Retake Data Mining
Date: 2-1-2019
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: True or False? (20 points)

State whether the following claims are true (A) or false (B).

1. (Frequent item set mining) An item set has the same closure as any of its supersets with the same support.
2. (Frequent sequence mining) "AI" occurs 2 times as a subsequence of "A GIANT MIND".
3. (Missing data) We have data on income and gender. Income is missing for 20% of the males and for 5% of the females. The probability that income is missing depends *only* on gender, which is completely observed. Claim: the average income of the *persons with income observed* is a correct reflection ("unbiased") of the average income of *all persons* contained in the data set.
4. (Classification trees) If we use resubstitution error as impurity measure, then the impurity reduction of the worst possible split may be negative.
5. (Bayesian networks) Two directed independence graphs are (Markov) equivalent if they have the same skeleton and the same immoralities (v-structures).
6. (Link-based classification) In link-based classification, the label of each node is assumed to be independent of the labels of the other nodes.

7. (Active Learning) Active learning is especially useful if unlabeled data is cheap, but obtaining class labels is relatively expensive.

8. (Random forests) Each split in a random forest is allowed to use only a subset of the features.

9. (Bias-Variance decomposition of prediction error) As the training set size increases, the variance component of expected prediction error decreases.

10. (Classification Trees) In a binary classification problem, the gini-index in a node t is equal to the variance of a Bernoulli random variable with probability of success equal to the proportion of cases with class 0 in t.

Question 2: Classification Trees (20 points)

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

x	8	8	10	10	10	10	14	15	15	16
y	0	1	0	0	1	1	1	1	1	1

The class label can take on two different values, coded as 0 and 1.

(a) Suppose we use the *gini-index* as impurity measure.
Give the best split(s) on x, and the corresponding impurity reduction.

(b) Suppose we use *resubstitution error* as impurity measure.
Give the best split(s) on x, and the corresponding impurity reduction.

Question 3: Multinomial Naive Bayes for Text Classification (15 points)

You are given the following collection of computer science course evaluations:

evaluationID	words in evaluation	class label
e1	`good teacher interesting lectures`	Positive
e2	`good lectures excellent course`	Positive
e3	`bad teacher discontinue course`	Negative
e4	`boring lectures teacher incompetent`	Negative

(a) Estimate $P(\texttt{good} \mid \text{Positive})$ and $P(\texttt{good} \mid \text{Negative})$ according to the multinomial naive Bayes model. Use Laplace smoothing.

(b) Assume the multinomial naive Bayes model is trained with Laplace smoothing on the give data set. Give the probability of the Positive class according to this model for the evaluation text: `very good teacher`.

Question 4: Undirected Graphical Models (25 points)

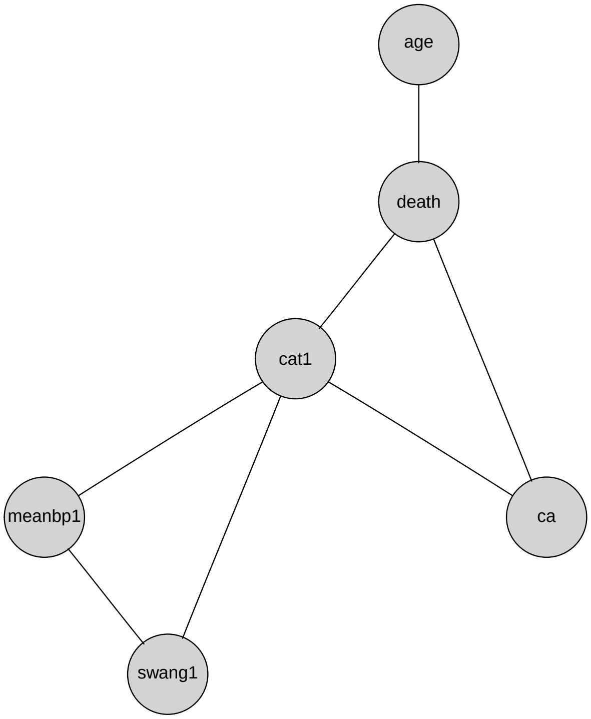
We have a data set containing data on 5,735 critically ill adult patients receiving care in an Intensive Care Unit (ICU) for 1 of 9 pre-specified disease categories. The data was collected in five US teaching hospitals between 1989 and 1994.

The objective of the original study that used (a superset of) this data was to examine the association between the use of right heart catheterization (RHC) during the first 24 hours of care in the ICU and subsequent survival.

This subset contains the following variables:

1. cat1: disease category (9 different values)
2. death: did the patient die within 180 days after admission? (yes/no)
3. swang1: was right heart catheterization performed within first 24 hours? (yes/no)
4. ca: cancer status (yes/no/metastatic)
5. age: age of patient divided into 5 categories
6. meanbp1: mean blood pressure of patient divided into 2 categories

Consider the graphical log-linear model with the following independence graph:



(a) According to the given model, is age independent of disease category (age $\perp\!\!\!\perp$ cat1)? Explain your answer.

(b) According to the given model, are performance of right heart catheterization (swang1) and death independent given disease category (swang1 $\perp\!\!\!\perp$ death | cat1)?
Explain your answer.

(c) Give the formula for the maximum likelihood fitted counts of this model.

(d) How many parameters (u-terms) does the given model have?

(e) Is the model obtained by adding an edge between age and swang1 decomposable? Explain.

Question 5: Bayesian Networks (20 points)

The table below shows the numbers of successes and failures for minor and major operations in two hospitals: one academic hospital and one local hospital. The total number of operations is $n = 2900$.

n(operation, hospital, result)		result	
operation	hospital	success	failure
minor	academic	685	15
	local	584	16
major	academic	1425	75
	local	93	7

We perform a greedy hill-climbing search to find a good Bayesian network structure. Neighbour models are obtained by adding a single edge to the current model, deleting a single edge, or turning a single edge around. We start the search process from the empty graph (the mutual independence model).

(a) Compute the change in log-likelihood score if we add the edge *operation* → *hospital.*

(b) Consider the model where the edge *operation* → *hospital* has been added. Call this the current model. List all the neighbours of the current model and indicate which neighbours are equivalent to each other, and which neighbours are equivalent to the current model.

The academic hospital has a success rate of $685/700 = 97.86\%$ for minor operations, while the local hospital scores a bit worse: $677/700 = 97.33\%$. Likewise, the academic hospital has a success rate of $1425/1500 = 95\%$ for major operations, while the local hospital scores $93/100 = 93\%$. But overall, the academic hospital has a success percentage of $2110/2200 = 95.9\%$, while the local hospital scores a bit better, namely $677/700 = 96.7\%$.

(c) Explain how it is possible that the academic hospital scores better for both minor and major operations, but overall the local hospital scores better.