Medical Data Science, WS 2021/2022
Prof. Dr. Nico Pfeifer
Chair for Methods in Medical Informatics
University of Tuebingen

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

2021-12-07

Assignment 4

Deadline: Tuesday, December 21, 9:59 p.m.

This problem set is worth 50 points. You can submit in groups of two people or alone. Submit your solutions digitally by uploading to the ILIAS page (none of the other students can see the files you upload). Just upload a zipped folder containing all necessary files and name the folder by your last names. The folder should be named according to the following scheme:

`[MDS][Assignment4]_lastname1_lastname2`

Problem 1 (T, 19 Points)

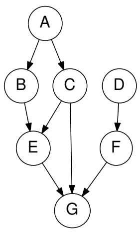
Graphical models and Causality.

(a) (3P) What makes a model (such as a Bayesian Network) generative? How can dependencies be taken into account in such networks?

(b) (6P) Draw a Bayesian network (2P) that represents the joint probability $p(q, r, s, t_1, t_2, ..., t_{25})$ given by $p(q)p(r|q)p(s|q, r)\prod_{i=1}^{25} p(t_i|r)$. Use the plate-notation from the lecture (p.7ff). Are q and s independent given the empty set? Are they independent given r? Argue using the d-separation criterion (2P). Expand the graph such that all nodes (i.e. $(q, r, s, t_1, t_2, ..., t_{25})$) become independent. You can include as many nodes/edges between the existing nodes as you want. You don't have to proof independence but explain briefly how you proceeded (2P).

(c) (3P) True or false? Explain briefly or give counterexamples:
a) If there are many paths between two nodes we always have to test every single path to say whether the two nodes are d-separated.
b) If A is d-separated from B, B is d-separated from A.
c) If A is d-separated from B and B is d-separated from C, A is d-separated from C.

(d) (4P) Consider the following DAG G.



Can you show conditional independence with the help of d-separation for the following examples? Keep in mind the results from (c) and write down how many paths you need to test to show independence.

- $a \perp\!\!\!\perp B$
- $A \perp\!\!\!\perp G\,|C, E, B$
- $D \perp\!\!\!\perp C\,|E, G$
- $G \perp\!\!\!\perp B\,|A, E$

(e) (2P) Explain why Simpson's Paradox occurs in your own words. Why is it problematic and how can we handle it?

(f) (1P) What's the difference between $P(A|B)$ and $P(A|do(B))$?

Problem 2 (T, 10 Points)

Evaluate the distributions $p(c)$, $p(b|c)$, and $p(c|a)$ corresponding the joint distribution given in Table 1. Hence show by direct evaluation that $p(a, b, c) = p(a)p(c|a)p(b|c)$. You don't need to give counter probabilities.

a	b	c	$p(a, b, c)$
0	0	0	0.162
0	0	1	0.144
0	1	0	0.074
0	1	1	0.220
1	0	0	0.194
1	0	1	0.062
1	1	0	0.044
1	1	1	0.100

Table 1: The joint distribution over three binary variables.

Problem 3 (P/T, 21 Points)

Consider the Similarity Network Fusion (SNF) method with number of neighbors $k = 2$ (remember that the first neighbor of a node is the node itself). Note: if you provide runnable code for a) and b), you can use c) to solve a) and b).

(a) (7P) Given matrices

$$\mathbf{W}^{(1)} = \begin{pmatrix} 1.00 & 0.50 & 0.30 & 0.10 & 0.10 \\ 0.50 & 1.00 & 0.40 & 0.10 & 0.10 \\ 0.30 & 0.40 & 1.00 & 0.30 & 0.30 \\ 0.10 & 0.10 & 0.30 & 1.00 & 0.50 \\ 0.10 & 0.10 & 0.30 & 0.50 & 1.00 \end{pmatrix}$$

and

$$\mathbf{W}^{(2)} = \begin{pmatrix} 1.00 & 0.20 & 0.50 & 0.10 & 0.10 \\ 0.20 & 1.00 & 0.30 & 0.10 & 0.10 \\ 0.50 & 0.30 & 1.00 & 0.30 & 0.30 \\ 0.10 & 0.10 & 0.30 & 1.00 & 0.50 \\ 0.10 & 0.10 & 0.30 & 0.50 & 1.00 \end{pmatrix}$$

provide $\mathbf{P}^{(1)}$, $\mathbf{P}^{(2)}$, $\mathbf{S}^{(1)}$, and $\mathbf{S}^{(2)}$.

(b) (7P)Perform two steps of the similarity network fusion method (i.e., compute $\mathbf{P}_1^{(1)}$, $\mathbf{P}_1^{(2)}$, $\mathbf{P}_2^{(1)}$, and $\mathbf{P}_2^{(2)}$ as well as the corresponding $\mathbf{P}^{(c)}$s).

(c) (7P) Implement the SNF starting from the similarity matrices $\mathbf{W}^{(i)}$ in Python with the convergence criterion ϵ as described in the supplement of the paper (VPN necessary to access this page) and check whether the graph structure of the $\mathbf{P}^{(c)}$s changes for $t > 2$ for the above described data.