[20241107] INFOMDM - Data mining - 1-GS - USP

Course: BETA-INFOMDM Data Mining (INFOMDM)

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

[20241107] INFOMDM - Data mining - 1-GS - USP

Course: Data Mining (INFOMDM)

Number of questions: 10

1 TRUE OR FALSE?

Answer the following TRUE/FALSE questions:

2 pt. **a.** (Frequent sequence mining) “ai” occurs 2 times as a subsequence of “a giant mind” (there are 2 different mappings).

- **a.** TRUE
- **b.** FALSE

2 pt. **b.** (Classification trees) In a binary classification problem, if we use resubstitution error as impurity measure, then the impurity reduction of a split is zero when both child nodes have the same majority class.

- **a.** TRUE
- **b.** FALSE

2 pt. **c.** (Bayesian networks) Two directed independence graphs are (Markov) equivalent if they have the same skeleton and the same moral graph.

- **a.** TRUE
- **b.** FALSE

2 pt. **d.** (Random forests) Each tree in a random forest is allowed to use only a subset of the features.

- **a.** TRUE
- **b.** FALSE

2 pt. **e.** (Text Mining) In the bag-of-words representation, the order of words is ignored.

- **a.** TRUE
- **b.** FALSE

2 pt. **f.** (Frequent Tree Mining) In frequent tree mining with the FREQT algorithm, a candidate frequent tree may contain an infrequent subtree.

- **a.** TRUE
- **b.** FALSE

2 pt. **g.** (Frequent item set mining) All maximal frequent item sets are closed.

- **a.** TRUE
- **b.** FALSE

2 pt. **h.** (Link-based classification) In link-based node classification, the label of each node is assumed to be independent of the labels of the other nodes.

a. TRUE

b. FALSE

2 pt. **i.** (Frequent Sequence Mining) Suppose we adjust the subsequence relation in the GSP algorithm by adding a maximum gap constraint. We say r (a sequence of length m) is a subsequence of s (a sequence of length $n \geq m$) if there is a mapping φ from r to s that satisfies the usual conditions, but we add the condition: $\varphi(i+1) - \varphi(i) \leq 2$, for $i = 1,...,m-1$. Support is defined as usual in the GSP algorithm.

Claim: the resulting subsequence relation is anti-monotone.

a. TRUE

b. FALSE

2 pt. **j.** (Link Prediction) In the article: *Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki: Link Prediction Using Supervised Learning, SDM Workshop on Link Analysis, 2006,* link prediction is modeled as a binary classification problem.

a. TRUE

b. FALSE

2 Classification Trees: Computing Splits

5 pt.

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

x	4	4	8	12	16	16	20	26
y	0	0	0	1	0	0	1	1

We use the gini-index as impurity measure.

If we use the segment borders algorithm to determine the best split on x, we need to compute the impurity reduction of the following splits (1 or more answers may be correct):

a. $x \leq 6$

b. $x \leq 10$

c. $x \leq 14$

d. $x \leq 18$

e. $x \leq 23$

3
5 pt.

Classification Trees: Cost-Complexity Pruning

We are pruning a tree T that has been grown on n=100 training examples. The class variable has 3 possible values, denoted by A, B and C.

Consider a node t in this tree, with the following class distribution:

class	A	B	C
number of examples	8	2	6

The branch T_t of tree T has 5 leaf nodes that are all pure.

What is the critical alpha value g(t) for node t? (round your answer to two decimal places)

The critical value g(t) for node t is: .....

4
5 pt.

Classification Trees: Cost-Complexity Pruning

Consider a classification problem with 3 possible class labels. We compare a tree that consists of just the root node, with the tree T_{max} that has only pure leaf nodes.

For which values of α is the root node *guaranteed* to have the same or lower total cost than T_{max}?

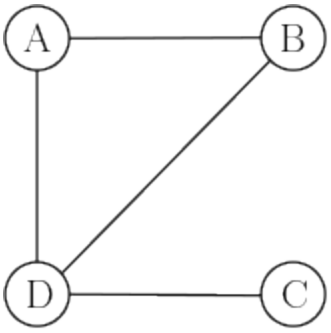
Hint: pick the most unfavorable case for the root node, and the most favorable case for T_{max}.

Round your answer to two decimal places.

The root node is guarantueed to have the same or lower total cost than T_{max} for α ≥ .....

5 Undirected Graphical Models

Consider a graphical log-linear model on binary variables A,B,C and D, 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. $\frac{n(A,B,D)n(C,D)}{n(D)}$

b. $\frac{n(A,B)n(A,D)n(B,D)n(C,D)}{n(D)}$

c. $\frac{n(A,B,D)n(C)}{n(D)}$

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.)

