# Data Mining Homework Set 4

## Course: BETA-INFOMDM Data Mining (INFOMDM)

**Number of questions:** 4

# Data Mining Homework Set 4

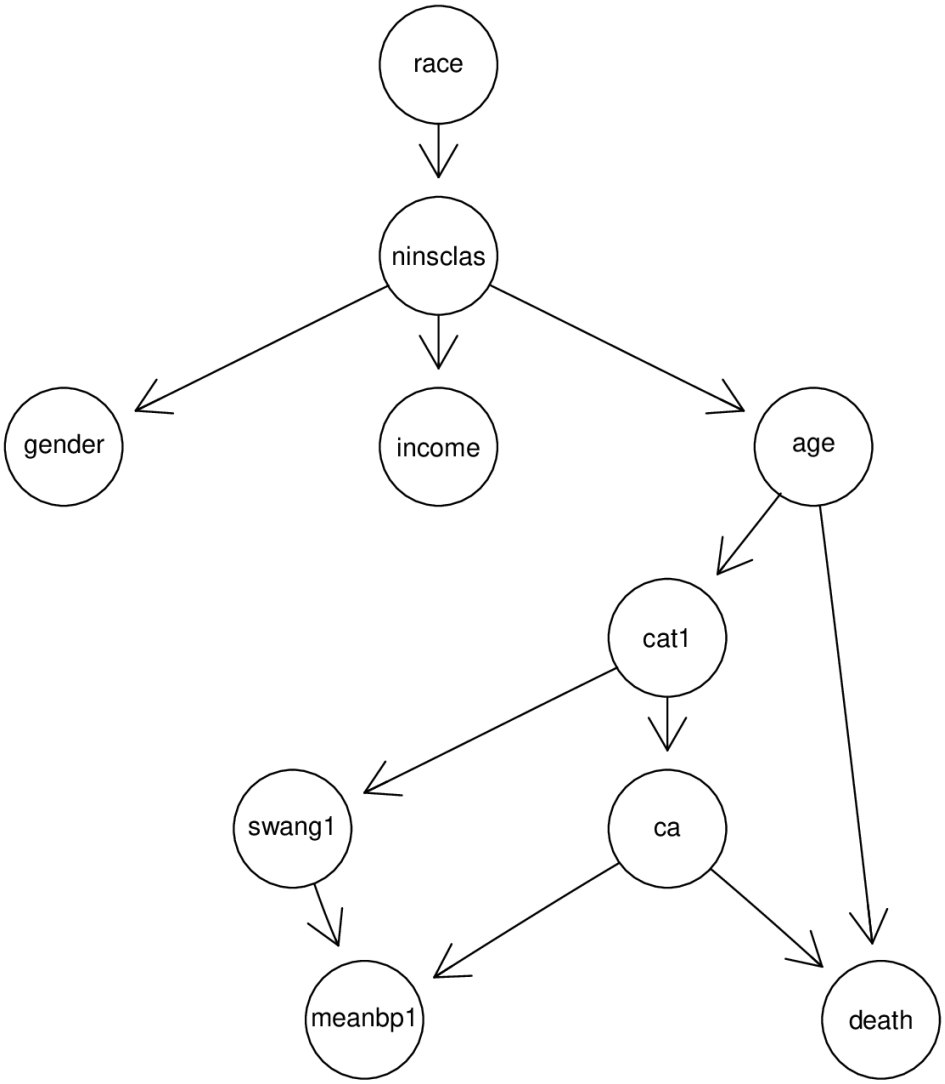
## Course: Data Mining (INFOMDM)

Welcome!

**Number of questions:** 4

**1** Consider the following Bayesian network constructed on data from an intensive care unit:

2 pt.




Which of the following conditional independence properties are true according to the given model? (0 or more answers may be correct)

**a.** gender ⊥ income | ninsclas

**b.** ca ⊥ age | cat1

**c.** ca ⊥ age | {cat1, death}

**d.** meanbp1 ⊥ cat1 | {swang1, ca}

**2** The table below shows the number of successes and failures for minor and major operations in two hospitals: one academic hospital and one local hospital. The total number of operations is n = 2900.

| | | RESULT | RESULT |
|---|---|---|---|
| OPERATION | HOSPITAL | success | failure |
| minor | academic | 685 | 15 |
| | local | 584 | 16 |
| major | academic | 1425 | 75 |
| | local | 93 | 7 |

The Bayesian network we want to fit to the data contains the following edges: OPERATION → RESULT, and HOSPITAL → RESULT.

Compute the maximum likelihood estimates of the following parameters (round your answer to two decimal places):

P(OPERATION = minor): **a.** ..... (1 pt.)

P(RESULT = success | OPERATION = major, HOSPITAL = academic): **b.** ..... (1 pt.)

The total number of parameters of the given Bayesian network is: **c.** ..... (1 pt.)

**3** For the data, please refer to exercise 2 of this set.

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 answers to two decimal places.

The contribution of the node RESULT to the log-likelihood score of the current (mutual independence) model is: **a.** ..... (1 pt.)

The change in log-likelihood score (Δ score) if we add the edge OPERATION → RESULT to the current model is: **b.** ..... (1 pt.)

The change in BIC-score if we add the edge OPERATION → RESULT to the current model is: **c.** ..... (1 pt.)

**4** We perform a greedy hill-climbing search to find a good Bayesian network structure on 4 variables
2 pt. denoted A,B,C, and D. Neighbour models are obtained by deleting an edge from the current graph (addition and reversal are not allowed). We start the search process from the saturated model (the full graph), where the alphabetical ordering is used to order the variables (so A is a source and D is a sink)

In step 1 of the search we find that deleting the edge A → D gives the biggest improvement in the BIC score. Assume that Δ scores of operations computed in previous iterations that are still valid are not recomputed, but are retrieved from memory. All other Δ scores need to be computed!

For which operations do we need to compute the Δ score in step 2 of the search? (0 or more answers may be correct)

**a.** Delete C → D

**b.** Delete A → B

**c.** Delete A → C

**d.** Delete B → D

Thank you. Goodbye!