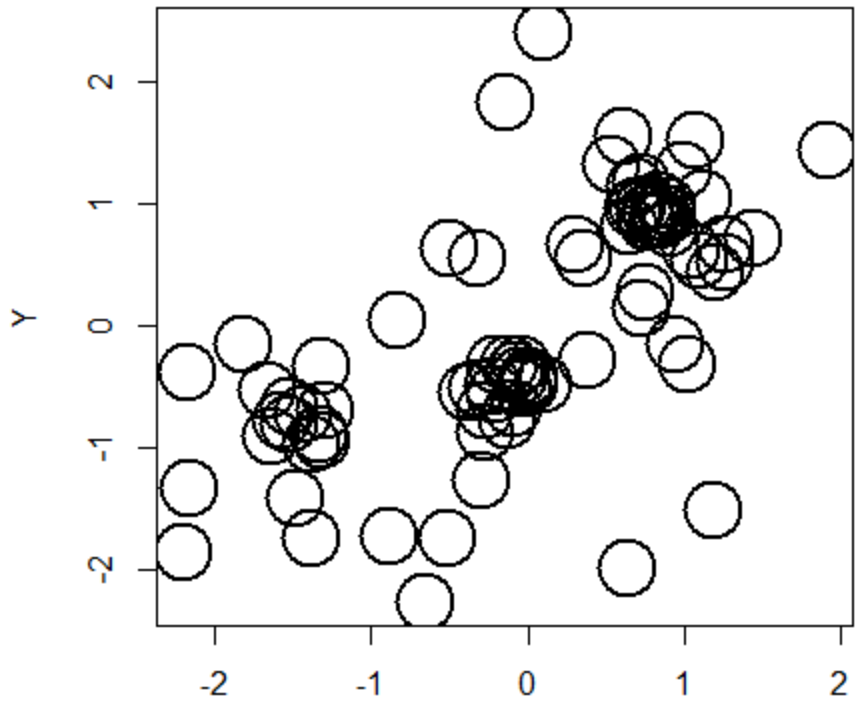
Introduction to K-means Clustering

K-means clustering: Algorithm

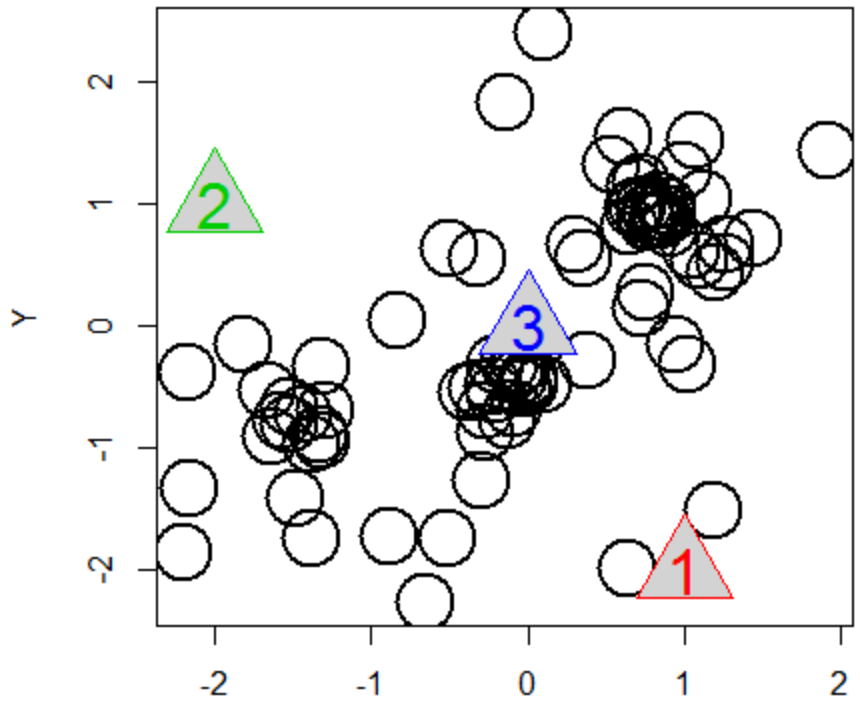
- Pre-requisites
  1. Get points in multi-dimensional space.
     - table, matrix, rectangular dataset
  2. Specify the number of clusters
     - Weakest point in algorithm
     - Get a random center for each cluster (makes algorithm non-deterministic)
     - Another weak point in the algorithm
- Repeat until convergence:
  1. For each point, determine its closest cluster center and assign that point to that cluster
  2. Determine the centroid (mean) for each cluster of points

