# Written Exam Data Mining
# Date: 6-11-2012
# Time: 9.00-12.00

## General Remarks

1. You are allowed to consult 1 A4 sheet with notes written on both sides.
2. You are allowed to use a pocket calculator. Use of mobile phones is not allowed.
3. Always show how you arrived at the result of your calculations.
4. You may answer in Dutch or English.
5. There are five questions, for which you can score a total of 100 points.

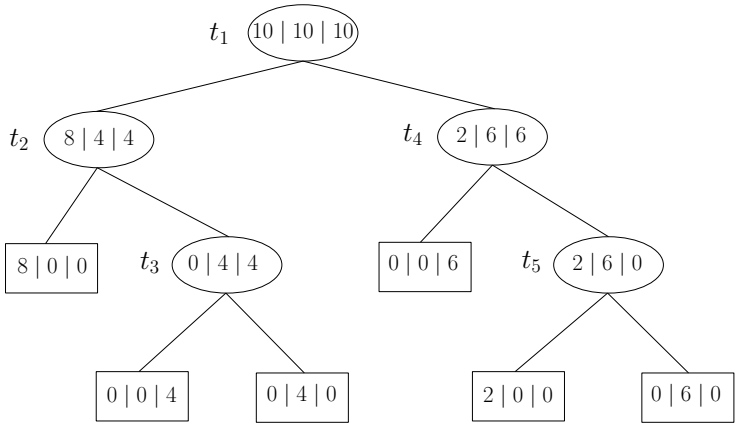
## Question 1: Classification Trees (15 points)

(a) Consider the following data on numeric attribute $x$ and binary class label $y$:

| $x$ | 8 | 11 | 12 | 12 | 12 | 15 | 15 | 15 | 18 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |

We use the gini-index as impurity measure. What is the optimal split on $x$, and what is the impurity reduction of that split?

(b) The tree given below, denoted by $T_{\max}$, has been constructed on the training sample:




In each node, the number of observations with class 1 is given in the left part, the number of observations with class 2 is given in the middle part, and the number of observations with class 3 is given in the right part. The leaf nodes have been drawn as rectangles.

Compute the cost-complexity pruning sequence $T_1 > T_2 > \ldots > \{t_1\}$, where $T_1$ is the smallest minimizing subtree for $\alpha = 0$. Also give the corresponding sequence of $\alpha$ values.

# Question 2: Undirected Graphical Models (25 points)

Consider the graphical log-linear model $M$:

$$\log P(x_1, x_2, x_3) = u_\emptyset + u_1 x_1 + u_2 x_2 + u_3 x_3 + u_{23} x_2 x_3,$$

where $x_i \in \{0, 1\}$ for $i = 1, 2, 3$. All questions are about model $M$.

(a) Draw the independence graph.

(b) Use the factorization criterion for independence to show that $X_1 \perp\!\!\!\perp (X_2, X_3)$ (or if you prefer coordinate projection notation: $X_{\{1\}} \perp\!\!\!\perp X_{\{2,3\}}$).

(c) Prove that the maximum likelihood fitted counts are given by

$$\hat{n}(x_1, x_2, x_3) = \frac{n(x_1) n(x_2, x_3)}{N},$$

where $N$ denotes the total number of observations. Clearly justify each step of your proof.

(d) We observe the following data:

| $n(x_1, x_2, x_3)$ | | $x_3$ | |
|---|---|---|---|
| $x_1$ | $x_2$ | 0 | 1 |
| 0 | 0 | 12 | 20 |
| | 1 | 32 | 6 |
| 1 | 0 | 8 | 10 |
| | 1 | 8 | 4 |

Compute the deviance of $M$ on this data set.

(e) Test $M$ against the saturated model using $\alpha = 0.05$. Use the following table with critical values:

| degrees of freedom ($\nu$) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| critical value ($\chi^2_{\nu;0.05}$) | 3.84 | 6.00 | 7.82 | 9.50 | 11.1 | 12.6 | 14.1 | 15.5 | 17.0 | 18.3 |

Clearly state whether or not $M$ is rejected, and explain how you made that decision.

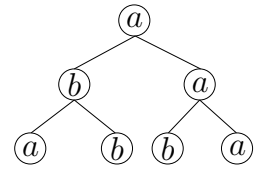
# Question 3: Frequent Pattern Mining (25 points)

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

| tid | items |
|---|---|
| 1 | $AB$ |
| 2 | $ADE$ |
| 3 | $CE$ |
| 4 | $BCD$ |
| 5 | $ABDE$ |

(a) Use the Apriori algorithm to compute all frequent item sets, and their support, with minimum support 2. Clearly indicate the steps of the algorithm, and the pruning that is performed.