6 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	BC
7	CDE
8	D

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
- **g.** D

7 Text Classification: Multinomial Naive Bayes

You are given the following collection of hotel reviews and corresponding sentiment:

Words in review	Sentiment
large room clean good service	Positive
good service excellent breakfast	Positive
dirty bathroom bad service	Negative
disgusting breakfast noisy	Negative

Answer the following questions:

5 pt. **a.** (a) The estimate of P(good | Positive) according to the multinomial naive Bayes model with Laplace smoothing is:

- **a.** 1/9
- **b.** 1/7
- **c.** 3/25
- **d.** 3/19

5 pt. **b.** (b) According to the multinomial naive Bayes model with Laplace smoothing estimated on the given training set, the probability P(Positive | good noisy) is given by:

- **a.** 1/294
- **b.** 3/625
- **c.** 361/655
- **d.** 1587/2837

8 Logistic Regression

How do people choose between taking the car and taking public transport to go to work? This might be explained by the difference in travel time. Let $y_i = 1$ if person i takes the car to work, and $y_i = 0$ if person i travels to work by public transport. Furthermore, let x_i be the travel time with public transport minus the travel time by car (in minutes) for person i. We use the method of maximum likelihood to estimate a logistic regression model, with the result:

$\beta_0 = -0.4$ and $\beta_1 = 0.05$,

where β_0 is the intercept, and β_1 the coefficient of x.

Answer the following questions:

For questions (a) and (b), round your answer to two decimal places. For question (c), round your answer to a whole number.

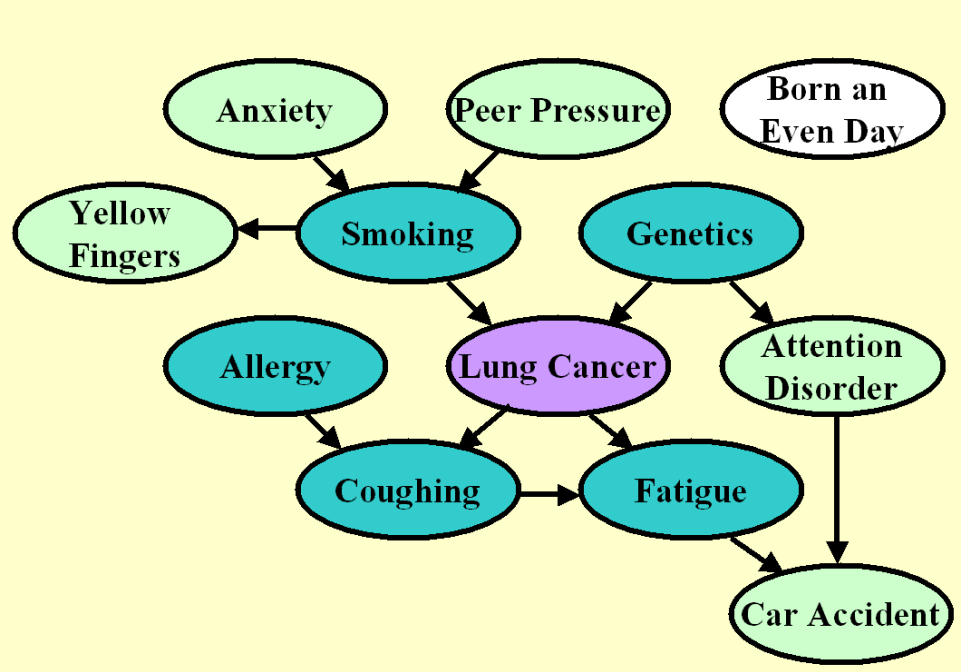
(a) If the travel time by car and public transport is the same, the probability that someone chooses public transport is: **a.** ..... (5 pt.)

(b) If the difference in travel time x is equal to **b.** ..... (5 pt.) minutes, then people are indifferent between traveling by car and traveling by public transport.

(c) If x increases with 10 minutes, the odds of taking the car increase with **c.** ..... (5 pt.) percent.

9 Bayesian Networks

Consider a greedy hill-climbing search for a good Bayesian network structure in a medical domain. The current model in the search is given in the graph below:



All the variables are binary, with values coded as 0 and 1. Furthermore, the following relevant data is given:

	Genetics	Smoking	Lung Cancer	Count
1	0	0	0	321
2	1	0	0	7
3	0	1	0	229
4	1	1	0	0
5	0	0	1	108
6	1	0	1	59
7	0	1	1	1063
8	1	1	1	213
Total (n)				2000

In this table, each row specifies a possible value combination of the variables Genetics, Smoking, and Lung Cancer respectively. The final column contains the number of times this value combination occurs in the data.

Answer the following questions:

Use the *natural* logarithm in your computations. This is the "ln" button on your calculator.

(a) What is the contribution of the node Genetics to the loglikelihood-score of the current model? Round your final answer to two decimal places.

The contribution of the node Genetics to the loglikelihood-score is: **a.** ..... (5 pt.)

(b) What is the contribution of the node Genetics to the loglikelihood-score after we add an edge from Smoking to Genetics? Round your final answer to two decimal places.

The contribution of the node Genetics to the loglikelihood-score after adding an edge from Smoking to Genetics is: **b.** ..... (5 pt.)

(c) What is the size of the penalty for the number of parameters in the BIC-score of the current model (the whole model, as shown in the picture)? Round your final answer to two decimal places.

The size of the penalty for the number of parameters of the whole model in the BIC score is: **c.** .... (5 pt.)

10
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 full graph (the saturated model). All edges are directed according to alphabetical order, that is, from A to B, from A to C, from A to D, from B to C, from B to D, and finally from C to D. Neighbor models are obtained by removing an edge from 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 remove (B → D) is largest and positive, so in iteration 1, we remove the edge from B to D. 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. remove (B → C)

b. remove (A → D)

c. remove (A → B)

d. remove (C → D)

Don't forget to fill in the course evaluation in Caracal! Thank you.