K-means

Mirco Musolesi
Department of Geography, UCL
m.musolesi@ucl.ac.uk

Clustering

- In general, data sets usually show some forms of "clusters":
  - Millionaires live in some parts of London.
  - Datasets of customers of a supermarket or voters can be "clustered" in groups.
- There are *many* algorithms for clustering.
- Clustering does not provide a "label" for each cluster we extract, but it is possible to look at the content of each cluster in order to assign a label to a cluster.

Definition of a Distance

- In order to cluster together different people we need a measure of distance:
  - It can be defined as a measure of difference between two entities (example: you look at the number of common products bought by two customers in a supermarket, two users are more similar if they bought the same products, i.e., they are less “distant”).
  - In geographic terms, you can use geodesic distance (or Euclidean, Haversine, etc.).
- For simplicity in this module, we will perform clustering of geographic points.

K-means Algorithm

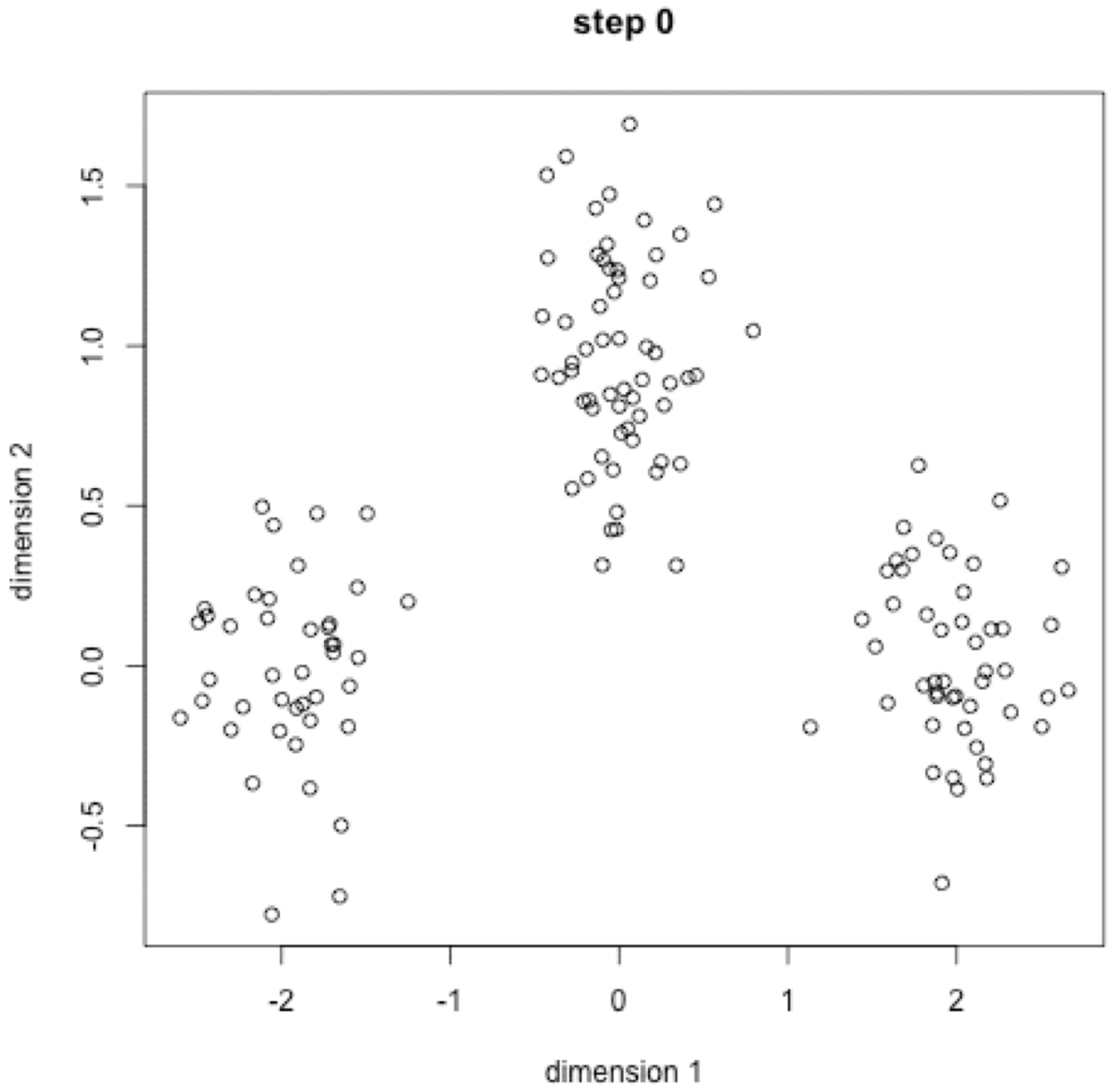
- This is one of the simplest, but at the same time very popular algorithms for clustering.
- In the *k*-means algorithm, the number of clusters (equal to *k*) is chosen in advance.
- The goal is to partition the inputs into sets S_1,…S_2, ...S_k in a way that the minimises total sum of the squared distances from each point to the mean of its assigned cluster.

K-means Algorithm

- There are many ways of assigning *n* points to *k* clusters.
- Optimal clustering is a very hard problem.
- We will discuss a possible implementation, where the number of the *k* clusters is an input of the algorithm, i.e., it is defined a priori.

K-means Algorithm

- We will use an *iterative* algorithm composed of these 4 steps:
  1) Start with a set of *k*-means, which are points for example in a 2-dimensional space (but it can also be d-dimensional!).
  2) Assign each point to the mean to which it is closest.
  3) If no point's assignment has changed, stop and keep the clusters.
  4) If some point's assignment has changed, recompute the means and return to step 2.



Credit: Simply Statistics

Practical: Clustering of Location Data

- Implement the algorithm using the data you have trying to modify the value of *k.*
- The code will be provided in the lab
- **Additional exercise**: you might want to explore the algorithm using a set of points loaded from a text file generated by you.

References

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning. Springer. 2014
  - Chapter 1: “Statistical Learning”
  - Chapter 10: “Unsupervised Learning”

The book is freely available online at:

http://www-bcf.usc.edu/~gareth/ISL/

References

- A discussion of the k-means algorithm in Python can be found in:

  Joel Grus. Data Science from Scratch. O'Reilly 2015. Chapter 19.

The e-book is available for free for UCL students (follow the link from the Library Catalogue).

References

- The k-means algorithm was described for the first time in:

  J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297.

[This is a very mathematical paper: I am listing it just for reference – not required for the exam]