Clustering

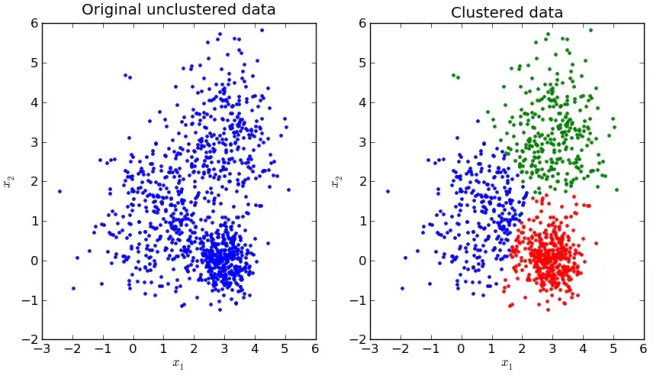
Lluís Talavera, 2022

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

ETSEIB

Clustering

- It does not requires labeled data (unsupervised learning).
- Goal: discover *meaningful* or *natural* groups (clusters).
- Meaningful/natural are vague terms; in practice clustering algorithms make assumptions about what it is considered a good partition.
- It is useful to think as points in a hyperspace where members of a cluster are more similar between them and different from the members of the other clusters.



k-means

- A centroid-based model. Each cluster is represented by a *centroid* or *center*, a kind of summary where the value of each feature f_i is the average of the values of f_i across all the examples of the cluster.
- Finds the centroids which minimize the within-cluster distance or within-cluster sum squared error (SSE).
- Requires a distance metric (same as k-NN).

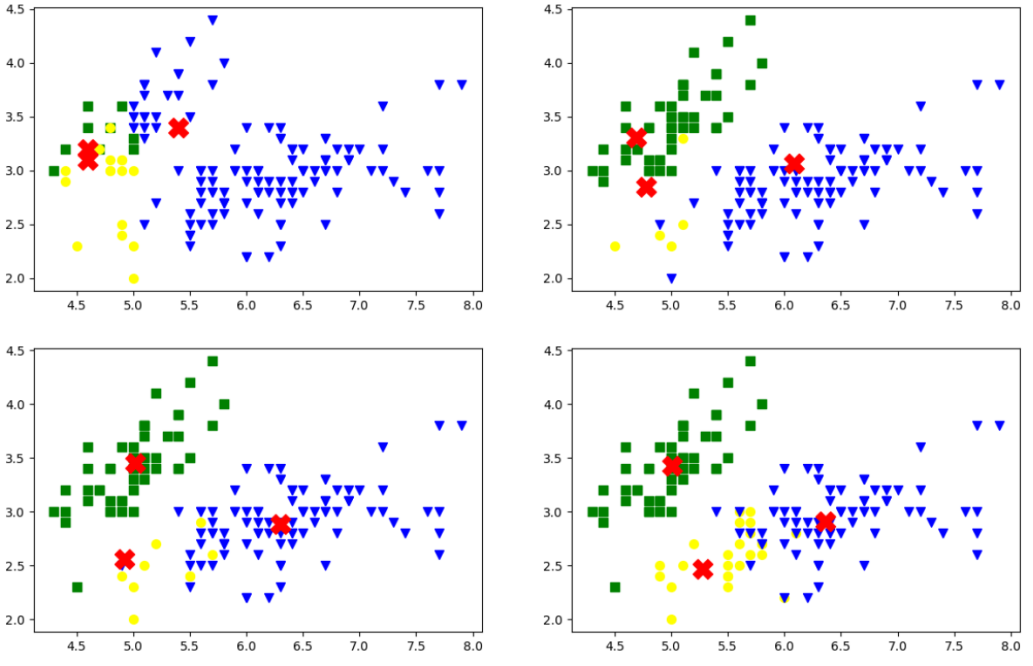
Algorithm:

Input: a set of examples, k number of clusters

1. Randomly choose k examples as initial centroids
2. Assign the examples to their closest centroid (cluster)
3. Compute the new centroid of each cluster
4. Go back to step 2 until the assignment does not change or a number of iterations is reached

