Online Exam Data Mining
January 19, 2022, 17.30-20.30 hrs

General Remarks

1. You are allowed to consult all of the required literature of the course.
   It is not allowed to consult any other materials or persons.
2. You are allowed to use a computer for performing calculations.
3. Always show how you arrived at the result of your calculations.
4. This exam contains six questions for which you can earn 100 points.

Question 1: Multiple Choice (16 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 bagging and random forests are true? (any number from 0 to 4 can be true!)
   (a) In bagging, we split the data into a number folds, grow a tree on each fold, and take the majority vote of their predictions.
   (b) In random forests, at each split we randomly sample `nfeat` features *with replacement* from the feature set.
   (c) The purpose of bagging is to reduce the variance component of error.
   (d) Random forests always have lower prediction error than bagging, because they reduce the correlation between the predictions of the individual trees.

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

x_1 \ x_2	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?

(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.

4. 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.

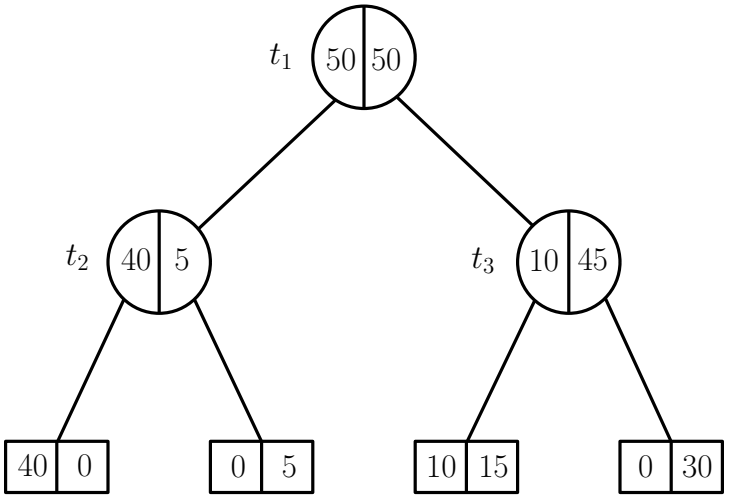
(b) "AI" occurs 4 times as a subsequence in "DATA MINING" (there are 4 different mappings).

(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) The FREQT algorithm for mining frequent induced subtrees may produce candidates for level $k+1$ that have a level k induced subtree that is infrequent.

Question 2: Classification Trees (28 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. The leaf nodes have been drawn as rectangles.

(a) Give the impurity reduction of the split performed in t_3 using the gini index as impurity measure.

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

The following are general questions about classification trees and cost-complexity pruning.

(c) Let $p(j|t), j = 1, \ldots, J$, denote the relative frequency of class j in node t, computed on the training data. It is common to predict the majority class in a leaf node, which gives a probability of making a wrong prediction of $1 - \max_j p(j|t)$. Suppose that, instead, we predict class j with probability $p(j|t), j = 1, \ldots, J$. 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? Give an interpretation of the resulting expression.

(d) In cost-complexity pruning, show that for every internal (non-terminal) node $t \in T$: $g(t) \leq R(t)$.

(e) In cost-complexity pruning, show that for all $t' \in T_t : R(t') \leq R(t)$.

(f) In cost-complexity pruning we usually start the pruning sequence at $\alpha_1 = 0$.
Instead, suppose we specify a value $\alpha_{min} > 0$, and want to determine a pruning sequence with T_1 the smallest minimizing subtree of T_{max} for $\alpha = \alpha_{min}$. Show that this information can be exploited when growing the tree by not expanding further nodes t with $R(t) < \alpha_{min}$.

Question 3: Frequent Item Set Mining (16 points)

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

tid	items
1	ABC
2	ABC
3	ABC
4	BCD
5	BCD
6	DE

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.

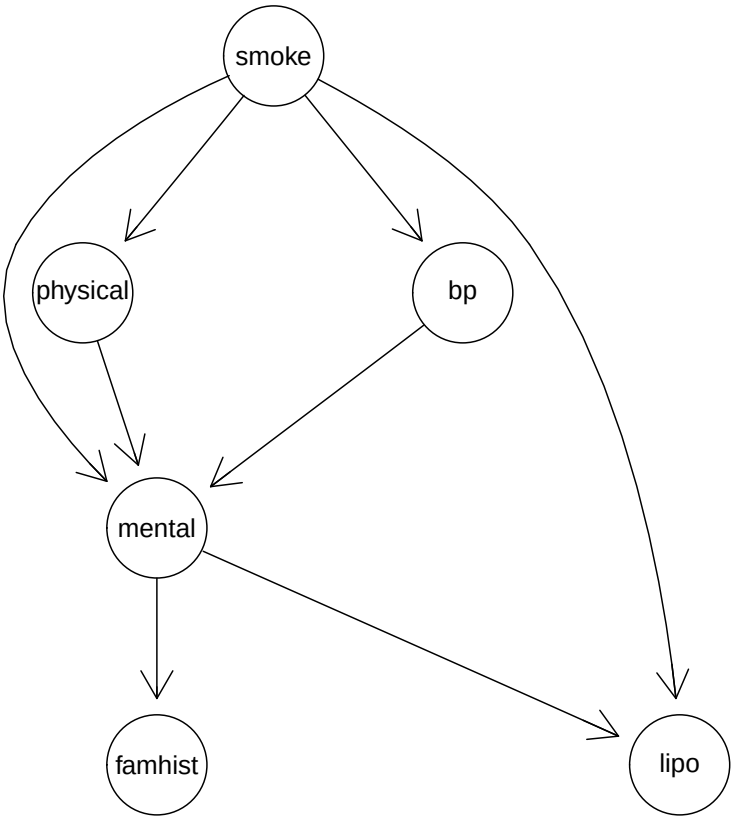
(b) List the generators found under (a), and compute their closure to obtain the set of closed frequent item sets. Also give the support for each closed frequent item set.

Question 4: Bayesian Networks (20 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:

mental \ lipo	≤ 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 (10 points)

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

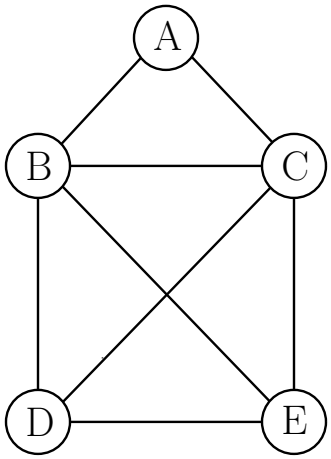
evaluationID	words in evaluation	class label
e1	`good teacher interesting lectures`	Positive
e2	`good lectures excellent course`	Positive
e3	`bad teacher discontinue course`	Negative
e4	`boring lectures teacher incompetent`	Negative

(a) Estimate $P($`good` | Positive), $P($`good` | Negative), $P($`teacher` | Positive), and $P($`teacher` | 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 given data set. Give the probability of the Positive class according to this model for the evaluation text: `very good teacher`.

Question 6: Undirected Graphical Models (10 points)

Consider the undirected graphical model with independence graph as given below:



(a) Derive a formula for the maximum likelihood fitted counts using the applicable margin constraints and the conditional independencies encoded in the graph. Clearly justify each step in your proof. Note: the proof should be from first principles; application of the RIP ordering algorithm does not count as a proof.

(b) Suppose we remove the edge $B - C$ from the graph. Give a formula for the maximum likelihood fitted counts of the resulting graphical model, if such a formula exists (no proof from first principles required). Otherwise, explain why there is no such formula.