ALGONQUIN COLLEGE

CST8502
MACHINE LEARNING

Week 6
Clustering

Professor: Dr. Anu Thomas

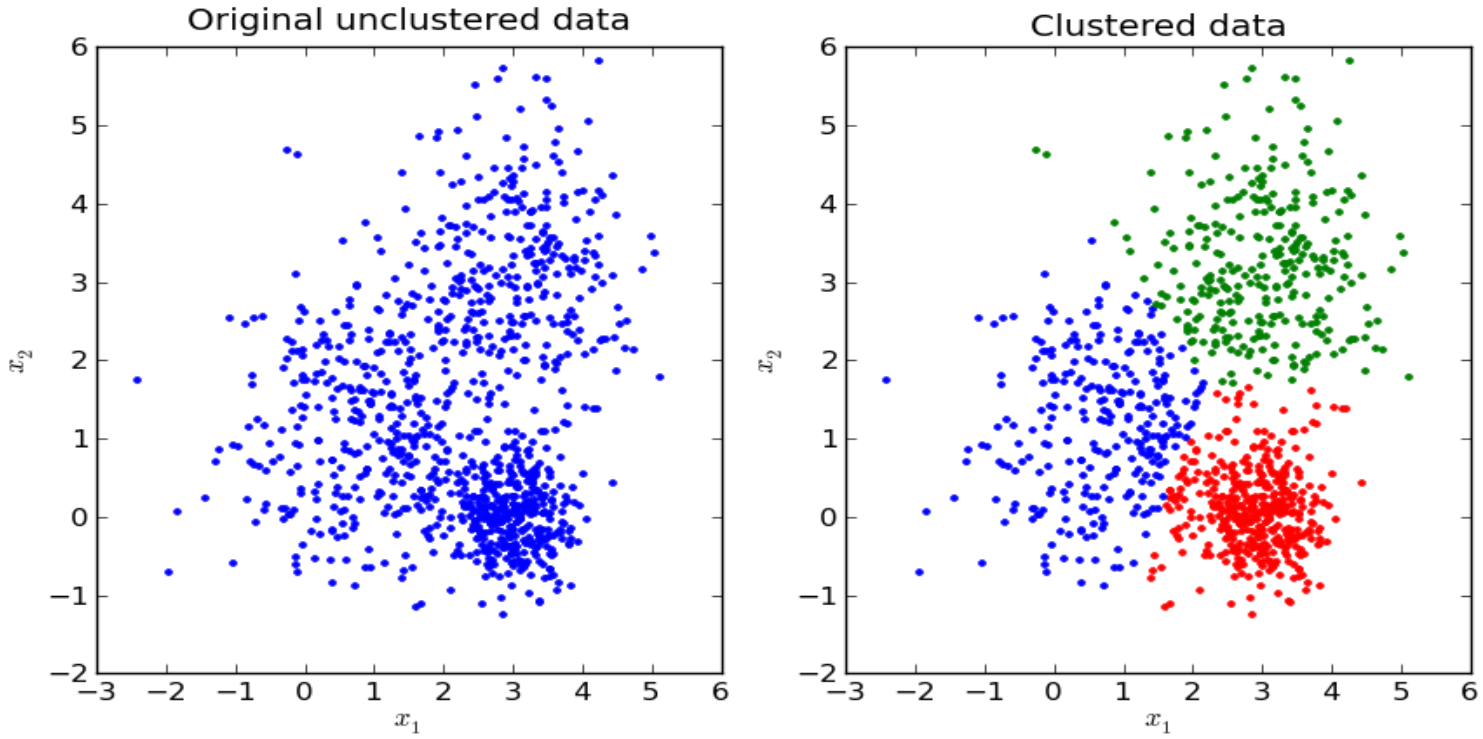
Learning - Recap

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

Clustering

- Cluster – a collection of data objects
  - Similar instances grouped in the same cluster
  - Dissimilar ones in other clusters
- Unsupervised technique

