Exam Data Mining
Date: 10-11-2016
Time: 13.30-16.30

General Remarks
1. You are allowed to consult 1 A4 sheet with notes written (or printed) on both sides.
2. You are allowed to use a (graphical) calculator. Use of mobile phones is not allowed.
3. Always show how you arrived at the result of your calculations. Otherwise you can not get partial credit for incorrect answers.
4. This exam contains five questions for which you can earn 100 points.

Question 1: Short Questions (20 points)

Answer the following questions:

(a) In frequent tree mining, let $T_1 = ab \uparrow c$ and $T_2 = abb \uparrow a \uparrow cb \uparrow ac \uparrow a$ be labeled rooted ordered trees. How many times does $T_1$ occur as an embedded subtree of $T_2$? Give the corresponding matching functions.

(b) We specify a hierarchical log-linear model by listing its highest order interaction terms. Consider the following hierarchical log-linear model on the variables $A, B,$ and $C$:

$$[AB], [AC], [BC]$$

where the interaction terms are given between square brackets.

(1) Draw the independence graph of this model.

(2) Is this a graphical model? Explain why or why not.

(c) In Bayesian network classifiers, is smoothing of parameter estimates more likely to be beneficial for naive Bayes or for TANs? Explain why.

(d) What is the difference between link-based classification and “ordinary” classification when it comes to predicting the class labels of a collection of new cases? How is this issue addressed in the paper of Lu and Getoor? (Qing Lu and Lise Getoor Link-based Classification Proceedings of ICML, 2003).
Question 2: Classification Trees (20 points)

Consider the following data on numeric attribute \(x\) and class label \(y\). The class label can take on two different values, coded as A and B.

<table>
<thead>
<tr>
<th>(x)</th>
<th>8</th>
<th>8</th>
<th>12</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>16</th>
<th>18</th>
<th>20</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

We use the gini-index as impurity measure. The optimal split is the one that maximizes the impurity reduction.

(a) Which candidate split(s) do we have to evaluate to determine the optimal one? (don’t list any more splits than strictly necessary)

(b) What is the optimal split on \(x\), and what is the impurity reduction of that split?

Question 3: Frequent Pattern Mining (25 points)

Consider the following database of sequences, containing orders in which different people have watched movies with Rocky Balboa as the main character:

<table>
<thead>
<tr>
<th>id</th>
<th>viewing sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ROKR</td>
</tr>
<tr>
<td>2</td>
<td>ROCY</td>
</tr>
<tr>
<td>3</td>
<td>OCKR</td>
</tr>
<tr>
<td>4</td>
<td>YROKR</td>
</tr>
<tr>
<td>5</td>
<td>RORK</td>
</tr>
</tbody>
</table>

(R = Rocky, O = Rocky II, . . ., Y = Rocky V).

(a) Use the GSP algorithm to find all frequent sequences with minimum support of 3. For each level give a table with the candidate frequent sequences, their support, and a check mark \(\checkmark\) if the sequence is frequent. Don’t list as candidates sequences that do not need to be counted on the database.

(b) Suppose we don’t care about the order in which the movies have been watched. We only care about whether or not someone has seen a particular movie. Use Apriori to find all frequent item sets with minimum support of 2. For each level give a table with the candidate frequent item sets, their support, and a check mark \(\checkmark\) if the item set is frequent. Don’t list as candidates item sets that do not need to be counted on the database. To generate the candidates, use the alphabetical order on the items (CKORY).
**Question 4: Undirected Graphical Models (20 points)**

Consider the undirected graphical model on discrete variables \((A, B, C, D, E)\) with independence graph:

![Graphical Model](image.png)

(a) Does \((A, B) \perp \perp (D, E) \mid C\) hold in this model? Explain why or why not.

(b) Give the “observed = fitted” margin constraints that must be satisfied by the maximum likelihood fitted counts.

(c) Give the formula for the maximum likelihood fitted counts \(\hat{n}(A, B, C, D, E)\) in terms of margins of the observed counts \(n(A, B, C, D, E)\).

**Question 5: Bayesian Networks (15 points)**

Consider a greedy hill-climbing search for a Bayesian network on 4 variables, denoted \(A, B, C, \) and \(D\). The search process is started from the empty graph. A neighbor model is obtained by adding, deleting, or reversing and edge. For each such operation, we determine the change in BIC score caused by it; we call this the delta score of the operation.

In step 1 of the search it is found that the best operation is to add an edge from \(A\) to \(B\), that is, \(\text{add}(A \to B)\). Assume that delta scores of operations computed in previous iterations that are still valid are not recomputed. For all other operations the delta scores have to be computed, that is, we perform no other optimizations.

For which (0 up to 6) of the following operations do we have to compute the delta score in the second step?

(a) \(\text{add}(C \to B)\)

(b) \(\text{add}(B \to A)\)

(c) \(\text{reverse}(A \to B)\)

(d) \(\text{add}(A \to C)\)

(e) \(\text{delete}(A \to B)\)

(f) \(\text{add}(B \to D)\)