|  |
| --- |
| Data Analysis and Knowledge Discovery |
| Exercise work pt.1 |
| Documentation |

# Plotting a histogram of each feature

First of all we need to read file *wine.data*. We read the data into two variable x and y. X consists list of labels (*1*, *2* and *3*) and variable y consists feature data in matrix.

We use method defaultHistogram – method to plot a histogram of each feature. Method gets an attribute: matrix. Matrix is a variable y consisting feature data. Method add one feature of every sample into one plot and shows it. This is done for every 13 feature.

Most of histograms has one common factor. Most of the values are near average at the distribution.

# Plotting histograms with visualizing each classes

To plot histograms with visualizing each differ cultivation, we need to separate samples by their cultivation. Variable x has the information about the class. We need three subplots, because we have three differ cultivation. Same as above, method add one feature of every sample of a cultivation into one subplot. Method shows for every 13 feature a histogram including three subplots.

Visualizing each cultivation separately shows that cultivations differs from one another.

# Making a parallel plot of the features

In parallel plot we wanted to show every cultivations with different color. Thus Class1 has red, Class2 green and Class3 blue. Method parallelFeatures makes three subplots which have every sample of each cultivations separately.

The parallel plot doesn’t give much additional information because one of the features has significantly larger values than others. Therefore it dominates the plot over other features. With zooming and scaling the plot, we could get better information.

# Calculating the correlation coefficients

There are two interesting correlations which have correlation factor higher than 0.75. First correlation is with feature 5 (total phenols) and feature 6 (flavanoids) having correlation factor 0,86. The second correlations is with feature 6 (flavonoids) and feature 11 (OD280/OD315 of diluted wines) having correlation factor 0,79.

# Visualizing interesting correlations with a scatter plot

As said above, we have two interesting correlations (feature5/feature6 and feature6/feature11). Method visualizeAsScatterPlot gets strong correlations from previous method which calculates correlation coefficients and returns interesting correlations. Interesting correlations are those, which have factor higher than 0,75.

# Calculating the eigenvalues and eigenvectors of the covariance matrix

# PCA