Uppsala University
Department of Information Technology
Kjell Orsborn

# Final Exam 2012-10-17
# DATA MINING I - 1DL360

```
Date ..................................... Wednesday, October 17, 2012
Time ..................................................... 08:00-13:00
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 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**. The examiner reserves the right to lower these numbers.
- You are allowed to use dictionaries to and from English, a calculator but **no other material**.

1. **Data mining:** 6 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) 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 wouldnt consider it to be data mining.

(c) 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.

(d) 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.

(e) Extracting the frequencies of a sound wave.

Answer: No. This is signal processing.

(f) Monitoring and predicting failures in a hydropower plant.

Answer: Yes. In this case, ...

2. **Classification:** 8 pts

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

Table 1:

| Instance | $a_1$ | $a_2$ | $a_3$ | Target Class |
|---|---|---|---|---|
| 1 | T | T | 1.0 | + |
| 2 | T | T | 6.0 | + |
| 3 | T | F | 5.0 | - |
| 4 | F | F | 4.0 | + |
| 5 | F | T | 7.0 | - |
| 6 | F | T | 3.0 | - |
| 7 | F | F | 8.0 | - |
| 8 | T | F | 7.0 | + |
| 9 | F | T | 5.0 | - |

(a) The entropy is given by: $Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$, where $c$ is the number of classes and $p(j|t)$ is the relative frequency of class $j$ at node $t$. What is the entropy of this collection of training examples with respect to the positive class? (1 pt)

Answer:
The entropy is given by: $Entropy(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$, where $c$ is the number of classes and $p(j|t)$ is the relative frequency of class $j$ at node $t$.
There are four positive examples and ve negative examples. Thus, $P(+) = 4/9$ and $P(-) = 5/9$. The entropy of the training examples is $-4/9 \log_2(4/9) - 5/9 \log_2(5/9) = 0.9911$.

(b) What are the information gains of $a_1$ and $a_2$ relative to these training examples? (3 pts)

Answer: For attribute $a_1$, the corresponding counts and probabilities are:

| $a_1$ | + | - |
|---|---|---|
| T | 3 | 1 |
| F | 1 | 4 |

The entropy for $a_1$ is:
$\frac{4}{9}\left[-(3/4)\log_2(3/4) - (1/4)\log_2(1/4)\right]$
$+\frac{5}{9}\left[-(1/5)\log_2(1/5) - (4/5)\log_2(4/5)\right] = 0.7616$.
Therefore, the information gain for $a_1$ is $0.9911 - 0.7616 = 0.2294$.

For attribute $a_2$, the corresponding counts and probabilities are:

| $a_2$ | + | - |
|---|---|---|
| T | 2 | 3 |
| F | 2 | 2 |

The entropy for $a_2$ is:
$\frac{5}{9}\left[-(2/5)\log_2(2/5) - (3/5)\log_2(3/5)\right] +$
$\frac{4}{9}\left[-(2/4)\log_2(2/4) - (2/4)\log_2(2/4)\right] = 0.9839$.
Therefore, the information gain for $a_2$ is $0.9911 - 0.9839 = 0.0072$.

(c) For $a_3$, which is a continuous attribute, compute the information gain for every possible split. (3 pts)

Answer:

| $a_3$ | Class label | Split point | Entropy | Info Gain |
|---|---|---|---|---|
| 1.0 | + | 2.0 | 0.8484 | 0.1427 |
| 3.0 | - | 3.5 | 0.9885 | 0.0026 |
| 4.0 | + | 4.5 | 0.9183 | 0.0728 |
| 5.0 | - | | | |
| 5.0 | - | 5.5 | 0.9839 | 0.0072 |
| 6.0 | + | 6.5 | 0.9728 | 0.0183 |
| 7.0 | + | | | |
| 7.0 | - | 7.5 | 0.8889 | 0.1022 |

The best split for aoccurs at split point equals to 2.

(d) What is the best split (among $a_1$ , $a_2$, and $a_3$) according to the information gain? (1 pt)

Answer:
According to information gain, $a_1$, produces the best split.

3. **Clustering:** 8 pts

(a) Explain the definition of a core point in DBSCAN. (1 pt)

(b) Explain the definition of a centroid in k-means. (1 pt)

(c) The following is a set of one-dimensional points: $\{1, 1, 2, 3, 5, 8, 13, 21, 33, 54\}$. Perform two iterations of k-means on these points using the two initial centroids 0 and 11. (2 pts)

(d) Assume that you have to explore a large data set of high dimensionality. You know nothing about the distribution of the data. In text of no more than one page, discuss the following. (4 pts)

i. How can k-means and DBSCAN be used to find the number of clusters in that data?

