ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 4

Clustering

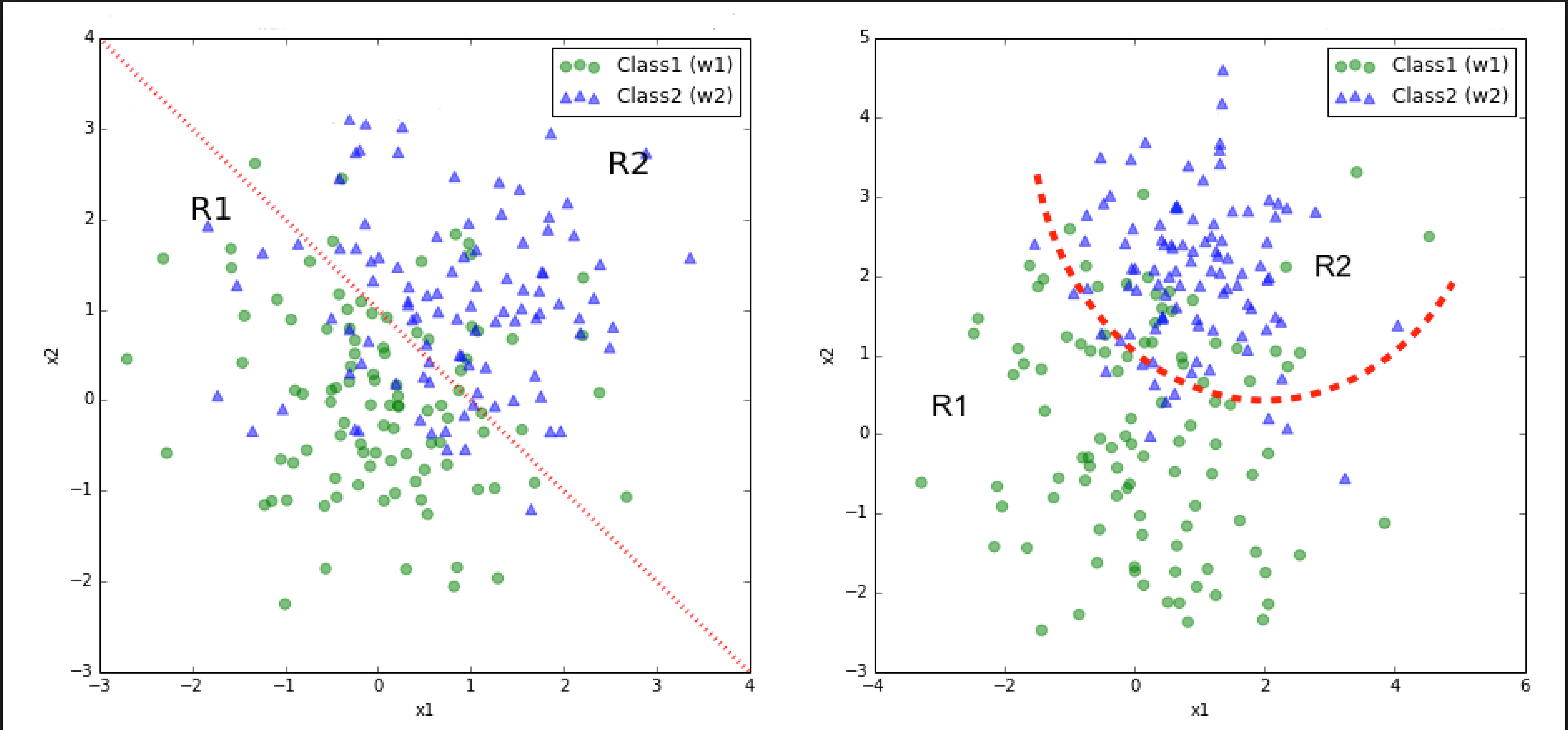
Professor: Dr. Anu Thomas

Learning

- Supervised learning – classification, regression
- Unsupervised learning – clustering, outlier detection
- Semi-supervised

