Unsupervised Learning: K-means algorithm

Dr. Chandranath Adak

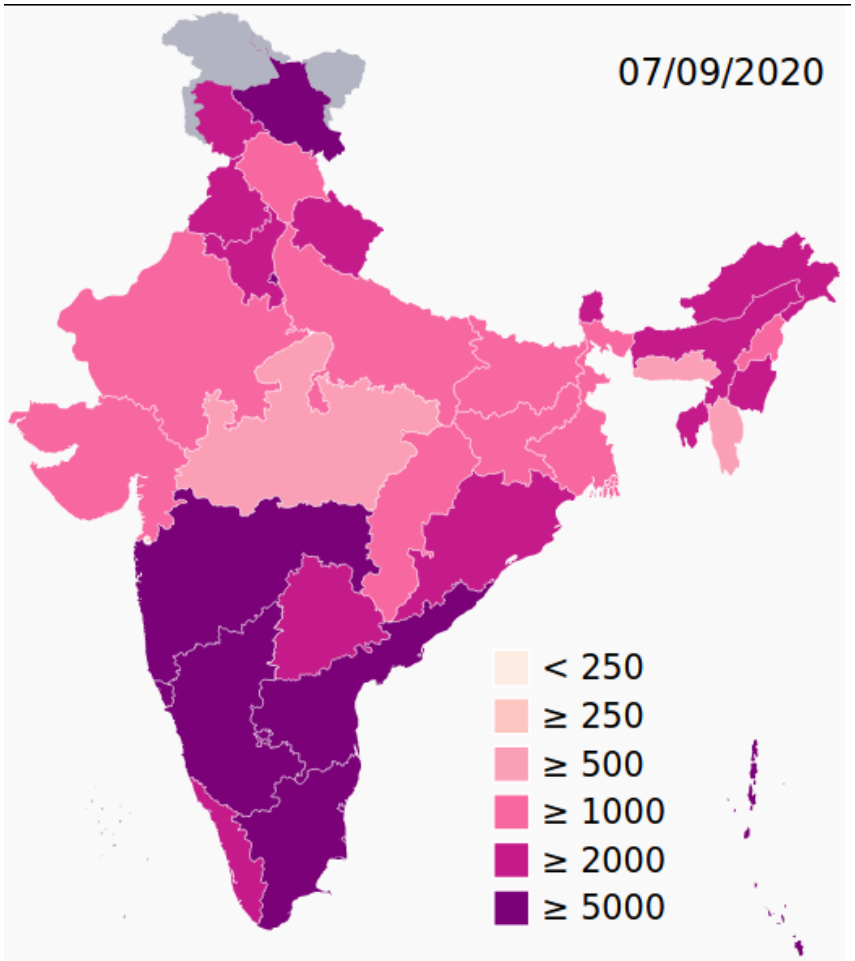
Dept. of CSE, Indian Institute of Technology Patna

October 9, 2025

Introduction to unsupervised learning

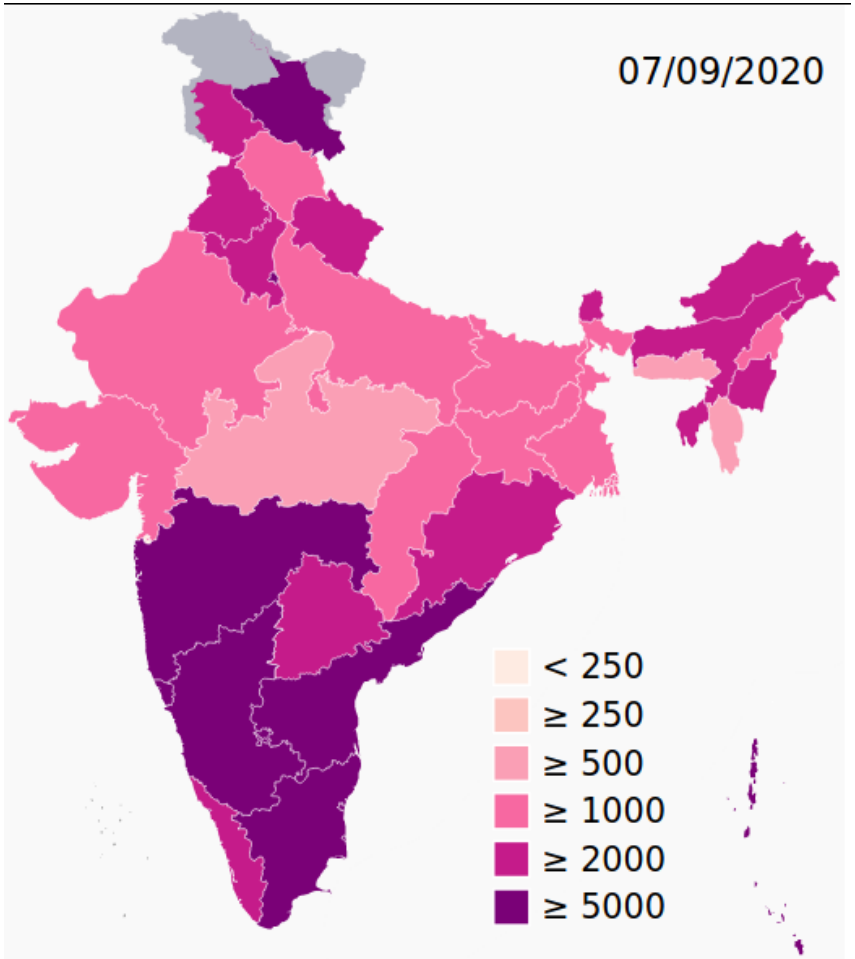
- Unsupervised Clustering
  - Requires a minimum human supervision
  - Does not require any knowledge of human-labeled data
- A two different types of unsupervised learning
  - Clustering
  - Dimensionality reduction
- Examples
  - *Market research*: For differentiating groups of customers based on certain attributes
  - *Recommender systems*: To recommend customers
    - Online shopping purchase suggestions
    - Netflix movie matches
    - News feed suggestion
  - *Medical imaging*: For distinguishing different kinds of tissues / diseases

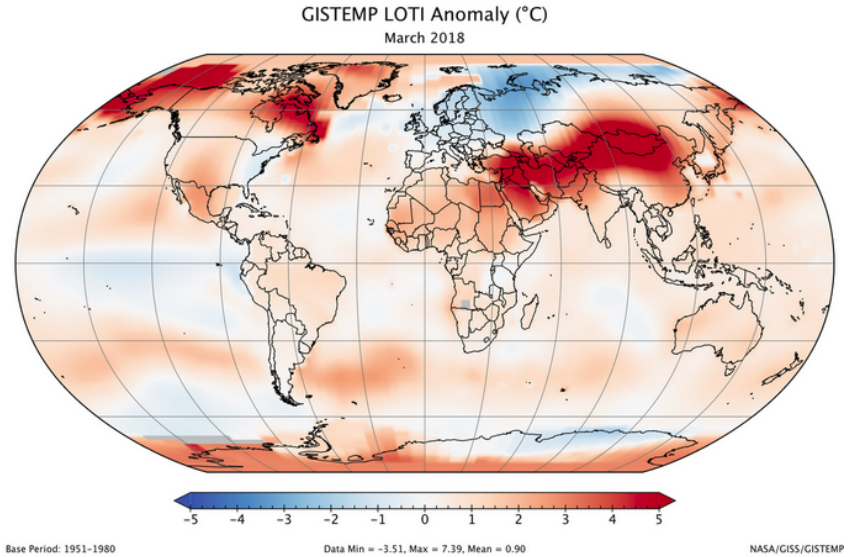
Introduction to clustering



https://en.wikipedia.org/wiki/COVID-19_pandemic_in_India
https://climate.nasa.gov/news/2714/
march-2018-was-one-of-six-warmest-marches-on-record/

