Exam Data Mining
November 9, 2022, 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 (25 points)

1. (Text Mining) Which of the following statements about text mining are true? (any number from 0 to 4 can be true!)
   (a) In the bag-of-words representation, the order of words is ignored.
   (b) In the bag-of-words representation, the frequency of words is ignored.
   (c) A sentence with n words, all of them different, contains $n-1$ bigrams.
   (d) By looking at the *sign* of the mutual information we can determine whether a word points towards the positive or towards the negative class.

2. Which of the following statements about frequent pattern mining are true? (any number from 0 to 4 can be true!)
   (a) Every embedded subtree of tree T is also an induced subtree of T.
   (b) "AI" occurs 4 times as a subsequence in "DATA MINING" (there are 4 different mappings).
   (c) In frequent sequence mining with the GSP algorithm, if there is only one frequent sequence of length k, then there are no candidates for level $k+1$.
   (d) Every maximal frequent item set is closed.

3. (Cost-complexity pruning) In a binary classification problem, what is the smallest value of α for which the root node is guaranteed to be the smallest minimizing subtree?

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 of the following two models has the best AIC score on this data?

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

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

Explain your answer.

5. (Logistic Regression) In a logistic regression with hours of study as the only predictor variable, and success on the exam as the class variable (pass=1; fail=0) we find that the maximum likelihood estimate of the coefficient of the predictor variable is $\hat{\beta}_1 \approx 0.215$. Hence, according to the fitted model:

(a) Every extra hour of study increases the probability of passing with about 21.5%.

(b) Every extra hour of study increases the probability of passing with about 24%.

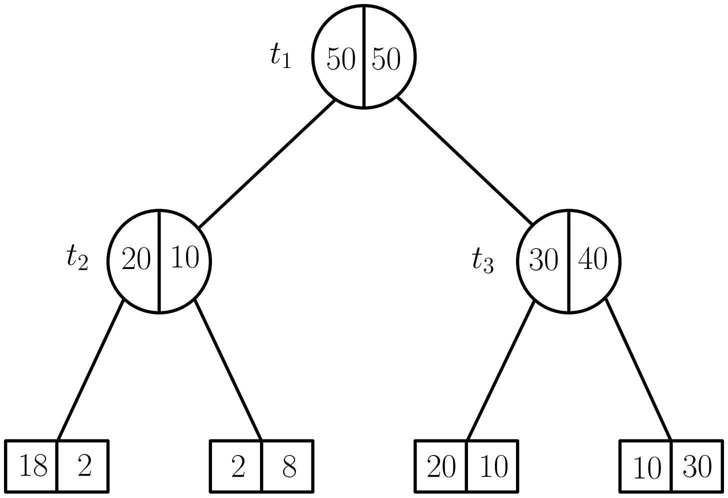
(c) Every extra hour of study increases the odds of passing with about 21.5%.

(d) Every extra hour of study increases the odds of passing with about 24%.

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 $T_1 > T_2 > \ldots > \{t_1\}$.

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	ABC
4	BCD
5	BCD
6	DE
7	B
8	C

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.

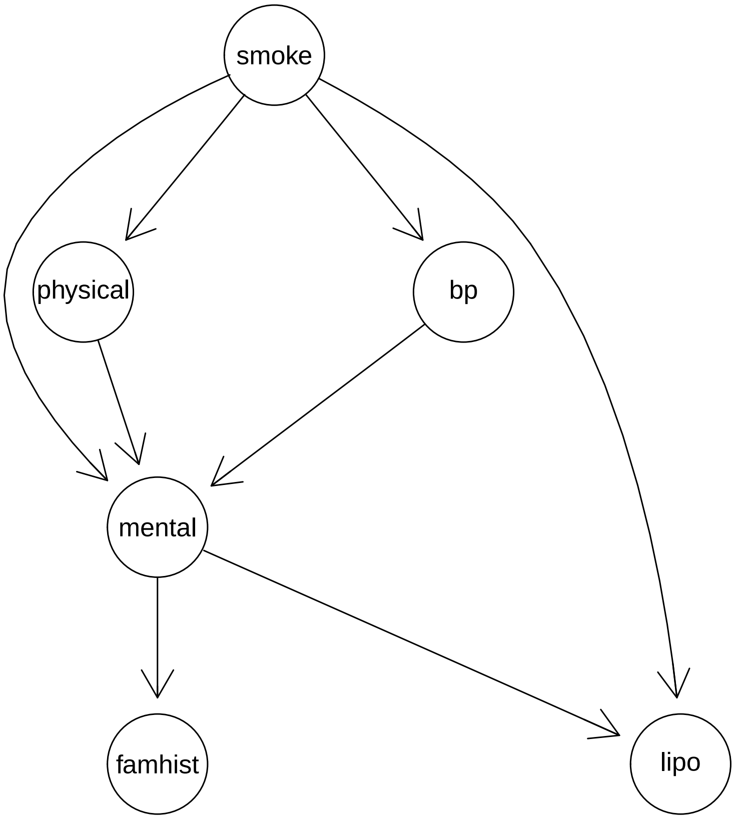
(b) List the generators found under (a), and compute their closure to obtain the set of closed frequent item sets. Also give the support for each closed frequent item set.

Question 4: Bayesian Networks (25 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 $\leq$ 3?
famhist	Is there a family history of coronary heart disease?

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



(a) Does the conditional independence {physical} $\perp\!\!\!\perp$ {bp} | {smoke} hold in the given model? Motivate your answer.

(b) Does the conditional independence {physical} $\perp\!\!\!\perp$ {bp} | {smoke, mental} hold in the given model? Motivate your answer.

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

lipo mental	≤ 3	> 3	Total
no	724	406	1130
yes	337	374	711
Total	1061	780	1841

Use the natural logarithm (ln) in your computations.

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

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

Question 5: Multinomial Naive Bayes for Text Classification (15 points)

You are given the following collection of hotel reviews and corresponding sentiment:

reviewID	words in review	sentiment
r1	`large room clean good service`	Positive
r2	`good service excellent breakfast`	Positive
r3	`dirty bathroom bad service`	Negative
r4	`disgusting breakfast noisy`	Negative
r5	`good riddance noisy`	?

Here r1-r4 are the training examples, and r5 is a test example with unknown class label.

(a) Use r1-r4 to estimate $P(\texttt{good} \mid \text{Positive})$, $P(\texttt{good} \mid \text{Negative})$, $P(\texttt{noisy} \mid \text{Positive})$, and $P(\texttt{noisy} \mid \text{Negative})$ according to the multinomial naive Bayes model with Laplace smoothing.

(b) Compute $P(\text{Positive} \mid \text{r5})$ according to the multinomial naive Bayes model with Laplace smoothing.

(c) What is the purpose of Laplace smoothing?