Exam Advanced Data Mining
Date: 11-11-2010
Time: 13.30-16.30

General Remarks

1. You are allowed to consult 1 A4 sheet with notes written on both sides.
2. Always show how you arrived at the result of your calculations.
3. If you are a native speaker, answers in Dutch are preferred.
4. There are six questions, for which you can score a total of 100 points.

Question 1 Multiple Choice (16 points)

For the following questions, zero or more answers may be true.

a) Which of the following statements are true?

1. An important difference between induction and deduction is that deductive reasoning is truth-preserving, while in an inductive argument, the conclusion may be false even though the premises are true.
2. In data mining we primarily deal with observational data rather than experimental data.
3. If the data to be analysed has missing values, it is justified to exclude the incomplete observations from the analysis, as long as there are enough complete observations left.
4. In data mining projects, the analysis phase is generally the most time consuming.

b) Which of the following statements about classification trees are true?

1. The resubstitution error of a tree never goes up when we split one of its leaf nodes.
2. In growing a tree, we can always continue splitting until each leaf node contains examples of a single class, but the resulting tree will be overfitted.

3. If both children produced by a split have the same majority class, then the impurity reduction of the split is zero.
4. For classification problems with more than two classes, in order to determine the optimal split for a categorical attribute with L distinct values, we have to compute $2^{L-1} - 1$ possible splits.

c) Which of the following statements about frequent pattern mining are true?

1. If all the proper subsets of an itemset are frequent, then the itemset itself must also be frequent.
2. All maximal frequent itemsets are closed.
3. If we only know the maximal frequent itemsets and their support, we can infer from that *all* frequent itemsets and their support.
4. For an association rule, if we move one item from the right-hand-side to the left-hand-side of the rule, then the confidence will never go down.

d) In Pattern Set Mining which of the following are general criteria to determine the quality of a pattern set?

1. Surprisingness.
2. Complexity.
3. Descriptiveness.
4. Understandability.

Question 2 Frequent Itemset Mining (25 points)

Given are the following five transactions on items $\{A, B, C, D, E\}$:

tid	items
1	$\{A, B\}$
2	$\{A, B, D\}$
3	$\{B, D, E\}$
4	$\{B, C, D, E\}$
5	$\{A, B, C\}$

a) Use the Apriori algorithm to compute all frequent itemsets, and their support, with minimum support 2. It is important that you clearly indicate the steps of the algorithm.

b) Use the A-close algorithm to compute all *closed* frequent itemsets, with minimum support 2. It is important that you clearly indicate the steps of the algorithm.

c) Use all closed frequent itemsets computed at (b) to construct a KRIMP codetable, and compute how often each itemset in the codetable is used in covering the database. Also compute the optimal code length for each itemset in the codetable.
Note: Here you should use logarithms with base 2.

d) Suppose that in addition to the transactions, we also have information about the price of each item. Consider constraints of the type

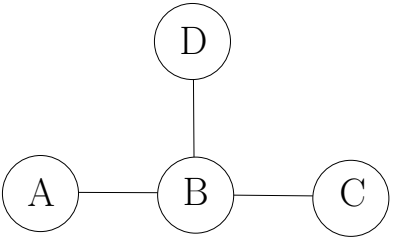
$$\max(I.\text{price}) - \min(I.\text{price}) \leq c$$

where $\max(I.\text{price})$ denotes the maximum of the prices of the items in itemset I, and c is some positive constant. Suppose that we want to find all frequent itemsets that also satisfy a constraint of this type.

Can we use this constraint in an Apriori style levelwise search to further prune the search space, while still guaranteeing completeness of the results? Explain your answer.

Question 3 Undirected Graphical Models (15 points)

Let M be an (undirected) graphical model on discrete variables (A, B, C, D) with independence graph:



a) Give all pairwise conditional independencies that hold for this model.

b) Give the *observed = fitted* margin constraints that hold for the maximum likelihood fitted counts of M.

c) Use the constraints of (b) together with the conditional independence properties of M, to find an expression for the maximum likelihood fitted counts $\hat{n}(A, B, C, D)$ in terms of margins of the observed counts $n(A, B, C, D)$.

