Exam Data Mining
Date: 10-11-2021 Time: 19.00-22.00

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 final answers.
4. This exam contains five questions for which you can earn 100 points.

Question 1: Multiple Choice (20 points)

1. (Bias-Variance decomposition) Which of the following statements about bias and variance are true? (any number from 0 to 4 can be true!)
   (a) A pruned classification tree has smaller bias than an unpruned one.
   (b) The variance component of error gets smaller as the training set size increases.
   (c) Bagging reduces the variance.
   (d) The bias component of error gets smaller as the training set size increases.
2. (Graphical Models) Consider the undirected graphical model with independence graph $A - B - C - D$. The formula for the maximum likelihood fitted counts is:
   (a) $\hat{n}(A, B, C, D) = n(A, B, C, D)$.
   (b)
   $$\hat{n}(A, B, C, D) = \frac{n(A, B)n(B, C)n(C, D)}{n(B, C)}.$$
   (c)
   $$\hat{n}(A, B, C, D) = \frac{n(A, B)n(B, C)n(C, D)}{n(B)n(C)}.$$
   (d) This model has no closed-form solution for the maximum likelihood fitted counts.

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

x_1 \ x_2	0	1	Total
0	10	10	20
1	10	10	20
Total	20	20	40

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) The saturated model.

4. (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%.

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

(a) All maximal frequent item sets are closed.

(b) Every frequent item set is a subset of a maximal frequent item set.

(c) Every database transaction, regarded as an item set, is closed.

(d) An item set has the same closure as any of its subsets with the same support.

Question 2: Classification Trees (20 points)

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

x	2	3	5	5	6	8	8	8	9	9
y	0	0	0	0	0	0	1	1	1	1

The class label can take on two different values, coded as 0 and 1. We use the gini-index as impurity measure. The best split is the one that maximizes the impurity reduction.

(a) List all the splits on x that are allowed by the algorithm that was discussed during the course.

(b) For which of the splits that you listed under (a) do we need to compute the impurity reduction in order to determine the best split? (don't list any more splits than strictly necessary)

(c) Give the impurity reduction of the best split.

(d) In the lectures on classification trees we have discussed algorithms that only make binary splits. Some classification tree algorithms, however, make a separate child node for each possible value of a categorical variable. The impurity reduction of a split s in node t generalizes in the obvious way:

$$\Delta i(s,t) = i(t) - \sum_{j=1}^{J} \pi(j) i(j),$$

where $\pi(j)$ is the proportion of cases in node t with $X = j$, and $i(j)$ is the impurity of the child node for the cases in node t with $X = j$. Argue that such a split has an impurity reduction at least as high as the impurity reduction of any binary split in node t on the same attribute. Assume we use the gini-index to measure the impurity of a node. You may use $J = 3$ to make your argument.

Question 3: Frequent Item Set Mining (20 points)

