Exam Data Mining
Date: 5-11-2015
Time: 13.30-16.30
Answer sketch

Question 1 Short Questions (20 points)

(a) By definition

$$\text{conf}(X \to Y) = \frac{s(X \cup Y)}{s(X)}.$$

When we move an item from the right-hand side to the left-hand side, the denominator ($s(X)$) will decrease, and the numerator ($s(X \cup Y)$) doesn't change. Hence, the confidence will increase.

(b) Counterexample: take two graphs on three nodes, one the full graph, the other a v-structure. Both have the same moral graph, but they are not equivalent.

(c) An induced subtree preserves the parent-child relationship, an embedded subtree only preservers the ancestor-descendant relationship.

(d) The edges between A and B, and between B and C become bi-directional. The other two edges remain as they are.

Question 2: Classification Trees (25 points)

(a) $i(t_1) = \frac{9}{10} \times \frac{1}{10} = \frac{9}{100}$; $i(t_2) = \frac{30}{35} \times \frac{5}{35} = \frac{6}{49}$; $i(t_3) = \frac{60}{65} \times \frac{5}{65} = \frac{12}{169}$.

(b)

$$\Delta i = \frac{9}{100} - \left(\frac{35}{100} \times \frac{6}{49} + \frac{65}{100} \times \frac{12}{169} \right) \approx 0.001$$

(c) $T_1 = \{t_1\}$.

(d) $\{t_1\}$ is the smallest minimizing subtree for $\alpha \in [0, \infty)$.

Question 3: Frequent Sequence Mining (15 points)

We present the answer in tables, like in Apriori.

Level 1:

candidate	support	frequent?
A	3	✓
B	3	✓
C	1	✗
D	1	✗

Level 2:

candidate	support	frequent?
AA	3	✓
AB	2	✓
BA	3	✓
BB	2	✓

Level 3:

candidate	support	frequent?
AAA	1	✗
AAB	1	✗
ABA	2	✓
ABB	2	✓
BAA	2	✓
BAB	1	✗
BBA	1	✗
BBB	0	✗

There are no level 4 candidates, i.e. all level 4 pre-candidates we can make by combining 2 level 3 frequent sequences contain an infrequent subsequence. E.g., pre-candidate $ABAA$ contains infrequent sequence AAA.

Question 4: Undirected Graphical Models (25 points)

(a) $\hat{P}(S = 1|B = 1) = \frac{39}{59} \approx 0.66$ and $\hat{P}(S = 1|B = 0) = \frac{16}{41} \approx 0.39$.

(b) Yes, probability of getting sick when you have eaten a Berehap is bigger than when you have not eaten a Berehap.

(c) Graph: $B - F - S$.

(d) The fitted counts are:

$\hat{n}(B,F,S)$		S	
B	F	0	1
0	0	22.29	3.71
0	1	3.46	11.54
1	0	7.71	1.29
1	1	11.54	38.46

(e) The deviance is 0.22. Since $0.22 < \chi^2_{2;0.05} = 6$, the model is not rejected.

Question 5: Bayesian Networks (15 points)

(a) Every operation that changes the parent set of D: add($C \to D$), add($B \to D$), delete($A \to D$), and reverse($A \to D$).

(b) We only look one step ahead. Deleting the edge may be bad, so we never get the opportunity to add it in the opposite direction.