Exercises: Clustering

1. Download the two-dimensional ”halfmoon.RData” dataset from our website. Plot it with and without visualizing the given labels.

2. Perform a 2-means clustering (k-means) of this data and plot the result. Repeat this several times. What is your conclusion?

3. Perform a hierarchical clustering (hclust) using Euclidean distance and

   (a) average linkage

   (b) single linkage

   (c) complete linkage.

   Visualize the trees constructed by the respective methods (plot).

4. Cut the trees obtained in the previous exercise at the root two obtain two clusters (cutree). Compare the clustering results with the given labels in a cross-table (table). Which performs best, and why?

5. Increase the number of clusters to see how additional segmentation of the data is done for k-means and for single linkage clustering.

6. Implement ”k-medoids” clustering and apply it to the halfmoon data. Note that one can choose different distance. Try Manhattan distance (L1) instead of the Euclidean distance (L2). How is the first implementation different from the ”k-medians” algorithm?