Introduction to clustering





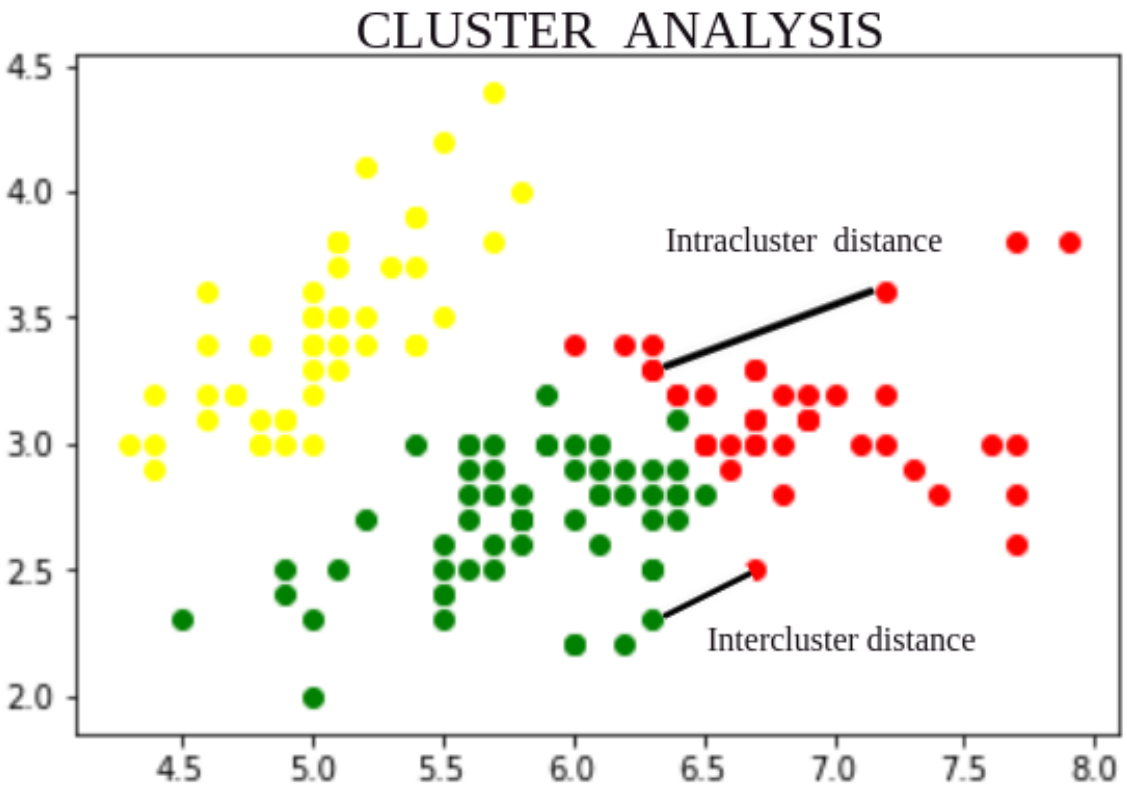
https://en.wikipedia.org/wiki/COVID-19_pandemic_in_India
https://climate.nasa.gov/news/2714/march-2018-was-one-of-six-warmest-marches-on-record/

Introduction to clustering

- Clustering is an unsupervised learning technique
- Objective of clustering
  - To discover overall distribution patterns
  - Correlations among the data attributes

- Given the features of the sample data to be clustered, objective is
  - To put similar samples in the same clusters
  - The similarity between two samples is measured by distance metric
  - The similarity of two samples is more
    - If the distance between them is less
  - A clustering algorithm is considered to be good, if
    - Intercluster distance between different clusters is more
    - Intracluster distance of same cluster is less

Intercluster distance vs Intracluster distance



[1]

[1] Source: https://www.geeksforgeeks.org/ml-intercluster-and-intracluster-distance/

Intercluster distance

- The distance between two sample data belonging to two different clusters

Different measures

- **Single Linkage Distance**: The minimum distance between two sample data belonging to two different clusters

$$\delta(C_1, C_2) = \min_{x \in C_1, y \in C_2} d(x, y)$$

- **Complete Linkage Distance**: The maximum distance between two sample data belonging to two different clusters

$$\delta(C_1, C_2) = \max_{x \in C_1, y \in C_2} d(x, y)$$

Intercluster distance

Different measures

- **Average Linkage Distance**: The average distance between all sample data belonging to two different clusters

$$\delta(C_1, C_2) = \frac{1}{|C_1| \times |C_2|} \sum_{x \in C_1, y \in C_2} d(x, y)$$

- **Centroid Linkage Distance**: The distance between two centroids belonging to two different clusters

$$\delta(C_1, C_2) = d(\mu_1, \mu_2)$$

where $\mu_1 = \frac{1}{|C_1|} \sum_{x \in C_1} x$

$$\mu_2 = \frac{1}{|C_2|} \sum_{y \in C_2} y$$

Intercluster distance

Different measures

- **Average Centroid Linkage Distance**: The average distance between the centroid of a cluster and all the objects belonging to a different cluster

$$\delta(C_1, C_2) = \frac{1}{|C_1| + |C_2|} \sum_{x \in C_1} d(x, \mu_2) + \sum_{y \in C_2} d(\mu_1, y)$$

where $\mu_1 = \frac{1}{|C_1|} \sum_{x \in C_1} x$

$$\mu_2 = \frac{1}{|C_2|} \sum_{y \in C_2} y$$

Intracluster distance

- The distance between two sample data belonging to the same cluster

Different measures

- **Complete Diameter Distance**: The furthest distance between two sample data belonging to the same cluster

$$\Delta(C_1) = \max_{(x,y)\in C_1, x\neq y} d(x,y)$$

- **Average Diameter Distance**: The average distance between each pair of sample data belonging to the same cluster

$$\Delta(C_1) = \text{avg}_{(x,y)\in C_1, x\neq y} d(x,y)$$

- **Centroid Diameter Distance**: The average distance between each sample data and the centroid of the cluster

$$\Delta(C_1) = 2 \times \text{avg}_{x\in C_1} d(x,\mu_1)$$

Question

A good clustering algorithm should have:

- A) More complete diameter distance.
- B) Less complete linkage distance.
- C) Less average diameter distance.
- D) More centroid diameter distance.

K-means clustering

- Very popular clustering algorithm
- Requirements
  - K : The number of clusters
  - Sample data points represented by a set of features
- Objective
  - To minimize the distance between two sample data belonging to a cluster across all clusters
- Output
  - Return k clusters
  - Clusters are mutually exclusive and collectively exhaustive
- Iterative algorithm
- In each iteration
  - First update center of the clusters (also called centroid)
  - Assign cluster index for each sample data

Mathematical formulation of K-means clustering

- Inputs
  - Training data: $X = \{x_1, x_2, \ldots, x_n\}$; where $\forall x_i \in \mathbb{R}^m$
    (n training samples, each sample is an m dimensional feature vector)
  - k number of clusters $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$
- Output
  - $\forall x_i \in X$, (x_i, c_i)
    $c_i \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ is the cluster number of x_i
- Objective
  - $\mu_i \in \mathbb{R}^m$ is the centroid of $\mathcal{C}_i \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$
  - Optimization objective:
    Minimize $J(\underbrace{c_1, c_2, \ldots, c_n, \mu_1, \mu_2, \ldots, \mu_k}_{n+k \text{ parameters}})$

$$J(c_1, c_2, \ldots, c_n, \mu_1, \mu_2, \ldots, \mu_k) = \frac{1}{n} \sum_{i=1}^{n} ||x_i - \mu_{c_i}||^2$$

$\mu_{c_i} \in \{\mu_1, \mu_2, \ldots, \mu_k\}$ is the centroid of the cluster to which x_i belongs

Question

- Suppose x_1 belongs to Cluster 3, which of the following is true?
  - A) $c_1 = 3$
  - B) $c_3 = 1$
  - C) $\mathcal{C}_1 = 3$
  - D) $\mathcal{C}_3 = 1$

K-means clustering algorithm

- Inputs:
  - Training data: $X = \{x_1, x_2, \ldots, x_n\}$; where $\forall x_i \in \mathbb{R}^m$
  - k number of clusters $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$
- Randomly initialize k cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^m$

Repeat (until terminates){

/*Cluster assignment step*/

For all $x_i \in X$:

$c_i \leftarrow \underset{j \in \{1,2,\ldots,k\}}{\arg\min} \|x_i - \mu_j\|^2$;

