**Topics of exercises**

* Clustering

**Textbooks and readings**

**Exercise 4.1**

Suppose that the data mining task is to cluster the following eight points (with (x,y) representing location) into three clusters:

A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9).

The distance function is Euclidean distance. Suppose initially we assign A1 , B1 , and C1 as the centre of each cluster, respectively.

1. Use the **k-means** algorithm (**Don't use R!**) to show only
   * The three cluster centres after the first round execution,
   * The final three clusters.
2. Perform a hierarchical clustering of these data using **single link** method (**Don't use R!**). Build the dendrogram.
3. Use R to check your results.

**Exercise 4.2**

Use the **roses\_by\_day** data frame (see Exercise 2.1).

1. Partition **roses\_by\_day$number\_of\_bouquets** into 2 clusters using **k-means** algorithm.
2. Partition **roses\_by\_day$number\_of\_bouquets** into 3 clusters using **k-means** algorithm.
3. Compare the above results. Find the objects that belong to the cluster with the smallest number of elements.

**Exercise 4.3**

1. Download and unpack the [accidents.txt.zip](http://home.agh.edu.pl/~mszpyrka/lib/exe/fetch.php?media=lectures:dm:accidents.txt.zip) file.
2. Load the file as **accidents** data frame. Use the **head** and **summary** functions to become familiar with the data.
   * **vehicles** - number of vehicles involved in the crash
   * **pedestrians** - number of pedestrians involved in the crash
   * **road\_lanes** - number of lanes of the road the crash occurred on
   * **fatals** - number of people who died as a result of the crash
   * **drunk\_drivers** - number of alcohol influenced drivers involved in the crash
3. Partition the data into 5 clusters using **PAM** algorithm (**clustering 1**). Display medoids and find their characteristics.
4. Analyse the content of each cluster - display the contents of each cluster and compare objects with the medoid.
5. Copy the data frame into **accidents2**. Normalize each column of **accidents2** using min-max normalization with new max and min equal to 1 and 0 respectively.
6. Partition the **accidents2** data into 5 clusters using **PAM** algorithm (**clustering 2**).
7. Analyse the content of each cluster - display the contents of each cluster (use original (not normalized) data) and compare the cluster with the corresponding cluster belonging to **clustering 1** set of clusters. Which clustering results do you find more valuable?

**Exercise 4.4**

1. Partition the data from **Exercise 3.3** into 2 clusters. Use the dissimilarity matrix from **Exercise 3.3** and the **PAM** algorithm.
2. Partition the data from **Exercise 3.4** into 2 clusters. Use the dissimilarity matrix from **Exercise 3.4** and the **PAM** algorithm.
3. Compare the above clustering results.

**Exercise 4.5**

Use the **bouquets\_by\_province** data frame (see Exercise 2.4).

1. Perform a hierarchical clustering of these data using **single link**, **complete link**, **average link**, and **centroid link** methods. Display dendrogram representation of each of the results.
2. Use cophenetic correlation coefficient to check which of these methods best fits the data (see slide 22, part 4).
3. Using the best method partition the data into 3 clusters and analyse the content of each cluster.

**Exercise 4.6**

Perform a hierarchical clustering of the data from **Exercise 3.6** using **single link** and **complete link** methods. Compare the dendrograms.

**Exercise 4.7**

Execute the following script:

x <- c(1,1,2,2,2,2,2,3,3,3,3,3,3,3,4,4,4,5,5,5,

6,6,6,6,6,6,7,7,7,7,7,8,8,8,9,9,9,9,9,

10,10,10,10,10,11,11,11,11,12,12,12,12,12,

13,13,13,14,14,14,14)

y <- c(1,2,1,2,3,5,6,1,2,3,4,5,6,7,1,8,9,7,8,9,

1,2,6,7,8,9,1,2,5,6,7,1,6,7,1,3,7,8,9,

1,2,3,4,9,4,5,6,9,5,6,7,8,9,8,9,11,9,10,11,12)

xy <- data.frame(x, y)

plot(x, y, axes = FALSE, ann = FALSE, asp = 1, pch = 20)

axis(1, pos = 0, at = 0:14)

axis(2, pos = 0, at = 0:14)

abline(h = 1:15, col = "lightgrey", lty = 3)

abline(v = 1:15, col = "lightgrey", lty = 3)

