# Data Mining 2020
# Exercises Undirected Graphical Models

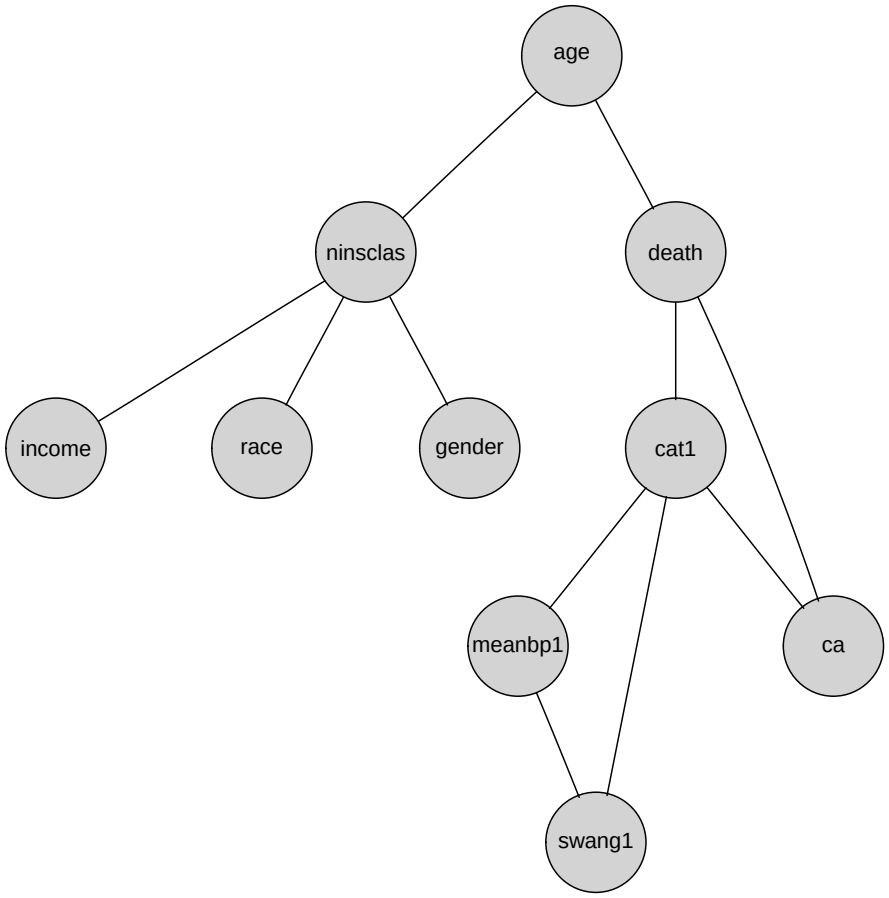
## Exercise 1

We have a data set containing data on 5,735 critically ill adult patients receiving care in an Intensive Care Unit (ICU) for 1 of 9 pre-specified disease categories. The data was collected in five US teaching hospitals between 1989 and 1994.

The objective of the original study that used (a superset of) this data was to examine the association between the use of right heart catheterization (RHC) during the first 24 hours of care in the ICU and subsequent survival (among others).
This subset contains the following variables:

1. cat1: disease category (9 different values)
2. death: did the patient die within 180 days after admission?
3. swang1: was right heart catheterization performed within first 24 hours?
4. gender: male/female
5. race: black/white/other
6. ninsclas: type of medical insurance of patient (six different values)
7. income: income of patient, divided into four categories
8. ca: cancer status (yes/no/metastatic)
9. age: age of patient divided into 5 categories
10. meanbp1: mean blood pressure of patient divided into 2 categories

Consider the following independence graph for this domain:




Use the notion of separation in the graph to determine if the following independencies hold:

(a) swang1 $\perp\!\!\!\perp$ death

(b) swang1 $\perp\!\!\!\perp$ death | cat1

(c) ca $\perp\!\!\!\perp$ death | cat1

(d) swang1 $\perp\!\!\!\perp$ death | {cat1, ca}

(e) death $\perp\!\!\!\perp$ {income, race, gender, ninsclas, meanbp1, swang1} | {cat1, age, ca}

(f) gender $\perp\!\!\!\perp$ race

(g) gender $\perp\!\!\!\perp$ race | ninsclas

Consider you answer to (f). Does it make sense?

# Exercise 2

Consider the variables Gender and Eye Color. The independence model assumes that Gender and Eye Color are independent, that is, Gender $\perp\!\!\!\perp$ Eye Color. The maximum likelihood fitted counts for the independence model are given by the formula:

$$\hat{n}(\text{gender}, \text{eye color}) = \frac{n(\text{gender})n(\text{eye color})}{N}, \tag{1}$$

where $N$ denotes the total number of observations in the data set.