/*Update centroids*/

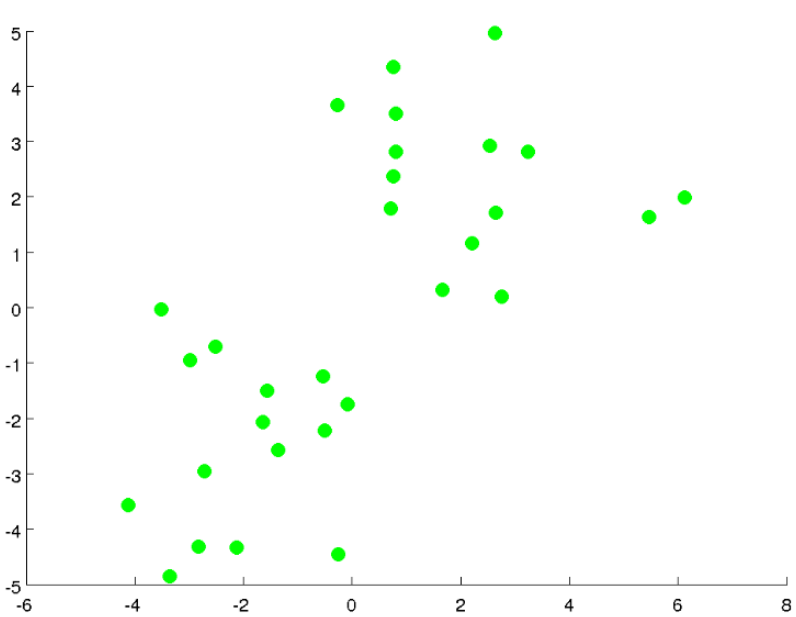
For all $\mu_i \in \{\mu_1, \mu_2, \ldots, \mu_k\}$:

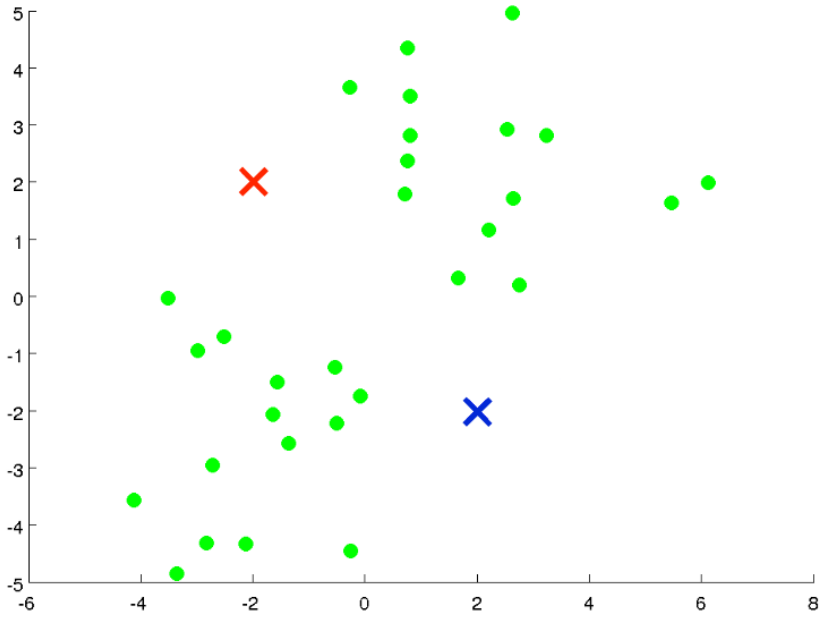
$\mu_i \leftarrow$ Update the value from newly generated cluster;