Clustering - Example



Clustering Algorithms

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

K-Means Clustering Algorithm

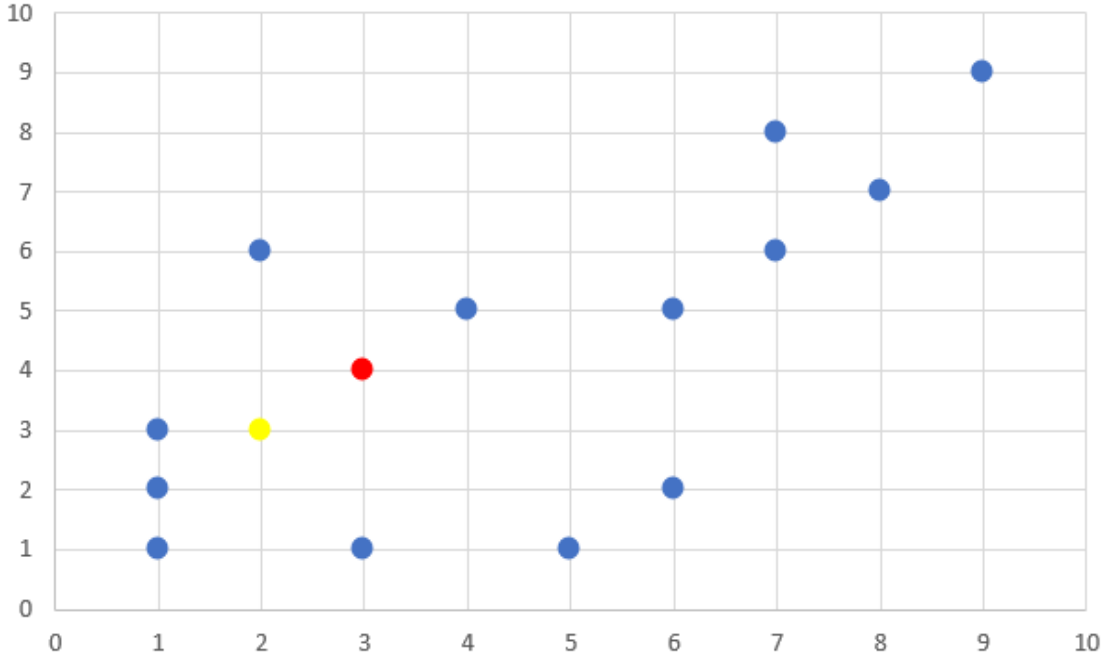
- Uses unlabeled numeric data. It automatically groups data elements into different groups.
- The parameter K refers to how many groups for the data.
- The data must be numeric because it calculates distances.

K-Means Clustering Algorithm

1. Decide k value (how many clusters you want to create)
2. Select k points **randomly** (initial centroids)
3. Find the distance between every point to the selected k centroids
4. Group each point based on the smallest distance (forming clusters)
5. Recalculate centroids (by taking the average values of x and y of the points in each group)
6. Repeat steps 3-5 until centroids converge.

