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

Question 1: Classification Trees (15 points)

(a) Knowledge of theory can save you a lot of computation! The optimal split can only occur between two segments, where a segment is a collection of consecutive x-values for which the class distribution is identical. Hence, only $x \leq 11.5$ and $x \leq 16.5$ can be optimal (we assume the split is always halfway between two consecutive x-values). The first split produces child nodes with class distributions $2 \mid 0$ and $4 \mid 4$. The second produces child nodes with class distributions $2 \mid 0$ and $6 \mid 2$. Obviously the second split is better. The impurity reduction achieved is:

$$\Delta i = \frac{24}{100} - \frac{8}{10} \cdot \frac{3}{16} = \frac{9}{100}$$

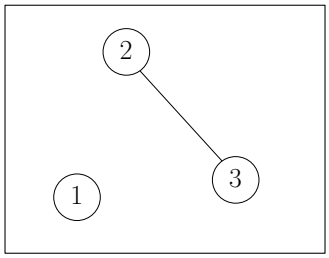
(b) Since we continued splitting until all leaf nodes were pure, we have: $T_1 = T_{\max}$. The subscript of g indicates the iteration of the pruning algorithm:

	t_1	t_2	t_3	t_4	t_5
g_1	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}^*$
g_2	$\frac{3}{20}$	$\boxed{\frac{2}{15}^*}$	$\boxed{\frac{2}{15}^*}$	$\frac{6}{30}$	–
g_3	$\frac{1}{6}^*$	–	–	$\boxed{\frac{6}{30}}$	–

We prune in the nodes with the starred values. The boxed values did not need to recomputed; they were copied from the previous iteration. Summarizing: T_2 is obtained from T_1 by pruning in node t_5. T_3 is obtained from T_2 by pruning in t_3 and t_2, and $T_4 = \{t_1\}$. The α- intervals are: $T_1 : [0, \frac{1}{15})$, $T_2 : [\frac{1}{15}, \frac{2}{15})$, $T_3 : [\frac{2}{15}, \frac{1}{6})$, and $T_4 : [\frac{1}{6}, \infty)$.

Question 2: Undirected Graphical Models (25 points)

(a) Independence graph:



(b) The factorization criterion states that random vectors X and Y are independent if and only if there are functions $g(x)$ and $h(y)$ such that for all values (x, y):

$$\log P(x, y) = g(x) + h(y)$$

To prove that $X_1 \perp\!\!\!\perp (X_2, X_3)$ we have to show that there are functions $g(x_1)$ and $h(x_2, x_3)$ such that

$$\log P(x_1, x_2, x_3) = g(x_1) + h(x_2, x_3)$$

Pick $g(x_1) = u_\emptyset + u_1 x_1$ and $h(x_2, x_3) = u_2 x_2 + u_3 x_3 + u_{23} x_2 x_3$.

(c) Proof:

$$\hat{P}(x_1, x_2, x_3) = \hat{P}(x_1)\hat{P}(x_2, x_3) \qquad (X_1 \perp\!\!\!\perp (X_2, X_3))$$

$$\hat{n}(x_1, x_2, x_3) = \frac{\hat{n}(x_1)\hat{n}(x_2, x_3)}{N} \qquad \text{(Multiply left and right by } N^2/N\text{; } \hat{n} = N \cdot \hat{P})$$

$$\hat{n}(x_1, x_2, x_3) = \frac{n(x_1)n(x_2, x_3)}{N} \qquad (\hat{n}_a(x_a) = n_a(x_a) \text{ for cliques } a)$$

(d) To compute the fitted counts, we need $n(x_1)$ and $n(x_2, x_3)$. We have $n_1(0) = 12 + 20 + 32 + 6 = 70$, $n_1(1) = 8 + 10 + 8 + 4 = 30$, $n_{23}(0, 0) = 12 + 8 = 20$, $n_{23}(0, 1) = 20 + 10 = 30$, $n_{23}(1, 0) = 32 + 8 = 40$, and $n_{23}(1, 1) = 6 + 4 = 10$.

The fitted counts therefore are:

$\hat{n}(x_1, x_2, x_3)$		x_3	
x_1	x_2	0	1
0	0	14	21
	1	28	7
1	0	6	9
	1	12	3

The deviance is

$$\begin{aligned}\text{dev}(M) &= 2\left[12\ln\frac{12}{14} + 20\ln\frac{20}{21} + 32\ln\frac{32}{28} + 6\ln\frac{6}{7} + 8\ln\frac{8}{6} + 10\ln\frac{10}{9} + 8\ln\frac{8}{12} + 4\ln\frac{4}{3}\right] \\ &= 3.569116\end{aligned}$$

(e) The appropriate degrees of freedom is 3, because 3 u-terms are set to zero (u_{12}, u_{13}, and u_{123}). Hence, the critical value is 7.82. The observed deviance is smaller than the critical value, so the model is not rejected.

Question 3: Frequent Pattern Mining (25 points)

(a) Level 1:

candidate	support	frequent?
A	3	Y
B	3	Y
C	2	Y
D	3	Y
E	3	Y

All 1-itemsets are frequent, so all 2-itemsets are candidates at level 2:

candidate	support	frequent?
AB	2	Y
AC	0	N
AD	2	Y
AE	2	Y
BC	1	N
BD	2	Y
BE	1	N
CD	1	N
CE	1	N
DE	2	Y

All subsets of ABD and ADE are frequent, so these are the candidates at level 3:

candidate	support	frequent?
ABD	1	N
ADE	2	Y

Note that e.g. ABE is not a level 3 candidate, because its level 2 subset BE is not frequent. There are no level 4 candidates.