The following questions are about frequent tree mining. Consider the labeled ordered tree $d_1$:



In the questions we use the following string representation of labeled ordered trees: we list the node labels according to pre-order (depth-first) traversal, and use the special symbol $\uparrow$ to indicate that we go up one level in the tree. For example, the string representation of $d_1$ is: $aba \uparrow b \uparrow\uparrow ab \uparrow a$.

(b) How many times does the tree $T = ab \uparrow a$ occur as an induced subtree in $d_1$? Give the rightmost occurrence list (RMO-list) of $T$ in $d_1$ as it is maintained by the FREQT algorithm.

(c) How many times does the tree $T = ab \uparrow a$ occur as an embedded subtree in $d_1$? Give the corresponding matching functions (copy the table below on your answer sheet and complete it; the nodes of $T$ have been named $w_1, w_2$ and $w_3$).

| | $w_1$ | $w_2$ | $w_3$ |
|---|---|---|---|
| $\phi_1$ | | | |
| etc. | | | |

(d) The FREQT algorithm uses the right-most extension technique to generate candidate $k+1$-trees from frequent $k$-trees. We use the following definition of support

$$\text{supp}(T, D) = \frac{|\{d \in D \mid T \preceq d\}|}{|D|}$$

Assume the label set is $\Sigma = \{a, b, c\}$, and assume that $d_1$ is frequent in $D$. How many candidate trees will FREQT generate from $d_1$? Explain your answer.

(e) In an alternative tree mining scenario we have one big data tree $d$, and the support of $T$ in $d$ is defined as the number of distinct occurrences of $T$ in $d$ (i.e. the number of distinct matching functions). For level wise search, the so-called anti-monotone property

$$T_1 \preceq T_2 \Rightarrow \text{supp}(T_1, d) \geq \text{supp}(T_2, d)$$

is essential. Assume we are looking for induced subtrees of labeled ordered trees. Does the anti-monotone property hold in this alternative tree mining scenario? If you answer yes, give a compelling argument. If you answer no, give a counterexample.

# Question 4: Bayesian Networks (20 points)

Consider the following data on whether a defendant is convicted to the death penalty ($d$), whether the victim is white ($w$), and whether the defendant is black ($b$).

| $n(w, b, d)$ | | $d$ | |
|---|---|---|---|
| $w$ | $b$ | no | yes |
| no | no | 1 | 1 |
| | yes | 19 | 5 |
| yes | no | 20 | 25 |
| | yes | 11 | 18 |

Consider a heuristic search for a Bayesian Network that maximizes the BIC score

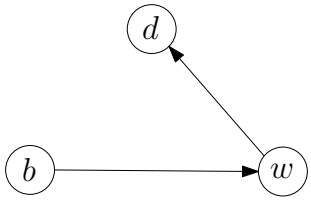
$$\mathrm{BIC}(M) = \mathcal{L}(M) - \frac{\ln N}{2}\dim(M).$$

Here ln denotes the natural logarithm, and $N$ denotes the total number of observations.

The algorithm performs a hill-climbing search where the neighbors of the current model are obtained by either:

1. removing an arrow from the current model
2. adding an arrow to the current model
3. turning an arrow of the current model around

The current model in the search is:




(a) Give all neighbors of the current model, and indicate which neighbors are equivalent to each other. Also indicate which neighbors are equivalent to the current model.

(b) Compute the contribution of node $d$ to the BIC score in the current model. Use the *natural* logarithm in your computations.

(c) Does the saturated model have a better BIC score than the current model? Justify your answer by showing the relevant calculations.

(d) Taking only the marginal table $n(w, d)$ into consideration, would you say there is evidence of racial discrimination in the application of the death penalty? (A qualitative argument is sufficient; you don't need to perform a statistical test).

# Question 5: Subgroup Discovery (15 points)

We are given the following data on age and gender of 10 accepted loan applicants. If the client defaulted on the loan this is coded as $y = 0$, if the client did not default this was coded as $y = 1$. We are using the PRIM algorithm to find groups of clients with a low default risk.

| Record | age | gender | $y$ |
|---|---|---|---|
| 1 | 22 | male | 0 |
| 2 | 46 | male | 0 |
| 3 | 24 | female | 0 |
| 4 | 23 | female | 0 |
| 5 | 29 | male | 0 |
| 6 | 45 | male | 1 |
| 7 | 63 | female | 1 |
| 8 | 36 | female | 1 |
| 9 | 25 | male | 1 |
| 10 | 50 | female | 1 |

(a) We are using the top-down peeling algorithm of PRIM with peeling fraction $\alpha = 0.3$ and minimum support $\beta_0 = 0.3$. For categorical attributes, $\alpha$ is ignored. Give all the peeling actions on age and gender that are considered by PRIM on the given data set. Which of these peeling actions will be chosen?

(b) Continue peeling until the stopping criterion of the top-down peeling algorithm is met. Report the resulting box, the average of $y$ in the box, and the support of the box.