}

- Termination criteria: When there is no updation of the centroids

Example of K-means algorithm

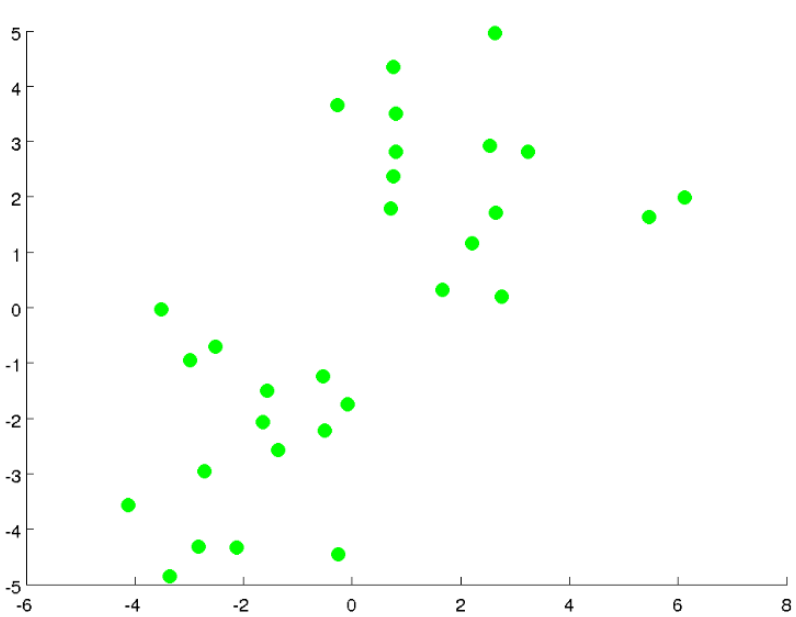


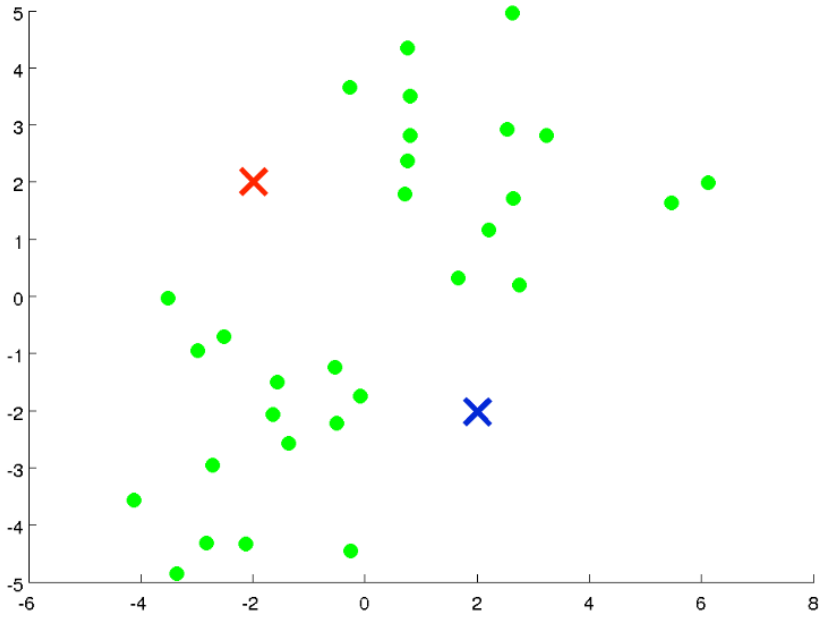


- Random initialization of cluster centroids

Taken from the slides of Prof. Andrew Ng

Example of K-means algorithm

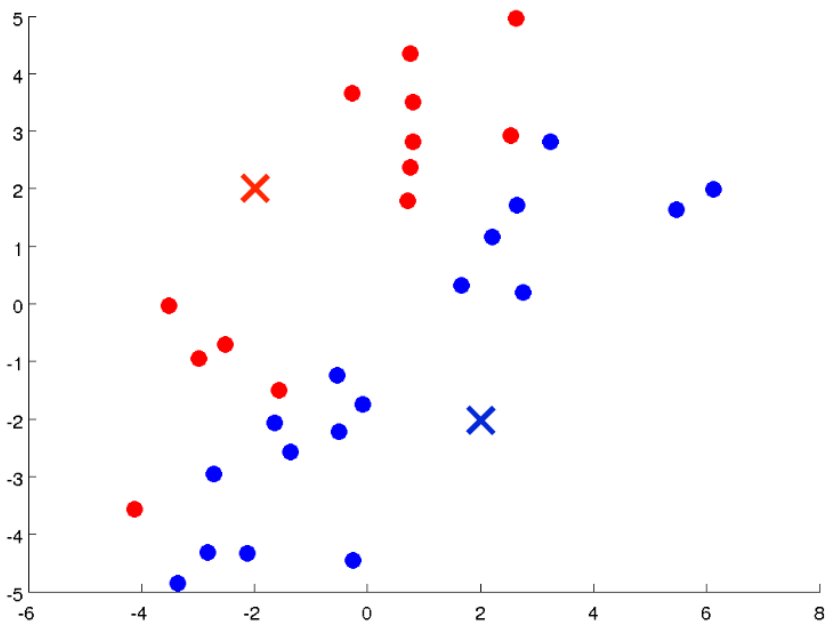


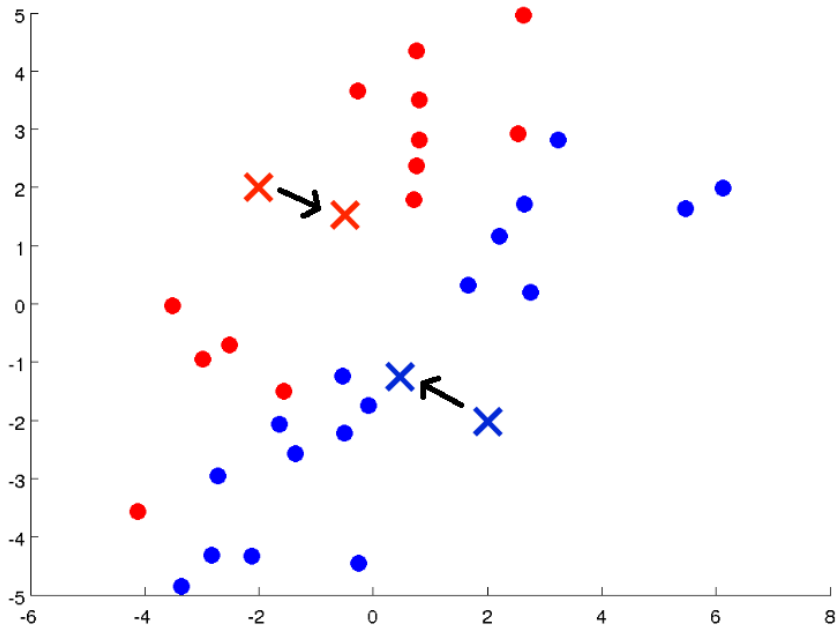


- Random initialization of cluster centroids

Taken from the slides of Prof. Andrew Ng

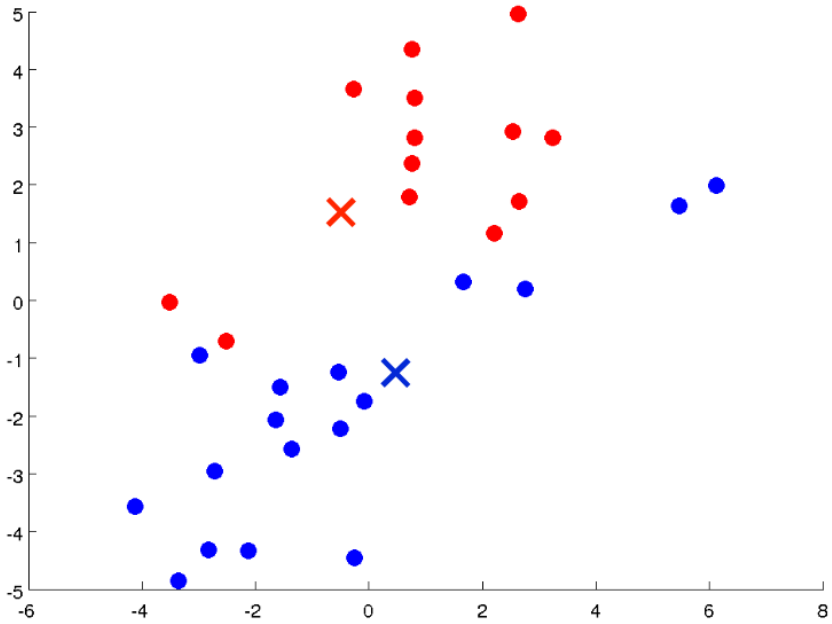
Example of K-means algorithm

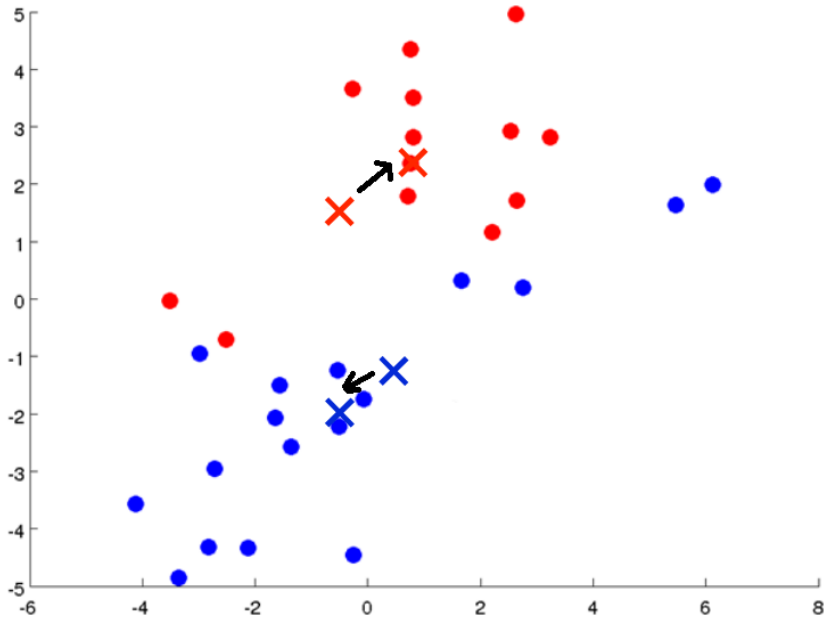




Taken from the slides of Prof. Andrew Ng

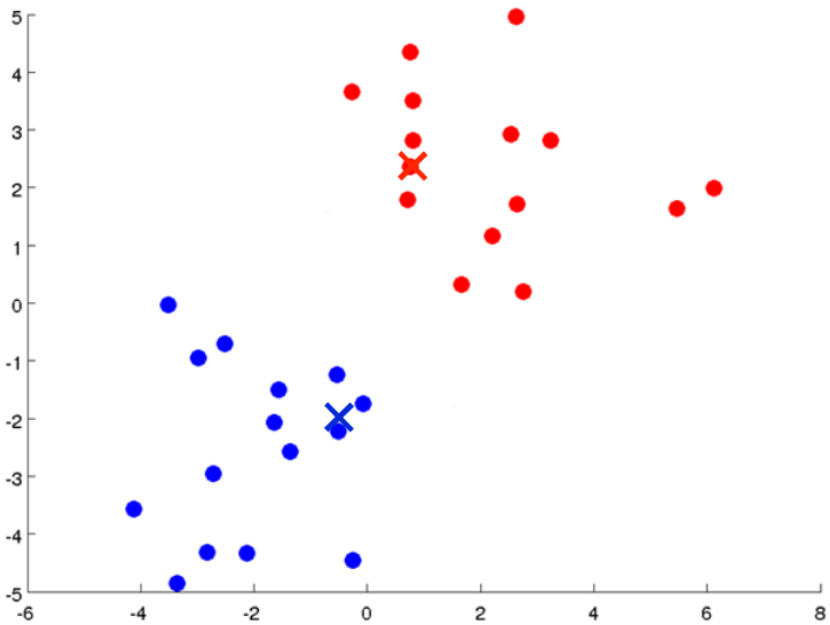