K-Means Clustering (1)



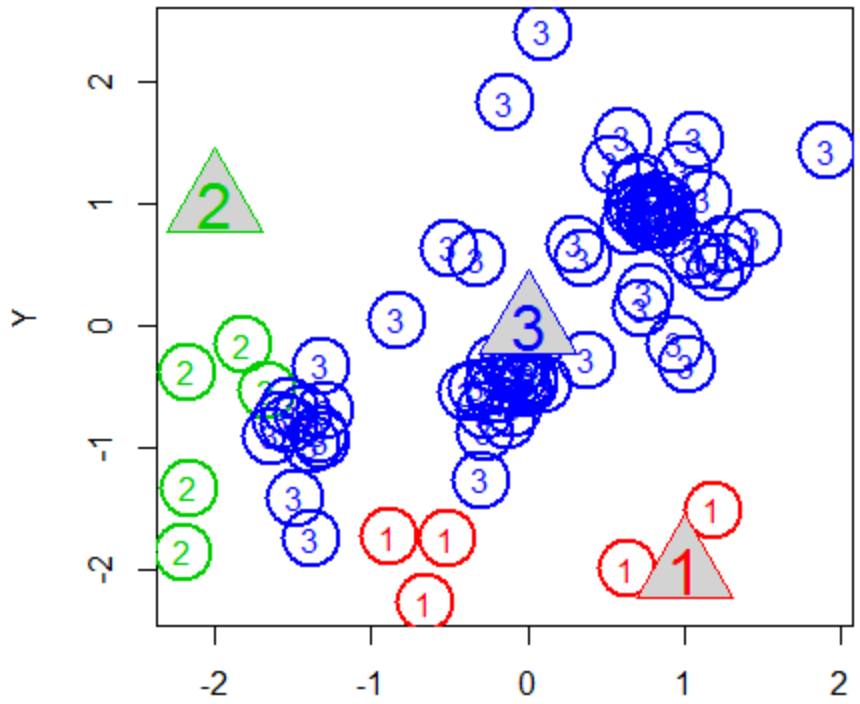
- Clustering starts by getting the data and representing the data as points in space. In this example the space is 2-dimensional.
- Each point describes an observation. An observation is an individual item.
- The dimensions are attributes that describe the item.

K-Means Clustering (2)



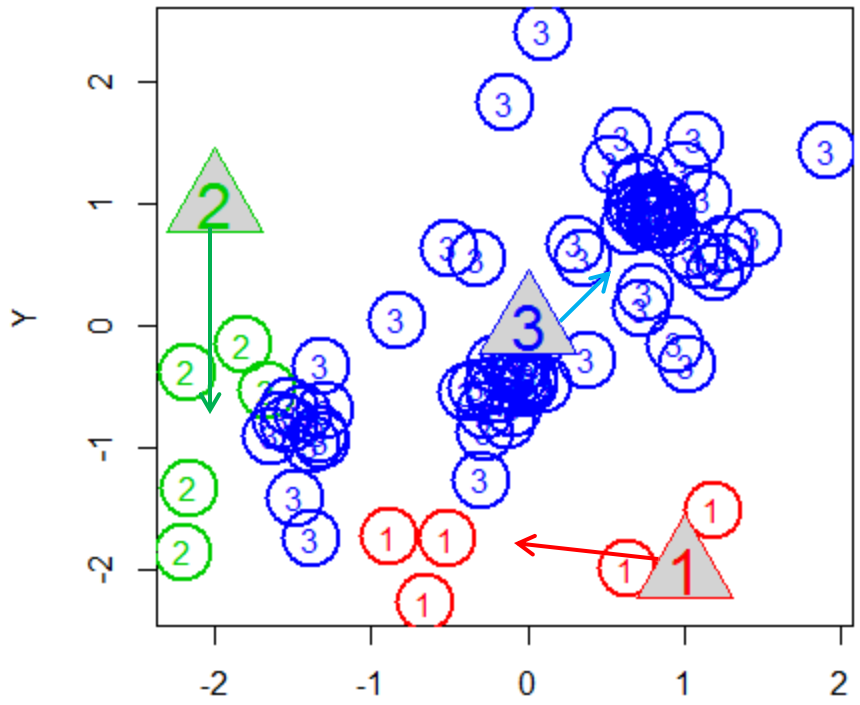
- Clustering continues by guessing or specifying a number of clusters (K). In this case we specify K = 3 by specifying 3 centroids.
- Each centroid represents a cluster.
- The centroid positions are determined randomly within the bounds of the points.