Supervised learning: Classification

- Data has class labels
- Based on the labels, classifiers are generated
- New data will be classified based on the generated classifier
- Predicts a **discrete** class label
- Example 1: Cancer dataset – Malignant and benign labels are present for each instance.
- Example 2: Iris dataset – data from 3 types of flowers – every instance has a class label



https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html

Unsupervised learning

- data has no class labels

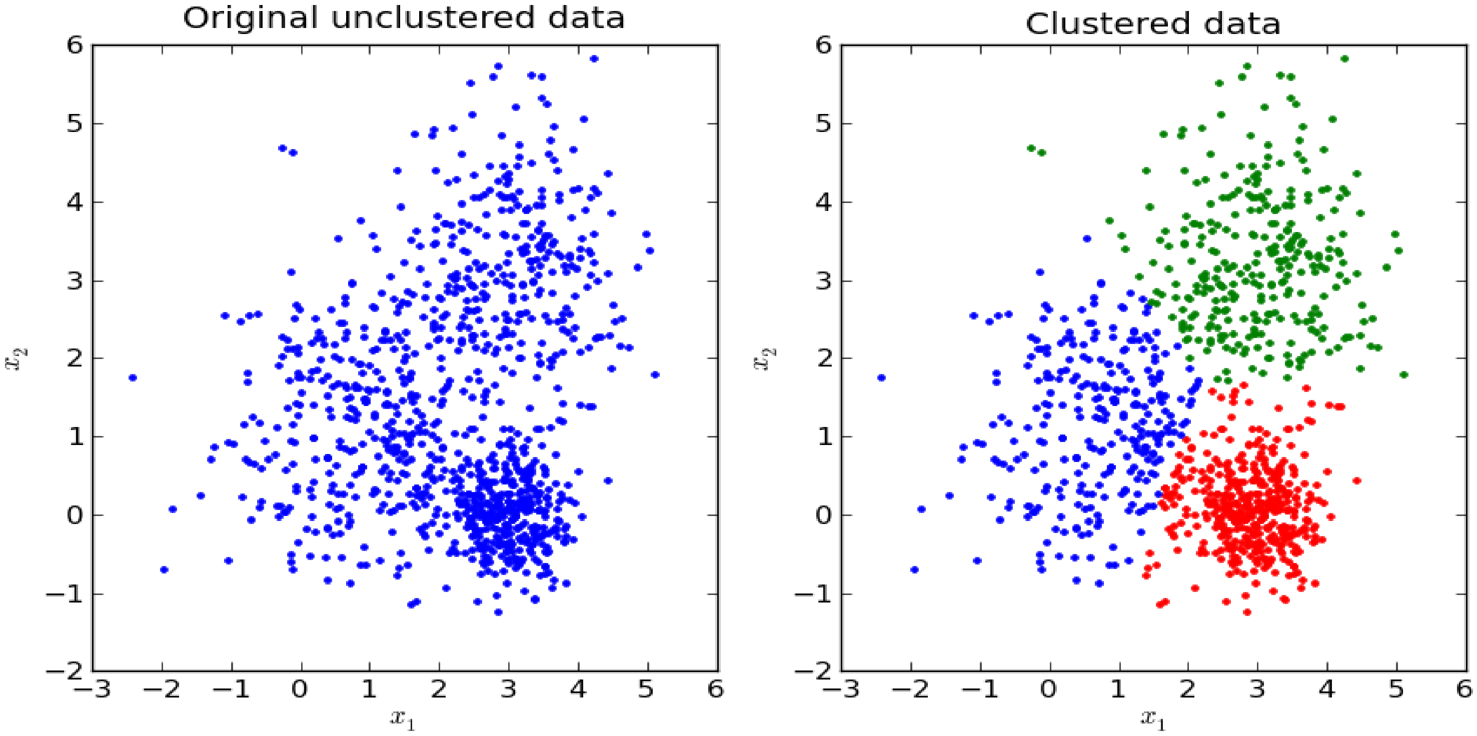
- The algorithm tries to identify the objects as being part of some group using a clustering algorithm. Similar instances grouped together to form clusters. (Ex. Insurance: Identifying groups of motor insurance policy holders with a high average claim cost)

