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 for 1DL105 (6hp) and 48 points for 1DL025 (7.5hp) 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 and 28 respectively. 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 single-sided paper with your hand-written notes that you have brought with you, but no other material.
1. Data in data mining: 8 pts

This question compares and contrasts some similarity and distance measures.

(a) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. 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

(b) Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but dont let this confuse you.) (2pts)

Answer:
The Hamming distance is similar to the SMC. In fact, SMC = 1 - (Hamming distance / number of bits). The Jaccard measure is similar to the cosine measure because both ignore 0-0 matches.

(c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.) (2pts)

Answer:
Jaccard is more appropriate for comparing the genetic makeup of two organisms; since we want to see how many genes these two organisms share.

(d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.) (2pts)

Answer:
Two human beings share >99.9% of the same genes. If we want to compare the genetic makeup of two human beings, we should focus on their differences. Thus, the Hamming distance is more appropriate in this situation.

2. Classification: 8 pts

Consider the following data set for a binary class problem given in Table 1.
Table 1:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Class Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
<td>+</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>+</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>+</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>+</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>-</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>-</td>
</tr>
</tbody>
</table>

(a) Calculate the information gain (based on entropy) when splitting on A and B. Which attribute would the decision tree induction algorithm choose? (3pts)

Answer:

The contingency tables after splitting on attributes A and B are:

<table>
<thead>
<tr>
<th>A = T</th>
<th>A = F</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>4</td>
</tr>
<tr>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B = T</th>
<th>B = F</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>3</td>
</tr>
<tr>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

The overall entropy before splitting is:

\[ E_{\text{orig}} = -0.4 \log 0.4 - 0.6 \log 0.6 = 0.9710 \]

The information gain after splitting on A is:

\[ E_{A=T} = -\frac{4}{7} \log \frac{4}{7} - \frac{3}{7} \log \frac{3}{7} = 0.9852 \]

\[ E_{A=F} = -\frac{3}{3} \log \frac{3}{3} - \frac{0}{3} \log \frac{0}{3} = 0 \]

\[ \Delta = E_{\text{orig}} - 7/10E_{A=T} - 3/10E_{A=F} = 0.2813 \]

The information gain after splitting on B is:

\[ E_{B=T} = -\frac{3}{4} \log \frac{3}{4} - \frac{1}{4} \log \frac{1}{4} = 0.8113 \]

\[ E_{B=F} = -\frac{1}{6} \log \frac{1}{6} - \frac{5}{6} \log \frac{5}{6} = 0.6500 \]

\[ \Delta = E_{\text{orig}} - 4/10E_{B=T} - 6/10E_{B=F} = 0.2565 \]

Therefore, attribute A will be chosen to split the node.
(b) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose? (3pts)

Answer:

The overall gini before splitting is:

\[ G_{\text{orig}} = 1 - 0.4^2 - 0.6^2 = 0.48 \]

The gain in gini after splitting on A is:

\[ G_A = T = 1 - \left( \frac{4}{7} \right)^2 - \left( \frac{3}{7} \right)^2 = 0.4898 \]

\[ G_A = F = 1 - \left( \frac{3}{3} \right)^2 - \left( \frac{0}{3} \right)^2 = 0 \]

\[ \Delta = G_{\text{orig}} - \frac{7}{10}G_A = T - \frac{3}{10}G_A = F = 0.1371 \]

The gain in gini after splitting on B is:

\[ G_B = T = 1 - \left( \frac{1}{4} \right)^2 - \left( \frac{3}{4} \right)^2 = 0.3750 \]

\[ G_B = F = 1 - \left( \frac{1}{6} \right)^2 - \left( \frac{5}{6} \right)^2 = 0.2778 \]

\[ \Delta = G - 4/10G_A - 6/10G_A = F = 0.1633 \]

Therefore, attribute B will be chosen to split the node.

(c) The entropy and the Gini index are both monotonously increasing on the range \([0, 0.5]\) and they are both monotonously decreasing on the range \([0.5, 1]\). Is it possible that information gain and the gain in the Gini index favor different attributes? Explain. (2pts)

Answer:

Yes, even though these measures have similar range and monotonous behavior, their respective gains, \(\Delta\), which are scaled differences of the measures, do not necessarily behave in the same way, as illustrated by the results in parts (a) and (b).

3. Clustering: 6 pts

Use the similarity matrix in Table 2 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram for each. The dendrograms should clearly show the order in which the points are merged.

Answer:

4. Association rules: 10 pts

The original association rule mining formulation uses the support and confidence measures to prune uninteresting rules.
Table 2:

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>1.00</td>
<td>0.10</td>
<td>0.41</td>
<td>0.55</td>
<td>0.35</td>
</tr>
<tr>
<td>p2</td>
<td>0.10</td>
<td>1.00</td>
<td>0.64</td>
<td>0.47</td>
<td>0.98</td>
</tr>
<tr>
<td>p3</td>
<td>0.41</td>
<td>0.64</td>
<td>1.00</td>
<td>0.44</td>
<td>0.85</td>
</tr>
<tr>
<td>p4</td>
<td>0.55</td>
<td>0.47</td>
<td>0.44</td>
<td>1.00</td>
<td>0.76</td>
</tr>
<tr>
<td>p5</td>
<td>0.35</td>
<td>0.98</td>
<td>0.85</td>
<td>0.76</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Chapter 8 Cluster Analysis: Basic Concepts and Algorithms

Table 8.1. Similarity matrix for Exercise 16.