Question 4 Bayesian Network Classifiers (20 points)

The Naive Bayes classifier makes the fundamental assumption that the attributes are independent within each class.

a) Explain why the Naive Bayes Classifier often performs quite well (in terms of the error-rate on a test sample), even when its independence assumption is not satisfied.

To relax the independence assumption, one can allow some (restricted) dependencies between the attributes. A well-studied example are the Tree Augmented Naive Bayes (TAN) classifiers.

b) For a problem with k binary attributes, and a class label with m possible values, how many parameters does a TAN classifier have?

c) Explain why a TAN classifier has the same independence properties as the undirected graph obtained by simply dropping the direction of all its edges.

d) In the article of Friedman et al. many different methods to use a Bayesian Network for classification are studied. One of them is to use a standard structure learning algorithm using BIC (MDL) as the score function, and then to use the Markov blanket of the class variable in the resulting network for classification. It is shown that for datasets with many attributes, this method tends to produce poor results. Explain why.

e) One of the advantages of restricting the structure on the attributes to trees is that there is a polynomial time algorithm to compute the optimal structure. Describe the steps of this algorithm.

Question 5 Bayesian Networks (10 points)

Consider a heuristic search for a Bayesian Network that maximizes the BIC score

$$\mathrm{BIC}(M) = \mathcal{L}(M) - \frac{\ln n}{2}\dim(M).$$

Here ln denotes the natural logarithm.

The algorithm performs a hill-climbing search where the neighbours of the current model are obtained by either

1. removing an arrow from the current model
2. adding an arrow to the current model
3. turning an arrow of the current model around

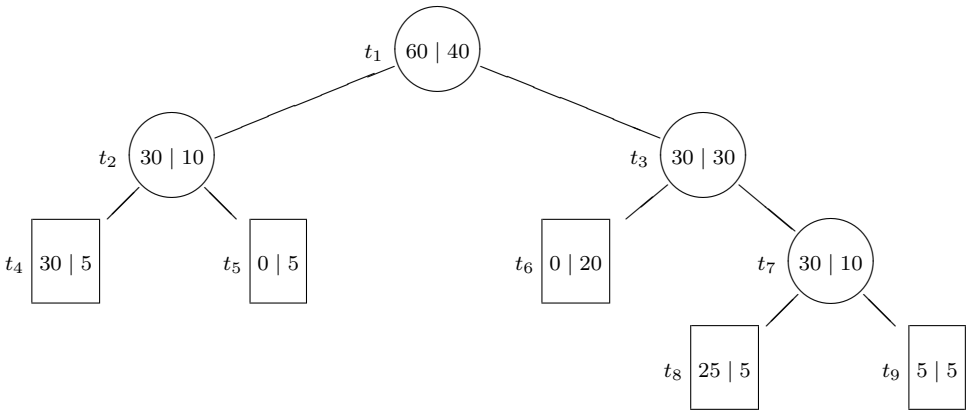
The current model on (X_1, X_2, X_3) is:

a) Give all neighbours of the current model, and indicate which neighbours are equivalent to eachother, in the sense that they encode the same independence properties.

b) Assume that each variable has 3 possible values, and that we have a data set with 100 observations. For each neighbour model that can be obtained by adding an arc, indicate how much the loglikelihood part of the BIC score should increase for it to have a better BIC score than the current model.

Question 6 Classification Trees (14 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 according to the gini-index. Give the impurity reduction achieved by the first split.

b) Compute the cost-complexity pruning sequence $T_1 > T_2 > \ldots > \{t_1\}$, where T_1 is the smallest minimizing subtree of $T_{\max}$ for $\alpha = 0$. For each tree in the sequence, give the interval of α values for which it is the smallest minimizing subtree of $T_{\max}$.