- Anomaly detection tries to find those instances which are distinct from the nature of the majority of instances. (Ex. Financial fraud detection)

Clustering

- Cluster – a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Clustering is an unsupervised classification method

Clustering - example



Clustering Algorithms

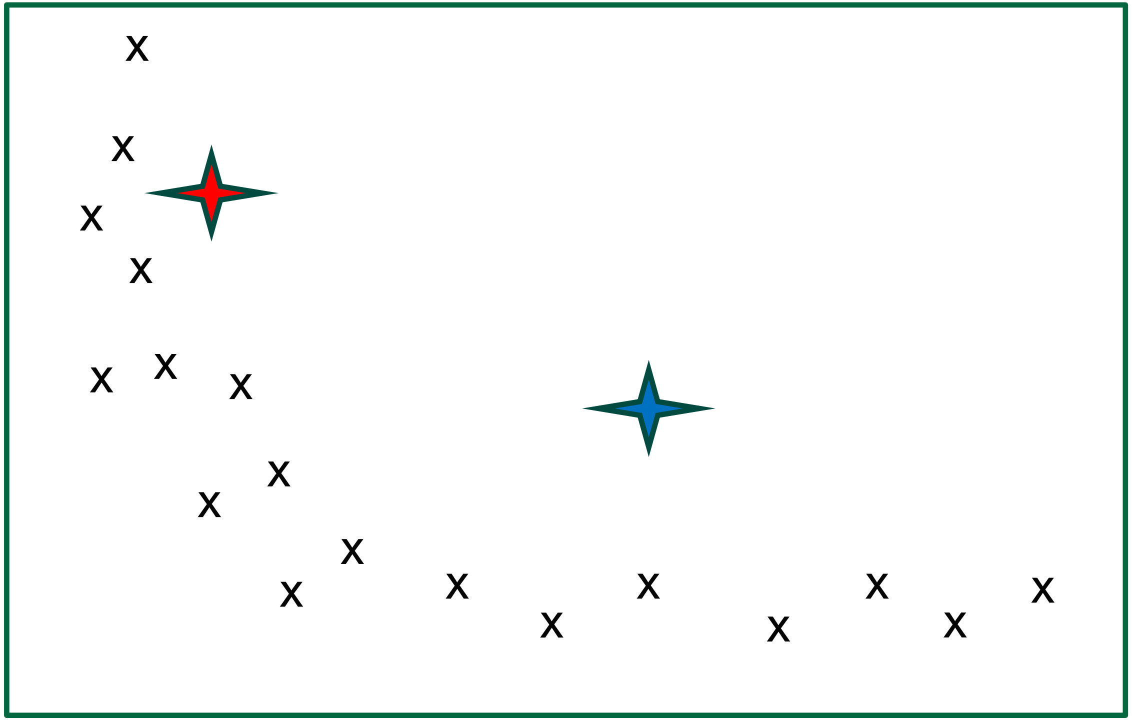
- K-Means
- Mean-shift
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Expectation-Maximization

K-Means Clustering Algorithm

- K-Means is an unsupervised learning algorithm. It uses unlabeled numeric data. It automatically groups data elements into different groups.
- The parameter K refers to how many groups for the data.
- The data must be numeric because it calculates distance, using square root. The square root of labels, like Hot and Cold, doesn't make sense.

