2 or larger) that has the largest support. (1pt)
Answer: Bread, Butter.

(e) Find a pair of items, a and b, such that the rules $\{a\} \rightarrow \{b\}$ and $\{b\} \rightarrow \{a\}$ have the same confidence. Any such pair will do. (1pt)

Answer: (Beer, Cookies) or (Bread, Butter).

5. **Sequence mining:** 8 pts

Consider the following frequent 3-sequences:

$< \{1,2,3\} >$, $< \{1,2\}\{3\} >$, $< \{1\}\{2,3\} >$, $< \{1,2\}\{4\} >$, $< \{1,3\}\{4\} >$, $< \{1,2,4\} >$, $< \{2,3\}\{3\} >$, $< \{2,3\}\{4\} >$, $< \{2\}\{3\}\{3\} >$, and $< \{2\}\{3\}\{4\} >$.

(a) List all the candidate 4-sequences produced by the candidate generation step of the Generalized Sequential Patterns (GSP) algorithm. (3pts)
Answer: $< \{1,2,3\}\{3\} >$, $< \{1,2,3\}\{4\} >$, $< \{1,2\}\{3\}\{3\} >$,$< \{1,2\}\{3\}\{4\} >$, $< \{1\}\{2,3\}\{3\} >$, $< \{1\}\{2,3\}\{4\} >$.

(b) List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming no timing constraints). (2pts)

Answer: When there are no timing constraints, all subsequences of a candidate must be frequent. Therefore, the pruned candidates are: $< \{1,2,3\}\{3\} >$, $< \{1,2\}\{3\}\{3\} >$, $< \{1,2\}\{3\}\{4\} >$, $< \{1\}\{2,3\}\{3\} >$, $< \{1\}\{2,3\}\{4\} >$.

(c) List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming $maxgap = 1$). (3pts)

Answer: With timing constraint, only contiguous subsequences of a candidate must be frequent. Therefore, the pruned candidates are: $< \{1,2,3\}\{3\} >$, $< \{1,2\}\{3\}\{3\} >$,$< \{1,2\}\{3\}\{4\} >$, $< \{1\}\{2,3\}\{3\} >$, $< \{1\}\{2,3\}\{4\} >$.

Good luck!

/ Kjell