Discovering properties of K-Means, comparing K-Means with K-Medoids, X-Means and EM-Clustering, applying K-Means
Agathe Merceron, Beuth HS, Berlin

Write a report with all your answers to these exercises.

Exercise 1: Calculate the distance between Flower 1 and Flower 2, and between Flower 1 and Flower 3. Look at 2.4.6 Dissimilarity for Attributes of Mixed Types p. 75 in the book. Show how you proceed.

Flower 1: 5.1, 3.5, 1.4, 0.2, Iris-setosa
Flower 2: 4.9, 3.0, 1.4, 0.2, Iris-setosa
Flower 3: 6.3, 3.3, 4.7, 1.6, Iris-versicolor

Exercise 2: Getting K-means and K-medoids work
1. Read the file points1.xls or points1.csv. With the scatter plot identify the objects. Report the natural groups, that you see.
2. Run K-means with the number k for the number of clusters that you have identified above. Draw a scatter-plot of the objects so that all objects in one cluster have the same color. Report this picture and comment whether these clusters coincide with the clusters identified above.
3. Report the centers of the clusters and describe qualitatively the clusters.
4. Vary K, the number of clusters, and find an optimal value for K. For this, plot SSE versus K; the optimal number corresponds to a sharp drop in the curve (elbow method). Report the value of SSE for this optimal K.
5. Use the K-medoids algorithm with the same dataset and the number K identified in question 4 above. Compare the clusters you get with the clusters obtained with K-means. What are the centers of the clusters? Explain briefly the main difference between K-means and K-medoids.
Report all your results.

Tipps if you use RapidMiner. In RapidMiner open Modeling, then Clustering and Segmentation and then double-click on K-Means.
Attach the output of read Excel to the example input of Clustering, and attach the output of Clustering to the ports of the canvas. Select Clustering and on the right enter the number k for the number of clusters you think is appropriate. Select NumericalMeasures for measure types (all attributes are numerical) and Euclidean Distance for numerical measure. Hit the run button and look at the results:

a. In plot view select a scatter plot and take cluster as the Color Column. Report your result. Does the clustering make sense?
b. In the Cluster Model look at Centroid Table and report it. What information does this table give?
c. Use the Loop Parameter operator with Performance (see the tutorial) to find an optimal value for K. Report your result.
d. Replace the K-means operator with the K-medoids operator and choose the optimal number of clusters as obtained above. Do you get the same clusters as with K-Means? What does the centroid table show? Report your results. Explain with you own words the main difference between K-means and K-medoids. K-medoids is explained in the book p. 454 10.2.2 k-Medoids: A Representative Object-Based Technique.

Tipps if you use Python. For an implementation of the K-medoid algorithm, see https://scikit-learn-extra.readthedocs.io/en/latest/generated/sklearn_extra.cluster.KMedoids.html or https://pyclustering.github.io/docs/0.9.0/html/d0/dd3/classpyclustering_1_1cluster_1_1kmedoids_1_1kmedoids.html#details

Exercise 3: Random repartition of the data

Write a small program to generate a data file containing 1000 points with random values. The number of points as well as the range of x- and y- coordinates should be easy to choose and, hence, parameters of your program. Save the coordinates of the points in a text file, each line containing the x-coordinate and the y-coordinate of one point separated by a comma. The first line of the file contains the names of the attributes.

Show a scatter-plot of your data. Do you see any natural clusters?

Run K-means with this file for values of K varying from 2 to 100. Plot the curve SSE (avg_withing_distance in RapidMiner) against K. What is its shape? Can you identify an optimal number of clusters?

Read 10.6.1 Assessing Clustering Tendency. There is a mistake on p. 485: “1. Sample n points, p1, . . . , pn, uniformly from D” should be “1. Sample n points, p1, . . . , pn, randomly from D”.

Exercise 4: Clustering the iris dataset

This exercise uses the iris data set. The aim of this exercise is to investigate whether the four attributes of this set, namely sepallength, sepalwidth, petallength and petalwidth are appropriate to determine the class of an iris using clustering algorithms. The classes present in the data set are: Iris-setosa, Iris-versicolor, and Iris-virginica. (Look in Internet for pictures of these flowers.)

1. Read the data. *Note RapidMiner:* the attribute `class` should have the role `Label`so that it is ignored by a clustering algorithm.

2. Report the summary statistics.

3. Explore your data with different visualizations. From this exploration, which assumption can you make concerning the four attributes to predict the class of an iris flower? Explain your answer.

4. Clustering as classification: Cluster the data into three clusters (you should to try out different clustering algorithms and different attribute selection / transformation) and report the accuracy (proportions of objects in the right clusters) for each algorithm and each selection / transformation of attributes that you have used. This kind of evaluation is called extrinsic: we have got the class as “ground truth” to compare the clusters with.

5. What is your answer: are the four attributes of this set, namely sepallength, sepalwidth, petallength and petalwidth are appropriate to determine the class of an iris?

Exercise 5: Wrap-up.

You are given a dataset and are asked whether clusters can be identified in the data. List the methodological steps that you should undertake to solve this task. Tipp: remember also the hands-on exercises on distance and k-means.