Figure 1: Dendrograms for problem 3

(a) Draw a contingency table for each of the following rules using the transactions shown in Table 3. (4pts)

Rules: \{b\} \rightarrow \{c\}, \{a\} \rightarrow \{d\}, \{b\} \rightarrow \{d\}, \{c\} \rightarrow \{e\}, \{c\} \rightarrow \{a\}.

Answer:

\[ \begin{array}{cccc}
    & c & \bar{c} \\
 b & 3 & 4 \\
\bar{b} & 2 & 1 \\
\end{array} \]

\[ \begin{array}{cccc}
    & d & \bar{d} \\
 a & 4 & 1 \\
\bar{a} & 5 & 0 \\
\end{array} \]

\[ \begin{array}{cccc}
    & d & \bar{d} \\
 b & 6 & 1 \\
\bar{b} & 3 & 0 \\
\end{array} \]

\[ \begin{array}{cccc}
    & c & \bar{c} \\
 e & 2 & 4 \\
\bar{e} & 3 & 1 \\
\end{array} \]

\[ \begin{array}{cccc}
    & a & \bar{a} \\
 c & 2 & 3 \\
\bar{c} & 3 & 2 \\
\end{array} \]

(b) Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures:

i. Support (2pts)

Answer:
Table 3:

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{a, b, d, e}</td>
</tr>
<tr>
<td>2</td>
<td>{b, c, d}</td>
</tr>
<tr>
<td>3</td>
<td>{a, b, d, e}</td>
</tr>
<tr>
<td>4</td>
<td>{a, c, d, e}</td>
</tr>
<tr>
<td>5</td>
<td>{b, c, d, e}</td>
</tr>
<tr>
<td>6</td>
<td>{b, d, e}</td>
</tr>
<tr>
<td>7</td>
<td>{c, d}</td>
</tr>
<tr>
<td>8</td>
<td>{a, b, c}</td>
</tr>
<tr>
<td>9</td>
<td>{a, d, e}</td>
</tr>
<tr>
<td>10</td>
<td>{b, d}</td>
</tr>
</tbody>
</table>

Rules | Support | Rank |
--- | --- | --- |
\(b \rightarrow c\) | 0.3 | 3 |
\(a \rightarrow d\) | 0.4 | 2 |
\(b \rightarrow d\) | 0.6 | 1 |
\(e \rightarrow c\) | 0.2 | 4 |
\(c \rightarrow a\) | 0.2 | 4 |

ii. Confidence (2pts)

Answer:

<table>
<thead>
<tr>
<th>Rules</th>
<th>Confidence</th>
<th>Rank</th>
</tr>
</thead>
</table>
\(b \rightarrow c\) | \(\frac{3}{7}\) | 3 |
\(a \rightarrow d\) | \(\frac{4}{5}\) | 2 |
\(b \rightarrow d\) | \(\frac{6}{7}\) | 1 |
\(e \rightarrow c\) | \(\frac{2}{6}\) | 5 |
\(c \rightarrow a\) | \(\frac{2}{5}\) | 4 |

iii. Interest \(X \rightarrow Y\) = \(P(X, Y)/(P(X)P(Y))\). (2pts)

Answer:

<table>
<thead>
<tr>
<th>Rules</th>
<th>Interest</th>
<th>Rank</th>
</tr>
</thead>
</table>
\(b \rightarrow c\) | 0.8571 | 3 |
\(a \rightarrow d\) | 0.8889 | 2 |
\(b \rightarrow d\) | 0.9524 | 1 |
\(e \rightarrow c\) | 0.6667 | 5 |
\(c \rightarrow a\) | 0.8000 | 4 |

5. Web Mining: (8 pts)

(a) In text of no more than one page present the main ideas and techniques of the PageRank algorithm used by Google. (4pts)

(b) In text of no more than one page present the main ideas and techniques of the Clever project: the ranking technique that is based on identifying authorities and hubs. (4pts)

NOTE! You SHOULD only solve the following problem 6 if you fulfil the following conditions:

You are a student registered to course 1DL025 - 7.5hp (but not to 1DL105 - 6hp) and you have NOT handed in AND passed the extra assignment on assignment 3, regarding direct hashing.
If you have handed in and passed the extra assignment on assignment 3 you will automatically get 8 points for this exam problem below and SHOULD NOT hand in any solution to problem 6.

6. Association analysis and hash trees: 8 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, 6\}, \{1, 3, 4\}, \{2, 3, 4\}, \{2, 4, 5\}, \{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. (4pts)

Answer:

![Hash tree for problem 6](image)

**Figure 2: Hash tree for problem 6**
(b) How many leaf nodes are there in the candidate hash tree? How many internal nodes are there? (1pt)

Answer: There are 5 leaf nodes and 4 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? (3pts)

Answer: 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.

Good Luck!

/ Kjell