ii. Explain how PCA can help find the dimensions where clusters separate.

iii. Explain why PCA might neglect cluster separation in some dimensions.

iv. Can k-means or DBSCAN be applied in a way that would help you find the dimensions in which the clusters separate?

Answer:

(a) core point in DBSCAN: a point surrounded by at least MinPts points in its Eps neighborhood.

(b) centroid: The average of all points in the cluster.

(c) **Practical k-means** Compute distances between each data point and each init centroid.

$1, 1, 2, 3, 5$ belongs to $c1$. $8, 13, 21, 33, 54$ belongs to $c2$.

New $c1 = 12/5 = 2.4$. New $c2 = 129/5 = 25.8$.

Compute distances between each data point and each init centroid.

$1, 1, 2, 3, 5, 8, 13$ belongs to $c1$. $21, 33, 54$ belongs to $c2$.

New $c1 = 33/7 = 4.7$. New $c2 = 108/3 = 36$.

(d) **Explanations** I would like to see at least the following:

i. Try k-means with increasing $k$. For which $k$ does MSE start to converge? Try DBSCAN with different parameter settings.

ii. PCA finds the dimensions with high variance. We will find the clusters that separate along those dimensions.

iii. If clusters separate in a dimension with low variance, PCA will overlook that separation.

iv. Perform k-means and DBSCAN with different subsets of the dimensions. If we find a set of dimensions d that don't separate any clusters, try to re-run the clustering on all dimensions but d. If we still get the same number of clusters, d are probably not significant.

4. **Association analysis:** 6 pts

Consider the data set shown in Table 2.

| Customer ID | Transaction ID | Items bought |
|---|---|---|
| 1 | 1001 | {i1, i4, i5} |
| 1 | 1024 | {i1, i2, i3, i5} |
| 2 | 1012 | {i1, i2, i4, i5} |
| 2 | 1031 | {i1, i3, i4, i5} |
| 3 | 1015 | {i2, i3, i5} |
| 3 | 1022 | {i2, i4, i5} |
| 4 | 1029 | {i3, i4} |
| 4 | 1040 | {i1, i2, i3} |
| 5 | 1033 | {i1, i4, i5} |
| 5 | 1038 | {i1, i2, i5} |

Table 2: Market basket transactions for question 4.

(a) Compute the support for itemsets $\{i5\}$, $\{i2, i4\}$, and $\{i2, i4, i5\}$ by treating each transaction ID as a market basket. (1 pt)

Answer: s( {i5}) = 8 / 10 = 0.8 s( {i2, i4}) = 2 / 10 = 0.2 s( {i2, i4, i5}) = 2 / 10 = 0.2

(b) Use the results in part 4a to compute the confidence for the association rules $\{i2, i4\} \longrightarrow \{i5\}$ and $\{i5\} \longrightarrow \{i2, i4\}$. (1 pt)

(c) Is confidence a symmetric measure? (1 pt)

Answer: $c(i2i4 \longrightarrow i5) = 0.2/0.2 = 100\%$ $c(i5 \longrightarrow i2i4) = 0.2/0.8 = 25\%$ No, condence is not a symmetric measure.

(d) Repeat part 4a by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.) (1 pt)

Answer: s( {i5}) = 4 / 5 = 0.8 s( {i2, i4}) = 5 / 5 = 1 s( {i2, i4, i5}) = 4 / 5 = 0.8

(e) Use the results in part 4d to compute the confidence for the association rules $\{i2, i4\} \longrightarrow \{i5\}$ and $\{i5\} \longrightarrow \{i2, i4\}$. (1 pt)

Answer: $c(i2i4 \longrightarrow i5) = 0.8/1 = 80\%$ $c(i5 \longrightarrow i2i4) = 0.8/0.8 = 100\%$

(f) Suppose $s_1$ and $c_1$ are the support and confidence values of an association rule $r$ when treating each transaction ID as a market basket. Also, let $s_2$ and $c_2$ be the support and confidence values of $r$ when treating each customer ID as a market basket. Discuss whether there are any relationships between $s_1$ and $s_2$ or $c_1$ and $c_2$. (1 pt)

Answer: There are no apparent relationships between s1 , s2 , c1 , and c2 .

5. **Association analysis and hash trees:** 6 pts

The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets.