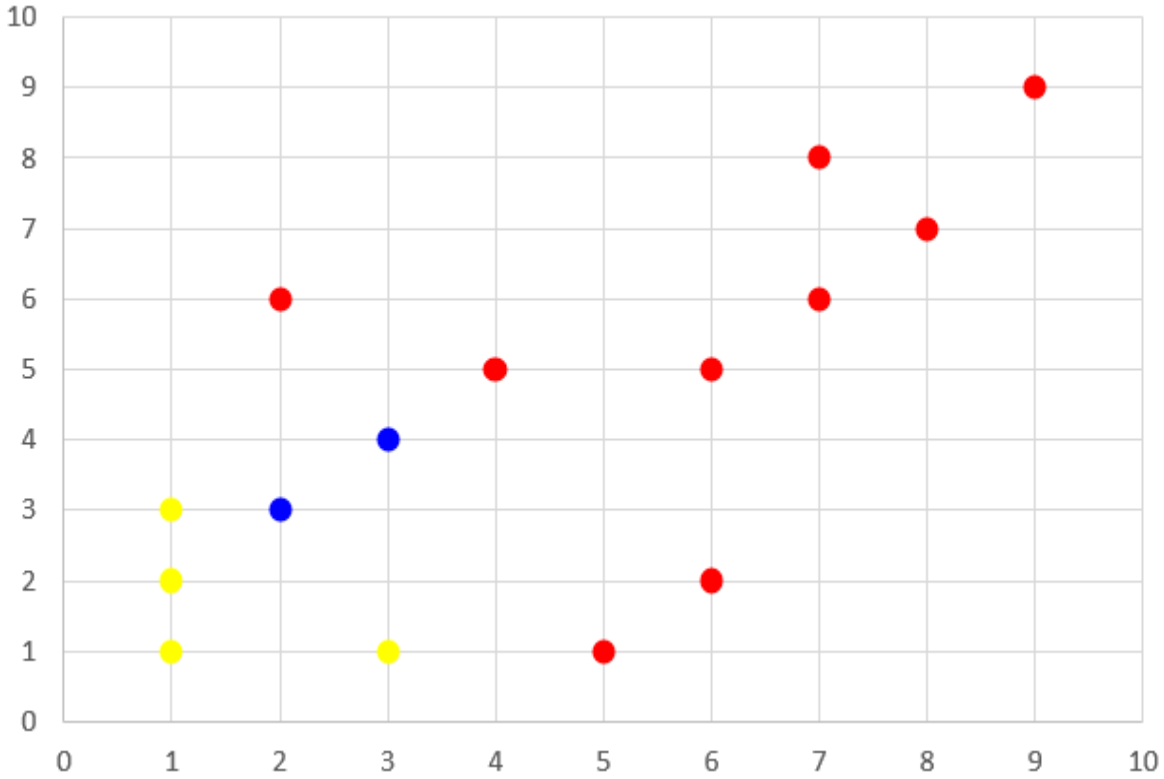
Example

Point	x	y
P1	2	3
P2	4	5
P3	1	2
P4	7	8
P5	6	5
P6	1	1
P7	3	4
P8	9	9
P9	8	7
P10	7	6
P11	1	3
P12	2	6
P13	5	1
P14	6	2
P15	3	1



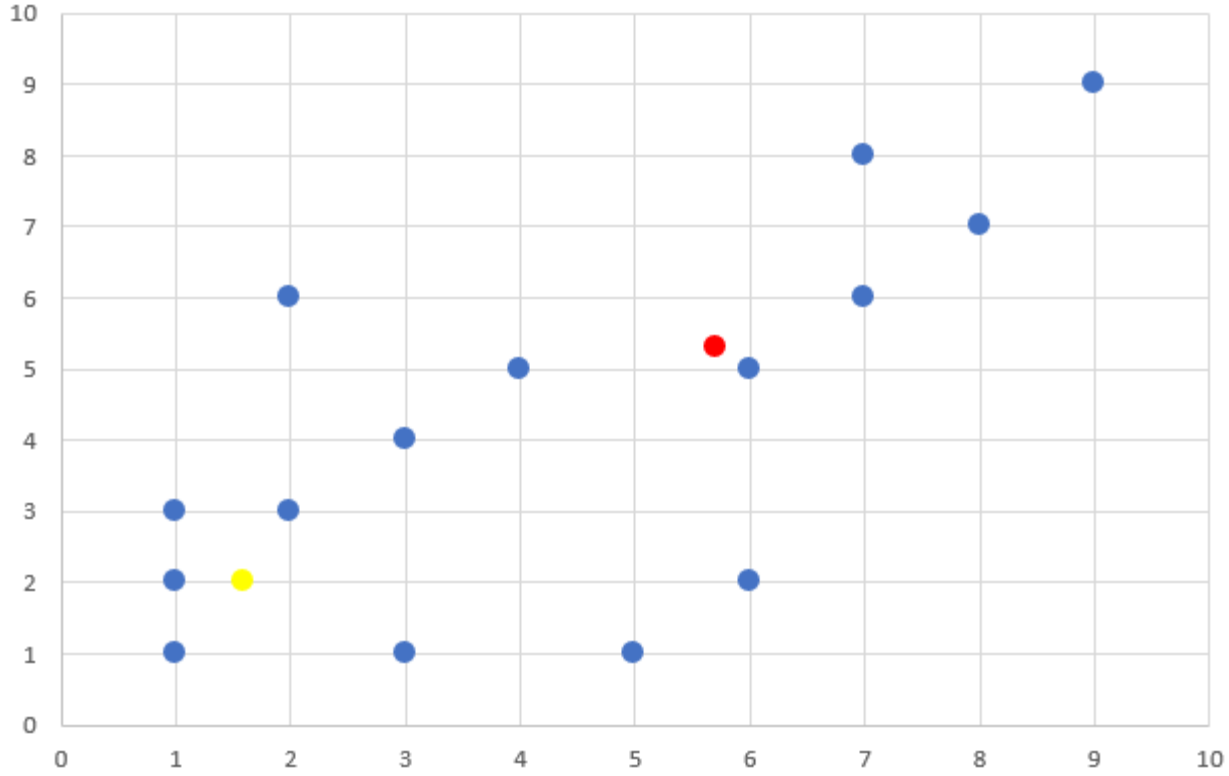
1. Decided k = 2
2. Selected 2 points **randomly**– P1, P7

