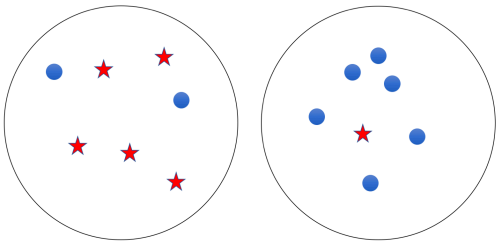
Unsupervised Learning and Evolutionary Computation Using R

Winter Term 2024/2025

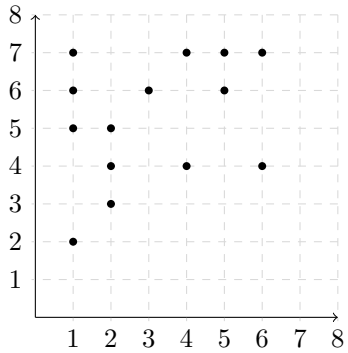
Exercise Sheet 5 (December, 11, 2024)

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!

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).

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.

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?