Exam Data Mining
January 11, 2023, 17.00-19.30 hrs

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.
3. Always show how you arrived at the result of your calculations.
4. This exam contains five questions for which you can earn 100 points.

Question 1: Mixed Short Questions (30 points)

1. Which of the following statements about bagging and random forests are true?
   (any number from 0 to 4 can be true!)

   (a) In bagging, we take a number of random permutations of the rows of the training data, grow a tree on each permutation, and take the majority vote of their predictions.

   (b) In random forests, for each tree we randomly sample a number of features from the feature set to construct the tree.

   (c) The purpose of bagging is to reduce the variance component of error.

   (d) In comparison to bagging, random forests tend to reduce the correlation between the predictions of the individual trees.

2. Which of the following statements about frequent pattern mining are true?
   (any number from 0 to 4 can be true!)

   (a) Every induced subtree of tree T is an embedded subtree of T.

   (b) "AI" occurs 4 times as a subsequence in "MACHINE LEARNING"
   (there are 4 different mappings).

   (c) In frequent tree mining with the FREQT algorithm, if there is only one frequent tree of size k, then there are no candidates for level $k + 1$.

(d) On dense (as opposed to sparse) data, with strong positive correlations between items, the Apriori-close (A-close) algorithm tends to be more efficient than Apriori.

3. (Classification Trees) Let $p(c \mid t), c = 1, \ldots, C$, denote the relative frequency of class c in node t, where C denotes the number of classes. It is common to predict the majority class in a leaf node, which gives a probability of making a wrong prediction of $1 - \max_c p(c \mid t)$. Suppose that, instead, we predict class c with probability $p(c \mid t)$, for all $c = 1, \ldots, C$. Give an expression for the probability of making a wrong prediction in a leaf node t when using this prediction rule. Does the expression look familiar? Explain.

4. (Graphical Models) Consider the table of counts on binary variables X_1 and X_2:

x_2 x_1	0	1	Total
0	18	42	60
1	12	28	40
Total	30	70	100

Which model has the best BIC score on this data? (choose 1 answer!)

(a) $\ln P(x_1, x_2) = u_\emptyset$.

(b) $\ln P(x_1, x_2) = u_\emptyset + u_1 x_1 + u_2 x_2$.

(c) $\ln P(x_1, x_2) = u_\emptyset + u_1 x_1 + u_2 x_2 + u_{12} x_1 x_2$.

(d) Both (b) and (c): they have the same BIC score.

Explain your answer.

5. (Logistic Regression) In a logistic regression with *months of programming experience* as the only predictor variable, and success on a programming assignment as the class variable (success=1; fail=0) we find that the maximum likelihood estimate of the coefficient of the predictor variable is $\hat{\beta}_1 \approx 0.14$. Hence, according to the fitted model (choose 1 answer!):

(a) Every additional month of programming experience increases the probability of success with about 14%.

(b) Every additional month of programming experience increases the probability of success with about 15%.

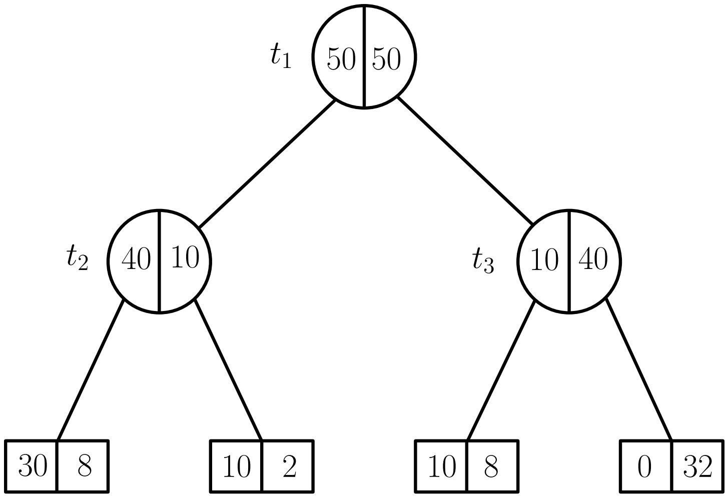
(c) Every additional month of programming experience increases the odds of success with about 14%.

(d) Every additional month of programming experience increases the odds of success with about 15%.

Show how you determined the answer.

Question 2: Classification Trees (20 points)

The tree T_{max} given below has been grown on the training sample.