Example of K-means algorithm

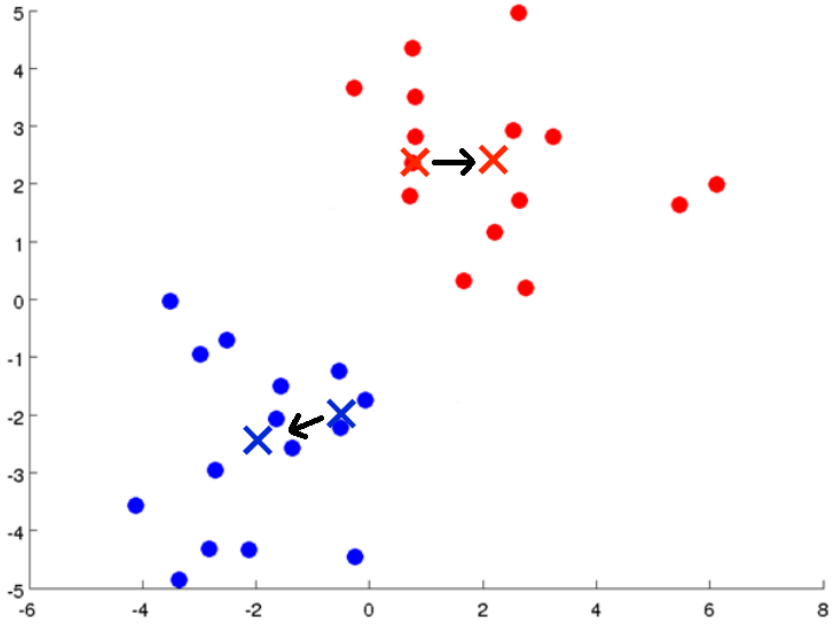




Taken from the slides of Prof. Andrew Ng

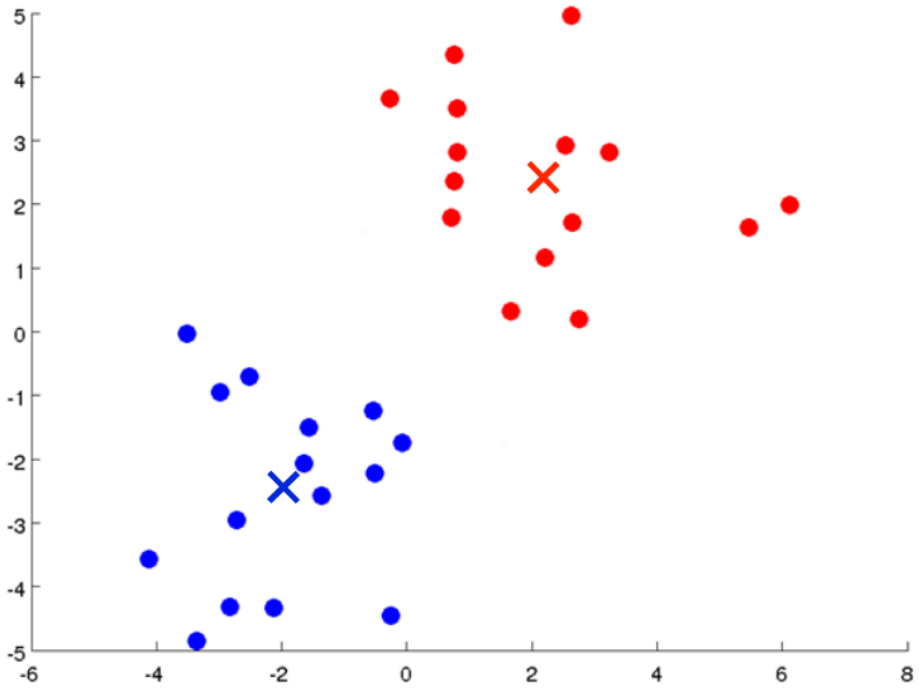
Example of K-means algorithm





Taken from the slides of Prof. Andrew Ng

Example of K-means algorithm



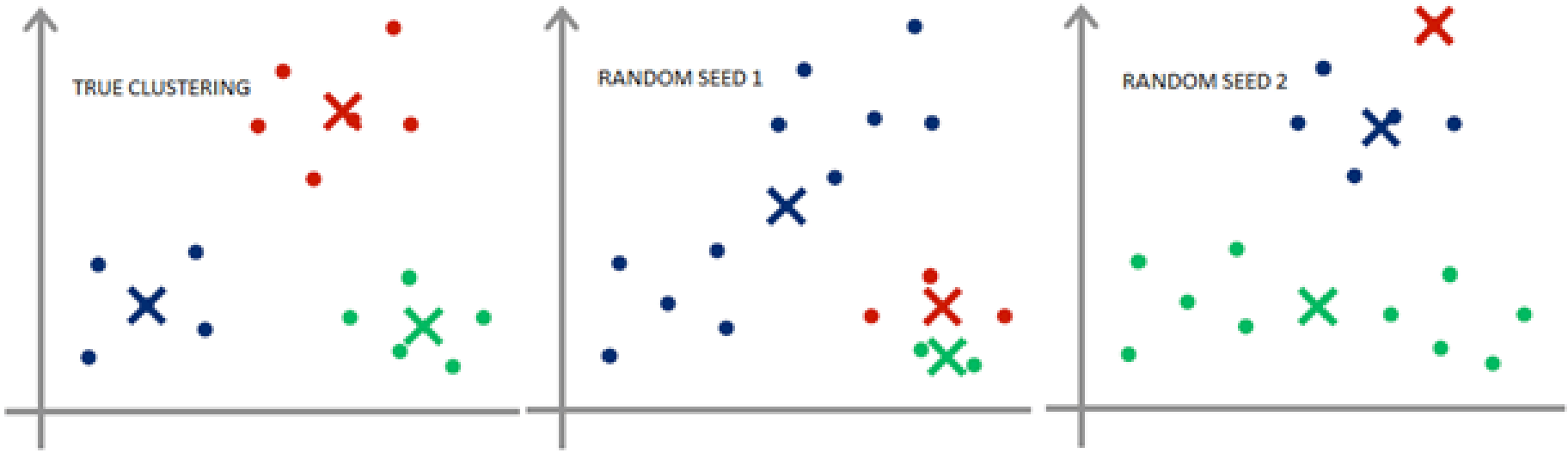
Taken from the slides of Prof. Andrew Ng

Advantages

- Easy to implement
- Easily adoptable for new problem
- Generalizes to clusters of different shapes and sizes
- Scales to large data sets

Random initialization of centroids

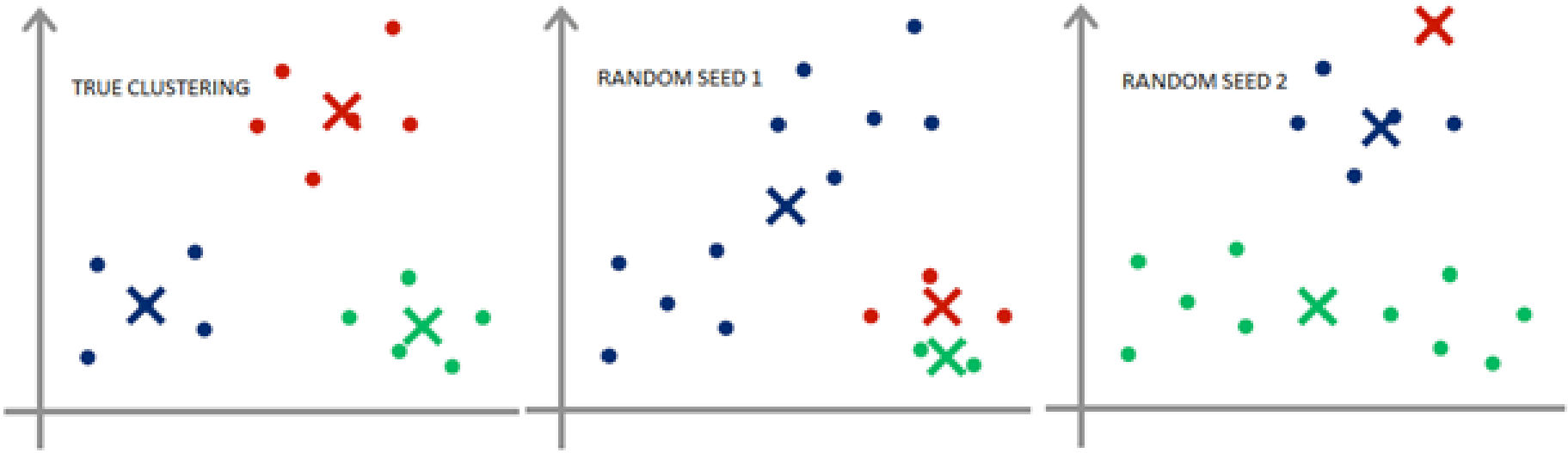
- Random initialization of the centroids has impact on the solution quality