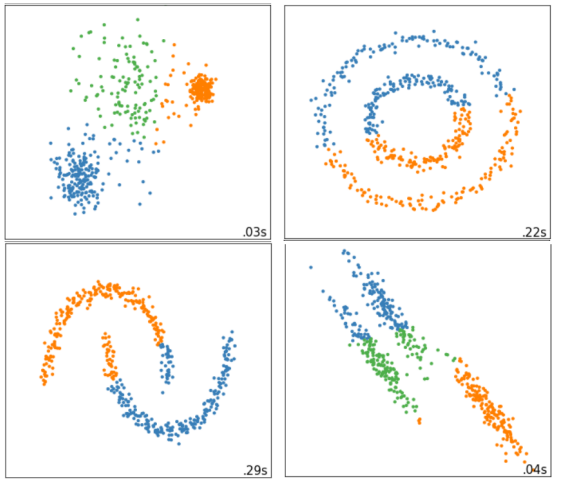
Ouput: k centroids and the cluster assignment of each example

k-means iterations



Some remarks about k-means

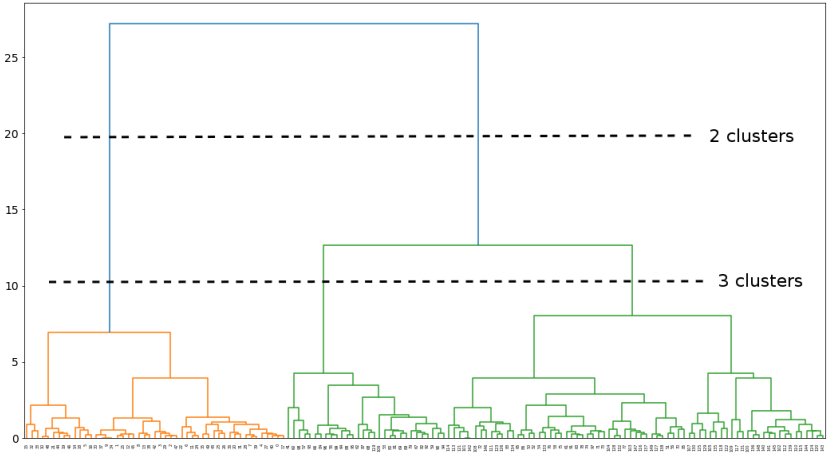
- Widely used and fast.
- Highly dependant on cluster initialization.
- Sensitive to scaling (distance-based).
- Sensitive to outliers.
- Not suitable for clusters with non-convex shapes.
- Assumes all clusters have equal, spherical variance, i.e., each cluster has roughly equal number of observations and the clusters tend to form a sphere shape.



Hierarchical clustering (I)

Start with each example in its own cluster and, at each step, merge the more similar pair of clusters until all the examples are in a single cluster (agglomerative clustering).

The result is a tree called dendrogram. Cutting the dendrogram we obtain flat clusters.



Hierarchical clustering (II)

Merging criteria:

- single linkage: minimum distance between members of the clusters.
- complete linkage: maximum distance between members of the clusters.
- average linkage: average distance between members of the clusters.
- centroid linkage: distance between centers.
- ward linkage: smallest increase in within cluster variance.

Some remarks:

- More informative structure, can help to decide the number of clusters.
- Greedy algorithm, cannot undo earlier mergings.
- Higher computation time.
- Sensitive to outliers.

DBSCAN

Spatial clustering. Assumes that clusters are continuous dense regions in the data space.

Two hyperparameters:

- `eps`: minimum distance to consider two points as neighbors. If the value is too small, a lot of data will be considered as outliers.
- `min_samples`: the minimum number of points clustered together (within eps distance) for a region to be considered dense.

Algorithm:

1. Select a random point among ponts not assigned to any cluster.
2. If there are at least `min_sample` points within a radius of `eps` to the point (core point) then we consider all these points to be part of the same cluster. If not, the point is considered noise.
3. All the points within the neighborhood of initial point become a part of the cluster. If these new points are also core points, the points that are in the neighborhood of them are also added to cluster. When no more points are available, go to 1.

Some remarks about DBSCAN

- Does not require to specify the number of clusters.
- Good values for the hyperparameters may not be easy to determine.
- Performs well with arbitrary shapes.
- Not sensitive to outliers.
- Performance can degrade for high dimensional data, the distance metric can become distorted.
- May not work well if clusters have very different densities.