Universiteit Utrecht

REVIEW MOMENT

Course: Data Mining (INFOMDM) | Test name: [20251106] INFOMDM - Data mining - 1-GS - USP | Status: Closed | Type: Exam | Time left: 116 min. and 30 sec.

No filters selected; all answers are displayed

Course
Data Mining (INFOMDM)

Blueprint name
[20251106] INFOMDM - Data mining - 1-GS - USP

Blueprint type
Exam

Status
Closed

Number of questions
10

Start/End time
November 6, 2025 from 1:30 PM to 3:39 PM

Score
78 (maximum score: 100)

Number of questions answered correctly: 6
Questions mostly answered correctly: 3
Questions mostly answered incorrectly: 1

Hide question text

Question 1 — **Answered on:** November 6, 2025 - 3:16 PM **Duration:** 22 min. and 20 sec. **Score:** 14 of 20 pts.

14 pts.

TRUE OR FALSE?

Indicate whether the following statements are true or false.

(Frequent Sequence Mining) "ai" occurs 2 times as a subsequence of "taai taai pop" (there are 2 different mappings) — 2 pts.

- ○ TRUE
- → ◉ FALSE — 2 pt. ✓

(Bayesian Networks) Two directed independence graphs are equivalent if they have the same skeleton and the same v-structures (immoralities). — 2 pts.

- → ◉ TRUE — 2 pt. ✓
- ○ FALSE

(Random Forests) Each split in a tree in a random forest is allowed to use only a random subset of the features. — 0 pts.

- ○ TRUE — 2 pt. ✓
- → ◉ FALSE — ✗

(Bias-Variance decomposition) As the training set size increases, the bias component of expected prediction error decreases. — 2 pts.

- ○ TRUE
- → ◉ FALSE — 2 pt. ✓

(Link-based classification) In link-based classification, we model the dependency between node labels and the labels of their neighboring nodes. — 2 pts.

- → ◉ TRUE — 2 pt. ✓
- ○ FALSE

(Gradient Boosting) In gradient boosting with trees, each additional tree tries to reduce the errors of the model constructed so far. — 2 pts.

- → ◉ TRUE — 2 pt. ✓
- ○ FALSE

(Cost-complexity pruning) In cost-complexity pruning, we prune in the nodes of the current pruned subtree of T_{max} that yield the smallest increase in resubstitution error per leaf reduction (i.e., per unit decrease in the number of leaves). — 2 pts.

- → ◉ TRUE — 2 pt. ✓
- ○ FALSE

(Bayesian networks) In the directed graphical model **A → B ← C**, A and C are independent given B. — 2 pts.

- ○ TRUE
- → ◉ FALSE — 2 pt. ✓

(Link prediction) In link prediction, we only choose pairs of nodes from the training interval that are not connected by an edge. The test interval is used to provide the class labels for those pairs. — 0 pts.

- ○ TRUE — 2 pt. ✓
- → ◉ FALSE — ✗

(Frequent pattern mining) Consider an alternative tree mining scenario with just a single data tree. In this scenario, the support of a pattern tree is equal to the number of distinct occurrences of the pattern tree in the data tree. Two occurrences are considered distinct if they correspond to mapping functions φ_1 and φ_2, where $\varphi_1(v_i) \neq \varphi_2(v_i)$ for at least one node v_i in the pattern tree.
Claim: the anti-monotonicity property between support and the induced subtree relationship holds in this scenario. — 0 pts.

- → ◉ TRUE — ✗
- ○ FALSE — 2 pt. ✓

Question 2 — **Answered on:** November 6, 2025 - 1:44 PM **Duration:** 4 min. and 24 sec. **Score:** 5 of 5 pts.

5 pts.

Classification Trees: Computing Splits

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	10	16	16	20	26
y	0	0	0	1	0	1	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): — 5 pts.

- ☐ x ≤ 6
- → ☑ x ≤ 9 ✓
- → ☑ x ≤ 13 ✓
- → ☑ x ≤ 18 ✓
- ☐ x ≤ 23

Question 3 — **Answered on:** November 6, 2025 - 1:47 PM **Duration:** 4 min. and 3 sec. **Score:** 5 of 5 pts.

5 pts.

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	4	6

The branch T_t of tree T has 3 leaf nodes that together make 2 errors.

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: 0.04 | 0.04 5 pt. — 5 pts.

Question 4 — **Answered on:** November 6, 2025 - 1:58 PM **Duration:** 13 min. and 54 sec. **Score:** 10 of 10 pts.

10 pts.

