

Unsupervised Learning and Evolutionary Computation Using R

Winter Term 2024/2025

Exercise Sheet 4 (November, 18, 2024)

Exercise 1 (Clustering by hand: k -means)

In this exercise, you will manually perform k -means clustering on a simple example. The figure below contains five points x_1, \dots, x_5 , as well as two cluster centres c_1, c_2 that have already been pre-determined. Using these cluster centres, perform one iteration of the Lloyd's k -means algorithm by assigning each point x_1, \dots, x_5 to a cluster and then updating the centres. Finally, assign clusters based on the updated centres. Report the location of both of the new centres and the final cluster assignments for all 5 points.

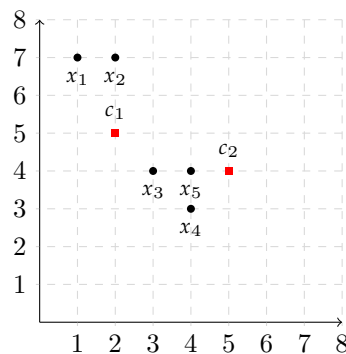


Figure 1: The starting position of the k -means algorithm for Exercise 1

Exercise 2 (Clustering by hand: Hierarchical Clustering)

Using the same 5 points from Exercise 1, manually perform agglomerative hierarchical clustering using single-linkage and the Manhattan distance as the distance measure. For each step, compute the distance matrix and draw a dendrogram showing the current clusters. After you have finished the clustering, cut the dendrogram at an appropriate point to obtain the final clusters.

Exercise 3 (Clustering in R: Implementing k -means)

In the this exercise, you will implement the k -means clustering algorithm discussed in the lectures, and evaluate it on multiple datasets

1. Implement the k -means algorithm in R that was discussed in the lectures. Implement at least two different initialisation methods.
2. Evaluate the implemented algorithm on the `languages.spoken.europe` dataset from the R package `cluster.datasets`. Explore the dataset and try to find two attributes that produce at least 3 distinct clusters.
3. Evaluate both initialisation method. Which one works better? Why?

Exercise 4 (Evaluating k -means on additional datasets)

So far, we have only used built-in datasets in R. In this exercise, you will explore some external datasets, learn how to use them in R, and evaluate the performance of the k -means algorithm on them.

1. For this exercise, we will be using datasets from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/datasets)¹. Take a look at the datasets and pick one from the classification tasks. Explore the set, and see if you can find a combination of attributes that you can cluster according to their class. While doing so, try to see if the algorithm you implemented in Exercise 1 works when clustering on more than 2 attributes.
2. Download the Forest Fires dataset, which contains data about forest fires in Portugal, including environmental conditions and the area of the fire. Plot the data based on the attributes *DC* (which measures the underground moisture) and *wind*, (which measures the wind speed). How many clusters do you see
3. Run the implemented k -means algorithm on the data. Does it correctly identify all of the clusters?

Exercise 5 (Clustering in R: Hierarchical Clustering)

In the final exercise, you will evaluate the performance of hierarchical clustering on the datasets that you used in the first two exercises. Use the R function *hclust* to perform hierarchical clustering using different linkage methods (at least "single" and "complete", but feel free to experiment with other methods provided by the function). Describe how the results of the hierarchical clustering differ from k -means. Are there datasets where it works better or worse?

¹<https://archive.ics.uci.edu/datasets>