Exam Data Mining
November 4, 2020, 17.00-20.00 hours
Short answers

Question 1: Mixed Short Questions (20 points)

(a) To prevent overfitting.

(b) A random forest picks the best split from a randomly selected subset of the attributes in each node. Bagging picks the best split from all available attributes.

(c) 2 times.

(d)
```
never say
say never
never again
```

(e) Problem: The link-attributes can not be computed because all node labels are unknown. Solution: Initial labels are predicted using only the object-attributes. Once we have predicted labels, the link-attributes can be computed. Label prediction and computation of link-attributes is iterated until convergence.

Question 2: Classification Trees (20 points)

(a)
$$i(t_1) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}, \quad i(t_2) = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}, \quad i(t_3) = \frac{4}{7} \times \frac{3}{7} = \frac{12}{49}$$

(b)
$$\Delta\, i = \frac{1}{4} - \left(\frac{3}{10} \times \frac{2}{9} + \frac{7}{10} \times \frac{12}{49}\right) = \frac{1}{84}$$

(c) Let SMS denote the smallest minimizing subtree.

1. $T_1 = T_{\max}$ is the SMS for $\alpha \in [0, 0.06)$.
2. Prune in t_2 to obtain T_2, which is the SMS for $\alpha \in [0.06, 0.17)$.
3. The root node is the SMS for $\alpha \geq 0.17$.

Question 3: Frequent Sequence Mining (20 points)

Level 1:

Candidate	Support	Frequent?
M	3	✓
N	3	✓
O	3	✓
R	1	✗

Level 2:

Candidate	Support	Frequent?
MM	0	✗
MN	3	✓
MO	3	✓
NM	0	✗
NN	0	✗
NO	1	✗
OM	0	✗
ON	3	✓
OO	3	✓

Level 3:

Candidate	Support	Frequent?
MON	3	✓
MOO	3	✓
OOO	0	✗
OON	2	✓

Level 4:

Candidate	Support	Frequent?
MOON	2	✓

Question 4: Undirected Graphical Models (20 points)

(a) The formula is:

$$\hat{n}(A, B, C, D, E) = \frac{n(A, C)n(B, C)n(D, C)n(E, C)}{n(C)^3}$$

(b) When no marrying of parents is required (there are no "immoralities" or "v-structures"), then the independence properties of the directed graph are identical to those of its undirected version. The directed graph specified doesn't have any v-structures, and its skeleton is identical to the given undirected graph. Hence, the two graphs express exactly the same independence properties.

(c) The BN-factorisation is:

$$P(A,B,C,D,E) = p(C)p(A|C)p(B|C)p(D|C)p(E|C)$$

Plugging in the ML estimates gives

$$\hat{P}(A,B,C,D,E) = \frac{n(C)}{N}\frac{n(A,C)}{n(C)}\frac{n(B,C)}{n(C)}\frac{n(D,C)}{n(C)}\frac{n(E,C)}{n(C)}$$

To obtain fitted counts, we multiply by N and obtain

$$\begin{aligned}\hat{n}(A,B,C,D,E) &= n(C)\,\frac{n(A,C)}{n(C)}\frac{n(B,C)}{n(C)}\frac{n(D,C)}{n(C)}\frac{n(E,C)}{n(C)}\\ &= \frac{n(A,C)n(B,C)n(D,C)n(E,C)}{n(C)^3}\end{aligned}$$

Question 5: Bayesian Networks (20 points)

(a) Adding the edge $A \to D$ changes the parent set of node D, so we need to compute the change in contribution of node D to the loglikelihood score. In the initial model its contribution is:

$$50 \log \frac{50}{100} + 50 \log \frac{50}{100} = -69.3$$

After adding $A \to D$, its contribution is:

$$40 \log \frac{40}{60} + 20 \log \frac{20}{60} + 10 \log \frac{10}{40} + 30 \log \frac{30}{40} = -60.7$$

Hence, the change in loglikelihoodscore is:

$$\Delta\,\mathcal{L} = -60.7 + 69.3 = 8.6$$

(b) Adding the edge $A \to D$ adds one extra parameter to the model. Each additional parameter costs

$$\frac{\log N}{2} = \frac{\log 100}{2} = 2.3$$

Hence, the change in BIC score is:

$$\Delta\ \mathrm{BIC} = 8.6 - 2.3 = 6.3$$

(c) 2 and 3