(b) Two times. The RMO-list is: $(5, 7)$.

(c) Six times. The matching functions are:

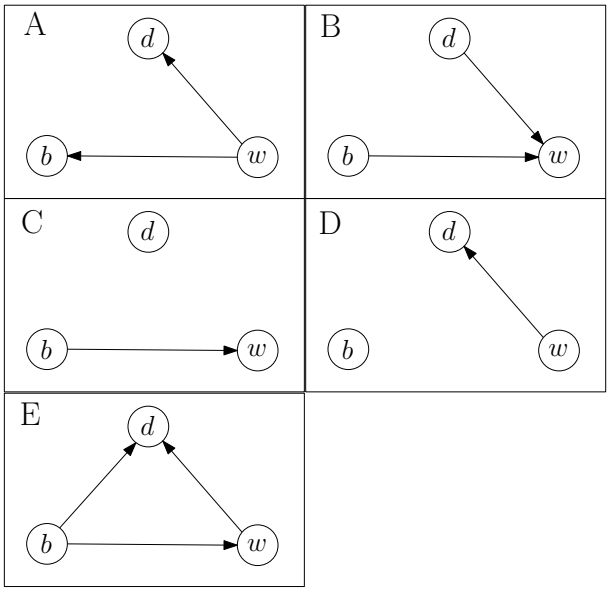
	w_1	w_2	w_3
ϕ_1	v_1	v_2	v_5
ϕ_2	v_1	v_2	v_7
ϕ_3	v_1	v_4	v_5
ϕ_4	v_1	v_4	v_7
ϕ_5	v_5	v_6	v_7
ϕ_6	v_1	v_6	v_7

(d) It depends on whether the label c is frequent or not (a and b must be frequent since d_1 is frequent). If c is frequent, then d_1 generates 9 candidates, otherwise 6 candidates. Each node on the path from the root to the rightmost leaf can get a node with any frequent label as its rightmost child.

(e) No, it does not. Consider a tree $d = ab \uparrow b$. We have $a \preceq ab$, but a has a support of 1 in d, and ab has a support of 2.

Question 4: Bayesian Networks (20 points)

(a) The neighbors are:



None of the neighbors are equivalent. Neighbor A is equivalent to the current model.

(b) The contribution of d to the loglikelihood score is:

$$6 \ln \frac{6}{26} + 20 \ln \frac{20}{26} + 43 \ln \frac{43}{74} + 31 \ln \frac{31}{74} = -64.36092$$

The number of parameters associated with d is 2 parent configurations times one probability per parent configuration is 2 in total. The penalty per parameter is $\frac{\ln 100}{2} = 2.302585$. So the total contribution of d to the BIC score is:

$$-64.36092 - 2 \times 2.302585 = -68.96609$$

(c) The saturated model is model E as given under (a). Compared to the current model, only the parent set of node d is different. The new contribution of d to the loglikelihood score becomes:

$$\ln \frac{1}{2} + \ln \frac{1}{2} + 5 \ln \frac{5}{24} + 19 \ln \frac{19}{24} + 25 \ln \frac{25}{45} + 20 \ln \frac{20}{45} + 18 \ln \frac{18}{29} + 11 \ln \frac{11}{29} = -63.82937$$

This is only a slight improvement compared to the current model, and clearly does not warrant the 2 extra parameters. For completeness, the BIC score of the saturated model is

$$-63.82937 - 4 \times 2.302585 = -73.03971$$

This is worse than the BIC score of the current model.

(d) $\hat{P}(d = \text{yes} | w = \text{no}) = \frac{6}{26} = 0.23$. $\hat{P}(d = \text{yes} | w = \text{yes}) = \frac{43}{74} = 0.58$. If the victim is white, the probability of the death penalty is much higher than if the victim is not white. Hence, one could argue that there is evidence for racial discrimination. Of course we ignored other variables that might explain this difference, so further analysis would be required to verify this claim.

Question 5: Subgroup Discovery (15 points)

(a) To support our calculations, we first construct the following table:

record	1	4	3	9	5	8	6	2	10	7
age	22	23	24	25	29	36	45	46	50	63
y	0	0	0	1	0	1	1	0	1	1

The possible peeling actions are:

Box b	$\bar{y}_{B-b}$
age < 25	$\frac{5}{7}$
age > 45	$\frac{3}{7}$
gender=male	$\frac{3}{5}$
gender=female	$\frac{2}{5}$

The best peeling action is age < 25 because it leads to the highest average in the remaining box.

(b) The possible peeling actions are:

Box b	$\bar{y}_{B-b}$
age < 36	$\frac{4}{5}$
age > 46	$\frac{3}{5}$
gender=male	$\frac{3}{3}$
gender=female	$\frac{2}{4}$

We peel gender=male. Now $\beta \leq \beta_0$, so the peeling stops.

The rule is: age ≥ 25 and gender = female ($\bar{y} = 1$, $\beta = 0.3$).