K-Means Clustering (3)



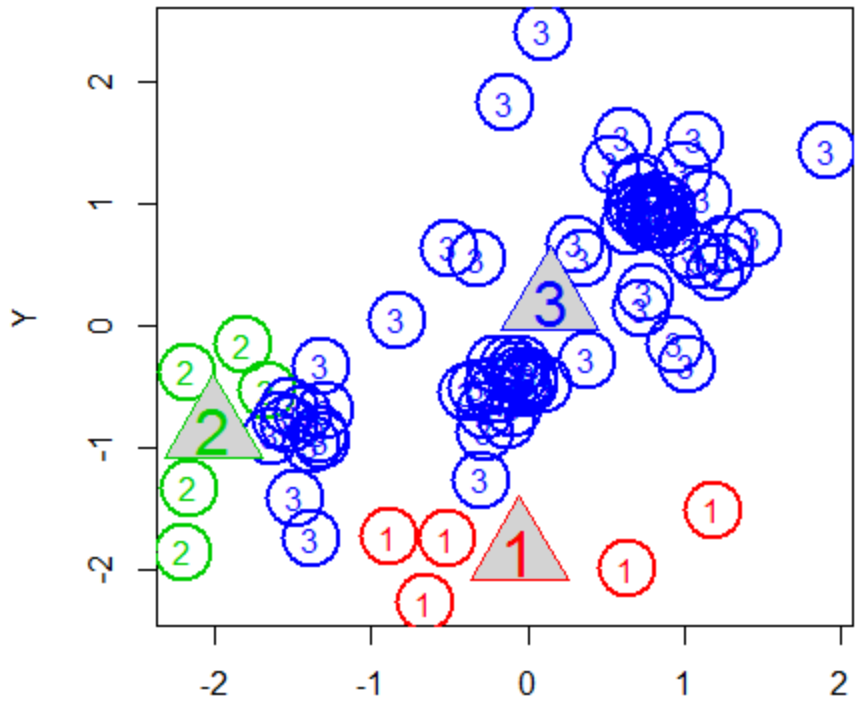
- Clustering continues by assigning each point to a cluster.
- For each point, the algorithm measures the distance to each centroid.
- For each point, the smallest distance to a centroid indicates the assignment.

K-Means Clustering (4)



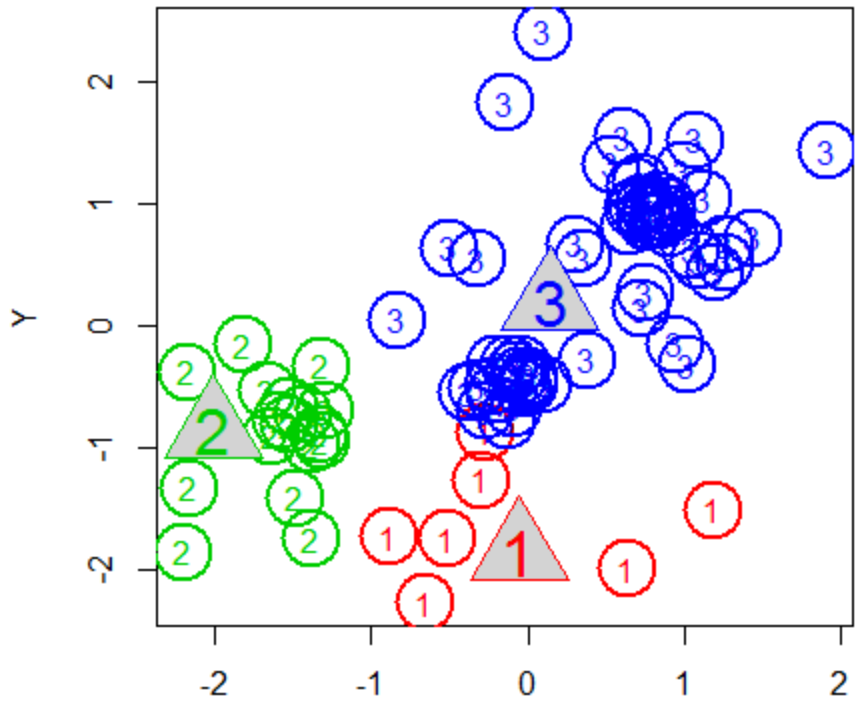
- Clustering continues by moving each centroid to the center of its cluster.

