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: Multiple Choice (18 points)

1. (Cost-complexity pruning) In computing the cost-complexity pruning sequence, we continue until:

   (a) The error on the test set goes up.
   (b) We have reached the root node of the tree.
   (c) The impurity reduction falls below a given threshold.
   (d) All leaf nodes are pure.

2. (Hierarchical Clustering) A naive implementation of hierarchical clustering of $N$ data points has computational complexity:

   (a) $O(2^N)$
   (b) $O(N^2)$
   (c) $O(N^3)$
   (d) $O(N \log N)$
3. (Association Rules) Consider an association rule $X \rightarrow Y$, where $X$ and $Y$ are disjoint closed itemsets. Suppose we move an item from the left hand size of the rule to the right hand side.

As a result, the confidence will (give the most specific correct answer):

(a) Go down.
(b) Go down or stay the same.
(c) Go up.
(d) Go up or stay the same.

4. (Frequent Itemset Mining) An important operation of the SON algorithm for frequent itemset mining is:

(a) Random sampling of baskets.
(b) Hashing items into buckets.
(c) Repeatedly reading small subsets of the baskets into main memory.
(d) Cross-validation.

5. The singular value decomposition (SVD) is primarily used for

(a) Dimensionality reduction.
(b) Decomposition of graphical models.
(c) Frequent item set mining.
(d) Pruning random forests.

6. (Frequent Itemset Mining) Which of the following statements about frequent itemset mining is false?

(a) All maximal frequent itemsets are closed.
(b) All subsets of a frequent itemset are frequent.
(c) All database transactions (baskets), regarded as itemsets, are closed.
(d) If all size $k - 1$ subsets of an itemset of size $k$ are frequent, then the itemset of size $k$ is frequent as well.
Question 2: Classification Trees (20 points)

Consider the following data on numerical attribute $x$ and class label $y$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>6</th>
<th>6</th>
<th>8</th>
<th>8</th>
<th>8</th>
<th>9</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The class label can take on two different values, coded as 0 and 1. We use the gini-index as impurity measure. The best split is the one that maximizes the impurity reduction.

(a) List all the splits on $x$ that are allowed by the algorithm that was discussed during the course.

(b) For which of the splits that you listed under (a) do we need to compute the impurity reduction in order to determine the best split? (don’t list any more splits than strictly necessary)

(c) Give the impurity reduction of the best split.

(d) (Cost-complexity pruning) Consider a training set with 40 examples of class 0, and 60 examples of class 1. We consider the pruning of a tree grown on this training set. Suppose that all leaf nodes of $T_1 = T(\alpha = 0)$ are pure, except for one leaf node that contains 5 examples of class 0, and 20 examples of class 1. Suppose furthermore that $T_1$ contains 5 splits in total. For what value of $\alpha$ does the tree obtained by pruning in the root node become better than $T_1$?

Question 3: Naive Bayes for Text Classification (20 points)

You are given the following collection of course evaluations with corresponding judgment (positive or negative):

<table>
<thead>
<tr>
<th>evalID</th>
<th>words in evaluation</th>
<th>Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>e1</td>
<td>great teacher interesting lectures</td>
<td>Positive</td>
</tr>
<tr>
<td>e2</td>
<td>great lectures</td>
<td>Positive</td>
</tr>
<tr>
<td>e3</td>
<td>great disappointment</td>
<td>Negative</td>
</tr>
<tr>
<td>e4</td>
<td>boring boring boring</td>
<td>Negative</td>
</tr>
<tr>
<td>e5</td>
<td>great lectures boring assignment</td>
<td>?</td>
</tr>
</tbody>
</table>

Here e1-e4 are the training examples, and e5 is a test example with unknown class label.

(a) Use e1-e4 to estimate $P(\text{great} \mid \text{Positive})$ and $P(\text{great} \mid \text{Negative})$ according to the multinomial Naive Bayes model. Use Laplace smoothing.

(b) Compute $P(\text{Positive} \mid e5)$ en $P(\text{Negative} \mid e5)$ according to the multinomial Naive Bayes model. Use Laplace smoothing.
Question 4: Undirected Graphical Models (20 points)

Consider the graphical log-linear model on binary variables A, B, C, and D, with independence graph:

We call this model $M_1$.

(a) Give the margin constraints that must be satisfied by the maximum likelihood fitted counts of $M_1$.

(b) Give the formula for the maximum likelihood fitted counts of $M_1$.

(c) Consider the model $M_0$ obtained by removing the edge between A and C from $M_1$. In a likelihood ratio test of $M_0$ against $M_1$ we find a deviance difference of 6.43. The significance level of the test is $\alpha = 0.05$ (see the table of critical values below). Clearly state whether or not $M_0$ is rejected in favour of $M_1$, and explain how you made that decision.

<table>
<thead>
<tr>
<th>degrees of freedom ($\nu$)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>critical value ($\chi^2_{\nu,0.05}$)</td>
<td>3.84</td>
<td>6.00</td>
<td>7.82</td>
<td>9.50</td>
<td>11.1</td>
<td>12.6</td>
<td>14.1</td>
<td>15.5</td>
</tr>
</tbody>
</table>

(d) Is $M_0$ decomposable? Explain.

Question 5: Bayesian Networks (22 points)

We perform a greedy hill-climbing search to find a good Bayesian network structure on 4 binary variables denoted $A, B, C,$ and $D$. Neighbour models are obtained by adding, deleting, or reversing an edge. We start the search process from the empty graph (the mutual independence model). In step 1 of the search we find that adding the edge $A \rightarrow D$ gives the biggest improvement in the BIC score. It is given that:

1. scores of operations (addition, deletion, or reversal of an edge) computed in previous iterations that are still valid are not recomputed, but retrieved from memory, and

2. scores of operations that produce a model that is equivalent to a model that has already been scored in a previous iteration are not recomputed, but are retrieved from memory as well.
Answer the following questions:

(a) For each of the following operations, state whether or not we need to compute (as opposed to retrieve from memory) the change in score in step 2 of the search, and give a short explanation:

1. add($B \rightarrow D$)
2. reverse($A \rightarrow D$)
3. add($D \rightarrow B$)
4. add($C \rightarrow D$)
5. delete($A \rightarrow D$)

Suppose the final model produced by the hill-climbing algorithm is:

![Diagram](image)

The following questions are all about this model:

(b) Give its essential graph.

(c) Does $A \perp \perp C \mid B$ hold? Explain why, or why not.

(d) How many parameters does it have?