https://www.edupristine.com/blog/k-means-algorithm

Random initialization of centroids

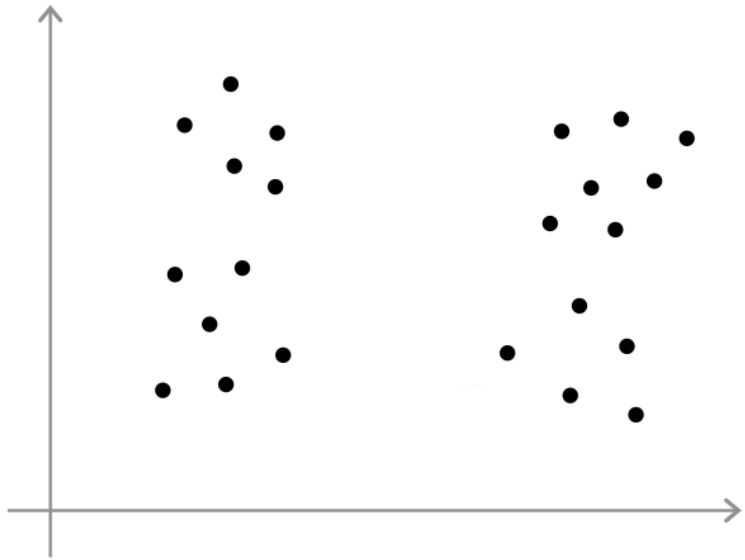
- Random initialization of the centroids has impact on the solution quality



- Suffers from local optima problem
- Run K-means algorithm multiple times with different random initialization of centroids and choose the best possible solution

https://www.edupristine.com/blog/k-means-algorithm

Question



- How many clusters are there?
  - A) 2
  - B) 4
  - C) 6
  - D) 11

Limitations

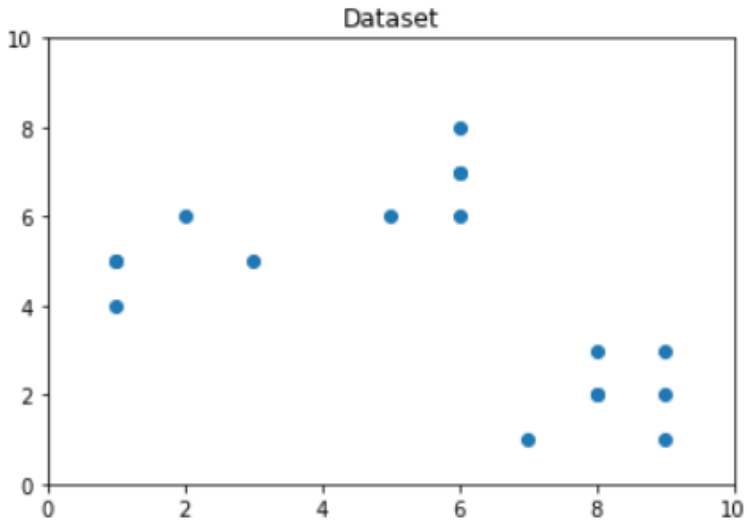
- Find value of k manually
- Highly dependent on initial values of the centroids
- Clustering outliers
- Scaling with number of dimensions

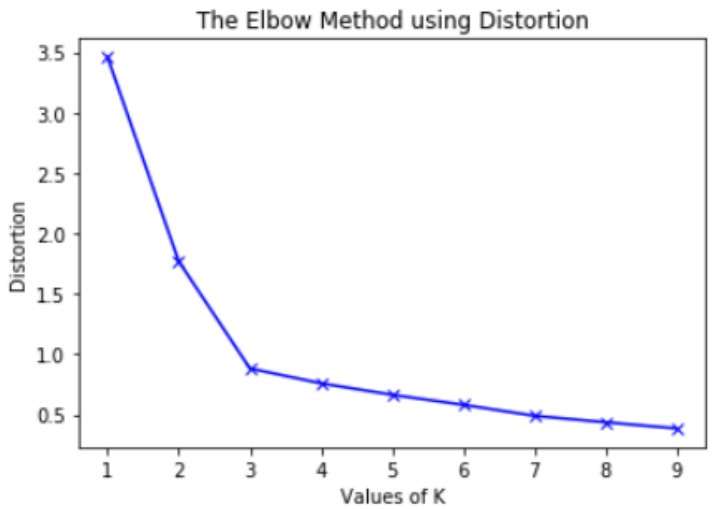
Elbow Method: To Decide the Value of K

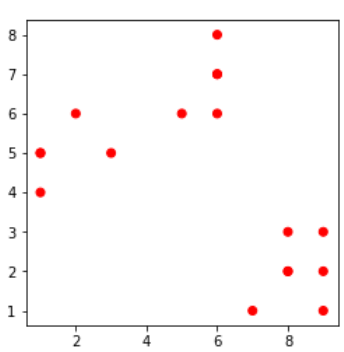
- In clustering, the number of clusters should be chosen such that adding another cluster doesn't give much better modeling of the data

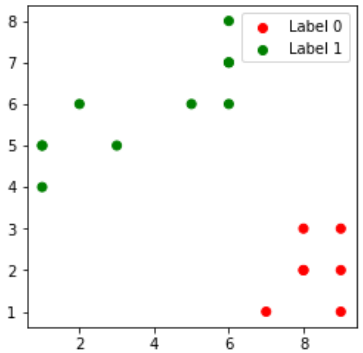
- Elbow Method
  - A **heuristic** used to determine the number of clusters in a data set
  - Consists of plotting the explained variation as a function of the number of clusters
  - The elbow of the curve is used as the number of clusters

