Ciências ULisboa | Informática

Advanced Machine Learning

Project 1 – Bayesian Networks

Introduction

Cardiovascular diseases are the number one cause of death worldwide; around 17.9 million people die each year, which represents 31% of all deaths worldwide. Cardiovascular diseases are a group of disorders of the heart and blood vessels and include coronary heart disease (CHD), cerebrovascular disease, rheumatic heart disease, and other conditions. Thus, CHD continues to be a leading cause of morbidity and mortality among adults. It is a disease of the blood vessels supplying the heart, i.e., the heart's blood supply is blocked or interrupted by a build-up of fatty substances in the coronary arteries. Hence, we intend to predict a 10-year risk of CHD in patients using Bayesian networks. The original data are from a large cohort from the BioLINCC website.

Dataset

In the late 1940s, the U.S. Government set out to better understand cardiovascular disease (CVD). So, the plan was to track a large cohort of initially healthy patients over time. The city of Framingham, Massachusetts, was selected as a site for study due to the appropriate size, stable population, and cooperative doctors and residents. The study began in 1948. The original cohort has 5209 patients aged between 30 and 59 years old. The patients underwent questionnaires and exams every two weeks to assess physical and behavioral characteristics.

The key to successful prediction of CHD is identifying important risk scores, which are variables that increase the chances of disease. So, we will investigate risk factors collected in the first data collection for the study, which is an anonymized version of the original data.

The current data for this project has 4240 patients with the following variables:

<u>Demographic risk factors:</u>

- *Sex*: sex of patient
- *age*: age in years at first examination

- *education*: some high school (1), high school/GED (2), some college/vocational school (3), college (4)

Behavioral risk factors:

- *currentSmoker*, *cigsPerDay*: smoking behavior

Medical history risk factors:

- *BPMeds*: on blood pressure medication at the time of first examination
- *prevalentStroke*: previously had a stroke
- *prevalentHyp*: currently hypertense
- *diabetes*: currently has diabetes

Risk factors from the first examination:

- *totChol*: total cholesterol (mg/dL)
- *sysBP*: systolic blood pressure
- *diaBP*: diastolic blood pressure
- *BMI*: body mass index, weight(kg)/height(m)^2
- *heartRate*: heart rate (beats/minute)
- *glucose*: blood glucose level (mg/dL)

The last variable in the dataset is the target variable (*TenYearCHD*) which tells us if the patient will have or not the disease.

Work plan

The main goal of this project is to implement a Bayesian network that guarantees the best results in predicting the risk of CHD. Thus, you need to learn a Bayesian network for the dataset described above using the techniques learned in lectures. You are free to explore more advanced techniques of Bayesian networks not taught in the lectures. However, you need to briefly explain the technique in the report.

The dataset has 15 features (without considering the target variable). If you are experiencing memory problems in learning your Bayesian network, you can do a feature selection and select the most relevant ones, i.e., select the k features with the highest score. For that, you may consider the `SelectKBest` function from `scikit-learn` with `chi2` as the score function. However, you are free to choose a different feature selection approach or criterion. Even if you

are not experiencing memory issues in learning your model, you are free to try feature selection to improve model performance.

This project has four main parts:

1. Prepare the data for the task.
2. Learn a Bayesian network from the data.
3. Assess the classification results when you predict the target variable.
4. Propose 3 examples of relevant queries (or questions) that might be answered with your model and present the corresponding result.

Submission

Each group should submit the report written in Jupyter Notebook with the name **AAA2324_P1_xx_yy.ipynb**, where xx and yy should be replaced by each of the student numbers of the group. The report should include (but not limited to):

- The identification of the members of the group
- Exploratory data analysis and preprocessing steps required for the project
- Learning of the Bayesian network with the corresponding parameters
- Validation of the results
- Discussion of the model and results obtained.

The report should also include your **explanations and justifications for your decisions**, as well as the **code and corresponding outputs** obtained.

Beware that **I will not run the code of all groups, but I might run some groups selected at random**, so I need to have explicit outputs in the report to confirm your conclusions.

Deadline

The deadline for this project is **October 22nd at 23:59 in Moodle**. The groups without access to Moodle to submit the project should send the project by email.