K-Means Clustering (5)



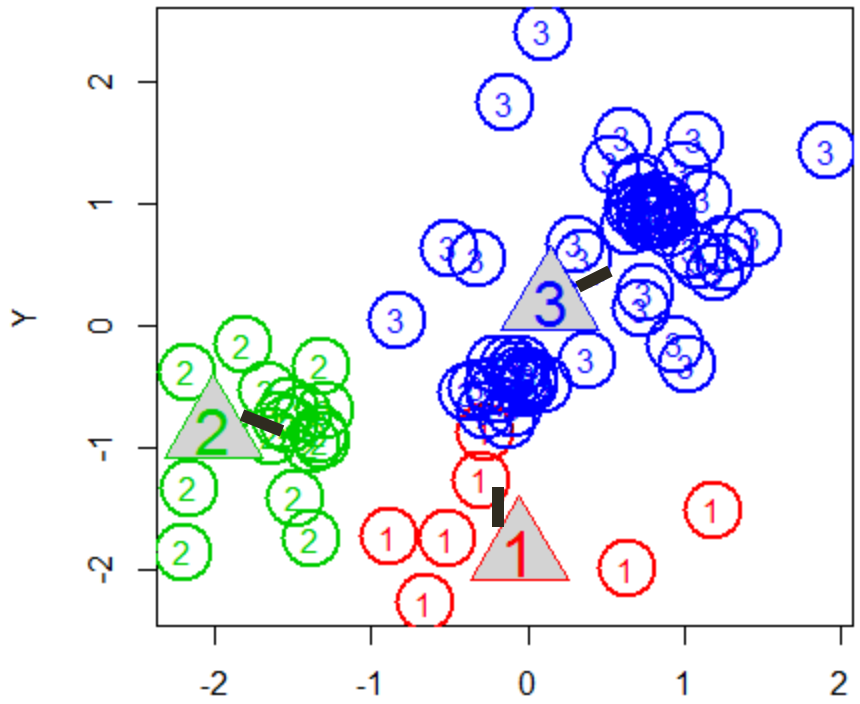
- Clustering continues by moving each centroid to the center of its cluster.

K-Means Clustering (6)



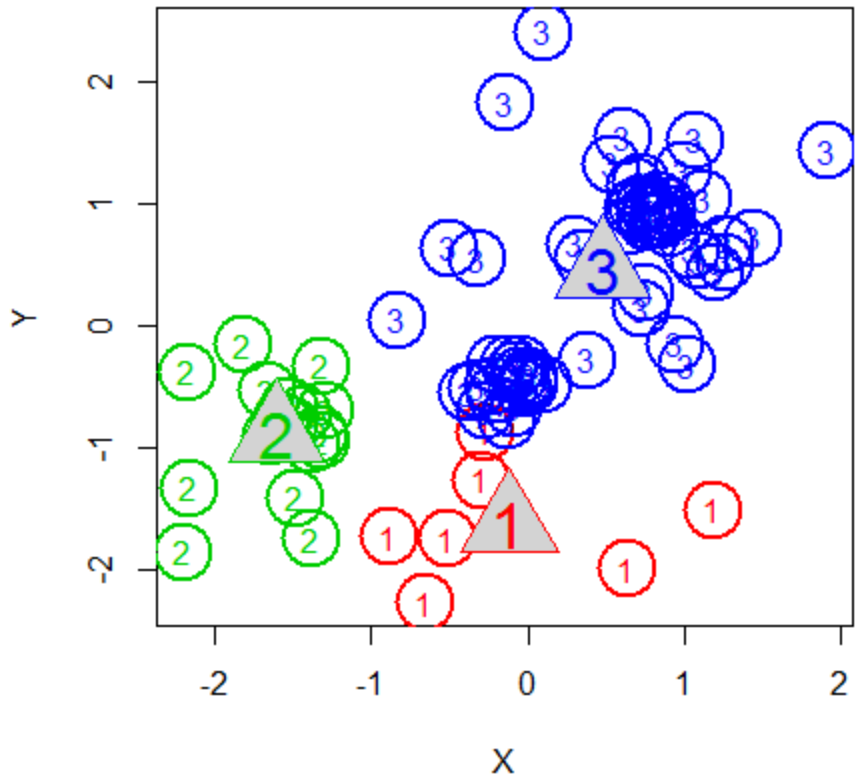
- Clustering continues by assigning each point to a cluster.
- For each point, the algorithm measures the distance to each centroid.
- For each point, the smallest distance to a centroid indicates the assignment.

