EBERHARD KARLS UNIVERSITÄT TÜBINGEN

2023-12-04

# Assignment 4

**Deadline:** Tuesday, December 19, 6:00 p.m.

This problem set is worth 50 points. You can submit in groups of two people. Submit your solutions digitally by uploading to the ILIAS page. 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`

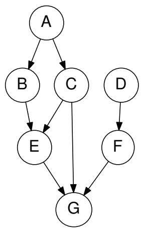
## Problem 1 (T, 11 Points)

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) (3P) True or false? Explain briefly or give counter examples:
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.

(c) (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.

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

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

## Problem 2 (P, 18 Points)

Evaluate the performance of different graph kernels using GraKeL (in case you have problems installing GraKel with python 3.8 or larger, please try python 3.7): link to website and the MUTAG data set from here: website with benchmark data sets.

- Compute the graphlet kernel using sampling for the graphlets of size 3 (1000 samples). Perform a 10-fold cross-validation for the binary classification problem using the kernel with an SVM. What is the accuracy for the best $C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}$? You can expect sth. larger than 0.8. Evaluating the accuracy for the best C value is sufficient, you do not have to evaluate the test accuracy separately.

How many samples would you need such that the deviation from the real distribution is less than 0.05 with probability larger than 0.9?

- Compute the Weisfeiler-Lehman subtree kernel for 4 iterations. Perform a 10-fold cross-validation for the binary classification problem using the kernel with an SVM. What is the accuracy for the best $C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}$? You can expect sth. larger than 0.85. Evaluating the accuracy for the best C value is sufficient, you do not have to evaluate the test accuracy separately.
- Compute the Weisfeiler-Lehman edge kernel (WLedge.m) for 1 iteration. Perform a 10-fold cross-validation for the binary classification problem using the kernel with an SVM. What is the accuracy for the best $C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$?

## 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 $\epsilon$ as described in the supplement of the paper (Uni Tuebingen VPN necessary to access the full paper, supplement is freely available) and check whether the graph structure of the $\mathbf{P}^{(c)}$s changes for $t > 2$ for the above described data.