Example – Iteration 1



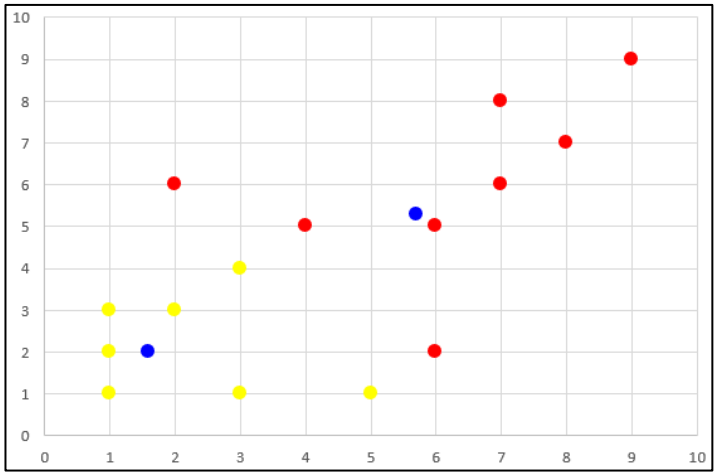
3. Calculated the distance between every point to the selected 2 centroids
4. Grouped each point with any of the centroids based on the smallest distance

Example – Recalculate centroids

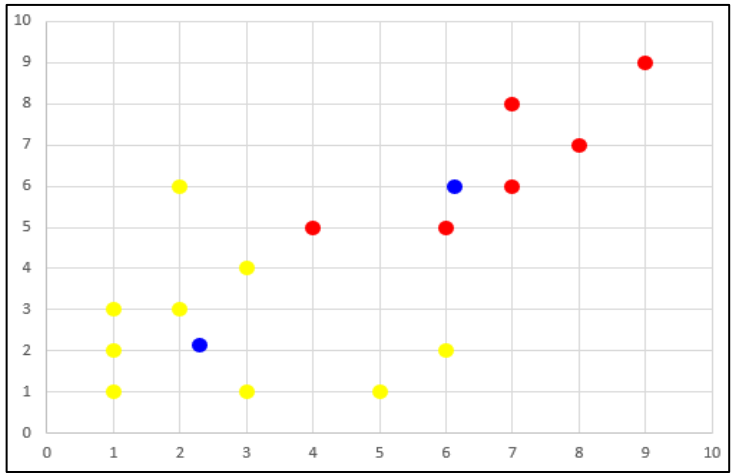


5. Recalculated centroids

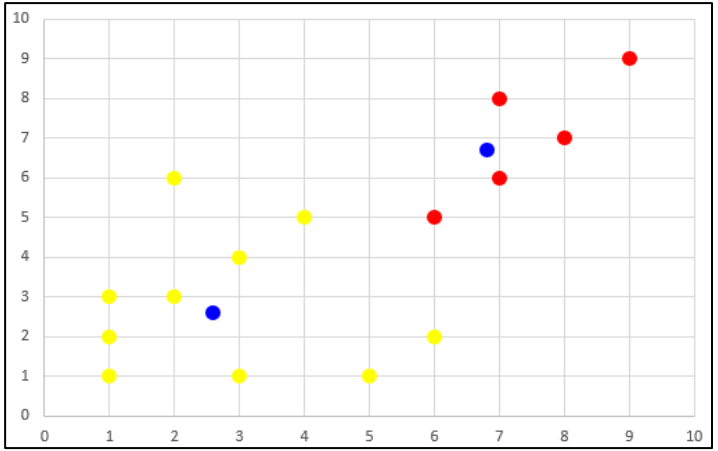
Iteration 2



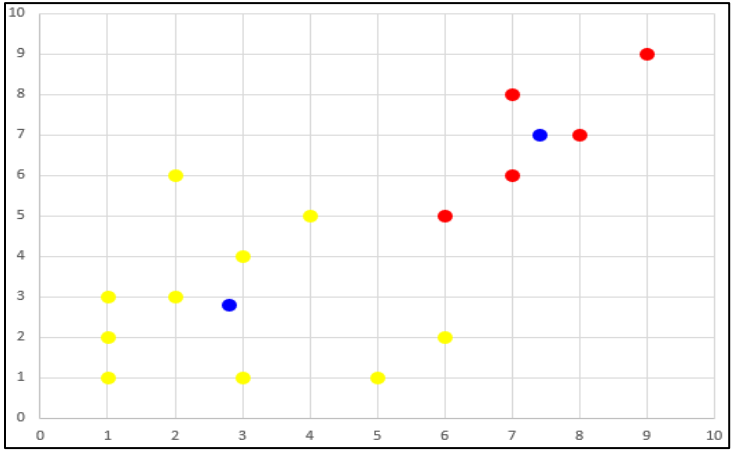
Iteration 4