K-Means Clustering (7)

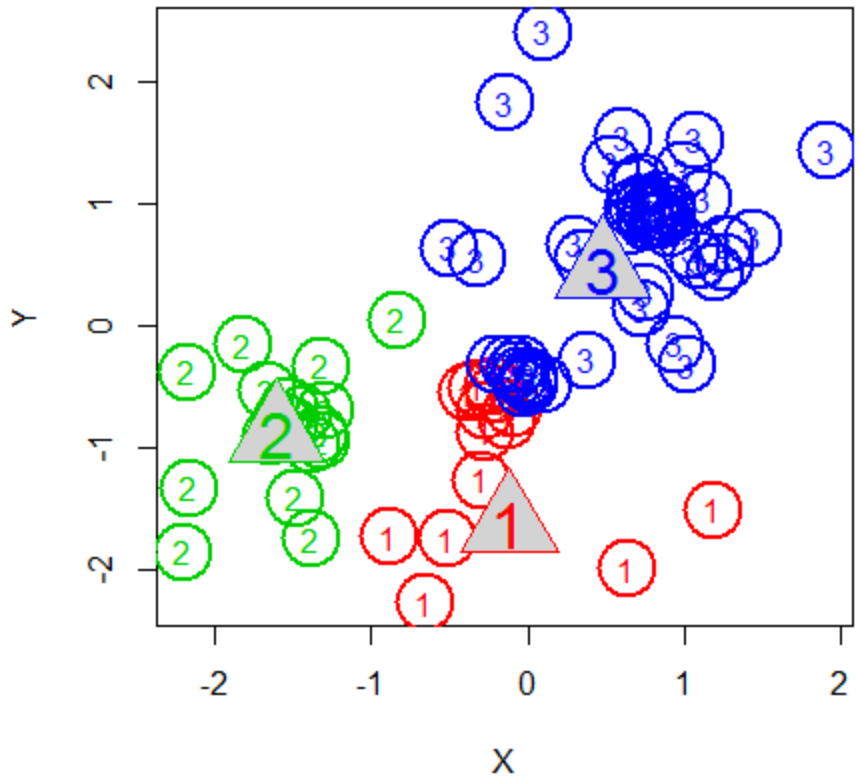


- Clustering continues by assigning each point to a cluster.
- For each point, the algorithm measures the distance to each centroid.
- For each point, the smallest distance to a centroid indicates the assignment.

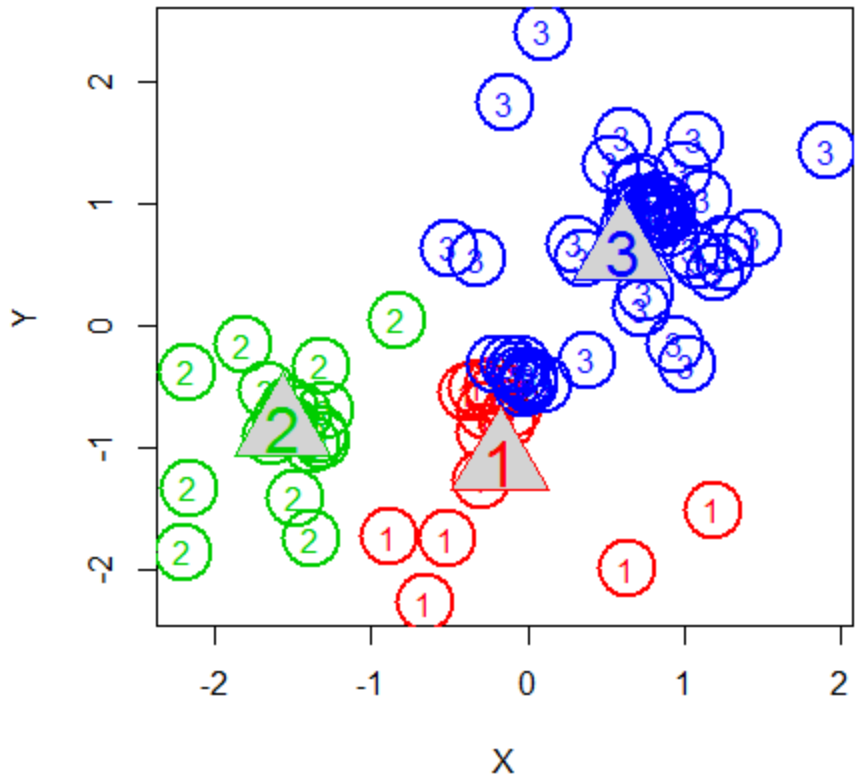
K-Means Clustering (8)