Consider the following set of candidate 3-itemsets:

{1, 2, 3}, {1, 2, 5}, {1, 2, 6}, {1, 3, 4}, {2, 3, 4}, {2, 4, 5}, {2, 4, 6}, {2, 4, 7}, {3, 4, 6}, {4, 5, 6}

(a) Construct a hash tree for the above candidate 3-itemsets. Assume the tree uses a hash function where all odd-numbered items are hashed to the left child of a node, while the even-numbered items are hashed to the right child. A candidate k-itemset is inserted into the tree by hashing on each successive item in the candidate and then following the appropriate branch of the tree according to the hash value. Once a leaf node is reached, the candidate is inserted based on one of the following conditions:

**Condition 1**: If the depth of the leaf node is equal to k (the root is assumed to be at depth 0), then the candidate is inserted regardless of the number of itemsets already stored at the node.

**Condition 2**: If the depth of the leaf node is less than k, then the candidate can be inserted as long as the number of itemsets stored at the node is less than maxsize. Assume maxsize = 2 for this question.

**Condition 3**: If the depth of the leaf node is less than k and the number of itemsets stored at the node is equal to maxsize, then the leaf node is converted into an internal node. New leaf nodes are created as children of the old leaf node. Candidate itemsets previously stored in the old leaf node are distributed to the children based on their hash values. The new candidate is also hashed to its appropriate leaf node. (3 pts)

Answer:

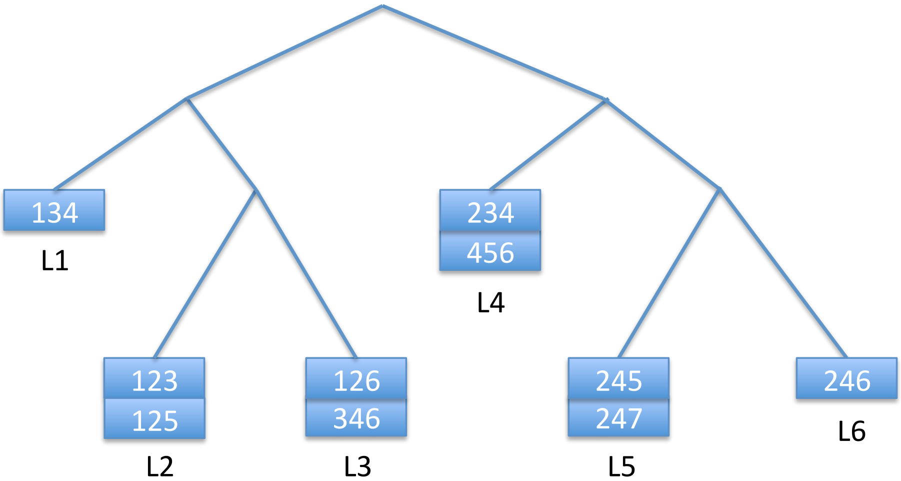

Figure 1: **Hash tree for problem 5**

(b) How many leaf nodes are there in the candidate hash tree? How many internal nodes are there? (1 pt)

Answer: There are 6 leaf nodes and 5 internal nodes.

(c) Consider a transaction that contains the following items:

{1, 2, 3, 5, 6}.

Using the hash tree constructed in part (a), which leaf nodes will be checked against the transaction? What are the candidate 3-itemsets contained in the transaction? (2 pts)

Answer: (to be updated!) The leaf nodes L1, L2, L3, and L4 will be checked against the transaction. The candidate itemsets contained in the transaction include 1,2,3 and 1,2,6.

6. **Data mining and privacy:** 6 pts

(a) Explain what deidentification of a data set in data mining is and its possible advantages or disadvantages. (2 pts).

Answer: Deidentification of a data object is the process to remove information of the data object that can reveal its identity and thereby also revealing potential private and sensible information. So the advantage is to protect privacy but this also means that one will lose some possibilities of retrieving detailed information and it can be difficult to accomplish a high privacy protection level without removing to much useful information. Even if all direct identification information is removed, there are possibilities to retrieve the identity by combining publicly available data with non-direct identifying attribute data from the data object and still be able to find the correct identity with a high probability. Techniques such as K-anonymity and others can prohibit this problem to some extent.

(b) Explain what restriction-based inference protection is and give an example of a specific type of restriction-based techniques. (2 pts).

Answer: Restriction-based techniques
Prevent queries for certain types of statistical queries

- query-set size control
- expanded query-set size control
- query-set overlap control
- audit-based control

(c) Explain what perturbation-based inference protection is and give an example of a specific type of perturbation-based techniques. (2 pts).

Answer: Perturbation-based techniques
Modies information that is stored or presented

- data swapping
- random-sample queries
- fixed perturbation
- query-based perturbation

- rounding (systematic, random, controlled)

---

Good Luck!

/ Kjell