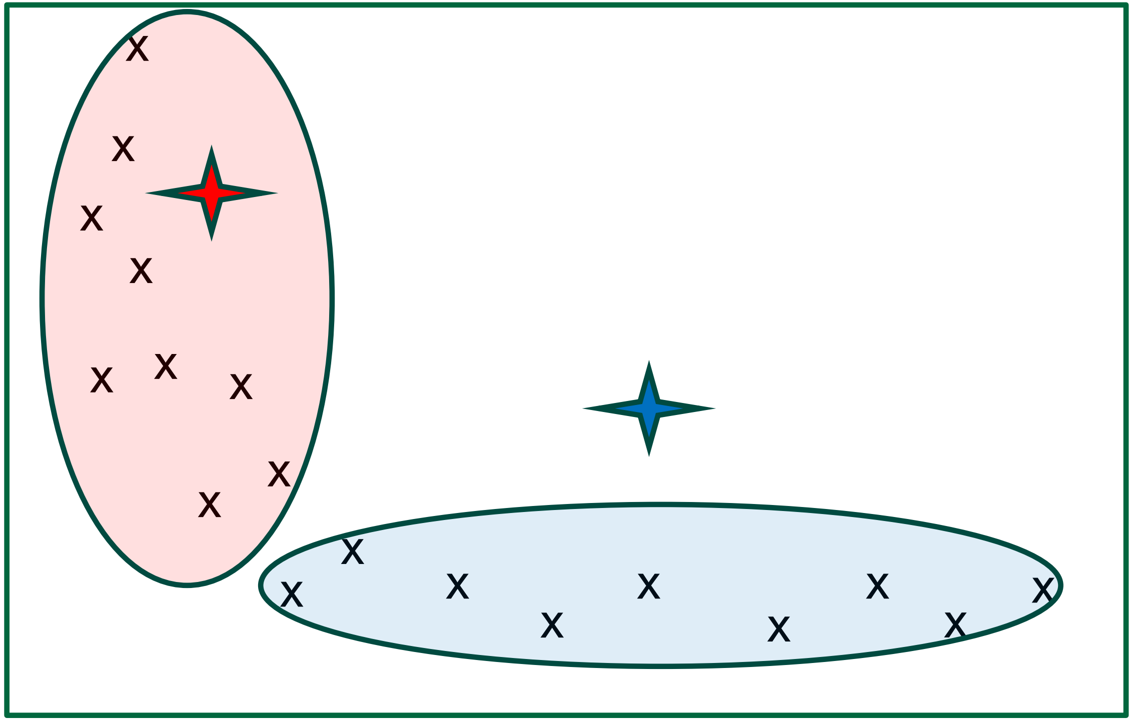
K-Means Clustering Algorithm

- Given a data set:
- Decide how many groups you want to calculate.
- You must know K in advance.
- Give each group a starting point (Centroid).

