Cezar Anicai

Arnaud Nguyen

I download the data and pass it through a variable.

1. We understand and visualize the data: the first rows and the statistical description for numerical data.

1. As everything is labialize, the aim is to classify the data. We want to develop models that are accurate classifying the data.

First we need to do more descriptive analysis and more visualization. For example the age and the profession determines the balance. The difficulty is to deal with categorical data. For example it is difficult to do feature selection or algo like k nearest.

Feature Selection

We want to reduce the data set.

Technics: PCA, LDA = (linear) Kernel PCA (non linear)

Statistics test with numerical data to be sure that each feature is different

Classification algo:

K nearest neighbor

Decision Tree

Naives Bayesian Classification

After we should try some NN and SVM

Then we have to combine classifier in order to find better model.