Uppsala University
Department of Information Technology
Kjell Orsborn

Final Exam 2007-04-16
DATA MINING - 1DL105, 1DL111

```
Date ..................................... Monday, Apr 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. **Similarity and distance measures:** 2 pts

   For binary data, the L1 distance corresponds to the Hamming distance. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors (2pts):

   ```
   x = 0101010001
   y = 0100011000
   ```

   Answer:

   Hamming distance = number of different bits = 3

   Jaccard Similarity = number of 1-1 matches /(number of bits - number 0-0 matches) = 2 / 5 = 0.4

2. **Classification:** 10 pts

   Consider the training examples shown in Table 1 for a binary classification problem.

   Table 1: Data set for binary classification problem.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

   (a) Compute the Gini index ($Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$) for the overall collection of training examples. (1pts)

   Answer: $Gini = 1 - 2 \times 0.5^2 = 0.5$.

   (b) Compute the Gini index for the Customer ID attribute. (1pts)

   Answer: The gini for each Customer ID value is 0. Therefore, the overall gini for Customer ID is 0.

   (c) Compute the Gini index for the Gender attribute. (2pts)

   Answer: The gini for Male is $1 - 2 \times 0.5^2 = 0.5$. The gini for Female is also 0.5. Therefore, the overall gini for Gender is $0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$.

(d) Compute the Gini index for the Car Type attribute using multiway split. (2pts)

Answer: The gini for Family car is 0.375, Sports car is 0, and Luxury car is 0.2188. The overall gini is 0.1625.

(e) Compute the Gini index for the Shirt Size attribute using multiway split. (2pts)

Answer: The gini for Small shirt size is 0.48, Medium shirt size is 0.4898, Large shirt size is 0.5, and Extra Large shirt size is 0.5. The overall gini for Shirt Size attribute is 0.4914.

(f) Which attribute is better, Gender, Car Type, or Shirt Size? (1pts)

Answer: Car Type because it has the lowest gini among the three attributes.

(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini. (1pts)

Answer: The attribute has no predictive power since new customers are assigned to new Customer IDs.

3. **Clustering:** 10 pts

Hierarchical clustering is sometimes used to generate K clusters, $K > 1$ by taking the clusters at the K^{th} level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.

The following is a set of one-dimensional points: $\{6, 12, 18, 24, 30, 42, 48\}$.

(a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids. (2pts)

i. centroids are 18 and 45

ii. centroids are 15 and 40

(b) Do both sets of centroids represent stable solutions; i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated? (2pts)

(c) What are the two clusters produced by single link? (2pts)

(d) Which technique, K-means or single link, seems to produce the "most natural" clustering in this situation? (For K-means, take the clustering with the lowest squared error.) (1pt)

(e) What definition(s) of clustering does this natural clustering correspond to? (Well-separated, center-based, contiguous, or density.) (1pt)

(f) What well-known characteristic of the K-means algorithm explains the previous behavior? (2pts)

4. **Association rules:** 6 pts

Consider the following set of frequent 3-itemsets: $\{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 3, 4\}, \{1, 3, 5\}, \{2, 3, 4\}, \{2, 3, 5\}, \{3, 4, 5\}$. Assume that there are only five items in the data set.

(a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

(b) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.

(c) List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

5. **Association rules:** 6 pts

Suppose we have market basket data consisting of 100 transactions and 20 items. If the support for item a is 25%, the support for item b is 90% and the support for itemset $\{a, b\}$ is 20%. Let the support and confidence thresholds be 10% and 60%, respectively.

(a) Compute the confidence of the association rule $\{a\} \rightarrow \{b\}$. Is the rule interesting according to the confidence measure? (2pts)

(b) Compute the interest measure for the association pattern $\{a, b\}$. Describe the nature of the relationship between item a and item b in terms of the interest measure. (2pts)

(c) What conclusions can you draw from the results of parts (a) and (b)? (2pts)

6. **Sequence mining:** 6 pts

Find all the frequent subsequences with *support* $\geqslant$ 50% given the sequence database shown in Table 2. Assume that there are no timing constraints (such as mingap, maxgap, maxspan, etc.) imposed on the sequences.

Table 2: Example of event sequences generated by various sensors.

Sensor	Timestamp	Events
S1	1	A, B
	2	C
	3	D, E
	4	C
S2	1	A, B
	2	C, D
	3	E
S3	1	B
	2	A
	3	B
	4	D, E
S4	1	C
	2	D, E
	3	C
	4	E
S5	1	B
	2	A
	3	B, C
	4	A, D

Good luck!

/ Kjell