Florian
Stanilewicz

MA 412 – Projet

Introduction :

We have a dataset with 18 features and 1874 samples, representing aircraft trajectories. Now, we would like to group using these features, to determine similar aircraft trajectories configurations.

Unsupervised classification methods, commonly referred to as clustering algorithms, aim to categorize data into groups or clusters based solely on the inherent structure or patterns within the data. Unlike supervised learning, where the algorithm learns from labelled data (input-output pairs), unsupervised learning operates on unlabelled data, seeking to discover the natural groupings or relationships among data points.

To test these different clustering methods, we have to score them on how they perform on our dataset. To score them, we will use the Silhouette Score and the Davies-Bouldin index scores.

Silhouette Score :

The Silhouette Score measures the quality of clustering by computing the average distance between a data point and all other points within the same cluster (intra-cluster distance) and comparing it to the average distance between that data point and all points in the nearest neighbouring cluster (inter-cluster distance).

The silhouette score ranges from -1 to +1, a score close to +1 indicates that the data point is well-clustered and lies far away from neighbouring clusters.

A score close to 0 indicates that the data point is close to the decision boundary between clusters.

A negative score suggests that the data point might have been assigned to the wrong cluster.

The average silhouette score for all samples is used as an overall measure of the clustering quality. A higher average silhouette score implies better-defined clusters.

Davies-Bouldin Index

The Davies-Bouldin Index measures the average similarity between each cluster and its most similar cluster, taking into account both intra-cluster and inter-cluster distances.

A lower Davies-Bouldin Index indicates better clustering; smaller values suggest better separation between clusters.

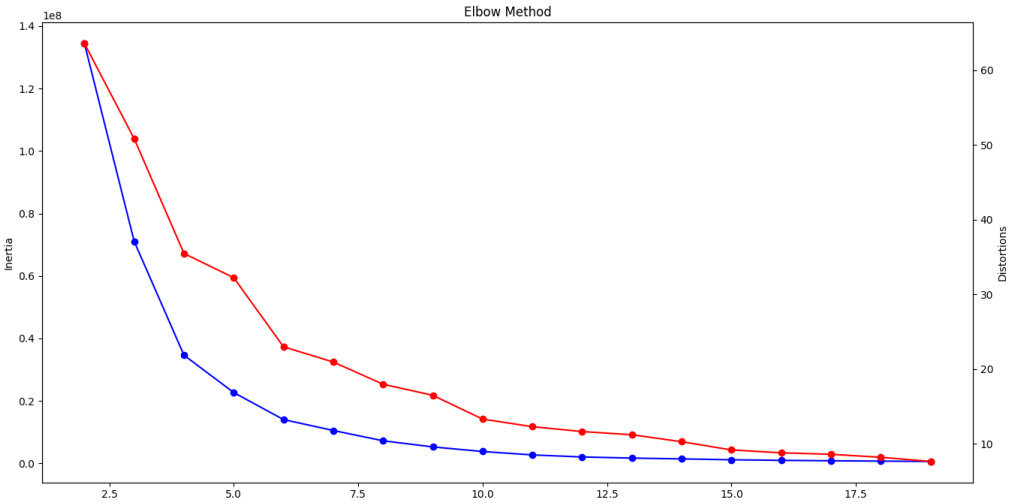
Both metrics are useful for evaluating different clustering algorithms or choosing the optimal number of clusters by comparing scores for various clustering configurations. The aim is to maximize the Silhouette Score and minimize the Davies-Bouldin Index for better-defined and well-separated clusters.

Determining the number of clusters :

For most clustering methods, determining the number of clusters that best describe the dataset is a crucial step before comparing these methods.

