Bayesian Networks
Assignment 1
Report: Heart Disease Research

Christoph Schmidl
s4226887
c.schmidl@student.ru.nl

Lisa Boonstra
s3018547
l.boonstra@student.ru.nl

November 24, 2017

Introduction

Cardiovascualar diseases is the leading cause of premature death in the modern world, including heart attacks, strokes and other circulatory diseases. The global number of deaths caused by cardiovascular diseases increased by 41% between 1990 and 2013 with 12.3 million to 17.3 million deaths. A number that is expected to grow to more than 23.6 million by 2030. The American Heart Association gauges the cardiovascular health of the nation by tracking seven key health factors and behaviours that increase risks for heart disease and stroke:

- Smoking
- Physical Activity
- Healthy Diet
- Cholesterol
- High Blood Pressure
- Blood Sugar/Diabetes
- Overweight/Obesity

Although these factors and behaviors can lead to cardiovascular diseases, they can be seen more as causes for specific indicators which can tell us if a patient is actually having some kind of heart disease. In this project, we will build on publicly available data from the Machine Learning Repository to build a Bayesian Network Model about the indicators of heart diseases.

Data

The project will use data from the "Machine Learning Repository" (`http://archive.ics.uci.edu/ml/datasets/heart+Disease`), which documents certain attributes of hospital patients who have been diagnosed with some kind of heart disease or the absence of it.
The original dataset consists of contributions from fours different institutes which produces 920 entries with 76 attributes. Each entry represents a patient who has been given a score for having some kind of heart disease: 0 for non-existent and 1 to 4 with increasing chance of having a heart disease. Prior published experiments refer to using only a subset of 14 of the original 76 attributes. For feasibility reasons and because prior research has shown that those 14 attributes are sufficient, our Bayesian Network will be build using the 14 attributes: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, height at peak exercise, inherited blood disorders, angiographic disease status.

Implementation Plan

We will implement our Bayesian Network as a linear model, where linear regression functions shall be used to represent the probability distributions at each node. We will implement our Bayesian Network in Python, using the pgmpy package.

Inference Problem

Unfortunately, the data contains missing values. The Bayesian Network will be used to predict and fill the missing values by using other available information. By providing information about a patient's health status which is mapped to the 14 attributes, the inference problem would be how likely it is that this patient has some kind of heart disease.