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 $x_1, ..., x_5$, 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 $x_1, ..., x_5$ 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.

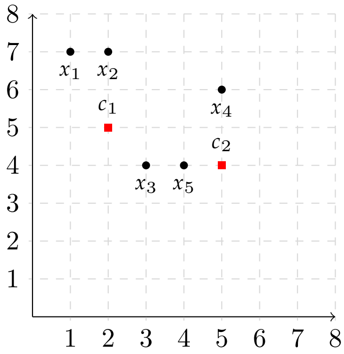

Figure 1: The starting position of the k-means algorithm for Exercise 1

Example solution: W e begin by assigning each point to its closest cluster. The distances for each point to each cluster are:
$d(x_1, c_1) = \sqrt{1^2 + 2^2} = \sqrt{5}$ $d(x_1, c_2) = \sqrt{4^2 + 3^2} = \sqrt{25}$
$d(x_2, c_1) = \sqrt{0^2 + 2^2} = \sqrt{4}$ $d(x_2, c_2) = \sqrt{3^2 + 3^2} = \sqrt{18}$
$d(x_3, c_1) = \sqrt{1^2 + 1^2} = \sqrt{2}$ $d(x_3, c_2) = \sqrt{2^2 + 0^2} = \sqrt{4}$
$d(x_4, c_1) = \sqrt{3^2 + 1^2} = \sqrt{10}$ $d(x_4, c_2) = \sqrt{0^2 + 2^2} = \sqrt{4}$
$d(x_5, c_1) = \sqrt{3^2 + 1^2} = \sqrt{10}$ $d(x_5, c_2) = \sqrt{1^2 + 0^2} = \sqrt{1}$
We assign each point to its closest cluster. x_1, x_2, x_3 are assigned to c1, and x_4, x_5 to c2. We then have to update the cluster centres to the mean of the points within the cluster.
$c_1(x) = (1/3)(1 + 2 + 3) = 2$, $c_1(y) = (1/3)(7 + 7 + 4) = 6$, $c_2(x) = (1/2)(4 + 5) = 4.5$, $c_2(y) = (1/2)(6 + 4) = 5$
Finally, we update the points to their closest centroids again in the same way we did at the start of the exercise. We see that x_3 has now shifted to cluster 2.

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.

Example solution:

We first calculate a distance matrix of all distances between points. Note that since distances are symmetric $d(x_i, x_j) = d(x_j, x_i)$ we only need to calculate one half of the matrix.

	x_1	x_2	x_3	x_4	x_5
x_1	-				
x_2	1	-			
x_3	5	4	-		
x_4	5	4	4	-	
x_5	6	5	1	3	-

The smallest distances are 1 between x_1 and x_2 and between x_4 and x_5. We link x_1 and x_2 first, and update the distance matrix. Since we are using single-linkage, the distance between clusters is the shortest distance between a pair of points:

	(x_1,x_2)	x_3	x_4	x_5
(x_1,x_2)	-			
x_3	4	-		
x_4	4	4	-	
x_5	5	1	3	-

We repeat this process, linking $x3, x5$ with a distance of 1:

	(x_1,x_2)	(x_3, x_5)	x_4
(x_1,x_2)	-		
(x_3, x_5)	4	-	
x_4	4	3	-

Finally, we link (x_3, x_5) with x_4, and then end by linking the two remaining clusters (x_1,x_2), (x_3, x_4, x_5). We have two sensible options for cutting the dendrogram: either at 2 (height 3) or at 3 (height 1) clusters. Given that the point x_4 is almost as close to x_2 as it is to x_5, having it as its own cluster likely makes more sense.

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[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