Exploratory data analysis: Visualization of multidimensional data

R. Benítez

Exercise 1: Load database:

Task 1: Obtain a multivariate dataset with numerical features for classification problems from the UCI Machine Learning Repository (n observations, d features).

Exercise 2: Data visualization

Task 1: Visualize one of the variables using the following representation methods: basic line plot, histogram, boxplot

Task 2: Obtain the summary statistics for the chosen attribute (mean, median, standard deviation, standard error of the mean ($\sigma/\sqrt{n}$), interquartile range, kurtosis, etc).

Task 3: Choose a pair of variables from the data set and visualize the observations using a scatter plot.

Task 4: Use quantile-quantile plot (q-q plot) to visualize if the two previous variables are equally distributed.

Task 5: Visualize all the variables in the data set using a scatter plot matrix. Compute the pairwise linear correlations between variables and represent the results as a correlation plot.

Task 6: Use the class labels of the observations in order to represent the data as a class-grouped scatter plot matrix.

Task 7: Apply Multidimensional Scaling (MDS) in order to project the d-dimensional data in a 2-d space.

Useful references

UCI Machine Learning Repository:

- MATLAB: `http://archive.ics.uci.edu/ml/`

Scatter correlation plot matrix:

- MATLAB: `https://es.mathworks.com/help/matlab/ref/plotmatrix.html?searchHighlight=plotmatrix&s_tid=doc_srchtitle`
- Python: `http://seaborn.pydata.org/examples/network_correlations.html`
- R: `https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html`

Multidimensional Scaling (MDS):

- MATLAB: `https://es.mathworks.com/help/stats/cmdscale.html`
- Python: `http://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html#sklearn.manifold.MDS`
- R: `https://www.r-bloggers.com/multidimensional-scaling-mds-with-r`