Data Mining Fall/Winter 2016
Instructor: Martin Ester, TA: Xin Shen
Programming Assignment 1

Due date
November 1, before class

Submission
Submit a hard copy of your code and a report in class.

Data
The data is a collection of the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. There are 178 instances in the dataset. In wine.csv, each instance has 13 numerical attributes and an instance ID. The numerical attributes are Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines and Proline. You can download the dataset from https://gist.github.com/tijptjik/9408623.

Programming environment
It is recommended, but not necessary, to download RStudio, the leading R programming environment, and use it for this and the following assignments.

Tasks
In this assignment, you will gain practical experience with K-means and hierarchical clustering algorithms. Follow the instructions and answer the questions using R. Most of the required functions are available in R by default, i.e. without loading further packages, but some functions require package "cluster".

1. Read in the dataset and standardize/normalize the data properly.

2. Use all attributes to conduct K-means clustering. Set the number of clusters to 3. Report the silhouette score of the result and plot your clustering result according to the attributes Alcohol and Phenols.

3. Find the best number of clusters between 2 and 10, i.e. the number of clusters that has the largest silhouette score. Report the best number of clusters and its silhouette score. You can use functions in the R package cluster to calculate the score.

4. Conduct hierarchical clustering analysis on the dataset using complete linkage, average linkage and single linkage separately. Cut all of the three trees resulting from your hierarchical clustering into groups by using the best number of clusters got in step 3.

Compare the partitions (when cutting the dendrograms) obtained by the dfferent algorithms. What do you observe?

5. Plot the dendrograms produced by the different algorithms in step 4. Discuss the differences between them and try to explain the reasons leading to those differences.