

Data Mining Homework Set 2, 2020

Cursus: BETA-INFOMDM Data Mining (INFOMDM)

Aantal vragen: 5

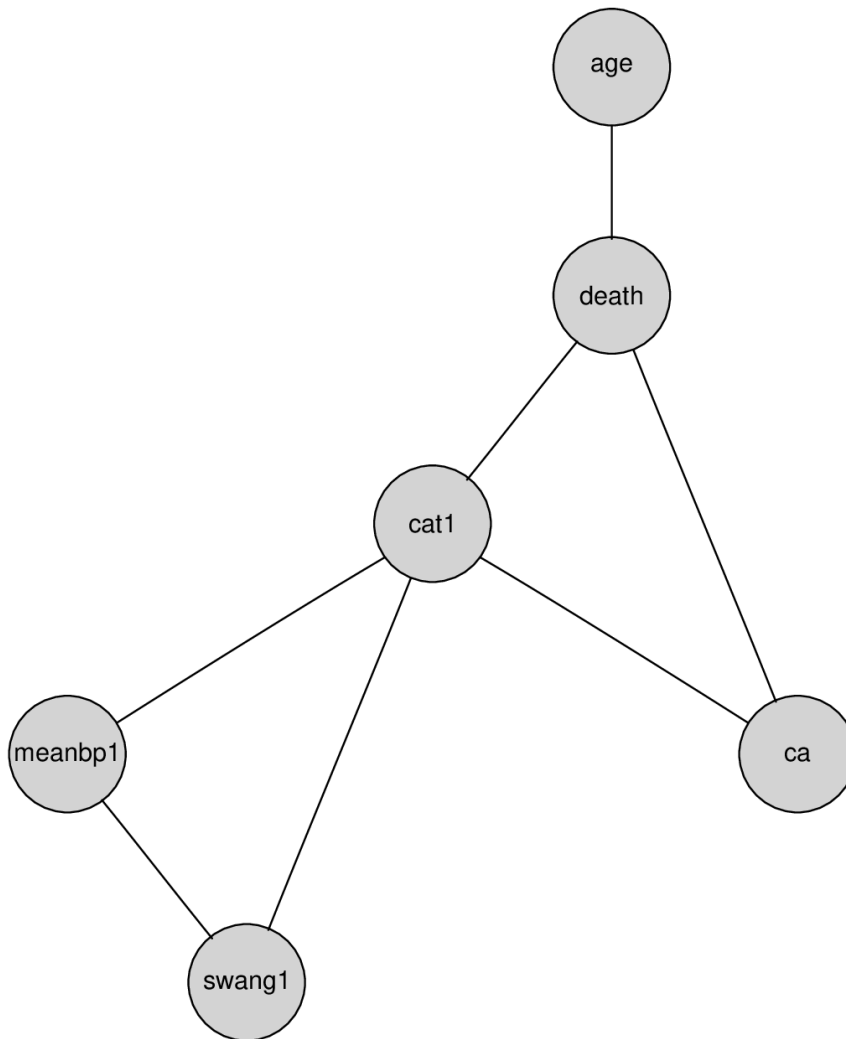
Data Mining Homework Set 2, 2020

Cursus: Data Mining (INFOMDM)

This is Homework Set 2 of Data Mining

Aantal vragen: 5

- 1** Consider the graphical log-linear model with the following independence graph:
2 pt.



Which of the following (conditional) independences hold in this model?

- a. $\text{age} \perp \text{cat1}$
- b. $\text{swang1} \perp \text{death} \mid \text{cat1}$
- c. $\text{age} \perp \text{swang1}$
- d. $\text{death} \perp \{\text{meanbp1}, \text{swang1}\} \mid \{\text{age}, \text{cat1}\}$
- e. $\text{death} \perp \text{ca}$
- f. $\text{death} \perp \{\text{meanbp1}, \text{swang1}\} \mid \text{cat1}$
- g. $\text{swang1} \perp \text{ca} \mid \text{meanbp1}$

- 2 Consider the following table of counts on binary variables x and y :

$n(x,y)$	$y=0$	$y=1$
$x=0$	80	20
$x=1$	40	60

Suppose we fit the independence model $x \perp y$ to this data. Give the fitted counts for:

$(x=0,y=0)$:

a. (0,5 pt.)

$(x=1,y=0)$:

b. (0,5 pt.)

$(x=0,y=1)$:

c. (0,5 pt.)

$(x=1,y=1)$:

d. (0,5 pt.)

- 3 Consider a graphical model M on three binary variables A, B , and C , with independence graph $G=(K,E)$ with $K = \{A,B,C\}$ and $E = \{\{B,C\}\}$.