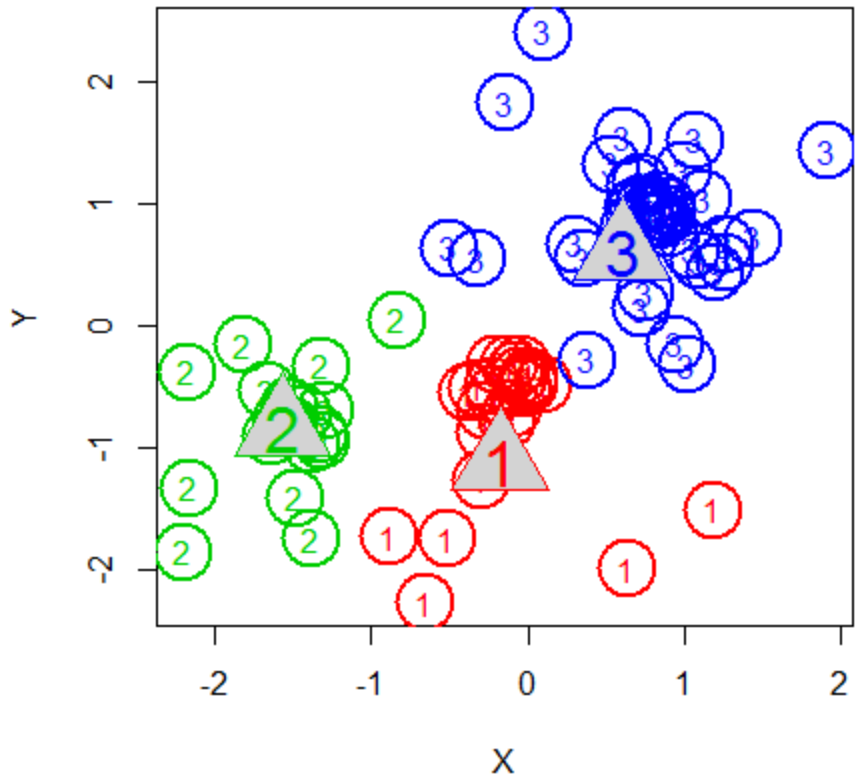
K-Means Clustering (9)



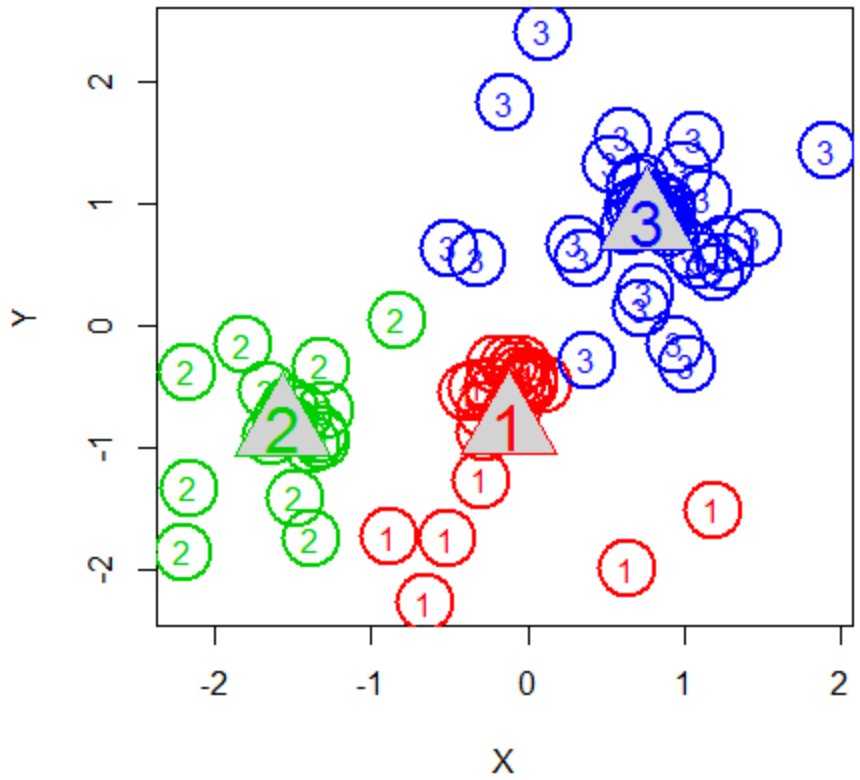
K-Means Clustering (10)



