BIG DATA STATISTICS FOR BUSINESS

AY 2023-24

Giovanni De Luca
Parthenope University of Naples

1

k-means algorithm

- The *k*-means algorithm is a clustering methodology that requires the specification of the number of clusters (*k*).
- It is an approach alternative to hierarchical clustering (that does not to specify *k*).

2

k-means algorithm

- The idea behind *K*-means clustering is that a good clustering is one for which the within-cluster dissimilarity is as small as possible (high intra-class similarity) and the objects from different clusters are as dissimilar as possible (low inter-class similarity).
- The within-cluster dissimilarity for cluster C_k is a measure of the amount by which the objects within a cluster differ from each other.
- The total within-cluster dissimilarity, summed over all *k* clusters, has to be as small as possible.
- Objective: to minimize the total within-cluster dissimilarity.

3

k-means algorithm

- The within-cluster dissimilarity for the *k*-th cluster is the sum of all the Euclidean distances between the objects in the cluster and the centroid of the cluster.
- The cluster centroid corresponds to the mean of points assigned to the cluster.
- Let's remind that the use of Euclidean distances implies that:

1. standardization has to be carried out in presence of differente scale of measurements;
2. categorical variables are not allowed.

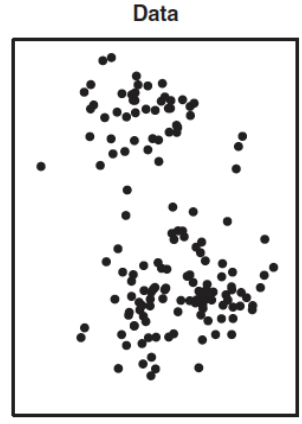
4

k-means algorithm

- The algorithm starts randomly assigning each of the objects to a cluster, from 1 to *k.*

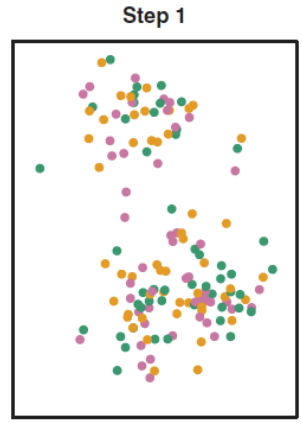
5

Example of *k*-means in two dimensions, *k*=3



6

Example of *k*-means in two dimensions, *k*=3



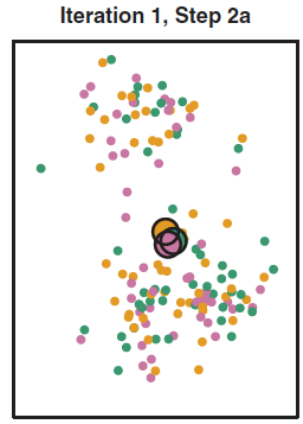
7

k-means algorithm

- The algorithm starts randomly assigning each of the objects to a cluster, from 1 to *k*.
- For each of the *k* clusters, the cluster center (centroid) is computed.

8

Example of *k*-means in two dimensions, *k*=3



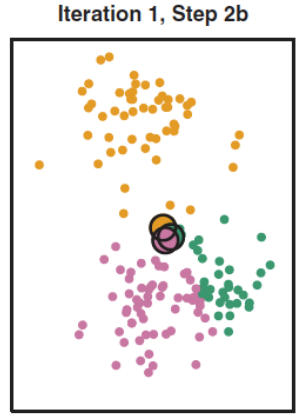
9

k-means algorithm

- The algorithm starts randomly assigning each of the objects to a cluster, from 1 to *k*.
- For each of the *K* clusters, the cluster centroid is computed.
- Then, each object is assigned to the cluster whose centroid is closest.

10

Example of *k*-means in two dimensions, *K*=3



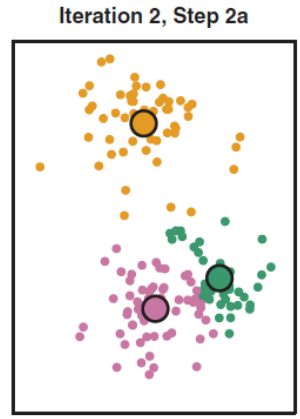
11

k-means algorithm

- The algorithm starts randomly assigning each of the objects to a cluster, from 1 to *k*.
- For each of the *K* clusters, the cluster centroid is computed.
- Then, each object is assigned to the cluster whose centroid is closest.
- A new centroid is computed for each cluster.

12

Example of *k*-means in two dimensions, *K*=3



13

k-means algorithm

- The algorithm starts randomly assigning each of the objects to a cluster, from 1 to *k.*
- For each of the *K* clusters, the cluster centroid is computed.
- Then, each object is assigned to the cluster whose centroid is closest.
- A new centroid is computed for each cluster.
- Each object is assigned to the cluster whose centroid is closest.

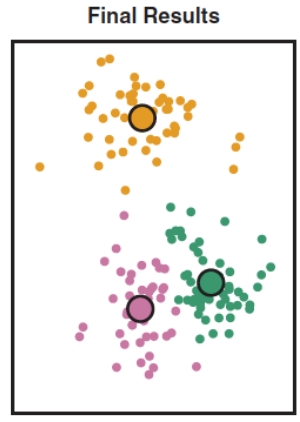
14

k-means algorithm

- The algorithm starts randomly assigning each of the objects to a cluster, from 1 to *k.*
- For each of the *K* clusters, the cluster centroid is computed.
- Then, each object is assigned to the cluster whose centroid is closest.
- A new centroid is computed for each cluster.
- Each object is assigned to the cluster whose centroid is closest.
- Iterate the last two steps until the cluster assignments stop changing (or the maximum number of iterations is reached; by default, *R* uses 10 as the default value for the maximum number of iterations).

15

Example of *k*-means in two dimensions, *K*=3



16

Conclusions

- The k-means clustering algorithm is straightforward and quick.
- It can handle extremely big data sets with efficiency.
- Nevertheless, it has certain drawbacks:

1. it assumes prior knowledge of the data and requires the analyst to choose the appropriate number of clusters (*k*) in advance
2. the final outcome can be sensitive to the initial random selection.

17

Mixed data

- The basic concept of *k*-means stands on mathematical calculations (Euclidean distances, means).
- If our data is non-numerical, the solution can be found in an extension of the *k*-means algorithm, the *k*-modes algorithm.
- Finally, for numerical and categorical data (mixed data), another extension of these algorithms exists, it is called *k*-prototypes.

18