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

tid	items
1	ABC
2	AB
3	BCD
4	CDE
5	CD
6	$ABCD$

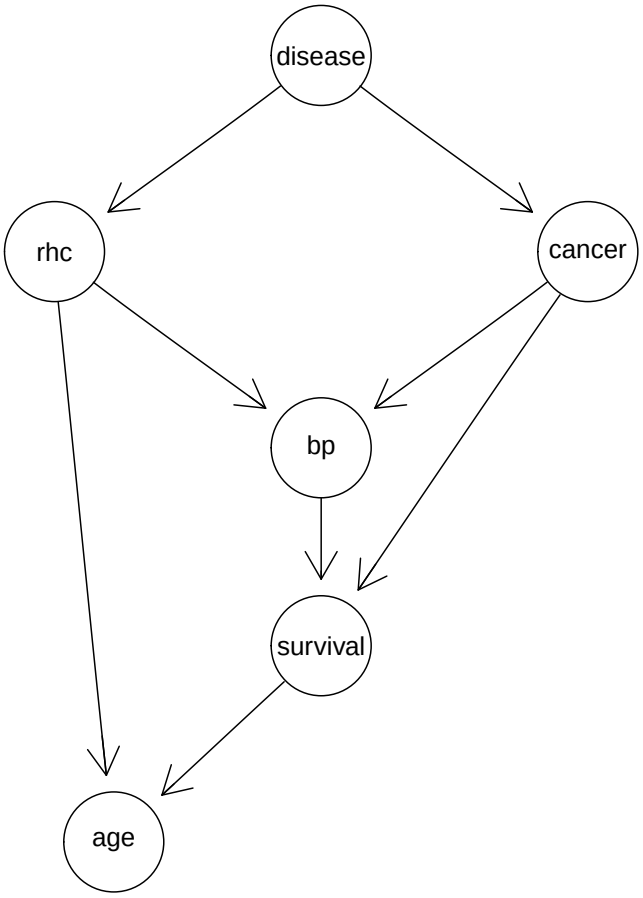
Use the Apriori-close algorithm to compute all closed frequent item sets, and their support, with minimum support 2. Do this in the following two steps:

(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 are learning the structure of a Bayesian network on the data of an intensive care unit. The variables are "disease" (9 disease categories), "cancer" (yes, no, or metastatic), "age" (5 categories), "bp" (blood pressure: ≤ 85 or > 85), "rhc" (whether or not the surgical procedure right heart catheterization was performed), and "survival" (yes/no). The current model in the search is given in the graph below:



(a) Does the conditional independence disease $\perp\!\!\!\perp$ survival | {cancer, bp, age} hold in the given model?

(b) Which edges of the given model become bi-directional in the essential graph? Turn the graph into blocks and cyclic.

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

disease	survival	rhc	cancer	age	bp
-9435	-3570	-3593	-3086	-8862	-3538

The counts on the training data for "cancer" (no, yes, metastatic), "bp" (≤ 85, > 85), and "rhc" (no, yes) are given in the following tables:

rhc = no	≤ 85	> 85	Total
no	1208	1444	2652
yes	218	420	638
metastatic	67	194	261
Total	1493	2058	3551

rhc = yes	≤ 85	> 85	Total
no	391	1336	1727
yes	63	271	334
metastatic	28	95	123
Total	482	1702	2184

Use the natural logarithm (ln) in your computations.

(c) What is the change in log-likelihood score if we delete the edge rhc $\rightarrow$ bp ? (round your answer to the nearest integer)

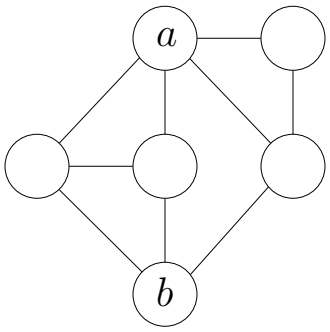
(d) How many parameters are removed from the model if we delete the edge rhc $\rightarrow$ bp ?

Question 5: Link Prediction (15 points)

In a friendship network, nodes represent persons, and two nodes are connected by an edge if the persons are friends on Facebook. We want to predict whether two persons will become Facebook friends in the future (yes=1, no=0). As a predictor we use the number of shared neighbors. A logistic regression fitted on the data has coefficient $\hat{\beta}_0 = -4$ and $\hat{\beta}_1 = 1.5$, where $\hat{\beta}_0$ is the intercept, and $\hat{\beta}_1$ the coefficient of "number of shared neighbors".

(a) Does the sign of $\hat{\beta}_1$ conform to commons sense? Explain.

Consider the mini friendship network given below:



(b) According to the fitted model, what is the probability that a and b will become Facebook friends in the future?

(c) Give one *path based* feature commonly used for link prediction. Compute its value for the pair (a, b).

(d) In link prediction, we make predictions for *pairs* (x, y) of nodes rather than for *single* nodes. Likewise, the features defined tend to measure some property for pairs of nodes. What could be the problem if some feature F is not symmetrical, that is, if $F(x, y) \neq F(y, x)$?