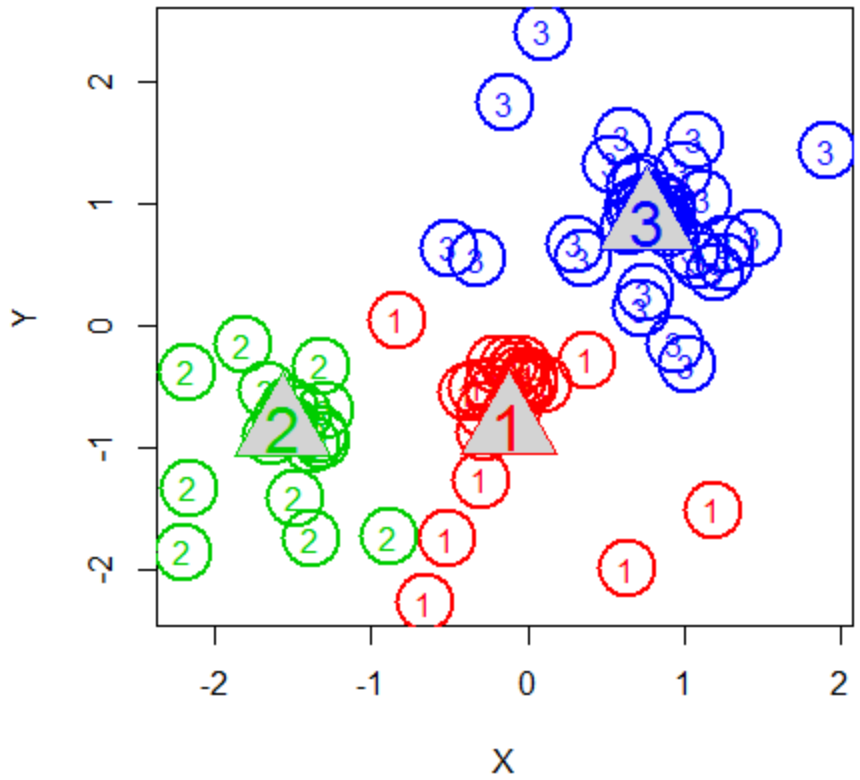
K-Means Clustering (11)



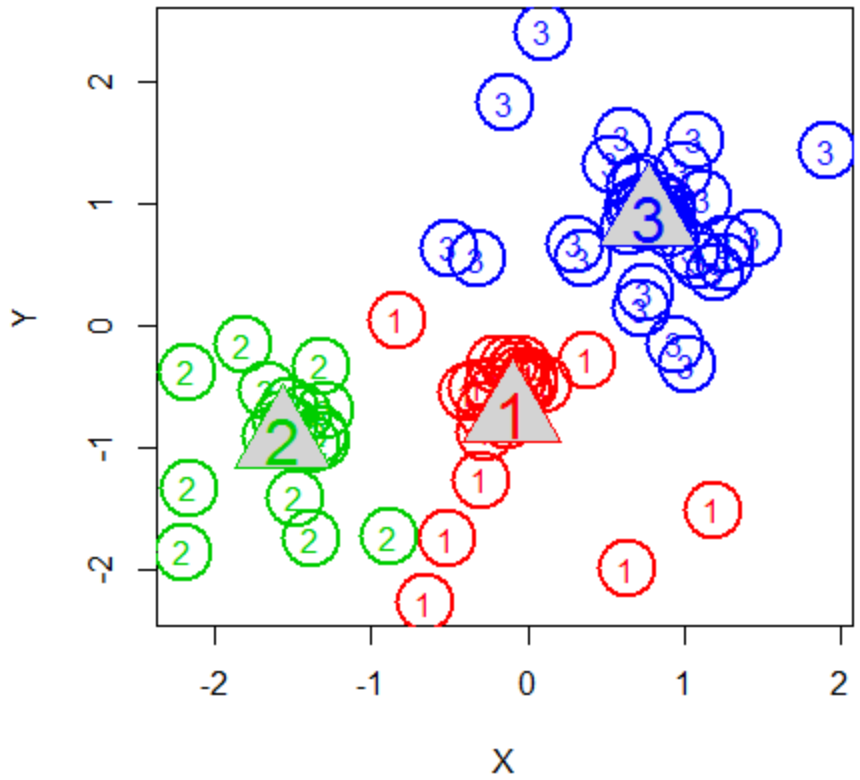
K-Means Clustering (12)