Frequent Itemset Mining: Closed Frequent Itemsets

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

tid	Items
1	ABC
2	ABC
3	AB
4	BCD
5	BCD
6	CD
7	C
8	BC

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

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

- ☐ AB
- → ☑ AC ✓
- ☐ AD
- → ☑ BC ✓
- → ☑ BD ✓
- ☐ CD

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

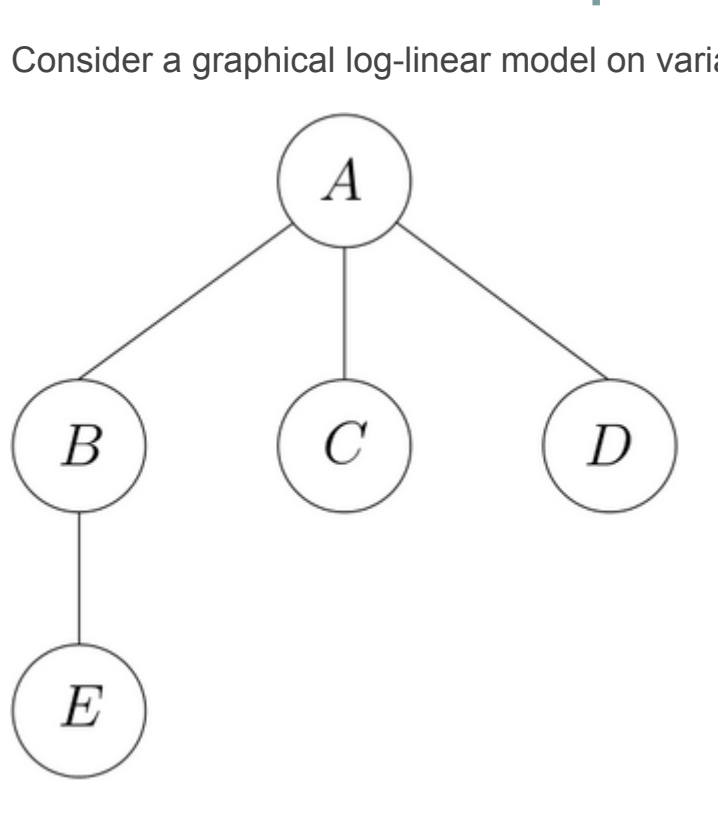
- → ☑ ABC ✓
- ☐ BD
- → ☑ BC ✓
- → ☑ CD ✓
- ☐ AC
- → ☑ BCD ✓
- ☐ D

Question 5 — **Answered on:** November 6, 2025 - 2:04 PM **Duration:** 5 min. and 31 sec. **Score:** 5 of 10 pts.

5 pts.

Undirected Graphical Models

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



A, B, C and D are binary variables, and E has 5 possible values.

Answer the following questions:

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

- → ◉ $\dfrac{n(A,B)n(A,C)n(A,D)n(B,E)}{n(A)^2 n(B)}$ — 5 pt. ✓
- ○ $\dfrac{n(A,B,E)n(A,C)n(A,D)}{n(A)^2}$
- ○ $\dfrac{n(A,B)n(A,C)n(A,D)n(B,E)}{2\,n(A)n(B)}$
- ○ 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:

10 | 16 5 pt. (correct) | 15 3 pt. — 0 pts.

Question 6 — **Answered on:** November 6, 2025 - 3:27 PM **Duration:** 16 min. and 29 sec. **Score:** 15 of 15 pts.

15 pts.

Response model changed by Ad Feelders on Nov 7, 2025, 10:04 AM for reason "On second thought both answers should be considered entirely correct."

Undirected Graphical Models

The following data concerns an outbreak of food poisoning after the traditional Christmas Lunch of the personnel of the Department of Information and Computing Sciences of our University. This time the theme was Dutch cuisine. Of the food eaten, interest focused on the “Berenhap” and “Frikandel”.

The variables are:
1. Berenhap (B) eaten (yes) or not eaten (no)
2. Frikandel (F) eaten (yes) or not eaten (no)
3. Sick (S) (yes) or not (no)
Questionnaires were completed by 100 of the 114 persons attending. The table of observed counts is given below.

Observed counts		Sick	
Berenhap	Frikandel	no	yes
no	no	22	4
	yes	3	12
yes	no	8	1
	yes	12	38

We fit the model B ⊥ S | F to this data and obtain fitted counts:

Fitted counts		Sick	
Berenhap	Frikandel	no	yes
no	no	22.29	3.71
	yes	3.46	11.54
yes	no	7.71	1.29
	yes	11.54	?

In your calculations, always use the natural logarithm. Round your final answers to two decimal places. Don't round intermediate results.

