Exam Data Mining
Date: 8-11-2017, Time: 17.00-20.00

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

Question 1 Short Questions (20 points)

Answer the following questions:

(a) In text mining, list the bigrams that occur in:

```
For your eyes only
```

(b) In frequent sequence mining, how many times does "AI" occur as a subsequence of "DATA MINING"? Give the corresponding mappings ϕ.

(c) In frequent tree mining, give the right-most occurrence (RMO) list of the FREQT algorithm for the occurrence of $T = ab \uparrow c$ as an induced subtree of $D = ab \uparrow ab \uparrow c \uparrow\uparrow c$.

(d) Lu and Getoor describe an algorithm for link-based classification. Their overall prediction rule is:

$$\hat{Y}(x, z) = \arg \max_{k \in 1,\ldots,K} \hat{P}(Y = k \mid x)\hat{P}(Y = k \mid z),$$

where x denotes the "object attributes", z the "link attributes", and K denotes the number of classes. To minimize misclassification error, one should use the rule:

$$\hat{Y}(x, z) = \arg \max_{k \in 1,\ldots,K} \hat{P}(Y = k \mid x, z).$$

Assuming the authors want to minimize misclassification error, can the rule actually used be justified by the two assumptions (1) $X \perp\!\!\!\perp Z$, and (2) $X \perp\!\!\!\perp Z \mid Y$? Explain your answer.

Question 2: Classification Trees (20 points)

Consider the following data on numeric attribute x and class label y. The class label can take on two different values, coded as A and B.

x	8	8	12	12	14	14	16	18	20	20
y	B	B	B	B	A	B	A	A	A	A

We use the gini-index as impurity measure. The optimal split is the one that maximizes the impurity reduction.

(a) Which split(s) do we have to evaluate to determine the optimal one(s)? (don't list any more splits than strictly necessary)

(b) List the optimal split(s). Give the corresponding impurity reduction.

In the seminal work *Classification and Regression Trees*, Breiman et al. state:

> A question that has been frequent among tree users is: which variables are the most important. The critical issue is how to rank those variables that, while not giving the best split of a node, may give the second or third best.

Consider the variable importance measure

$$I(x_j) = \sum_{t \in T} \Delta i(s_j^*, t),$$

where $\Delta i(s_j^*, t)$ denotes the impurity reduction of the best split on x_j in node t, and $I(x_j)$ denotes the overall importance of variable x_j. T denotes the optimal tree selected by a cross-validation or test sample procedure.

(c) Explain how this measure may overestimate the importance of a variable. (Hint: take into account the (dis)similarity of different splits, as used in determining surrogate splits).

Question 3: Frequent Item Set Mining (15 points)

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

tid	items
1	ABE
2	AD
3	ABD
4	AC
5	ACD
6	BCD

Use the Apriori algorithm to compute all frequent item sets, and their support, with minimum support 2. For each level give a table with the candidate frequent item sets, their support, a check mark ✓ if the item set is frequent, and a cross ✗ if the item set is not frequent. Don't list as candidates item sets that do not need to be counted on the database. To generate the candidates, use the alphabetical order on the items.

Question 4: Undirected Graphical Models (25 points)

Judges, probation officers and parole officers are increasingly using algorithms to assess a criminal defendants likelihood of becoming a recidivist, a term used to describe criminals who re-offend. We have data on 6,172 criminal defendants from Broward County, Florida[1]. The data contains information about gender (female or male) of the defendant, and whether or not the defendant re-offended within a two-year period after release. Below you find a cross-table with counts of gender and 2-year-recidivism:

$n(G, R)$	2-Year-Recidivism	
Gender	No	Yes
Female	762	413
Male	2601	2396

(a) Compute the maximum likelihood fitted counts for the independence model $G \perp\!\!\!\perp R$, where G denotes gender, and R denotes 2-year-recidivism.

(b) Test the independence model against the saturated model with significance level $\alpha = 0.05$. Use the natural logarithm in your computations. Clearly state whether or not the model is rejected, and explain how you made that decision. To perform the test, consult the following table with critical values:

degrees of freedom (ν)	1	2	3	4	5	6	7	8
critical value ($\chi^2_{\nu;0.05}$)	3.84	6.00	7.82	9.50	11.1	12.6	14.1	15.5

As it turns out, also the crime degree (felony or misdemeanor) has been recorded, where a felony is considered to be more serious than a misdemeanor. The table of observed counts is given below.

$n(C, G, R)$		2-Year-Recidivism	
Crime	Gender	No	Yes
Felony	Female	401	283
Felony	Male	1585	1701
Misdemeanor	Female	361	130
Misdemeanor	Male	1016	695

(c) Draw the undirected independence graph of the graphical model expressing the constraint $G \perp\!\!\!\perp R \mid C$, where C denotes Crime, and state the corresponding independence assumption(s) in words.

(d) Compute the fitted counts for the model given under (c).

(e) Test the model you fitted under (d) against the saturated model, using $\alpha = 0.05$. Clearly state whether or not the model is rejected, and explain how you made that decision.

[1] (see `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`)

Question 5: Bayesian Networks (20 points)

We perform a greedy hill-climbing search to find a good Bayesian network structure on 4 binary variables denoted A, B, C, and D. Neighbour models are obtained by adding, deleting, or reversing an edge. We start the search process from the empty graph (the mutual independence model). In step 1 of the search we find that adding the edge $A \to B$ gives the biggest improvement in the BIC score. It is given that:

1. scores of operations (addition, deletion, or reversal of an edge) computed in previous iterations that are still valid are not recomputed, but retrieved from memory, and
2. scores of operations that produce a model that is equivalent to a model that has already been scored in a previous iteration are not recomputed, but are retrieved from memory as well.

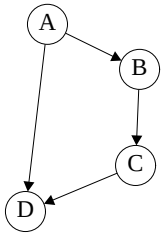
Answer the following questions:

(a) For each of the following operations, state whether or not we need to compute (as opposed to retrieve from memory) the change in score in step 2 of the search:

1. add($C \to B$)
2. add($B \to A$)
3. reverse($A \to B$)
4. add($B \to D$)
5. add($D \to B$)
6. remove($A \to B$)

(b) Pick one of the operations listed under (a) for which you answered "yes", and give the formula for computing the change in log-likelihood score for that particular operation. Write n for the number of observations, and (for example) $n(B = 0, D = 1)$ for the number of observations with $B = 0$ and $D = 1$.

(c) Suppose the final model produced by the hill-climbing algorithm is:



Give the essential graph of this model.

(d) Does $A \perp\!\!\!\perp C \mid B$ hold in the final model?

(e) How many parameters does the final model have?