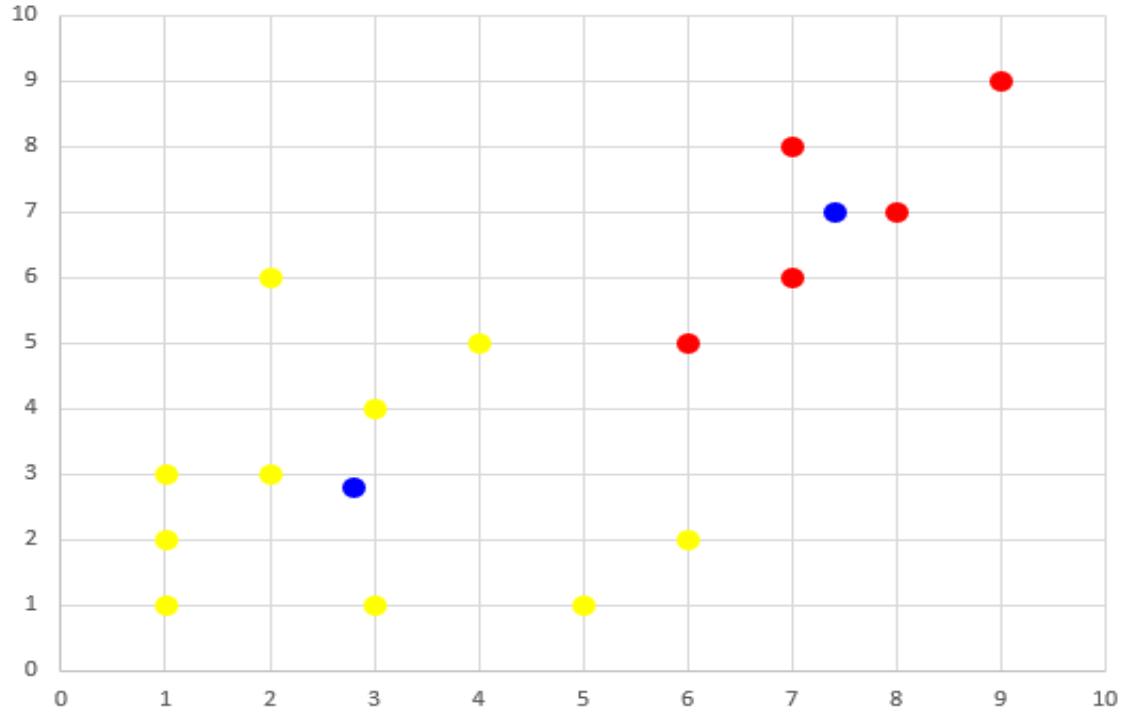
Iteration 3



Iteration 5

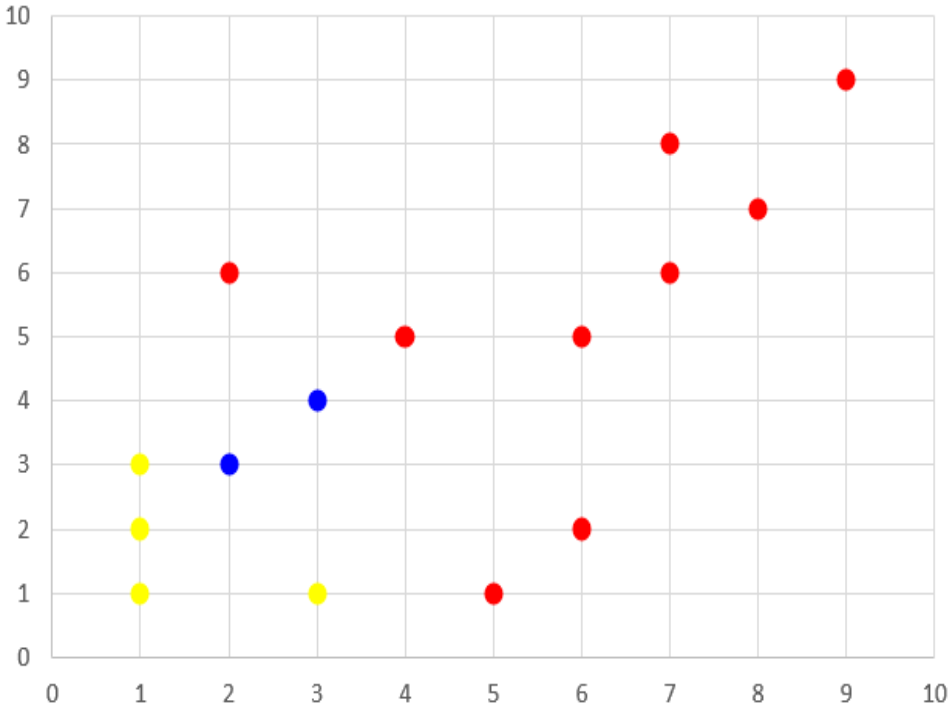


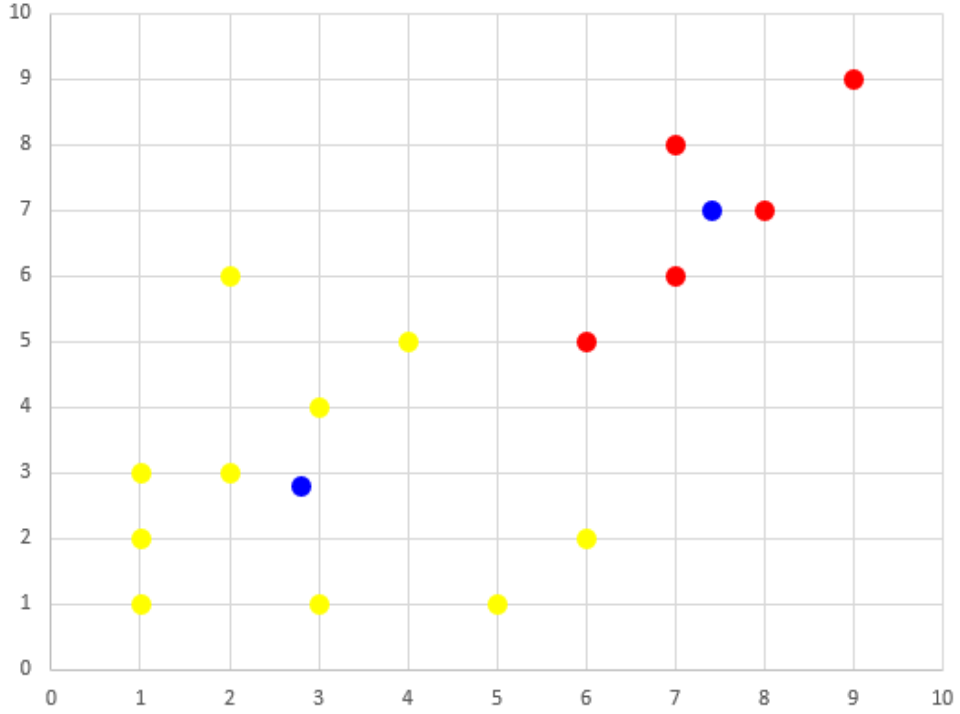
Example – Iteration 6



Iterations 5 & 6 have same centroids – centroids converged