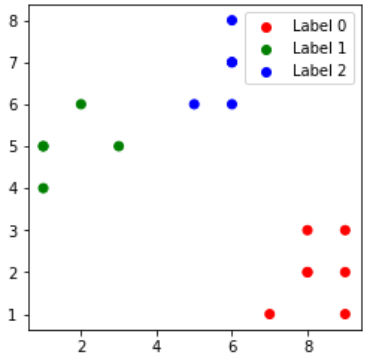
Elbow Method: To Decide the Value of K

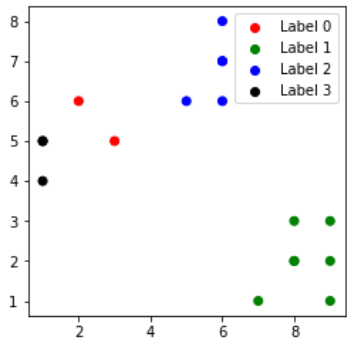








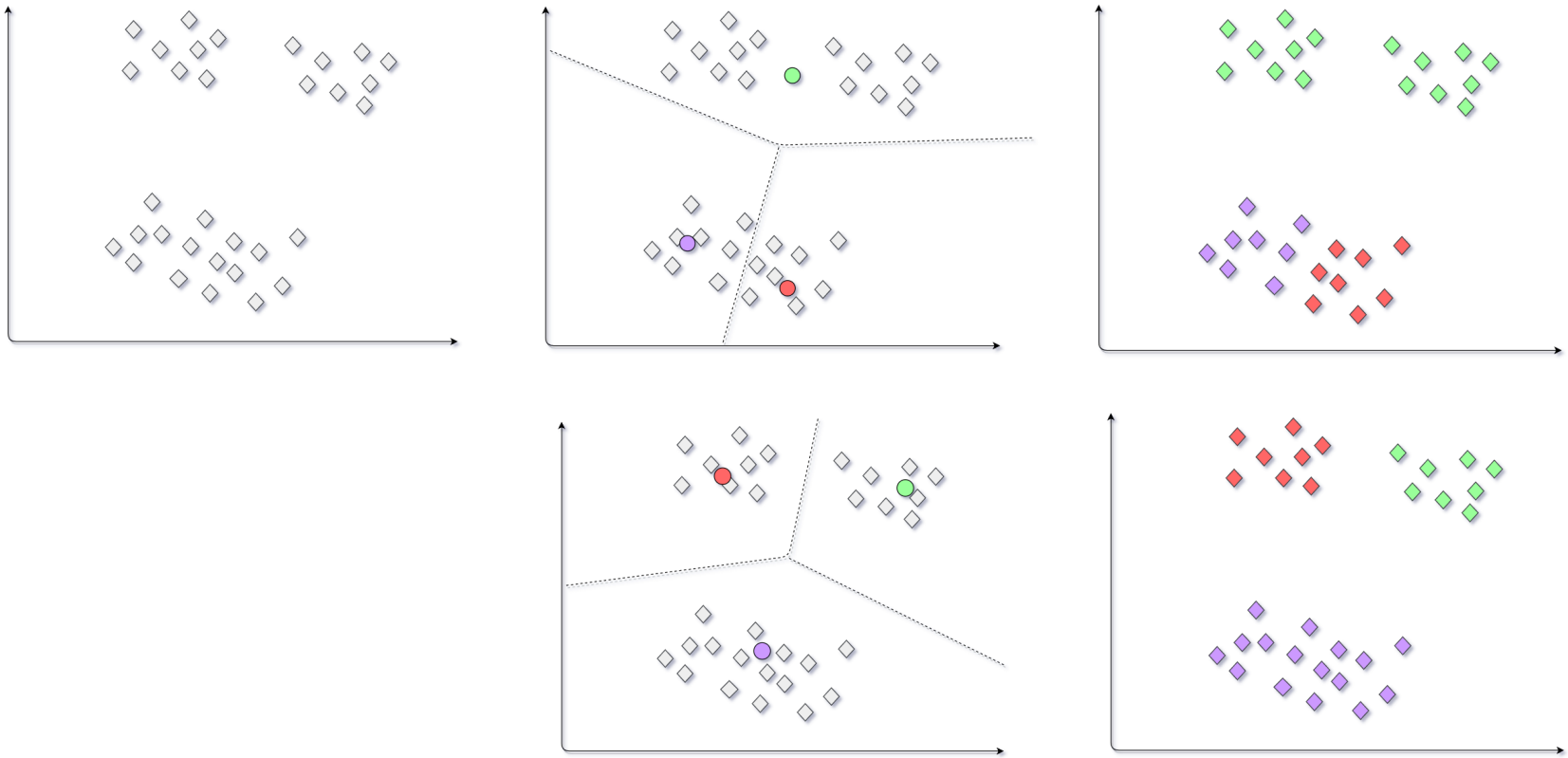




2

[2] Slide curtesy: `https://www.geeksforgeeks.org/`

Random initialization of centroids



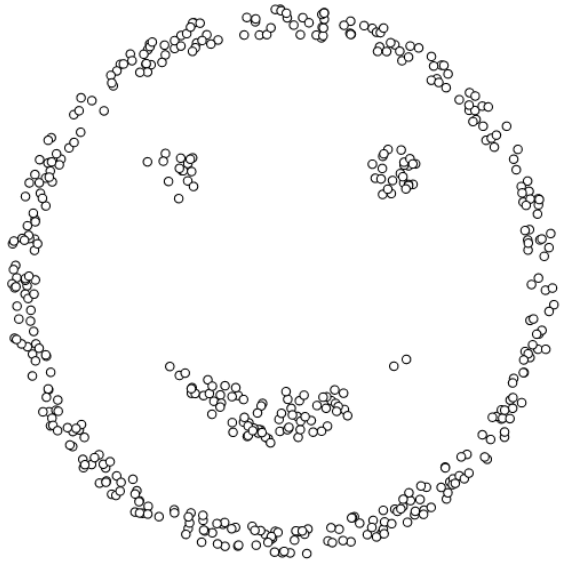
[3]

[3]Slide curtesy: https://www.geeksforgeeks.org/

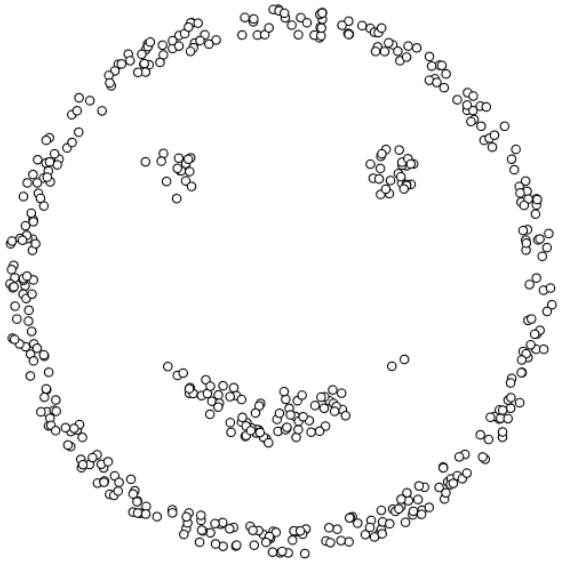
Random initialization of centroids

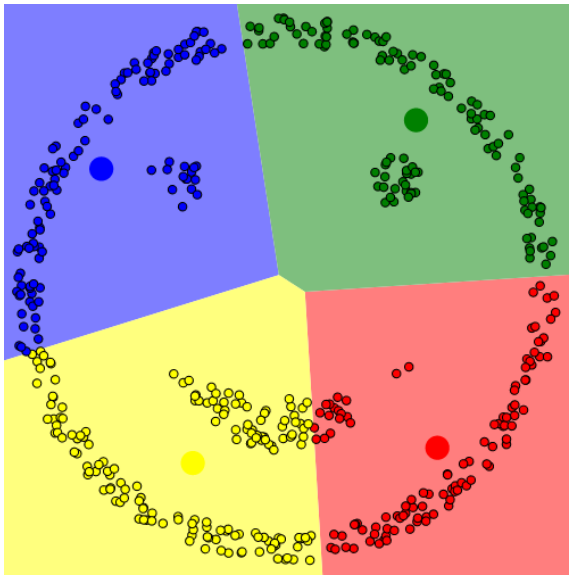
- Chooses any k distinct points from the data at random
- Furthest point initialization
- Kmeans++