The observed counts are given in the following table:

A	B	C	$n(A,B,C)$
1	1	1	40
1	1	0	10
1	0	1	5
1	0	0	50
0	1	1	30
0	1	0	5
0	0	1	20
0	0	0	40

Answer the following questions (do not round your answer):

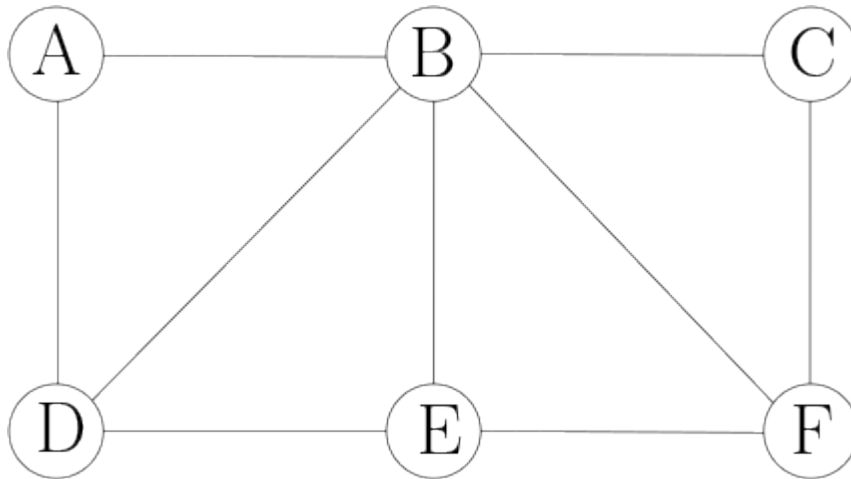
The fitted count $\hat{n}(1, 1, 1)$ according to model M is: **a.** ..(1 pt.)

The fitted count $\hat{n}(0, 1, 0)$ according to model M is: **b.** ..(1 pt.)

- 4** We are performing a hill-climbing search in the space of decomposable models. Neighboring models are obtained by either adding an edge to the current model, or removing an edge from the current model.

2 pt.

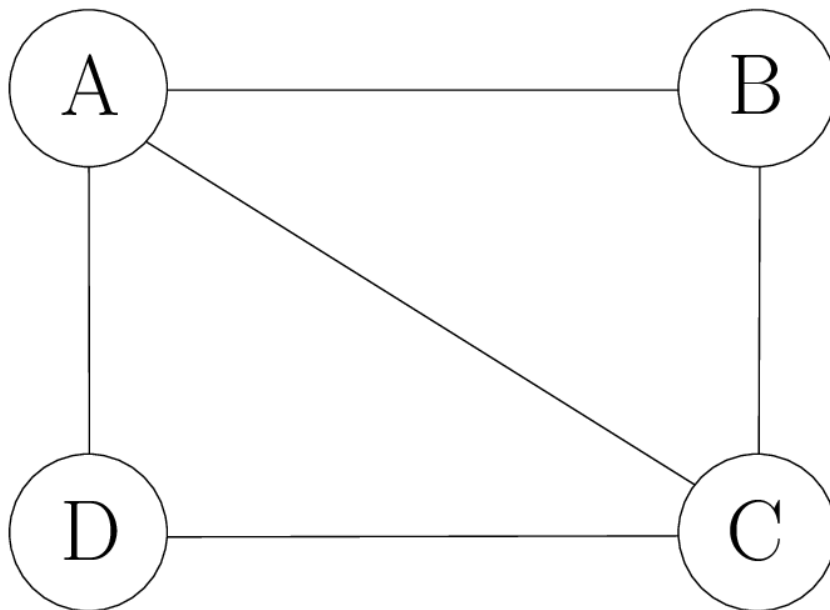
The current model is given in the following figure:



Which of the following operations produce a valid neighbor? (0 or more answers may be correct)

- a. Remove the edge between B and D
- b. Add an edge between A and E
- c. Add an edge between A and F
- d. Remove the edge between B and F
- e. Add an edge between C and D
- f. Remove the edge between A and D
- g. Remove the edge between B and E

- 5 Consider the graphical log-linear model M_1 on binary variables A,B,C, and D, with independence graph:



1 pt. a. The formula for the maximum likelihood fitted counts of M_1 is given by:

- a. $\frac{n(A, B, C)n(A, C, D)}{n(A)n(C)}$
- b. $\frac{n(A, B, C)n(A, C, D)n(A, C)}{n(A)n(C)}$
- c. $\frac{n(A, B, C)n(A, C, D)}{n(A, C)}$
- d. $\frac{n(A, B)n(B, C)n(A, C)n(C, D)n(A, D)}{n(A)n(B)n(C)n(D)}$

Consider the model M_0 obtained by removing the edge between A and C from M_1 . How many parameters (u-terms) are eliminated by this change?

The number of eliminated u-terms is: **b.** ..(1 pt.)

Thank you, goodbye!