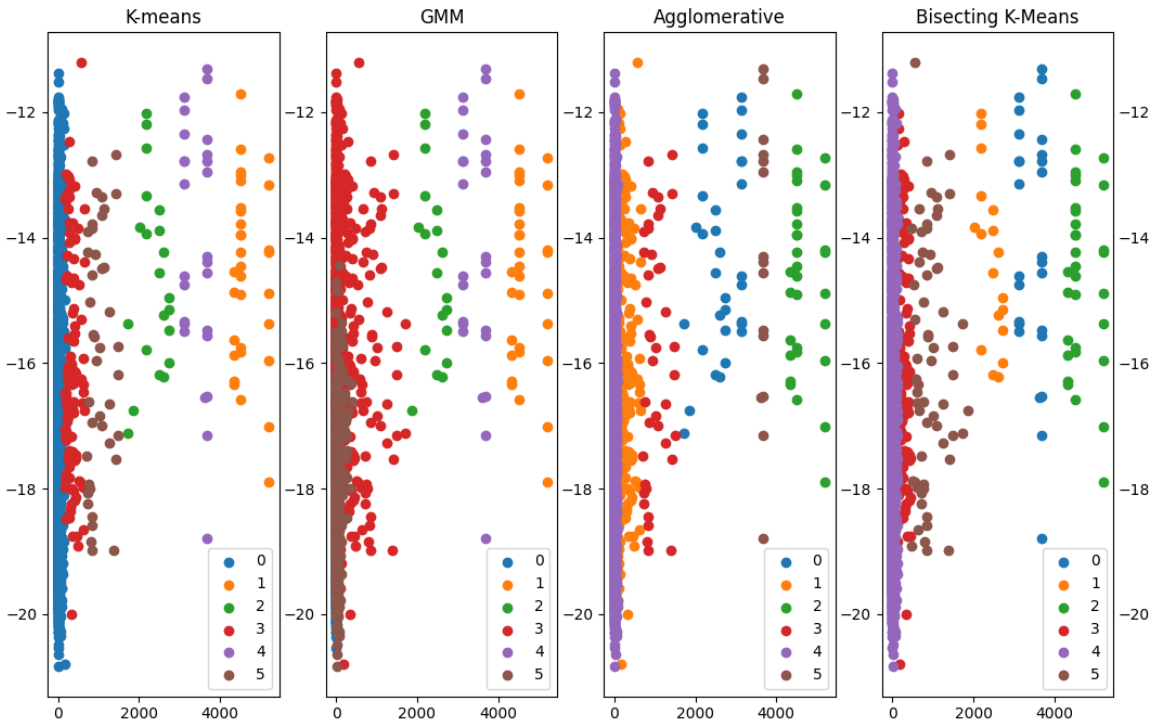
Florian
Stanilewicz

A common and simple method is to use the “Elbow Method”, which consists of looking at inertia and distortion for K-Means method for different numbers of centroids.



We plot the result of inertia and distortion in relation to the number of clusters. As expected, we find that the diminishing returns point (the ‘elbow’) of the curve would be at N = 6.

Results on the methods :



We can see that for a cluster size of 6, K-Means and the Agglomeration clustering methods perform the best on the dataset, compared to the Gaussian mixture model and bisecting K-Means methods.

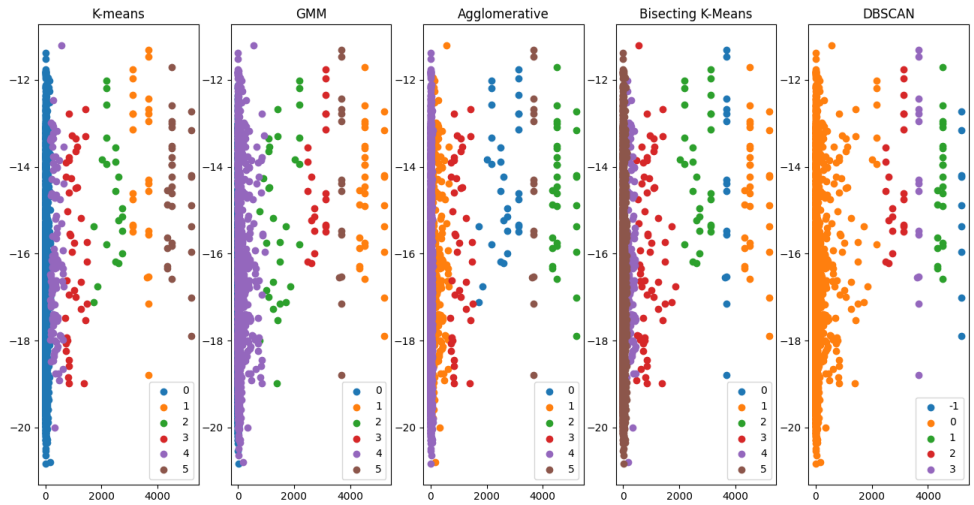
Florian
Stanilewicz

DBSCAN :

We will also test the use of DBSCAN, which doesn't have a fixed number of clusters (which includes noise) but uses density of the data. To tune it, a heuristic method used is to first defines the minimum points based on the number of dimensions + 1 (here 18).

Then, we try the clustering method on a range of epsilon from 2 to 600 to try to match the number of clusters found by the elbow method.

We find that at eps = 500, we have the numbers of clusters N = 4 (excluding noise).



Conclusion :

Using this method and comparing it to the others with its Silhouette and Davies-Bouldin, we find that the DBSCAN method is the best scoring method of them all, using only 4 clusters (so there is no overfitting).

Thus, we can conclude that it is the best method for our dataset.