K-means Clustering

Unsupervised Learning

- a type of unsupervised learning, which is used when you have unlabeled data.
- The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.
- The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

- Business uses: The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.
- Behavioral segmentation:
  - Segment by purchase history
  - Segment by activities on application, website, or platform
  - Define personas based on interests
  - Create profiles based on activity monitoring

Algorithm

- The algorithms starts with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:
- Data assigment step:
  - Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if ci is the collection of centroids in set C, then each data point x is assigned to a cluster based on

$$\underset{c_i \in C}{\arg\min}\ dist(c_i,\ x)^2$$

  - Let the set of data point assignments for each ith cluster centroid be Si.
  - Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means:

$$S_i^{(t)} = \big\{x_p : \big\|x_p - m_i^{(t)}\big\|^2 \le \big\|x_p - m_j^{(t)}\big\|^2\ \forall j, 1 \le j \le k\big\},$$

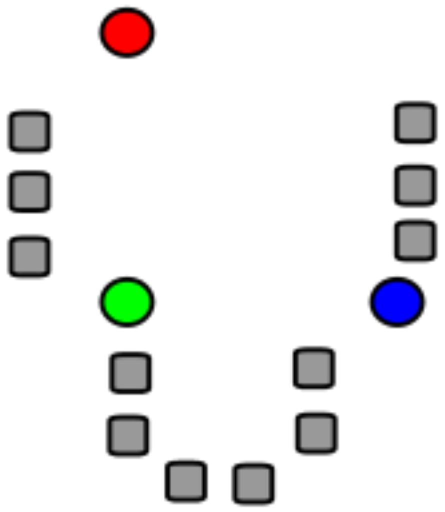
where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.

- Centroid update step:
  - In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

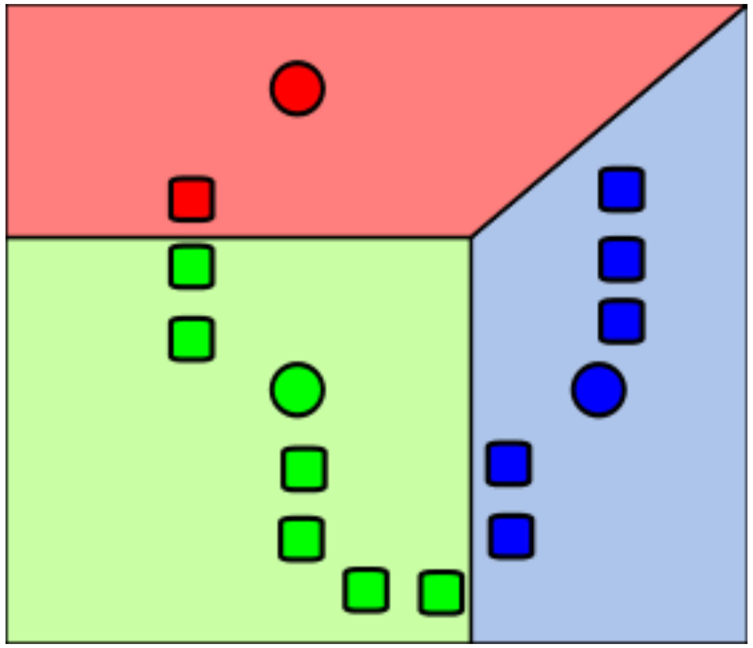
$$c_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} x_i$$

Algorithm

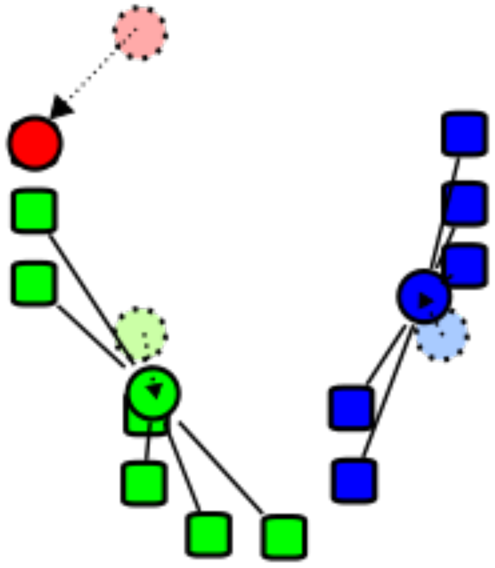
- The algorithm iterates between steps one and two until a stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).



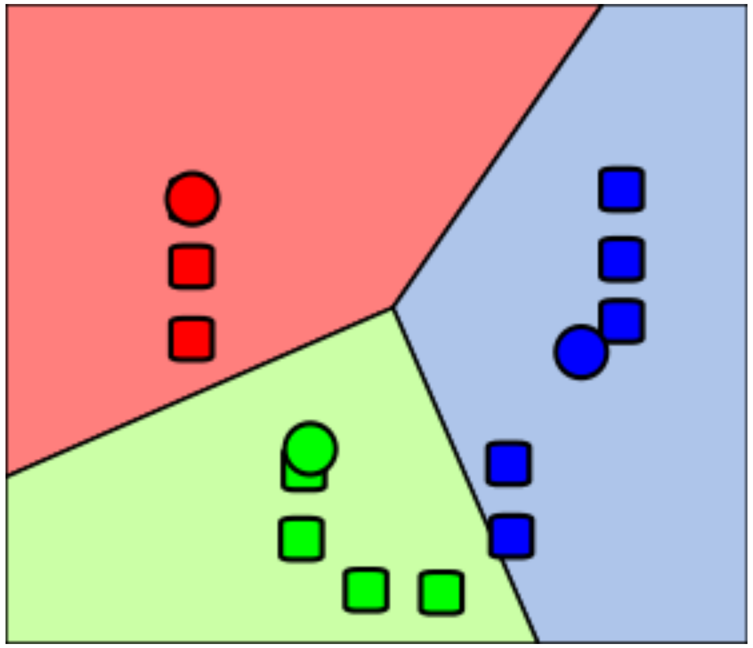
1. k initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).



