Online Exam Data Mining
January 19, 2022, 17.00-20.00 hrs
Short Answers

Question 1: Multiple Choice (16 points)

1. (4 pts)

   (a) In the bag-of-words representation, the order of words in a document is ignored.

   (c) A sentence with n words, all of them different, contains $n-1$ bigrams.

2. (4 pts)

   (c) The purpose of bagging is to reduce the variance component of error.

3. (4 pts)

   (b) $\ln P(x_1, x_2) = u_\emptyset + u_1 x_1 + u_2 x_2$.

4. (4 pts)

   (b) "AI" occurs 4 times as a subsequence in "DATA MINING" (there are 4 different mappings).

   (d) The FREQT algorithm for mining frequent induced subtrees may produce candidates for level $k+1$ that have a level k induced subtree that is infrequent.

Question 2: Classification Trees (28 points)

(a) (4 pts) $i(t_3) = \frac{10}{55} \times \frac{45}{55} = \frac{18}{121}$. $i(\ell) = \frac{10}{25} \times \frac{15}{25} = \frac{6}{25}$. $i(r) = 0$.
$\Delta i = \frac{18}{121} - \frac{25}{55} \times \frac{6}{25} \approx 0.04$.

(b) (8 pts) $T_1 = T(\alpha = 0)$ is obtained by pruning T_{max} in t_3. Then we get $g(t_1) = \frac{2}{10}$ and $g(t_2) = \frac{1}{20}$, so we obtain T_2 by pruning T_1 in t_2. Then we get $g(t_1) = \frac{7}{20}$.

Final answer:
T_1 is the SMS for $\alpha \in [0, \frac{1}{20})$
T_2 is the SMS for $\alpha \in [\frac{1}{20}, \frac{7}{20})$
$T_3 = \{t_1\}$ is the SMS for $\alpha \in [\frac{7}{20}, \infty)$

(c) (4 pts) Answer:

$$\sum_{j=1}^{J} p(j|t)(1 - p(j|t))$$

This is the gini-index.

(d) (4 pts) $g(t)$ is defined as:

$$g(t) = \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}$$

Since $R(T_t) \geq 0$, the numerator is $\leq R(t)$. Furthermore, $|\tilde{T}_t| \geq 2$, so the denominator ≥ 1. It follows that $g(t) \leq R(t)$.

(e) (4 pts) P.M.

(f) (4 pts) Suppose we grow the tree until $R(T) = 0$ to obtain T_{max}. The SMS for $\alpha = \alpha_{min}$ is then obtained by continuing to prune as long as $\min g_k(t) \leq \alpha_{min}$, where we take the minimum over all non-terminal nodes of T_k.

Since $g(t) \leq R(t)$, if $R(t) < \alpha_{min}$ it follows that $g(t) < \alpha_{min}$, so we will have pruned in all nodes with $R(t) < \alpha_{min}$ by the time we arrive at the SMS for $\alpha = \alpha_{min}$. Hence, there is not point in expanding such nodes further when growing the tree.

Question 3: Frequent Item Set Mining (16 points)

(a) (10 pts)

```
Level 1:
A:3
B:5
C:5
D:3
E:1 pruned because not frequent

Level 2:
AB:3 pruned because A has same support
AC:3 pruned because A has same support
AD:0 pruned because not frequent
BC:5 pruned because B (and C) has same support
BD:2
```

```
CD:2

There are no level 3 candidates.
```

(b) (6 pts)

Generator	Closure	Support
A	ABC	3
B	BC	5
C	BC	5
D	D	3
BD	BCD	2
CD	BCD	2

Question 4: Bayesian Networks (20 points)

(a) (4 pts) Yes, in the moral graph of the induced subgraph on the ancestral set of {physical, bp, smoke}, the path from physical to bp is blocked by smoke.

(b) (4 pts) No, in the moral graph of the induced subgraph on the ancestral set of {physical, bp, smoke, mental}, physical and bp are directly connected.

(c) (8 pts) The log-likelihood score of lipo after removal of smoke as a parent becomes:

$$724 \log \frac{724}{1130} + 406 \log \frac{406}{1130} + 337 \log \frac{337}{711} + 374 \log \frac{374}{711} \approx -1230$$

Hence, the change in log-likelihood score is: $-1230 + 1208 = -22$.

(d) (4 pts) Removing smoke as a parent of lipo saves 2 parameters. Each parameter costs $\frac{\log 1841}{2} = 3.76$. So the change in BIC-score is: $-22 + 2 \times 3.76 \approx -14.5$.

Question 5: Multinomial Naive Bayes for Text Classification (10 points)

(a) (5 pts) $P(\texttt{good} \mid \text{Positive}) = \frac{3}{18}$, $P(\texttt{good} \mid \text{Negative}) = \frac{1}{18}$, $P(\texttt{teacher} \mid \text{Positive}) = \frac{2}{18}$, and $P(\texttt{teacher} \mid \text{Negative}) = \frac{3}{18}$.

(b) (5 pts) `very` should be ignored as it does not occur in the training set.

$$P(\text{Positive} \mid \texttt{good teacher}) = \frac{\frac{3}{18} \times \frac{2}{18} \times \frac{1}{2}}{\frac{3}{18} \times \frac{2}{18} \times \frac{1}{2} + \frac{1}{18} \times \frac{3}{18} \times \frac{1}{2}} = \frac{2}{3}$$

Question 6: Undirected Graphical Models (10 points)

(a) (6 pts) The margin constraints are (fitted=observed for margins corresponding to cliques):

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

Furthermore, we will use the conditional independence:

$$A \perp\!\!\!\perp D, E \mid B, C$$

The derivation:

$$\begin{aligned}
\hat{P}(A, B, C, D, E) &= \hat{P}(A \mid B, C, D, E)\hat{P}(B, C, D, E) && \text{(product rule)} \\
&= \hat{P}(A \mid B, C)\hat{P}(B, C, D, E) && \text{(conditional independence)} \\
&= \frac{\hat{P}(A, B, C)\hat{P}(B, C, D, E)}{\hat{P}(B, C)} && \text{(product rule)}
\end{aligned}$$

Finally, multiply both sides by $n = n^2/n$ to get fitted counts, and use the margin constraints to replace fitted counts by observed counts on the right-hand side:

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

(b) (4 pts) Removing $B - C$ creates a chordless 4-cycle, so the resulting model does not have a closed form solution.