Assignment 1

Building a Bayesian Network

Deadline November 9th, 2020

Handout for the *Bayesian Networks* lecture, September 24th, 2020

Johannes Textor, Ankur Ankan

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 an inference task using a Bayesian network

The goal of this assignment is to build your own Bayesian network in a problem domain of your own choice. We will proceed in two phases. First you will make an "exposee", in which you explain the problem domain, an avaibable data set, and you outline some specific questions for the Bayesian network. In the second phase, you construct the actual network, fit some of its parameters and test some of its predictions, and perform some kind of inference on the network (such as determining a marginal probability, performing a prediction, or comparing the strengths of different causal effects to each other).

Note that "performing inference" does **not** imply that you have to implement an inference algorithm yourself! It is just as fine if you use one from an existing software package, such as the ones we use in the exercises.

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 9th, 2020**.
3. Build a Bayesian Network that solves the inference problem you described in your Exposee.
4. Write a brief report (about 5-6 pages) that documents your network. The **deadline** for the final report is **November 9th, 2020**.
5. One of you will present the Bayesian Network to the other students in a 2-minute "flash presentation", using at most 2 slides. The flash presentations will be given during the lecture **on December 15th, 2020**.

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

Set out the problem domain, in about 200 words.

Data

Explain which dataset you are going to use, 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 income" or "census income" datasets.**
- 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.

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.

Implementation Plan

Give as many details as you can, given your current knowledge, about what programming languages/packagess you will use for your project.

Application

Explain what you want to do/know/compute once your network is built. At least one kind of "inference" problem, where the task is to input some data and to get some information back, should be proposed.

Report

The report can be an extension of your exposee (see above), and should cover roughly 5-6 A4 pages. It should contain the following items:

- An explanation of the problem domain (like in the exposee)
- 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)
- The network itself (as a figure)
- A description of how the network has been built – how did you proceed to organize the variables and edges, was any kind of technique used to manage its complexity?
- A description of how the network was implemented (Programming language / packages)
- Information on how the network structure has been tested
- A description of your application (e.g. inference, assessment of the causal structure)
- A discussion in which you reflect on the success of your project
- References