[20231106] INFOMDM - Data mining - 1-GS - USP

Course: BETA-INFOMDM Data Mining (INFOMDM)

Duration:	2 hours and 30 minutes
Number of questions:	10

[20231106] INFOMDM - Data mining - 1-GS - USP

Course: Data Mining (INFOMDM)

Number of questions: 10

1 Classification Trees: Computing Splits

As we are growing a classification tree, we encounter a node that contains the following data on categorical attribute x and class label y:

x	A	A	B	B	B	B	C	D	D	D
y	1	1	0	0	0	1	0	0	1	1

We use the gini-index as impurity measure.

Answer the following questions:

8 pt. **a.** (a) If the algorithm uses the optimization that is possible for binary classification problems, then for which of the splits below does it compute the impurity reduction? (1 or more answers may be correct)

a. $x \in \{A\}$

b. $x \in \{A,B\}$

c. $x \in \{B\}$

d. $x \in \{A,B,D\}$

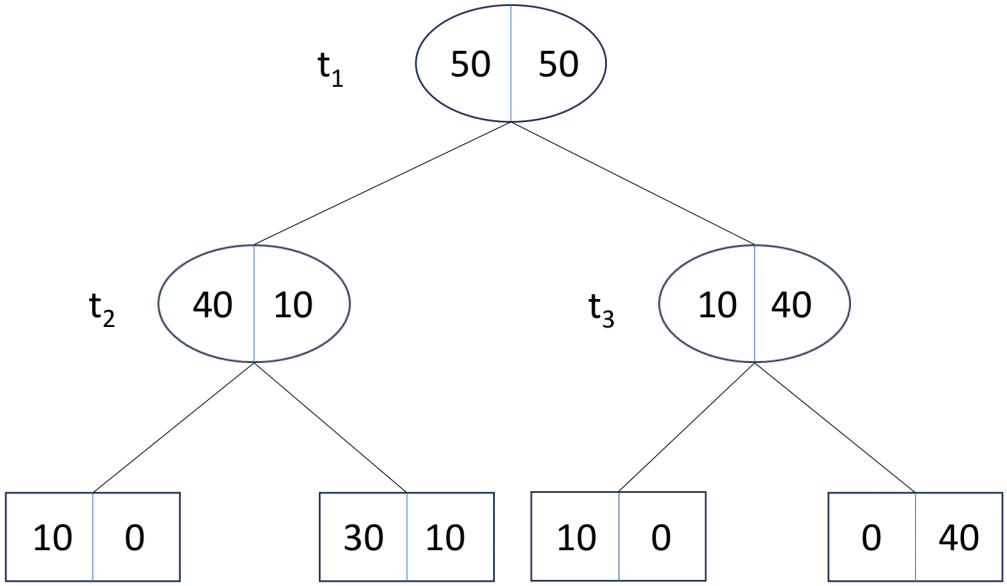
e. $x \in \{A,D\}$

(b) Consider the split $x \in \{A,D\}$. The impurity reduction of this split is:

b. ..... (8 pt.)

2 Classification Trees: Cost-complexity Pruning

Consider the classification tree T_{max} given below. The number of training cases with class A are given in the left part of each node, and the number of training cases with class B in the right part. Leaf nodes are displayed as rectangles.



Answer the following questions:

4 pt. **a.** (a) The smallest minimizing subtree of T_{max} for $\alpha_1 = 0$ is obtained by pruning in:

a. t_1

b. t_2

c. t_3

d. t_2 and t_3

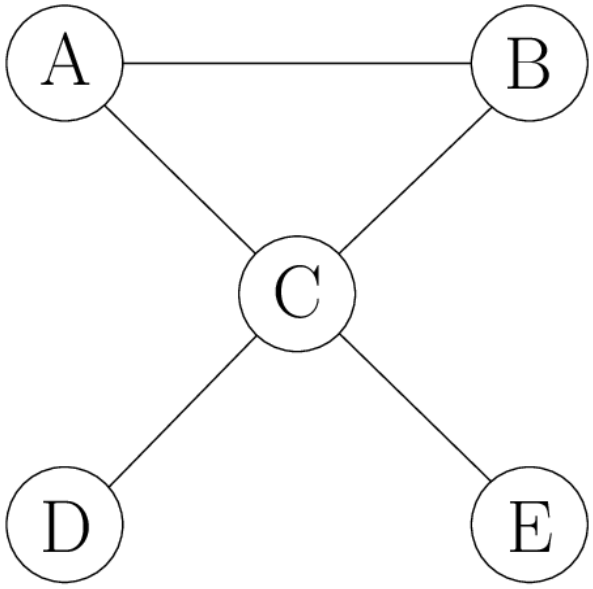
(b) The value of α_2 = **b.** ..... (6 pt.)

(c) In a binary classification problem, what is the smallest value of α for which the root node is guaranteed to be the smallest minimizing subtree?

c. ..... (5 pt.)

3 Undirected Graphical Models

Consider a graphical log-linear model on binary variables A,B,C,D, and E, with the following independence graph:



Answer the following questions:

5 pt. **a.** (a) The maximum likelihood fitted counts for this model are given by:

a. $n(A,B)n(A,C)n(B,C)n(C,D)n(C,E) / n(C)^3$

b. $n(A,B,C)n(C,D)n(C,E) / n(C)^2$

c. $n(A,B,C)n(C,D)n(C,E) / 2n(C)$

d. This model does not have a closed form solution for the maximum likelihood fitted counts.

(b) The number of u-terms of this model is:

b. ..... (5 pt.)

4
5 pt.

