Exploratory data analysis: Clustering Methods

R. Benítez

Exercise 1: Generate synthetic data (2D multivariate normal):

Task 1: Generate two 2D data vectors of size $N = 10^3$ following two multivariate normal distributions with parameters $(\vec{\mu}_1, \Sigma_1)$ and $(\vec{\mu}_2, \Sigma_2)$ respectively.

Task 2: Represent the data using a scatter plot.

Exercise 2: Compare different clustering methods

Task 1: Apply a k-means clustering algorithm with $k = 2$ and represent the results as a scatter plot with different colors for the two classes.

Task 2: Apply a hierarchical clustering algorithm with $k = 2$ and represent the results as a scatter plot with different colors for the two classes. Compare agains the results in the previous task.

Task 3: Apply a Gaussian Mixture Model clustering algorithm with $k = 2$ and represent the results as a scatter plot with different colors for the two classes. Compare agains the results in the previous tasks.

Task 4: Use the Bayesian Information Criterion (BIC) and the Akaike Criterion in order to choose the optimal Gaussian Mixture Model to fit the data.

Task 5: Change the centroids of the two data clusters in order to get them closer. Repeat the previous tasks in this exercise and comment the results. Which clustering algorithm is more robust to overlap between observations of different classes?.

Task 6: Increase the number of classes from 2 to 5. Comment on the performance of the different clustering methods when the number of classes increase.

Exercise 3: Application to real data:

Task 1: Obtain a dataset from the UCI Machine Learning Repository (keep only features, disregard the class labels).

Task 2: Apply GMM clustering and determine the optimum number of clusters using BIC or AIC. Compare the clustering results with the ones obtained with kmeans and hierarchical clustering using this number of clusters.

Useful references

Generate random samples from a multivariate normal distribution:

- MATLAB: `https://es.mathworks.com/help/stats/mvnrnd.html`
- Python: `https://docs.scipy.org/doc/numpy/reference/generated/numpy.random.multivariate_normal.html`
- R: `https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/mvrnorm.html`

UCI Machine Learning Repository:

- MATLAB: `http://archive.ics.uci.edu/ml/`

Clustering algorithms:

- MATLAB (k-means): `https://es.mathworks.com/help/stats/kmeans.html`
- MATLAB (hierarchical): `https://es.mathworks.com/help/stats/hierarchical-clustering.html`
- MATLAB (Gaussian Mixture Models): `https://es.mathworks.com/help/stats/clustering-using-gaussian-mixture-models.html`
- Python (k-means): `http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`
- Python (hierarchical): `http://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering`
- Python (Gaussian Mixture Models): `http://scikit-learn.org/stable/modules/mixture.html`
- R (k-means): `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html`
- R (hierarchical): `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html`
- R (Gaussian Mixture Models): `http://www.stat.washington.edu/mclust/`