In each node, the number of observations with class A is given in the left part, and the number of observations with class B in the right part.

(a) Give the impurity reduction of the split performed in the root node, using the gini-index as impurity measure.

(b) Give 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}.

(c) If we use cross-validation to select a tree from the pruning sequence under (b), what is the representative complexity value for tree T_2?

Question 3: Closed Frequent Item Set Mining (15 points)

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

tid	items
1	ABC
2	ABC
3	BCD
4	BCD
5	CDE

Use the Apriori-close (A-close) algorithm to compute all closed frequent item sets, and their support, with minimum support 2. Do this in the following two steps:

(a) For each level, list the candidate generators, their support, and whether or not they turn out to be generators. Use the alphabetical order on the items to generate candidates. Explain the pruning that is performed.

(b) List the generators found under (a), and compute their closure to obtain the set of closed frequent item sets. Also give the support for each closed frequent item set.

Question 4: Bayesian Networks (15 points)

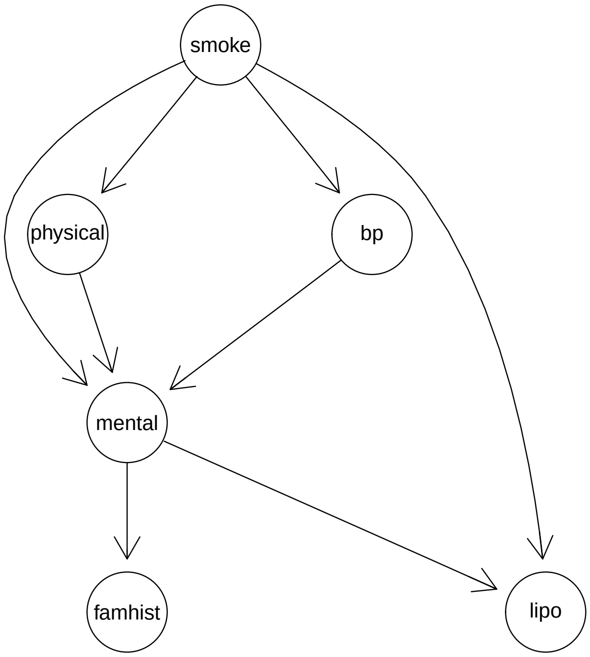
We analyze a data set concerning risk factors for coronary heart disease.
For a sample of 1841 car-workers, the following information was recorded:

Variable	Description
smoke	Does the person smoke?
mental	Is the person's work strenuous mentally?
physical	Is the person's work strenuous physically?
bp	Systolic blood pressure $\leq$ 140mm?
lipo	Ratio of beta to alfa lipoproteins ≤ 3?
famhist	Is there a family history of coronary heart disease?

The contribution of each node to the log-likelihood score of the current model (rounded to the nearest integer) is given below:

smoke	mental	physical	bp	lipo	famhist
-1274	-910	-1262	-1251	-1208	-744

The current model in the search is given in the graph below:



Answer the following questions:

(a) Is the model that we get by reversing the edge between "smoke" and "bp" equivalent to the current model? Explain

The counts on the training data for "smoke" and "lipo" are given in the following table:

lipo smoke	≤ 3	> 3	Total
no	598	363	961
yes	463	417	880
Total	1061	780	1841

Use the natural logarithm (ln) in your computations.

(b) What is the change in the log-likelihood score if we delete the edge mental → lipo ? (round your answer to the nearest integer)

(c) What is the change in the BIC-score if we delete the edge mental → lipo ?

Question 5: Multinomial Naive Bayes for Text Classification (20 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 $P(\texttt{good} \mid \text{Positive})$, $P(\texttt{good} \mid \text{Negative})$, $P(\texttt{teacher} \mid \text{Positive})$, and $P(\texttt{teacher} \mid \text{Negative})$ according to the multinomial naive Bayes model, using Laplace smoothing.

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

The next question is a general question about the multinomial naive Bayes model, not about the specific data set of questions (a) and (b).

(c) Construct a training set and test document `testdoc` for which maximum likelihood estimation of the multinomial naive Bayes model on the training set will result in $\hat{P}(\texttt{testdoc} \mid A) = 0$, and $\hat{P}(\texttt{testdoc} \mid B) = 0$, where A and B are the class labels. Make the training set and test document as small as possible, but non-empty.