RECSM Summer School: Machine Learning for Social Sciences

Session 3.3: K-Means Clustering

Reto Wüest

Department of Political Science and International Relations
University of Geneva

Clustering

Clustering

- Clustering refers to a set of techniques for finding subgroups, or clusters, in a data set.
- The goal is to partition the observations of a data set into distinct groups so that the observations within each group are similar to each other, while the observations in different groups are different from each other.
- This is an unsupervised problem because we are trying to discover structure (distinct clusters) on the basis of a data set.

Clustering Versus PCA

- Both clustering and PCA seek to simplify data via a small number of summaries.
- However, their mechanisms are different:
    - PCA tries to find a low-dimensional representation of the observations that explains a large fraction of the variance;
    - Clustering tries to find homogeneous subgroups among the observations.

K-Means Clustering and Hierarchical Clustering

- There are many clustering methods; K-means clustering and hierarchical clustering are the two best-known approaches.
- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.
- In hierarchical clustering, we do not know in advance how many clusters we want.
- We can cluster observations on the basis of the features in order to identify subgroups among the observations; or we can cluster features on the basis of the observations in order to discover subgroups among the features.

Clustering

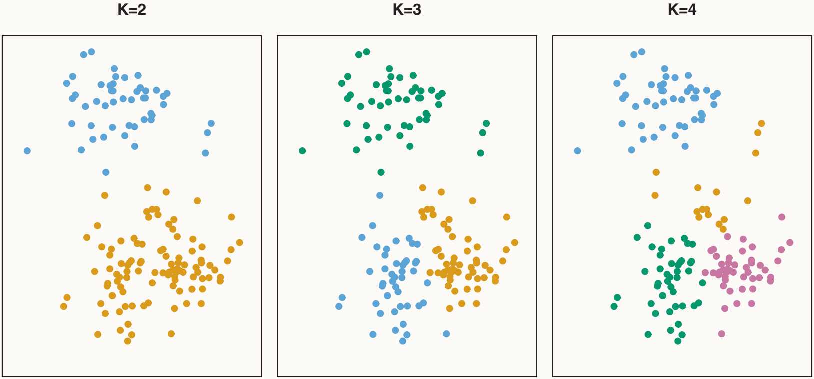
K-Means Clustering

K-Means Clustering

- K-means clustering partitions a data set into K distinct, non-overlapping clusters.
- We must first specify the desired number of clusters K.
- The K-means algorithm then assigns each observation to exactly one of the K clusters.

K-Means Clustering – Example

Simulated data set with 150 observations in two-dimensional space



(The colors of the observations are the output of the clustering algorithm: they indicate the cluster to which each observation was assigned by K-means clustering. Source: James et al. 2013, 387)

Details of K-Means Clustering

- Let $C_1, \dots, C_K$ denote sets containing the indices of the observations in each cluster.
- These sets satisfy two properties:
  1. $C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, n\}$. In other words, each observation belongs to at least one of the K clusters.
  2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, no observation belongs to more than one cluster.
- The goal is to find a good clustering, i.e., one for which the within-cluster variation is as small as possible.

Details of K-Means Clustering

- The within-cluster variation $W(C_k)$ is a measure of the amount by which the observations within cluster C_k differ from each other.
- We want to partition the observations into K clusters such that the sum of the within-cluster variation is as small as possible:

$$\underset{C_1,\dots,C_K}{\arg\min}\left\{\sum_{k=1}^{K} W(C_k)\right\}. \tag{3.3.1}$$

- To solve (3.3.1), we need to define the within-cluster variation $W(C_k)$.

Details of *K*-Means Clustering

- The most common definition of $W(C_k)$ is

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \qquad (3.3.2)$$

  where $|C_k|$ is the number of observations in cluster C_k.

- Combining (3.3.1) and (3.3.2) gives the optimization problem in K-means clustering:

$$\underset{C_1,\ldots,C_K}{\arg\min} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \qquad (3.3.3)$$

Details of *K*-Means Clustering

- Solving (3.3.3) is a very difficult problem, since there are many(!) ways to partition n observations into K clusters (unless K and n are small).
- However, the following algorithm can be shown to provide a local optimum to the K-means optimization problem.

Clustering

Algorithm for K-Means Clustering

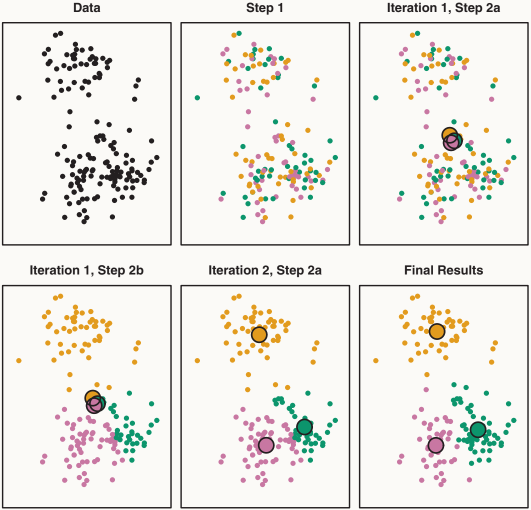
Algorithm for K-Means Clustering

Algorithm: K-Means Clustering

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance, i.e., the "straight-line" distance between two points).

Algorithm for K-Means Clustering

K-means algorithm run on the simulated data set with 150 observations ($K = 3$)



(Source: James et al. 2013, 389)

Algorithm for K-Means Clustering

- Because the K-means algorithm finds a local rather than a global optimum, the results obtained will depend on the initial random cluster assignments in Step 1 of the algorithm.
- Therefore, it is important to run the algorithm multiple times with different random initial values.
- Then one selects the best solution, i.e., that for which the objective (3.3.3) is smallest.

Algorithm for *K*-Means Clustering

Local optima obtained by running K-means clustering six times using different initial cluster assignments

(Above each plot is the value of the objective (3.3.3). Source: James et al. 2013, 390)