Exam Data Mining
Date: 5-11-2015
Time: 13.30-16.30

General Remarks

1. You are allowed to consult 1 A4 sheet with notes written on both sides.

2. You are allowed to use a pocket 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 if the final answer is incorrect.

4. There are five questions, with which you can earn 100 points.

Question 1 Short Questions (20 points)

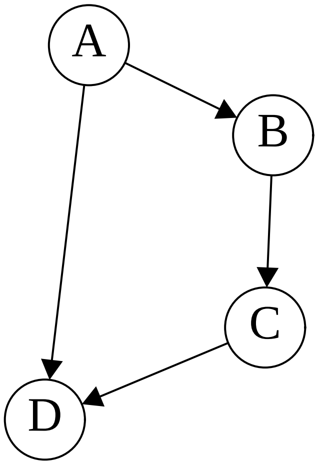
Answer the following questions:

(a) Consider the association rule $X \rightarrow Y$, where X and Y denote disjoint item sets. Prove that if we move an item from Y to X, then the confidence of the rule either increases or stays the same, but it can not decrease.

(b) Consider the following claim: *Two directed independence graphs are equivalent if they have the same moral graph.* Show that this claim is incorrect by giving a counterexample.

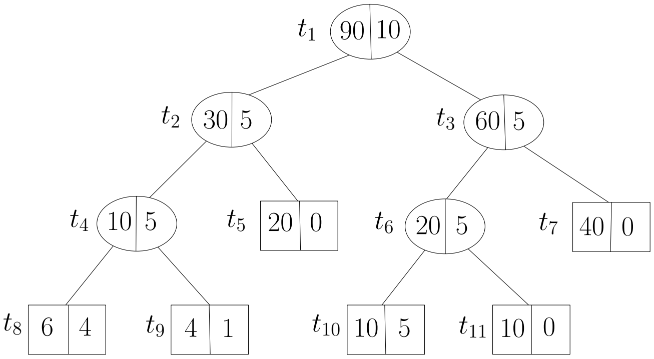
(c) In frequent tree mining, what is the difference between an induced subtree and an embedded subtree?

(d) Give the essential graph of the following directed independence graph:



Question 2: Classification Trees (25 points)

The tree given below, denoted by $T_{\max}$, has been constructed on the training sample:



In each node, the number of observations with class 0 is given in the left part, and the number of observations with class 1 in the right part. The leaf nodes have been drawn as rectangles.

(a) Compute the impurity of nodes t_1, t_2 and t_3 using the gini-index.

(b) Give the impurity reduction achieved by the first split.

(c) Compute T_1, the smallest minimizing subtree of $T_{\max}$ for $\alpha = 0$.

(d) Compute the cost-complexity pruning sequence $T_1 > T_2 > \ldots > \{t_1\}$. For each tree in the sequence, give the interval of α values for which it is the smallest minimizing subtree of $T_{\max}$.

Question 3: Frequent Sequence Mining (15 points)

Consider the following database of sequences:

sid	sequence
1	$ABBA$
2	$ABACAB$
3	$BADAD$

Use the GSP algorithm to find all frequent sequences with minsup=2. Visualize the search process as a prefix tree. Write the support between brackets next to a candidate sequence if and only if it needs to be counted on the database.
(It is advised to rotate your answer sheet 90° to draw the tree in landscape mode.)

Question 4: Undirected Graphical Models (25 points)

The following data concerns an outbreak of food poisoning after the traditional Christmas Lunch of the personnel of the Department of Information and Computing Sciences of our University. This time the theme was Dutch cuisine. Of the food eaten, interest focused on the "Berenhap" and "Frikandel". The variables are:

1. Berenhap eaten (1) or not eaten (0) (B)
2. Frikandel eaten (1) or not eaten (0) (F)
3. Sick (1) or not (0) (S)

Questionnaires were completed by 100 of the 114 persons attending. The table of observed counts is given below.

$n(B,F,S)$		S	
B	F	0	1
0	0	22	4
0	1	3	12
1	0	8	1
1	1	12	38

For example, $n(0,1,1) = 12$ is the number of people that did not have a Berenhap, but did eat the Frikandel, and became sick.

(a) Estimate $P(S = 1|B = 1)$ and $P(S = 1|B = 0)$.

(b) Based on the estimates computed at (a), would you say there is an association between eating a "Berenhap" and becoming sick? Explain.

(c) Draw the undirected independence graph of the graphical model expressing the constraint $B \perp\!\!\!\perp S \mid F$, and state the corresponding independence assumption(s) in words.

(d) Compute the fitted counts $\hat{n}(B, F, S)$ for the model given under (c).

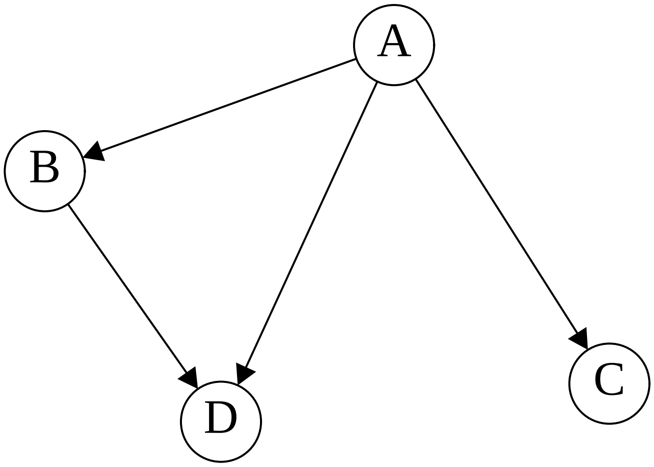
(e) Perform a statistical test to check whether the model you fitted under (d) gives an adequate fit of the data, using $\alpha = 0.05$. To perform the test, you may consult the following table with critical values:

degrees of freedom (ν)	1	2	3	4	5	6	7	8
critical value ($\chi^2_{\nu;0.05}$)	3.84	6.00	7.82	9.50	11.1	12.6	14.1	15.5

Clearly state whether or not the model is rejected, and explain how you made that decision.

Question 5: Bayesian Networks (15 points)

We perform a greedy hill-climbing search to find a good Bayesian network structure on 4 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 following initial graph:



In step 1 of the search we find that deleting the edge $B \rightarrow D$ gives the biggest improvement in the BIC score.

(a) For which operations (addition, deletion, reversal of an edge) do we need to compute the change in score in step 2 of the search? Note: assume that scores of operations computed in previous iterations that are still valid are not recomputed.

(b) Why do we have the reversal operator, even though the same change to the model could be achieved by first deleting the edge, and subsequently adding the edge in the opposite direction in the next step?