Assignment 1

Building a Bayesian Network

Deadline Exposee: October 4th; Assignment: November 15th.

Handout for the *Bayesian Networks and Causal Inference* lecture, September 2nd, 2024

Johannes Textor, Moabi Mokhoro

Objectives of This Exercise

1. *Construct* a Bayesian network for a problem domain of your choice.
2. *Test* the structure of a Bayesian network against a real dataset.
3. Perform a *causal inference* task using a Bayesian network

The goal of this assignment is to formulate a research question, and use a Bayesian network to help answer this research question. The research question should fall into the domain of causal inference, as discussed during the lecture and Exercise of the first week.

We will proceed in two phases. First you will make an "exposee", in which you explain the problem domain, an available data set, and you define one or more specific, causal research questions you would like to answer.

In the second phase, you construct the Bayesian actual network model, fit some of its parameters and test some of its predictions, and perform some kind of causal inference on the network (such as determining the causal effect of one variable on another using covariate adjustment, or fitting path coefficients in a linear model).

Tasks

There are four specific tasks that you need to complete for this assignment.

1. Form teams of three people for your assignment and register your team on Brightspace.
2. Write an exposee (see below) and have it approved by us. The **deadline** for submitting your exposee on Brightspace is **October 4th, 2024**.
3. Build a Bayesian Network that models the problem domain that you described in your Exposee.
4. Write a brief report (about 5-6 pages) that documents your network. There are two deadlines for your report. The first deadline, **October 18th, 2024**, is *optional*, but if you submit a version of your report by that deadline, we'll provide feedback that you can use to improve your report. The *mandatory* deadline for the final report is **November 15th, 2024**.

Exposee

The exposee is a 1-page plan for your project. It will allow us to judge beforehand whether your project is heading in a good direction and we may be able to give you some advice, e.g. on where you could find more data. An example is available for download on Brightspace.

Problem domain

Explain the problem domain – the context of your research question – in about 200 words.

Research question

Formulate a (few) specific research question(s) within the problem domain that you would like to answer in your project. This should be a *causal* research question that makes your project a causal inference task, not a prediction task. (See Problem 1.1).

Dataset

Explain which dataset you are going to use to answer your research question, how you are going to access this dataset, and list the relevant variables that your network will include. (You can change this later on, for instance by including more variables or by summarizing some of these variables, but you should not end up with a much smaller network than the one you have originally proposed.)

Datasets can be found in various online resources. Some examples are:

- The UCI machine learning repository: http://archive.ics.uci.edu/ml/. **DO NOT use the "Adult", "Adult Income" or "Census Income" datasets; also DO NOT use the " Student Performance Data Set"**
- The Princeton Office for Population Research Data Archive: http://opr.princeton.edu/archive/ (requires registration)
- The American Psychological Association (APA) provides various links to free datasets at: http://www.apa.org/research/responsible/data-links.aspx
- The Kaggle platform has some relevant datasets as well, but it also requires registration. Note that most UCI datasets are also on Kaggle; please do not use the datasets explicitly mentioned above under "UCI".

Or, choose a topic that you find interesting and go search for data yourself. Many scientific papers nowadays make their data freely available for download.

Report

The report can be an extension of your exposee (see above), and should cover *roughly* 5-6 A4 pages (possibly excluding appendices such as literature references, detailed code etc.) It should contain the following items:

- An explanation of the problem domain (like in the exposee)
- Your research question(s)
- A description of the data in form of a table, in which each variable is described:
  - Variable name
  - Variable type (continuous, ordinal, categorical)
  - Number of levels (for ordinal/categorical), range (for continuous)
- A figure of your Bayesian Network model of the problem domain, including the variables in the dataset and other relevant ones
- A description of how the network has been built – how did you proceed to organize the variables and edges? Did you base your network mainly on theoretical knowledge or did you also use the data itself?
- Information on how the network structure has been tested – what was the result of your first test? Did you discover misspecifications? How did you address those misspecifications? Did you re-test the modified network?
- A description of your causal inference question, and potentially other applications of your network (e.g., prediction).
- A discussion in which you reflect on the success of your project.
- References
- An appendix containing description of implementation details (Programming language / packages used, high-level description of the source code, link to the source code)