Bayesian Networks

To find a good Bayesian network structure on four variables A, B, C, D we perform a hill-climbing local search starting from the empty graph (the mutual independence model). Neighbor models are obtained by adding an edge to the current model. Models are scored using BIC. In iteration 1 of the search we compute the Δ scores of all possible operations in the initial model.

Suppose we find that Δ add (B → C) is largest, so in iteration 1, we add an edge from B to C. Assume that Δ scores of operations that have been computed in previous iterations and that are still valid, are not recomputed, but retrieved from memory. All other Δ scores must be computed. For which of the following operations do we need to compute the Δ score in iteration 2? (1 or more answers may be correct)

a. add (A → C)

b. add (B → D)

c. add (A → B)

d. add (D → C)

5

Bayesian Networks

Consider the following table of observed counts on binary variables x and y.

n(x,y)	y=0	y=1
x=0	32	48
x=1	8	12

We fit the independence model $x \perp y$ to this data using maximum likelihood estimation.

(a) Fill in the fitted count for the cell x=0,y=0:

a. ..... (2 pt.)

Fill in the fitted count for the cell x=0,y=1:

b. ..... (2 pt.)

Fill in the fitted count for the cell x=1,y=0:

c. ..... (2 pt.)

Fill in the fitted count for the cell x=1,y=1:

d. ..... (2 pt.)

(b) Consider the Bayesian network x → y. What is the BIC score of that model minus the BIC score of the independence model? Use the natural logarithm in your computations, and round your final answer to one decimal position:

e. ..... (7 pt.)

6
10 pt.

Text Classification

You are given the following movie reviews with corresponding sentiment:

Words in review	Sentiment
good script great actors	Positive
beautiful images great soundtrack	Positive
great script beautiful costumes	Positive
bad script horrible dialogues	Negative
bad actors beautiful soundtrack	Negative

The estimate of P(beautiful | Positive) according to the multinomial naive Bayes model with Laplace smoothing is:

a. 3/23

b. 2/12

c. 3/19

d. 3/32

7
5 pt.

Frequent Itemset Mining: Apriori

During the execution of the Apriori algorithm we have the following level-2 frequent itemsets: AB, AC, AD, BD, BE, CD, DE.

Which of the following itemsets are level-3 candidates? (1 or more answers may be correct)

a. ABC

b. ACD

c. BDE

d. ABD

e. CDE

f. ADE

8 Frequent Itemset Mining: Closed Frequent Itemsets

Consider the following transactions on items {A,B,C,D,E}:

tid	Items
1	ABC
2	ABC
3	ABC
4	BCD
5	BCD
6	DE
7	B
8	C

We use the Apriori-Close (A-Close) algorithm to find all closed frequent itemsets with minimum support of 2.

5 pt. **a.** Which of the following itemsets are level-2 generators? (1 or more answers may be correct)

- **a.** AB
- **b.** AC
- **c.** AD
- **d.** BC
- **e.** BD
- **f.** CD

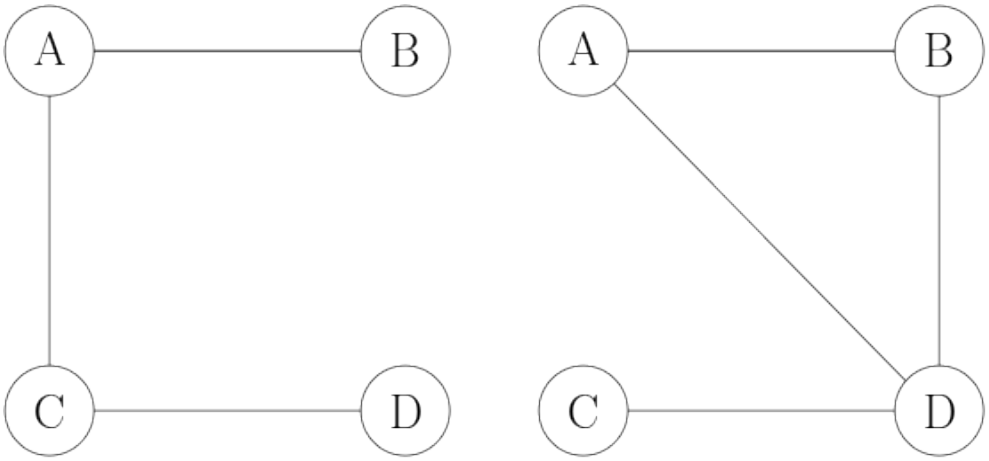
5 pt. **b.** Which of the following are closed frequent itemsets? (1 or more answers may be correct)

- **a.** ABC
- **b.** BD
- **c.** BC
- **d.** CD
- **e.** E
- **f.** BCD

9
6 pt.

Social Network Mining: Link Prediction

The network on the left represents authors (nodes) that wrote papers together (links) in 2021 (the training period), and the network on the right represents the same information for the year 2022 (the testing period). We model this as a binary classification problem according to the method described in: Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki:
Link prediction using supervised learning, SDM Workshop on Link Analysis, 2006, as discussed during the lecture on link prediction.



Which of the following node pairs become positive examples in the supervised learning problem? (1 or more answers may be correct)

a. AD

b. AB

c. BD

d. AC

10
8 pt.

Sequence Mining: Completeness of GSP

Suppose that the GSP algorithm for frequent sequence mining has found all frequent sequences of length up to and including k. Show that GSP candidate generation will not miss any frequent sequences of length k+1. Clearly indicate the steps in your argumentation, and clearly state which properties of frequent sequence mining and the GSP algorithm you use in your reasoning.