Visualization of massive data

Session 2

Contributors :

Clément Le Boëdec

Marc Jubault

Marvin Bellouard

Which constants allow to detect heart diseases ?

The nature of the data set :

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date.The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no/less chance of heart attack and 1 = more chance of heart attack

Attribute Information

1) age

2) sex

3) chest pain type (4 values)

4) resting blood pressure

5) serum cholestoral in mg/dl

6)fasting blood sugar > 120 mg/dl

7) resting electrocardiographic results (values 0,1,2)

8) maximum heart rate achieved

9) exercise induced angina

10) oldpeak = ST depression induced by exercise relative to rest

11)the slope of the peak exercise ST segment

12) number of major vessels (0-3) colored by flourosopy

13) thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

14) target: 0= less chance of heart attack 1= more chance of heart attack

Potential correlation between variables :

There seems not a high correlation between variables. That is, there will not be any multicollinearity problem in the model. Only a few variables seem to have a high positive relationship with others, so it might affect the model adequacy. Thus, some of the variables might be removed. Checking the relations of the target with other covariates, there seems a good positive correlation with many of the covariates. That means, covariates together or separately may conduct a proper model with high accuracy and have a considerable effect on heart attack risk.

Therefore we could observe minor correlations, for example when the thalach is higher than 140, people have a greater heart attack risk.

Explanations & observations about quantitative processing & visualization :

The goal of this notebook is to create a model that will be able to predict people with higher chance of heart attack, and find main features that impact possiblity of this. Before making any prediction, it is necessary to look at the data.

The dataset consists of 165 High Risk Class with only 138 observation on Low Risk Class. When the dataset seems tohave a higher rate of transaction in one class compared to other, there comes the bias in the resulting measures.

Comments on the results obtained :

We note a couple of things from this initial examination:

- The high risk portion has similar amount of men and women, but men are more prominent in the low risk portion. We should investigate the amount of men and women further to draw any conclusions from this observation.
- Max heart rate is higher among high risk patients. This intuitively seems plausible. Chest pain is more prominent among high risk patients. This also seems intuitively plausible.
- Slope for ST depression seems to be downsloping on average among high risk patients. On the other hand, the slope is flat on average for low risk patients.
- Number of major vessels is higher among low risk patients compared to high risk ones.

The age group 41-50 has the highest chance of getting a heart attack when compared to all other age groups

Work distribution :

Part 1 : Marc

Part 2 : Clément & Marvin

PDF : Clément & Marvin