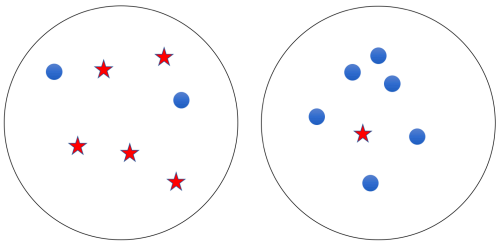
Unsupervised Learning and Evolutionary Computation Using R

Winter Term 2024/2025

Exercise Sheet 5 (December, 11, 2024)

Exercise 1 (Clustering: External Evaluation)



In the figure above you find the final clustering solution for a given problem. The circles indicate which observations are assigned to the same cluster. Luckily we also know the true labelling for this problem (same shape = same class). Therefore we are able to use external evaluation measures to determine the quality of our solution. Please calculate the RAND index and comment on the quality of the clustering solution!

Example solution:

$RI = \frac{TP+TN}{TP+FP+FN+TN} = \frac{TP+TN}{\binom{N}{2}}$

$TP = \binom{5}{2} + \binom{2}{2} + \binom{6}{2} = 26$

$TN = 5 \cdot 6 + 2 \cdot 1 = 32$

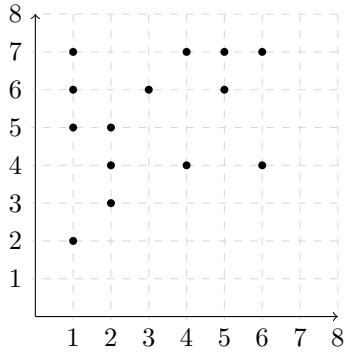
$FN = 1 \cdot 5 + 2 \cdot 6 = 17$

$FP = 5 \cdot 2 + 1 \cdot 6 = 16$

$RI = \frac{26+32}{91} \approx 0.637$

Solution is of medium quality.

Exercise 2 (DBSCAN Clustering By Hand)



In the above figure, you are given a set of points that need to be clustered using the DBSCAN algorithm. Use the parameters $minPts = 4$ and $\epsilon = \sqrt{2}$. For every point, either assign it to a cluster or mark it as noise as appropriate. Keep in mind that when counting points in order to determine core points, the epsilon neighbourhood includes the point itself (i.e., a point is a core point if it has minPts points in its epsilon neighbourhood, including itself).

Example solution:

- The core points are the ones at $(2, 5), (2, 4), (1, 5), (1, 6), (4, 7), (5, 6), (5, 7)$.
- The border points are $(1, 7), (2, 3), (6, 7), (3, 6)$
- noise points are $(1, 2), (4, 4), (6, 4)$
- From them, we will get two clusters, one on the top right and one of the left side
- The border point at $(3, 6)$ is a border point that could belong to either cluster, as it is connected to core points in both clusters. So it will be assigned to whatever cluster we start with

Exercise 3 (Silhouette Score)
In this exercise, you will evaluate the above clustering using the silhouette score. To make things a bit less time consuming, instead of calculating the score for every point, pick one of the clusters and calculate the silhouette score for just the points in the cluster. Interpret your findings: is the assignment found by the algorithm for this cluster good or not.

Example solution:
For an example, let's pick the point at $(2, 5)$ and calculate the Silhoutte score. To make things easier, we will use the Manhattan distance instead of the Euclidean distance
The average distance to its own cluster is $a = \frac{1}{6}(3 + 2 + 1 + 1 + 2 + 2) = 1.8$. The average distance to the closest other cluster is $b = \frac{1}{4}(4 + 5 + 4 + 6) = 4.75$
Based on the Silhouette definition, we will use the formula $1 - \frac{a}{b} = 1 - \frac{1.4}{4.75} = 0.62$, a fairly good result for this point.
Regardless of the cluster that we pick, most points will be clustered pretty well. The one exception will be the border point that could belong to either cluster, which should be at around 0.

Exercise 4 (`DBSCAN` Parameter Selection)
The `DBSCAN` exhibits the two parameters ϵ and `minPts`. Both have a significant impact on the clustering results. In general, the developers of `DBSCAN` recommend that `minPts` should be in the range of $[4, 2p]$ where p is the number of features in your dataset. For a given value of `minPts`, you can determine an optimal value for ϵ by using a knn-distance plot (see lecture). You will perform this in the following:

1. Create the data set with the following code using the package `mlbench`

```
library(mlbench)
data = as.data.frame(mlbench.spirals(500, cycles=2, sd=0)$x)
colnames(data) = c("x", "y")
```

2. Select the appropriate value for `minPts` based on the heuristic given above.
3. Determine the value of ϵ by creating a k-distance plot. Use the method `kNNdist` of the `dbscan` package.

4. Now, cluster the same dataset using k-Means with $k = 2$.
5. Plot both cluster assignments (of `DBSCAN` and k-Means) in a single plot using `facet_grid`.
6. Which algorithm succeds (in terms of correct cluster assignments), which not and what are possible reasons for failure?