The following data were collected from students enrolled in an introductory Statistics course:

| | Eye Color | | | | |
|---|---|---|---|---|---|
| Gender | blue | brown | green | hazel | Total |
| female | 370 | 352 | 198 | 187 | 1107 |
| male | 359 | 290 | 110 | 160 | 919 |
| Total | 729 | 642 | 308 | 347 | 2026 |

This data was taken from: Amy G. Froelich and W. Robert Stephenson, Does Eye Color Depend on Gender? It Might Depend on Who and How You Ask; *Journal of Statistics Education*, Volume 21, Number 2 (2013).

(a) (Just to get acquainted with the notation) Determine the values of $N$, $n(\text{female}, \text{brown})$, and $n(\text{hazel})$ for this data set.

(b) Use equation (1) to compute the table of fitted counts according to the independence model. You may round the fitted counts to two decimal places.

(c) Instead of using equation (1), we can also use the Iterative Proportional Fitting (IPF) algorithm to compute the fitted counts. Study the slides on IPF, and use it to fit the independence model to this data set. Start the iteration with a table $\hat{n}^{(0)}$ that has the same count in each cell. The algorithm has converged when all the margin constraints are (approximately) satisfied simultaneously. Again, you may round to two decimal places.

The deviance of the independence model is given by

$$2 \sum_{\substack{\text{gender} \\ \text{eye} \\ \text{color}}} \sum n(\text{gender, eye color}) \ln \frac{n(\text{gender, eye color})}{\hat{n}(\text{gender, eye color})} \approx 16.29,$$

where $\hat{n}$(gender,eye color) is given in equation (1).

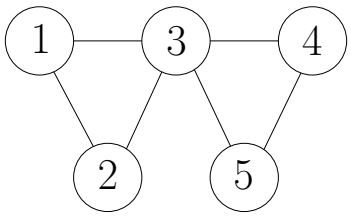
(d) Test the independence model against the saturated model at significance level $\alpha = 0.05$. To perform the test, you may consult the following table with critical values:

| degrees of freedom ($\nu$) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| critical value ($\chi^2_{\nu;0.05}$) | 3.84 | 6.00 | 7.82 | 9.50 | 11.1 | 12.6 | 14.1 | 15.5 |

Clearly state whether or not the model is rejected, and explain how you made that decision.

# Exercise 3

Consider the graphical model on variables $X_1, \ldots, X_5$ with the following independence graph:




(a) Use the property of separation in the graph to verify that the conditional independence

$$(X_1, X_2) \perp\!\!\!\perp (X_4, X_5) \mid X_3$$

holds.

(b) Which factorisation of $P(X_1, X_2, X_4, X_5 \mid X_3)$ does the conditional independence given under (a) allow?

A clique is a maximal complete subgraph, that is, a clique is a subset of the nodes such that every pair of nodes in the subset is connected by an edge. It is maximal in the sense that it has no superset that also has this property.

(c) Give the cliques of the graph, and the corresponding *observed = fitted* margin constraints that are satisfied by the maximum likelihood fitted counts.

(d) Give a formula for the maximum likelihood fitted counts in the terms of observed counts. The formula has to be derived from "first principles", that is, you are not allowed to use the formula based on a RIP-ordering of the cliques here.

Hint: Start with $\hat{P}(X_1, \ldots, X_5)$. The general strategy is to rewrite this into an expression containing only marginal distributions over cliques (or subsets of cliques). To achieve this goal, you need to make use of the conditional independencies that hold for the given model. You can use a conditional independency to simplify an expression in two basic ways; if $X \perp\!\!\!\perp Y \mid Z$, then

1. $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$, and
2. $P(X \mid Y, Z) = P(X \mid Z)$,

where $X$, $Y$ and $Z$ are arbitrary disjoint random vectors (or if you prefer: sets of random variables). You also often need the following law of probability: $P(X, Y) = P(X \mid Y)P(Y)$ where $X$ and $Y$ are random vectors. We refer to this law as the product law.

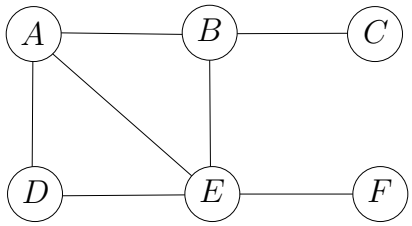
Once you have an expression containing marginal distributions over cliques, you manipulate the expression to get fitted counts rather than fitted probabilities, and finally you apply the margin constraints to replace fitted counts by observed counts.

# Excercise 4

(a) How many undirected graphs are there with $k$ (labeled) nodes?