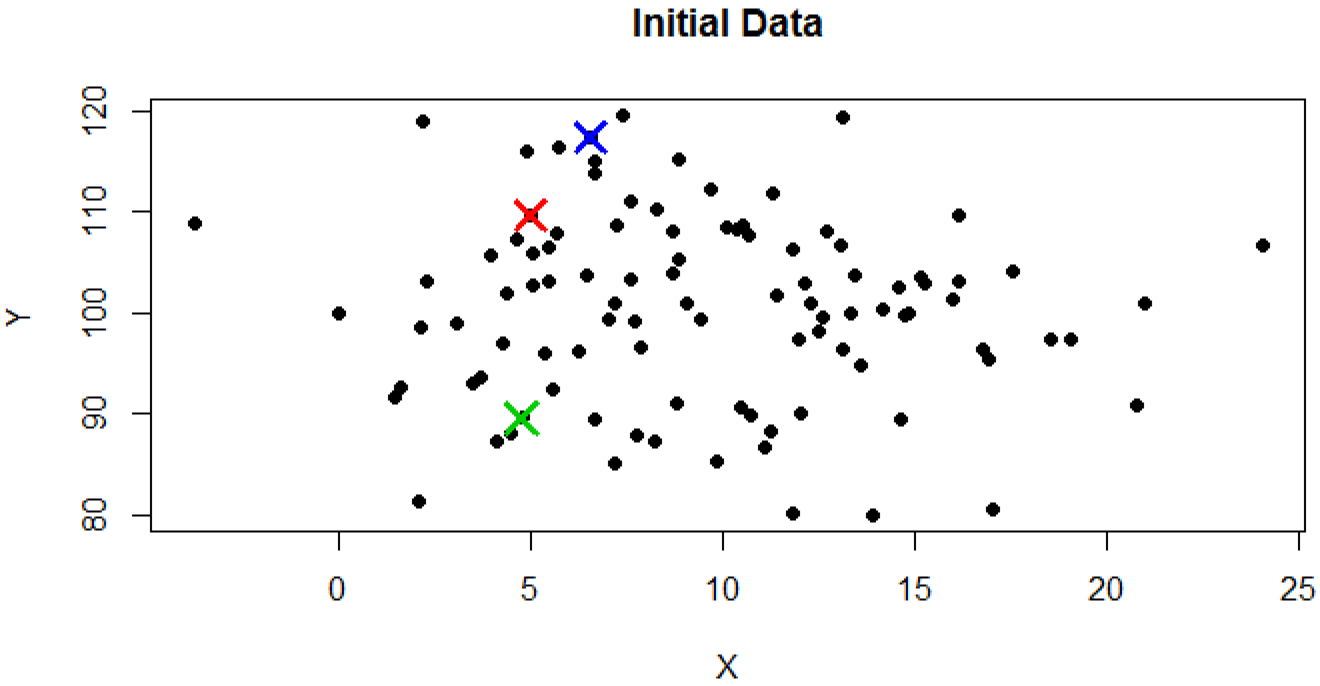


Repeat:

- For each data point, which is the nearest Centroid?
- Cluster items with the same Centroids.
- Re-compute Centroid location to middle of cluster until they stop moving.

