Exam Data Mining
Date: 5-11-2013
Time: 17.00-20.00

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.
4. There are five questions; you can score 20 points for each question.

Question 1 Short Questions

Answer the following questions:

(a) What is overfitting? Briefly describe one method to prevent overfitting in classification trees.

(b) Because many data mining algorithms cannot handle missing values, people sometimes remove all observations (rows) that contain missing values before the analysis. Give two potential disadvantages of this procedure.

(c) Describe the steps of the algorithm of Chow and Liu to learn a tree-structured Bayesian network that maximizes the log-likelihood score.

(d) In frequent item set mining, for what kind of data sets is the A-Close algorithm more efficient than Apriori?

(e) In link-based classification of objects in a (social) network, what problem do we run into when we want to classify the objects in the test set? How can this problem be solved?

Question 2: Classification Trees

Consider the following data on numeric attribute x and binary class label y:

x	8	11	12	14	14	15	15	17	18
y	0	0	0	0	1	0	1	1	1

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 than strictly necessary)

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

(c) Suppose that the optimal split is defined as the split that maximizes the impurity reduction among those splits that satisfy a minleaf constraint. Would your answer to (a) still be valid? Explain.

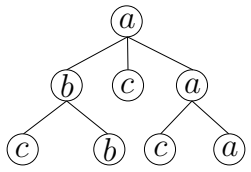
Question 3: Frequent Pattern Mining

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

tid	items
1	AB
2	AD
3	BCD
4	ACD
5	$ACDF$
6	ABE

(a) Use the Apriori algorithm to compute all frequent item sets, and their support, with minimum support 2. Clearly indicate the steps of the algorithm, and the pruning that is performed.

Consider the labeled ordered tree d_1:



In the questions we use the following string representation of labeled ordered trees: we list the node labels according to pre-order (depth-first) traversal, and use the special symbol $\uparrow$ to indicate that we go up one level in the tree. For example, the string representation of d_1 is: $abc \uparrow b \uparrow\uparrow c \uparrow ac \uparrow a$.

(b) How many times does the tree $T = ac \uparrow a$ occur as an induced subtree in d_1? Give the rightmost occurrence list (RMO-list) of T in d_1 as it is maintained by the FREQT algorithm.

(c) How many times does the tree $T = ac \uparrow a$ occur as an embedded subtree in d_1? Give the corresponding matching functions (copy the table below on your answer sheet and complete it; the nodes of T have been named w_1, w_2 and w_3).

	w_1	w_2	w_3
ϕ_1			
etc.			

(d) The FREQT algorithm uses the right-most extension technique to generate candidate $k+1$-trees from frequent k-trees. Assume the label set is $\Sigma = \{a, b, c\}$, and assume that d_1 is frequent. How many candidate trees will FREQT generate from d_1? Explain your answer.

Question 4: Iterative Proportional Fitting

Iterative Proportional Fitting (IPF) is an algorithm to compute the maximum likelihood fitted counts for hierarchical log-linear models.

We want to fit the independence model $X_1 \perp\!\!\!\perp X_2$ to the following table of observed counts on binary variables X_1 and X_2:

$n_{12}(x_1, x_2)$	$x_2 = 0$	$x_2 = 1$	$n_1(x_1)$
$x_1 = 0$	76	4	80
$x_1 = 1$	14	6	20
$n_2(x_2)$	90	10	100

All questions below are concerned with fitting the independence model to this data set.

(a) Which margin constraints have to be satisfied by the fitted counts?

(b) Compute the fitted counts using IPF, starting with:

$$\hat{n}^{(0)} = \begin{array}{c|cc|c} & 0 & 1 & \\ \hline 0 & 5 & 5 & 10 \\ 1 & 5 & 5 & 10 \\ \hline & 10 & 10 & \end{array}$$

Clearly show the steps of the algorithm.

(c) Compute the fitted counts using IPF, but this time starting with:

$$\hat{n}^{(0)} = \begin{array}{c|cc|c} & 0 & 1 & \\ \hline 0 & 15 & 1 & 16 \\ 1 & 3 & 1 & 4 \\ \hline & 18 & 2 & \end{array}$$

Clearly show the steps of the algorithm.

(d) Which solution, (b) or (c), is the correct one? How did you determine this?

Question 5: Bayesian Networks

Consider the following data on whether a cancer patient survived, the grade of the cancer (malignant or benign), and the location of the treatment center (Boston or Glamorgan).

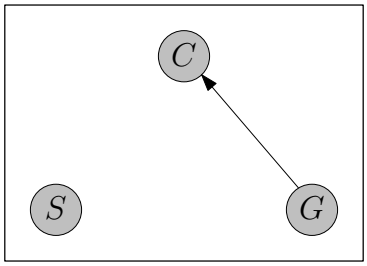
Boston	Malignant	Benign
Died	35	47
Survived	59	112

Glamorgan	Malignant	Benign
Died	42	26
Survived	77	76

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

$$\text{AIC}(M) = \mathcal{L}(M) - \dim(M).$$

The algorithm performs a hill-climbing search where the neighbors of the current model are obtained by either: removing an arrow from the current model, adding an arrow to the current model, or turning an arrow of the current model around.
The current model in the search is:



Here S represents Survival, G the grade of the cancer, and C the center of treatment.

(a) Give all neighbors of the current model, and indicate which neighbors are equivalent to each other. Also indicate which neighbors are equivalent to the current model.

(b) Compute the contribution of node G to the AIC score of the current model. Use the *natural* logarithm in your computations.

(c) Does the model obtained by adding an arrow from S to G have a better AIC score than the current model? Justify your answer by showing the relevant calculations.

(d) Using the relationship between directed and undirected independence graphs, state the independence assumption encoded by the model at (c) in a single sentence.