(b) Use your answer to (a) to compute the number of graphical models on 8 variables.

(c) Because the number of different graphical models becomes huge very fast, an exhaustive search to find the best model (according to some scoring function, for example, AIC) is not feasible. Therefore we typically apply some local search algorithm. Suppose the following model is the current model in a hill-climbing search:




Neighboring models are obtained by either removing an edge or adding an edge. How many neighboring graphical models does the current model have? And how many neighboring decomposable models?

(d) Determine the cliques of the graph given under (c), and find a RIP-ordering of those cliques. Then give a formula for the maximum likelihood fitted counts for this model.

# Exercise 5

Utrecht University is accused of discrimination against women in their admission policy for master programs. To check this claim, data has been gathered on the gender (G) of each applicant, together with the admission decision (A). The results are as shown in the table below:

| Gender | Admission | |
|---|---|---|
| | Yes | No |
| Male | 245 | 155 |
| Female | 75 | 125 |

(a) Compute the admission probability for males and females.

(b) Give the fitted cell counts according to the independence model $G \perp\!\!\!\perp A$.

(c) Compute the deviance of the fitted model (always use the natural logarithm).

(d) Test the independence model against the saturated model. Use $\alpha = 0.05$.

(e) Is there any evidence of discrimination against women? Explain.

# Exercise 6

It turns out that the table given in the previous exercise originated from two master programs, A and B. The three-way table is given below:

| Program | Gender | Admission | |
|---|---|---|---|
| | | Yes | No |
| A | Male | 25 | 80 |
| | Female | 35 | 115 |
| B | Male | 220 | 75 |
| | Female | 40 | 10 |

(a) Draw the independence graph of the model $G \perp\!\!\!\perp A \mid P$, where $P$ denotes the master program, and state the corresponding independence assumption(s) in words.

(b) Compute the table of fitted counts $\hat{n}(P, G, A)$ corresponding to the model specified under (a). What is the deviance of this model? Test it against the saturated model, using $\alpha = 0.05$.

(c) Is there any evidence of discrimination against women? Explain.

# Exercise 7

Use Iterative Proportional Fitting to fit the model $G \perp\!\!\!\perp A \mid P$ with observed data

| Program | Gender | Admission | |
|---|---|---|---|
| | | Yes | No |
| A | Male | 25 | 80 |
| | Female | 35 | 115 |
| B | Male | 220 | 75 |
| | Female | 40 | 10 |

To help you get started we have filled in the required marginal counts in a table with convenient structure to perform the iterations of the algorithm.

| Program | Gender | Admission | | |
|---|---|---|---|---|
| | | Yes | No | $n(P, G)$ |
| A | Male | | | 105 |
| | Female | | | 150 |
| | $n(P, A)$ | 60 | 195 | $n(P, G)$ |
| B | Male | | | 295 |
| | Female | | | 50 |
| | $n(P, A)$ | 260 | 85 | |

The initial table $\hat{n}^{(0)}$ is given by

| Program | Gender | Admission | | |
|---|---|---|---|---|
| | | Yes | No | $\hat{n}^{(0)}(P, G)$ |
| A | Male | 10 | 10 | 20 |
| | Female | 10 | 10 | 20 |
| | $\hat{n}^{(0)}(P, A)$ | 20 | 20 | $\hat{n}^{(0)}(P, G)$ |
| B | Male | 10 | 10 | 20 |
| | Female | 10 | 10 | 20 |
| | $\hat{n}^{(0)}(P, A)$ | 20 | 20 | |

## Exercise 8

We are given the following data on gender, treatment and outcome.
To get probabilities instead of counts, divide by 52.

| $n$(gender, treated, outcome) | | outcome | |
|---|---|---|---|
| gender | treated | neg | pos |
| female | no | 3 | 15 |
| | yes | 2 | 12 |
| male | no | 3 | 5 |
| | yes | 4 | 8 |

(a) Use the cross-product ratio to show that there is a positive association between treatment and outcome for both genders. For treatment, code "no" as zero, and "yes" as one. For outcome, code "neg" as zero, and "pos" as one.

(b) Use the cross-product ratio to show that treatment and outcome become independent if we collapse (sum) the table over gender.

(c) From (b) we conclude that treatment and outcome are marginally independent, and from (a) we learn that treatment and outcome are not independent given gender. Is there an undirected graph that captures this combination of constraints? If yes, draw the corresponding graph.

(d) Use common-sense knowledge to draw a causal picture of the situation. Draw an arrow from A to B if you think A could cause (or "influence") B. Don't use the data to determine this causal graph! Based on your causal model in combination with the data, would you say the treatment has a positive effect?