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

Question 1 Short Questions (20 points)

(a) Bigrams:

1. `For your`
2. `your eyes`
3. `eyes only`

(b) Four times. If you start numbering at 1, and count the space in "DATA MINING", then the mappings are:

	1	2
ϕ_1	2	7
ϕ_2	2	9
ϕ_3	4	7
ϕ_4	4	9

(c) The RMO-list is:(5,6).

(d) The rule can not be justified by the two assumptions given. Let's try:

$$\begin{aligned} P(Y \mid X, Z) &= \frac{P(Y, X, Z)}{P(X, Z)} && \text{(product rule)} \\ &= \frac{P(Y, X, Z)}{P(X)P(Z)} && (X \perp\!\!\!\perp Z) \\ &= \frac{P(Y, X)P(Y, Z)}{P(X)P(Z)P(Y)} && (X \perp\!\!\!\perp Z \mid Y) \\ &= \frac{P(Y \mid X)P(Y \mid Z)}{P(Y)} && \text{(product rule)} \end{aligned}$$

This is as far as we can get, so it appears we need the additional assumption that the class prior probabilities are equal in order to justify their prediction rule.

Question 2: Classification Trees (20 points)

(a) $x \leq 13$, $x \leq 15$.

(b) By argument of symmetry, the two splits have the same impurity reduction. Let's compute the impurity reduction of $x \leq 13$:

$$
\begin{aligned}
i(t) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\
i(\ell) &= 0 \times 1 = 0 \\
i(r) &= \frac{5}{6} \times \frac{1}{6} = \frac{5}{36} \\
\Delta i &= \frac{1}{4} - \frac{4}{10} \times 0 - \frac{6}{10} \times \frac{5}{36} = \frac{1}{6}
\end{aligned}
$$

(c) Suppose that at node t, the best split on x_j (s_j^*) has low similarity to the overall best split s^*, but does rank high on impurity reduction. Because of the low similarity, when t is split in a left and right child by s^*, it may happen that the best split on x_j in those child nodes is close to s_j^*, and has a high impurity reduction as well. Then, essentially the same split has contributed to the variable importance of x_j, not only at t, but also at its child nodes (and possibly their children as well). Thus the importance of x_j as measured by I will be misleadingly high.

Question 3: Frequent Item Set Mining (15 points)

Level 1:

item set	support
A	5 ✓
B	3 ✓
C	3 ✓
D	4 ✓
E	1 ✗

Level 2:

item set	support
AB	2 ✓
AC	2 ✓
AD	3 ✓
BC	1 ✗
BD	2 ✓
CD	2 ✓

Level 3:

item set	support
ABD	1 ✗
ACD	1 ✗

Question 4: Undirected Graphical Models (25 points)

(a) Fitted counts (rounded to the nearest integer):

$\hat{n}(G, R)$	2-Year-Recidivism	
Gender	No	Yes
Female	640	535
Male	2723	2274

(b) The deviance (rounded to the nearest integer) is

$$2\left[762\ln\left(\frac{762}{640}\right)+413\ln\left(\frac{413}{535}\right)+2601\ln\left(\frac{2601}{2723}\right)+2396\ln\left(\frac{2396}{2274}\right)\right]=64$$

The critical value for 1 degree of freedom is 3.84. Since the observed deviance is larger than the critical value, we reject the independence model.

(c) Graph: $G - C - R$. Words: 2-year-recidivism is independent of gender given the crime category.

(d) The fitted counts are (rounded to the nearest integer):

$\hat{n}(C, G, R)$		2-Year-Recidivism	
Crime	Gender	No	Yes
Felony	Female	342	342
Felony	Male	1644	1642
Misdemeanor	Female	307	184
Misdemeanor	Male	1070	641

(e) The deviance (rounded to the nearest integer) is 59. The critical value for 2 degrees of freedom is 6. Since the observed deviance is larger than the critical value, we reject the model $G \perp\!\!\!\perp R \mid C$.

Question 5: Bayesian Networks (20 points)

(a)
1. Yes.
2. No, this would create a cycle so it is not even allowed.
3. No, this model is equivalent to the current model.
4. No, the change in score for this operation is the same as in the previous iteration.

5. Yes.

6. No, this model is the same as the initial model.

(b) We pick: add($C \to B$). The old score is:

$$n_{ab}(0,0) \ln \frac{n_{ab}(0,0)}{n_a(0)} + n_{ab}(0,1) \ln \frac{n_{ab}(0,1)}{n_a(0)} + n_{ab}(1,0) \ln \frac{n_{ab}(1,0)}{n_a(1)} + n_{ab}(1,1) \ln \frac{n_{ab}(1,1)}{n_a(1)}$$

The new score is:

$$\begin{aligned}
&n_{abc}(0,0,0) \ln \frac{n_{abc}(0,0,0)}{n_{ac}(0,0)} + n_{abc}(0,1,0) \ln \frac{n_{abc}(0,1,0)}{n_{ac}(0,0)} \\
&+ n_{abc}(0,0,1) \ln \frac{n_{abc}(0,0,1)}{n_{ac}(0,1)} + n_{abc}(0,1,1) \ln \frac{n_{abc}(0,1,1)}{n_{ac}(0,1)} \\
&+ n_{abc}(1,0,0) \ln \frac{n_{abc}(1,0,0)}{n_{ac}(1,0)} + n_{abc}(1,1,0) \ln \frac{n_{abc}(1,1,0)}{n_{ac}(1,0)} \\
&+ n_{abc}(1,0,1) \ln \frac{n_{abc}(1,0,1)}{n_{ac}(1,1)} + n_{abc}(1,1,1) \ln \frac{n_{abc}(1,1,1)}{n_{ac}(1,1)}
\end{aligned}$$

I miscalculated the number of terms you had to write down. Sorry about that. I changed the notation to make it more compact.

(c) $A \to B$ and $B \to C$ become bidirectional, the other two edges are compelled (their direction is fixed in the equivalence class).

(d) Yes, the relevant moral graph is $A - B - C$, and in that graph A and C are separated by B.

(e) A has 1, B has 2, C has 2, and D has 4, so 9 parameters in total.