1. Identify the clusters in the figure using **DBSCAN** algorithm (**Don't use R!**), ε = 1, MinPts = 3.
2. Identify the clusters in the figure using **DBSCAN** algorithm (**Don't use R!**), ε = 1.42, MinPts = 5.
3. Use R to check your answers.
4. Use the **kNNdistplot** function to find the best value of ε parameter for the above clustering tasks.

**Exercise 4.8**

1. Download and unpack the [shapes\_2d.txt.zip](http://home.agh.edu.pl/~mszpyrka/lib/exe/fetch.php?media=lectures:dm:shapes_2d.txt.zip) file.
2. Load the file as **xy** data frame. Plot the points.
3. Use the **kNNdistplot** function to find the best value of ε parameter for the clustering the data.
4. Use the **DBSCAN** to find the clusters.
5. Plot the clustering results.

**Exercise 4.9**

1. Download and unpack the [shapes\_3d.txt.zip](http://home.agh.edu.pl/~mszpyrka/lib/exe/fetch.php?media=lectures:dm:shapes_3d.txt.zip) file.
2. Load the file as **xyz** data frame. Plot the points.
3. Use the **kNNdistplot** function to find the best value of ε parameter for the clustering the data.
4. Use the **DBSCAN** to find the clusters.
5. Plot the clustering results (use the **scatterplot3d** from the **scatterplot3d** or the **pairs** function).

**Exercise 4.10**

Use **DBSCAN** algorithm for clustering the data from **Exercise 3.8**. Analyse the content of each cluster. Can you find some outliers?

**Exercise 4.11**

1. Recall the dataset library and distinguish the **iris** dataset.
2. Visualize 2-dimentionally the categories **Sepal.Length** and **Petal.Width** playing the colors to distinguish the 'cluster-candidate'from each other.
3. Define **irisCluster** as **k-means** clusters (at first 3, the next 4 of them) for
   * **Sepal.Width** and
   * **Petal.Length**.
4. Consider, at first, 3 clusters, the next-4.
5. Compare the achieved means and the cluster sizes in both cases.

**Exercise 4.12**

Recall the **USArrests** data set.

1. Distinguish the first 4 rows of the data.
2. Define by **df** a result of a scaling of the data set.
3. Working in **set.seed(123)** environment define by **km.res** the **k-means** clustering with 4 clusters, starting from 25 as the number of the points 'population'. Consider also 30 and 40 as the size of the population.
4. Vizualize the **km.res** using the colors: **#2E9FDF,#00AFBB,#E7B800,FC4E07**and the ellipse type-visualisation.
5. Vizualize the **km.res** using the same colors with no declaration for the cluster forms, but with the added segments from centroids to items.

**Exercise 4.13**

Consider the **USArrests** data set and **df** data.

1. Create the diagram of the number of **k-means** clusters for that **df** data dependently on their size (in the sense of the measure method 'within sum of square').
2. Provide a convenient solution to estimate the optimal number of clusters establishing a vertical line in a point x=4.

**Exercise 4.14**

Return to **iris** data set and consider the categories: **Petal.Length** and **Petal.Width**.

1. Carry out the reasoning leading to detecting the optimal number of clusters in a sense of **k-means**. Do it with a support of the 'factoextra' package and without it. Compare the results.
2. Crate the hierarchical clusters for them and visualize them (Formulate a precondition of the operation performing).
3. Cut the obtained tree by a reduction of the cluster number up to 3 and visualize the results.