(a) The fitted count for B = yes, F = yes, S = yes is equal to: 38.46 | 38.46 (+/- 0.1) 5 pts. — 5 pts.

(b) The contribution of the cell B = no, F = no, S = no to the deviance of the model B ⊥ S | F is equal to: -0.58 | -0.58 (+/- 0.01) 5 pts. | -0.29 (+/- 0.01) 5 pts. — 5 pts.

It is given that the deviance of the model B ⊥ S | F is equal to 0.21, and the critical value of the χ^2 distribution with 2 degrees of freedom is equal to 9.21 for α = 0.01.

(c) Based on this data, the best supported conclusion is that people got sick because of eating: Frikandel | Frikandel 5 pt. — 5 pts.

Question 7 — **Answered on:** November 6, 2025 - 2:25 PM **Duration:** 13 min. and 50 sec. **Score:** 4 of 10 pts.

4 pts.

Text Classification: Multinomial Naive Bayes

You are given the following collection of song lyrics and corresponding music genre:

Words in lyrics	Genre
gone baby gone	Blues
woke up this morning	Blues
shake baby shake	Funk
shake ya fanky fanky ya ya	Funk

Answer the following questions:

(a) The estimate of P(baby | Blues) according to the multinomial naive Bayes model with Laplace smoothing is: — 4 pts.

- → ◉ 1/8 — 4 pt. ✓
- ○ 1/7
- ○ 2/23
- ○ 1/16

(b) According to the multinomial naive Bayes model with Laplace smoothing estimated on the given training set, the probability P(Blues | groovy baby) is given by: — 0 pts.

- ○ 1/18
- ○ 1/16
- ○ 9/17 — 6 pt. ✓
- → ◉ 324/580 — ✗

Question 8 — **Answered on:** November 6, 2025 - 2:37 PM **Duration:** 17 min. and 54 sec. **Score:** 5 of 10 pts.

5 pts.

Logistic Regression

We want to distinguish between fake and genuine hotel reviews using a bag-of-words representation of the reviews. Let $y_i = 1$ if review i is fake, and $y_i = 0$ if review i is genuine. Furthermore, let x_i be the number of times the word "location" occurs in review i. We use the method of maximum likelihood to estimate a logistic regression model, with the result:

$\beta_0 = 0.22$ and $\beta_1 = -1.1$,

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

Answer the following questions:

For question (a), round your answer to two decimal places, for example, 0.255 is rounded to 0.26.

For question (b), round your answer to a whole number, for example, 58.6 % is rounded to 59 %.

(a) For a review that contains the word "location" twice, the estimated probability that it is fake is: 0.12 | 0.12 (+/- 0.01) 5 pts. — 5 pts.

(b) For every additional occurrence of the word "location", the odds of the review being fake decrease with 64 | 67 (+/- 1) 5 pts. %. — 0 pts.

Question 9 — **Answered on:** November 6, 2025 - 2:55 PM **Duration:** 24 min. and 10 sec. **Score:** 10 of 10 pts.

10 pts.

Bayesian Networks

The table below shows the number of positive and negative outcomes for male and female patients that received one of two treatments. The total number of patients is n = 520.

Observed counts		Outcome	
Gender	Treatment	negative	positive
female	A	30	150
	B	20	120
male	A	30	50
	B	40	80

We perform a greedy hill-climbing search to find a good Bayesian network structure. Neighbour models are obtained by adding a single edge to the current model, deleting a single edge, or turning a single edge around. We start the search process from the empty graph (the mutual independence model).

In your calculations, always use the natural logarithm. Round your final answers to two decimal places. Don't round intermediate results.

(a) The contribution of the node OUTCOME to the log-likelihood score of the current (mutual independence) model is: -280.91 | -280.91 (+/- 1) 5 pts. | -122 (+/- 1) 3 pts. — 5 pts.

(b) The change in AIC-score (Δ AIC-score) if we add the edge TREATMENT → OUTCOME to the current model is (make sure you take into account the sign of the change): -1 | -1 (+/- 0.5) 5 pts. — 5 pts.

Question 10 — **Answered on:** November 6, 2025 - 2:58 PM **Duration:** 2 min. and 57 sec. **Score:** 5 of 5 pts.

5 pts.

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 model:

A → B → C ← D.

Neighbor models are obtained by adding an edge to the current model or deleting 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 add (A → C) is largest, so in iteration 1, we add an edge from A 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) — 5 pts.

- → ☑ delete (B → C) ✓
- ☐ add (B → D)
- ☐ delete (A → B)
- → ☑ delete (D → C) ✓

Show one question at a time