Iteration 1 vs Iteration 6

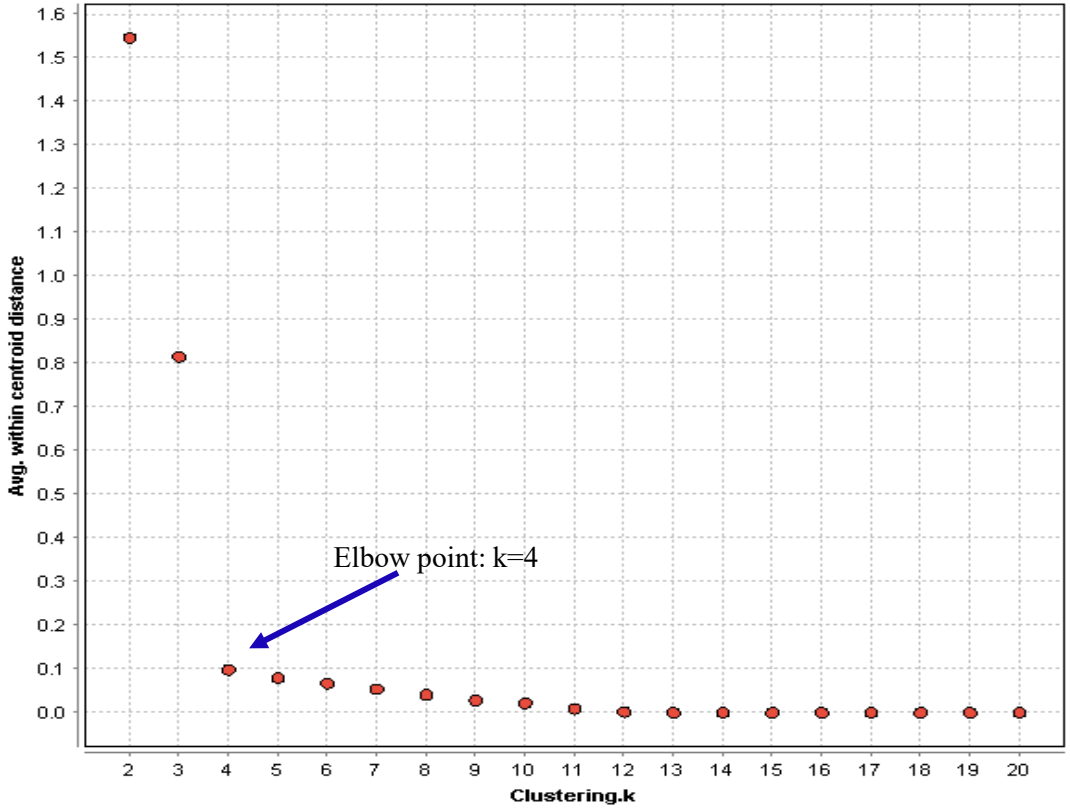




How do you choose the best K?

- 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 best value of K

How do you choose K?



RapidMiner Demo