Few cases where K-means fails

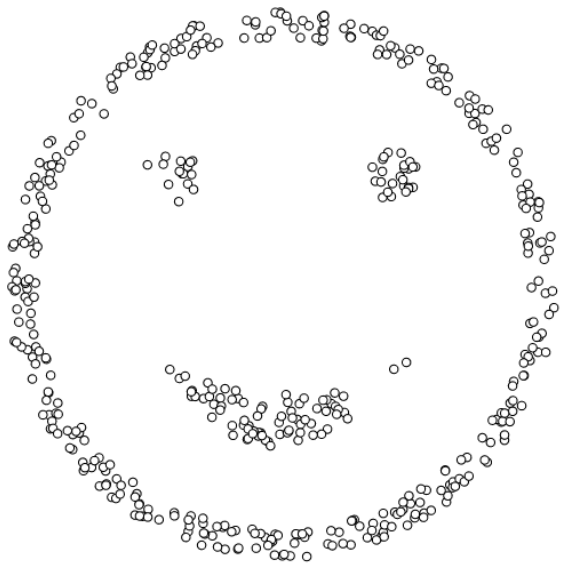


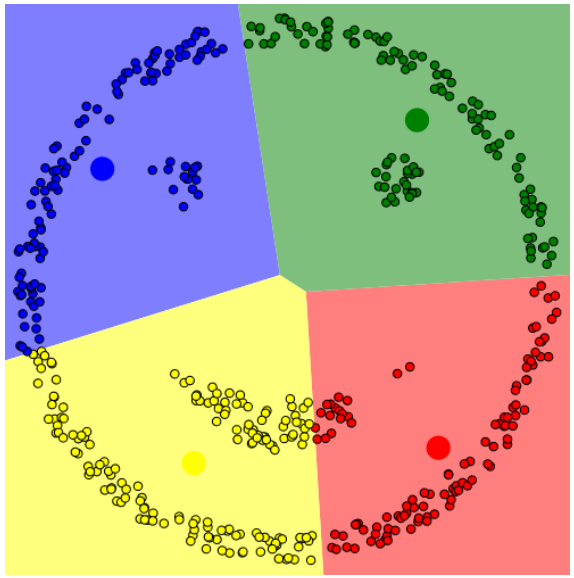
Few cases where K-means fails

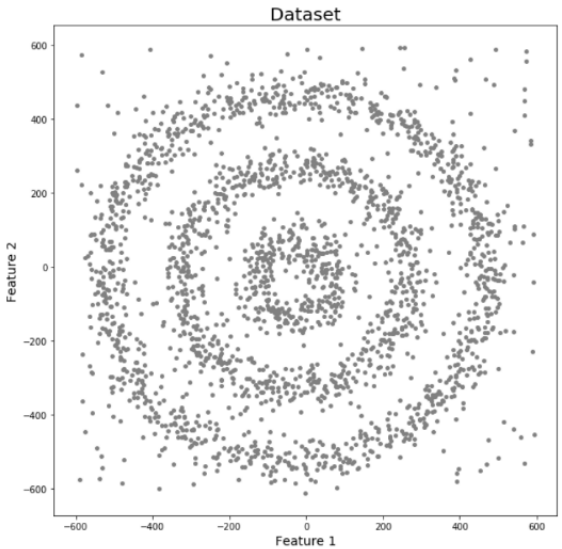




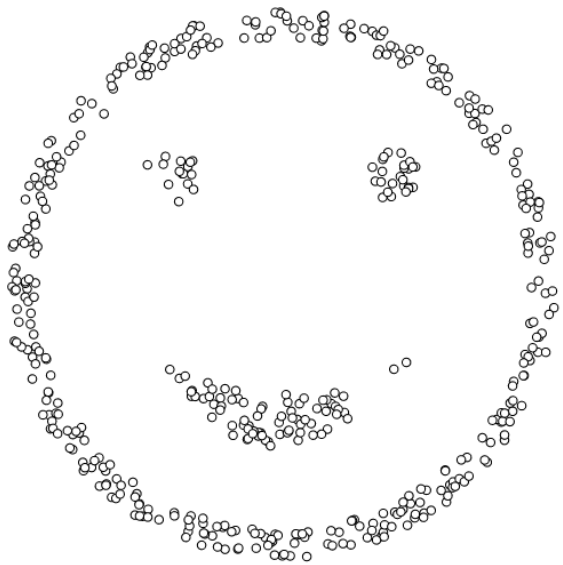
Few cases where K-means fails

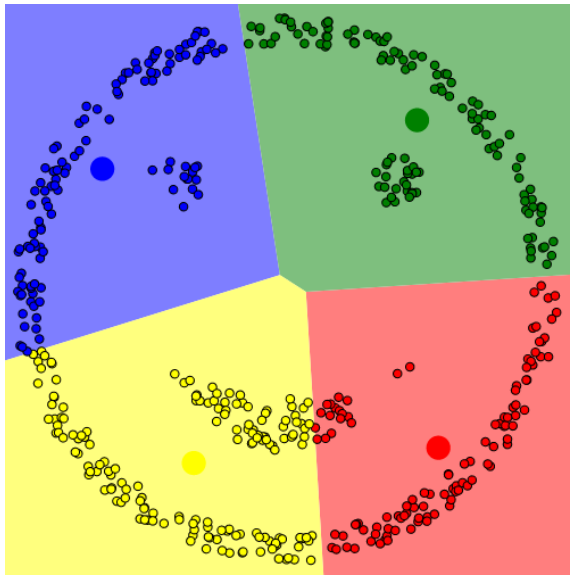


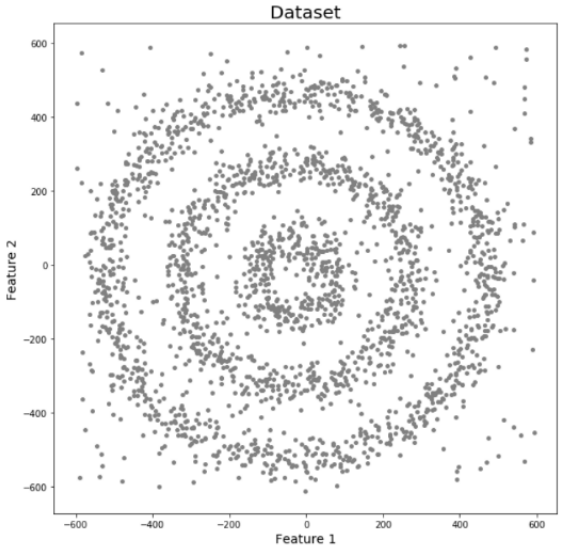


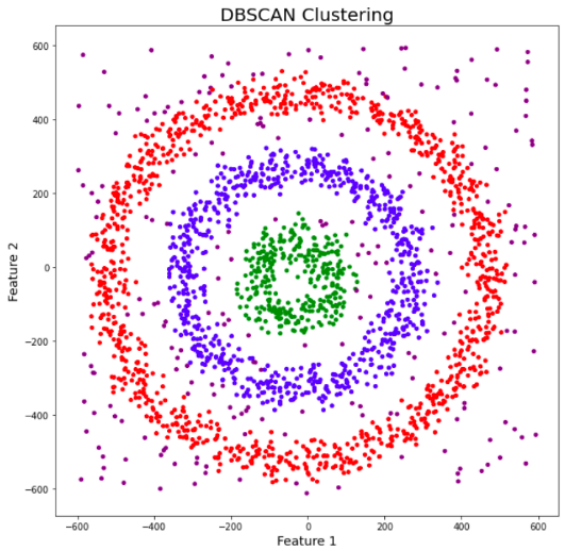


Few cases where K-means fails

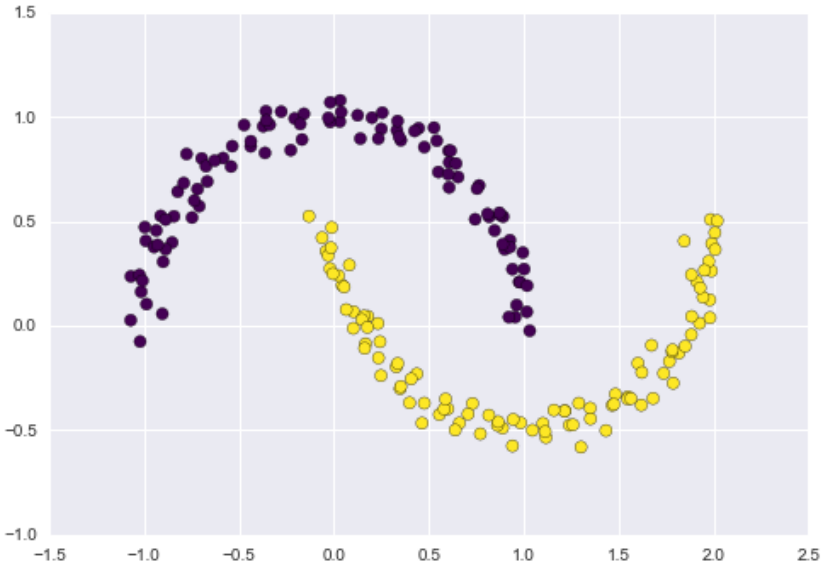


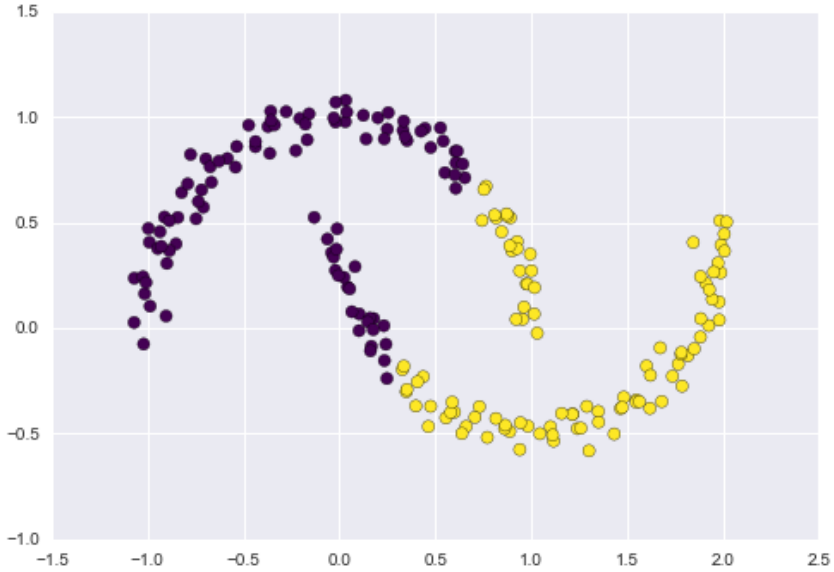






Few cases where K-means fails





If you are interested

- Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.

Thank You!