# Exam Data Mining
Date: 7-11-2018
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) All maximal frequent item sets are closed.
2. (Frequent sequence mining) "AI" occurs 7 times as a subsequence of "ARTIFICIAL INTELLIGENCE".
3. (Frequent item set mining) Every transaction, regarded as an item set, is closed.
4. (Classification trees) If we use the Gini-index as impurity measure, then the impurity reduction of the worst possible split may be negative.
5. (Bayesian networks) Two directed independence graphs are equivalent if they have the same skeleton and the same moral graph.
6. (Link-based classification) In link-based classification, we aim to exploit the fact that linked individuals tend to be similar to each other.

7. (Active Learning) Based on the data already labeled, the following conditional probabilities have been estimated

$$\hat{P}(Y = 1 \mid X = x_1) = 0.45 \quad \hat{P}(Y = 1 \mid X = x_2) = 0.60 \quad \hat{P}(Y = 1 \mid X = x_3) = 0.20,$$

   where $x_1$, $x_2$, and $x_3$ are three unlabeled points, and $Y$ is a binary class label.

   From these three points, confidence-based uncertainty sampling would pick $x_3$ as the next query point.

8. (Random forests) Each tree 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 bias component of expected prediction error decreases.

10. (Probability) Suppose that for each of the first 10 questions of this exam one tosses a fair coin, and answers "A" if the coin lands heads, and "B" if the coin lands tails. Let $X_i = 1$ if the answer to question $i$ is correct, and $X_i = 0$ otherwise. Let $T = \sum_{i=1}^{10} X_i$.

    If this experiment is repeated indefinitely, then the expectation of $T$, $\mathbb{E}(T) = 5$. Furthermore, the variance of $T$, $\mathbb{V}(T) = 2\frac{1}{2}$.

### Question 2: Classification Trees (20 points)

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

| $x$ | A | A | B | B | B | B | C | D | D | D |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |

The class label can take on two different values, coded as 0 and 1, and $x$ can take on four different values, coded as A, B, C, and D.

We use the gini-index as impurity measure. The optimal 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 optimal split?
(don't list any more splits than strictly necessary)

(c) What is the impurity reduction of the split that sends all cases with $x \in \{B, C\}$ to one child node, and all cases with $x \in \{A, D\}$ to the other child node?

## Question 3: Frequent Pattern Mining (20 points)

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

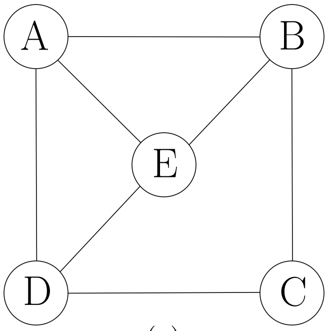
| tid | items |
|---|---|
| 1 | $AB$ |
| 2 | $ABC$ |
| 3 | $ABD$ |
| 4 | $ABE$ |
| 5 | $AE$ |
| 6 | $AC$ |
| 7 | $AD$ |
| 8 | $BC$ |

Use the Apriori algorithm to compute all frequent item sets, and their support, with minimum support 2. To generate the (pre-)candidates, use the alphabetical order on the items. For each level give:

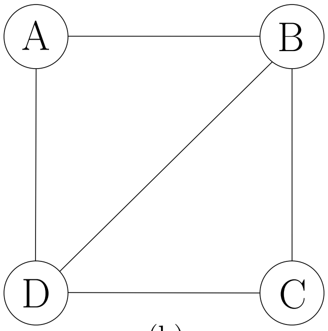
1. A table with the candidate frequent item sets, their support, and a check mark ✓ if the item set is frequent, and
2. A table with the pre-candidates that do not need to be counted on the database.

## Question 4: Undirected Graphical Models (20 points)

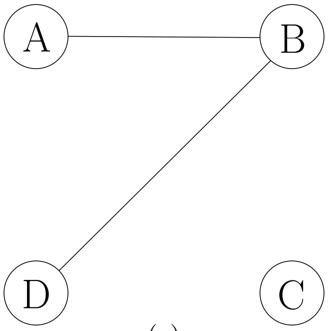
Consider the graphical log-linear models with the following independence graphs:



(a)



(b)



(c)

For each of the models (a), (b) and (c), answer the following two questions:

1. Give a formula for the maximum likelihood fitted counts, if one exists. You don't need to give a derivation of the formula. Otherwise, explain how you established that no such formula exists.
2. Assuming all variables are binary, how many parameters (u-terms) does the model contain?

**Question 5: Bayesian Networks (20 points)**

The table below shows the numbers of successes (recoveries) and failures for treatments involving both small and large kidney stones, where Treatment A includes all open surgical procedures and Treatment B only involves a small puncture. The total number of observations is $n = 700$.

| $n$(size, treatment, outcome) | | outcome | |
|---|---|---|---|
| size | treatment | success | failure |
| small | A | 81 | 6 |
| | B | 234 | 36 |
| large | A | 192 | 71 |
| | B | 55 | 25 |

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. It is not allowed to remove or reverse edges. We start the search process from the empty graph (the mutual independence model).

(a) Compute the change in log-likelihood score if we add an edge from “size” to “treatment”.

(b) Suppose that adding an edge from “size” to “treatment” gives the biggest increase in BIC score, so the search algorithm moves to this neighbour. For which operations do we have to recompute the change in score in the second iteration? (assume that scores from previous iterations that are still valid are not recomputed, but are retrieved from memory).

Treatment A has a success rate of 81/87=93% in the group of small kidney stones, while treatment B only scores 234/270=87%. Likewise, Treatment A has a success rate of 192/263=73% in the group of large kidney stones, while treatment B only scores 55/80=69%. But overall, treatment B has a success percentage of 289/350=85%, while treatment A scores only 273/350=78%.

(c) Which treatment do you consider to be more effective? Explain.