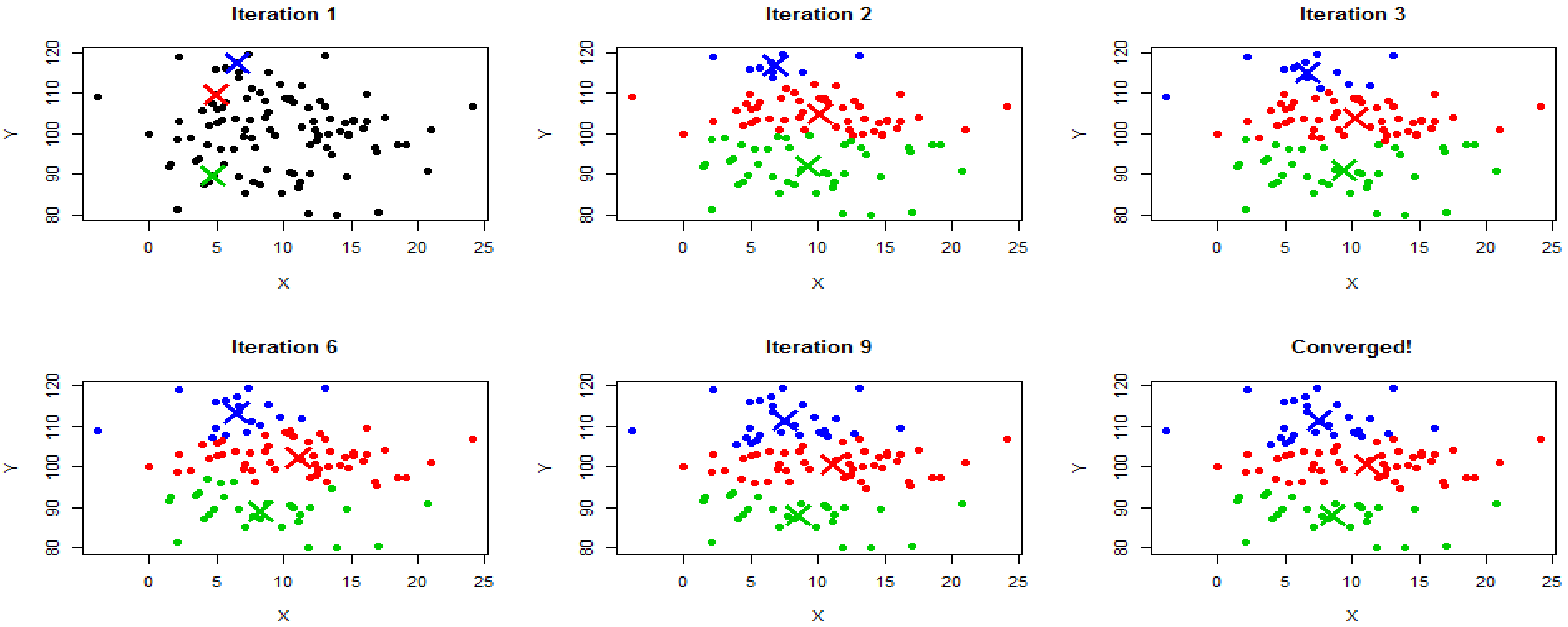


Example



Taken from: http://www.learnbymarketing.com/methods/k-means-clustering/

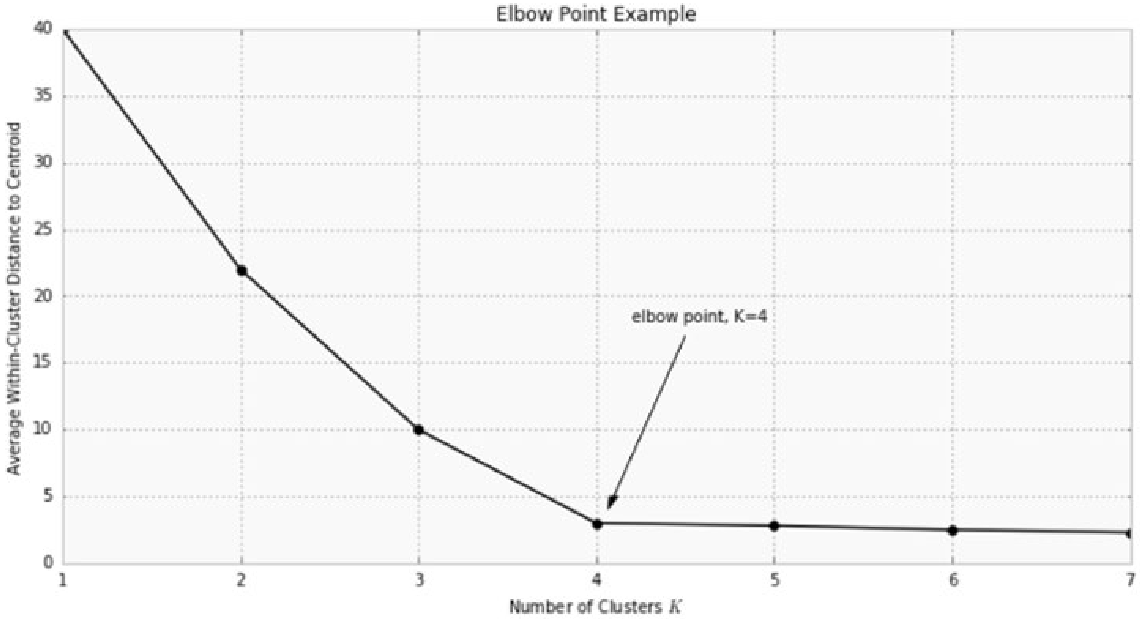
K-means clustering



How do you choose K?

- Run the algorithm with 2 centroids. Then calculate the average distance from each point to its nearest centroid.
- Repeat the steps with 3, 4, 5, … n centroids. If you plot the average within-cluster distance to the nearest centroid, you will see an "elbow point". That value should be the value of K, the number of groups in your data.
- The centroids are the average value for each cluster.

How do you choose K?



- https://www.datascience.com/blog/k-means-clustering

Demo – Calculations in Excel

Weka Demo

References

- https://www.datascience.com/blog/k-means-clustering
- Play with the examples:
- http://www.onmyphd.com/?p=k-means.clustering
- The 5 Clustering Algorithms Data Scientists Need to Know