MTH6101 Introduction to Machine Learning

Laboratory week five

The intention of this laboratory is to do an introduction to agglomerative clustering in `R`. The tasks involve the library `cluster`.

1. Open `RStudio` and a new `R` script. You need to install the library `cluster`.

2. Type the data of lectures in a matrix called `X`:

x_1	x_2
0	4
3	6
6	2
0	5
1	1

3. Load the library `cluster` and use the function `dist` to build a distance matrix. With this function, build distance matrices exploring different distances such as `manhattan`, `euclidean` and `minkowski` (with exponent equal to four). **Verify** that you can reproduce some the distances between points one and three. Continue exploring this function with different values of the **logical** input flags `diag` and `upper`.

4. Using **euclidean** distance and **complete** linkage, perform aglomerative clustering with these data. To this end, use the function `agnes` from the library you just loaded and save the output in an object termed `A`.

5. Use `plot(A,which.plot=2)` to plot the dendrogram of this analysis.

6. It is clear that the first cluster is formed at distance equal to one. By manipulation of the elements in the distance table, coerced to be a matrix, verify that the (complete linkage) distance between clusters '14' and '2' is precisely the one in the dendrogram. Recall that the distances in the dendrogram are retrieved by `A$height`.

7. Consider the dataset `ruspini` from the same library. Do a scatterplot of these data and use the `text` function to put individual labels in it.

8. Perform agglomerative clustering using **euclidean** distance and **average** linkage. Plot the cluster and try an interpret the dendrogram. Large stems suggest places to cut the diagram and detect clusters.

9. Compare with the dendrogram using the same distance and **single** linkage. Comment on the differences and similarities.