2. k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3. The centroid of each of the k clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

Algorithm

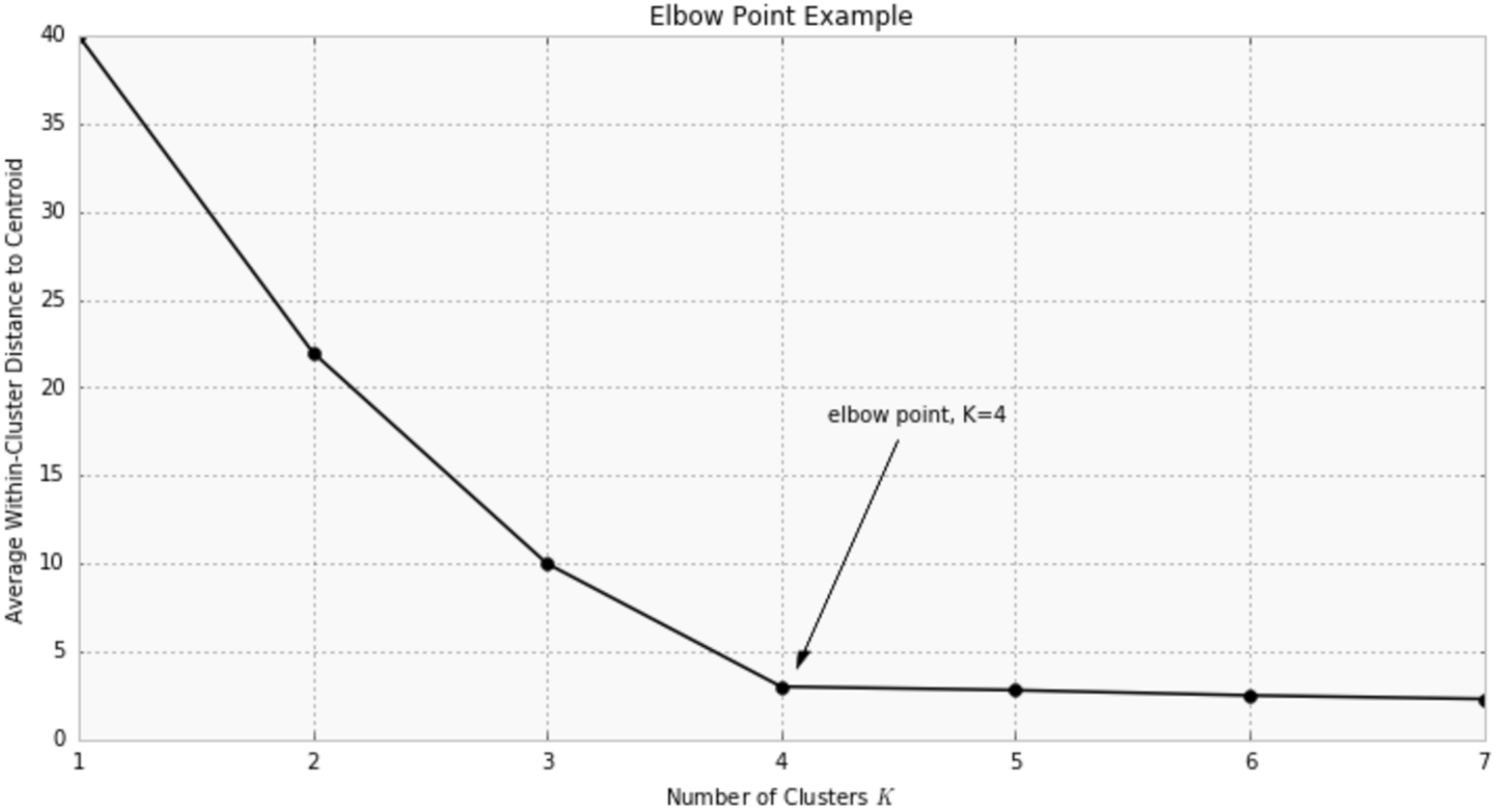
- The algorithm iterates between steps one and two until a stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).
- minimize the total within-cluster variation

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

- |Ck|: total number of observations in cluster k
- i: indices of observations in cluster Ck
- p: number of variables in the dataset

How to choose K?

- One of the metrics that is commonly used to compare results across different values of K is the mean distance between data points and their cluster centroid.
- Since increasing the number of clusters will always reduce the distance to data points, increasing K will always decrease this metric, to the extreme of reaching zero when K is the same as the number of data points.
- Mean distance to the centroid as a function of K is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine K.



K-means clustering advantages and disadvantages

Advantages of K-means clustering algorithm:

- It is simple to implement.
- If variables are huge, K-Means run most of the times computationally faster if we keep k smalls.

Limitation to K-nearest neighbors algorithm:

- Difficult to predict K-Value
- With global cluster, it didn't work well.
- Different initial partitions can result in different final clusters.
- It does not work well with clusters (in the original data) of different size and different density.