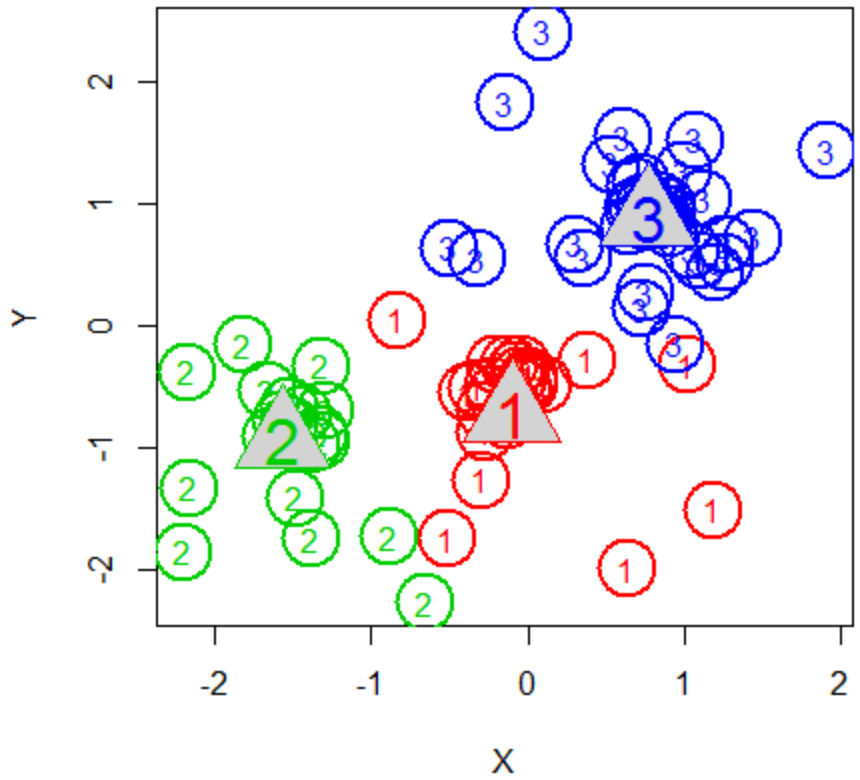
K-Means Clustering (13)



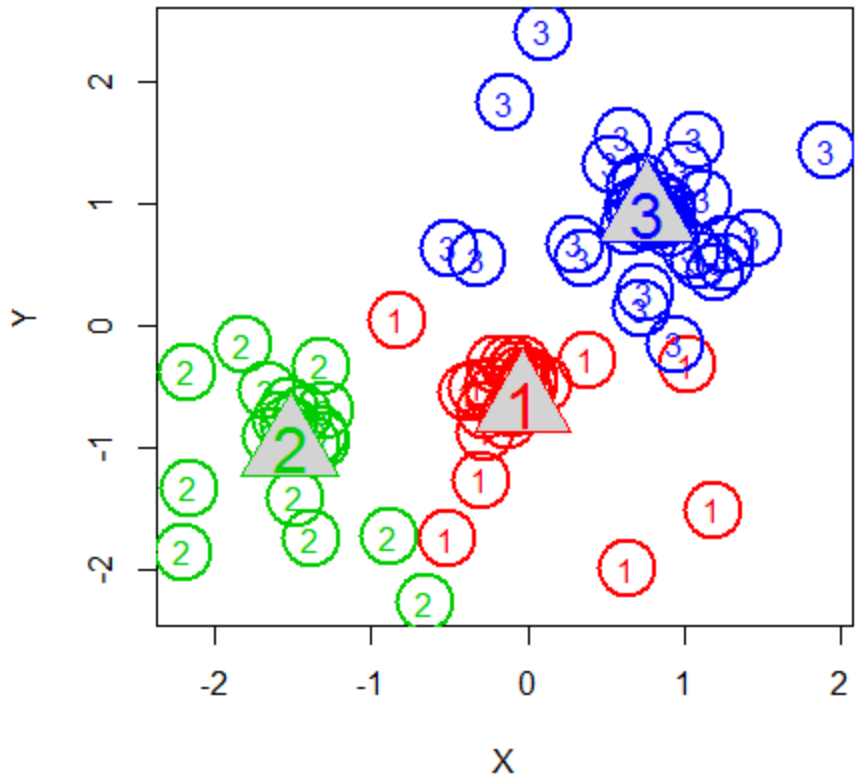
K-Means Clustering (14)



K-Means Clustering (15)



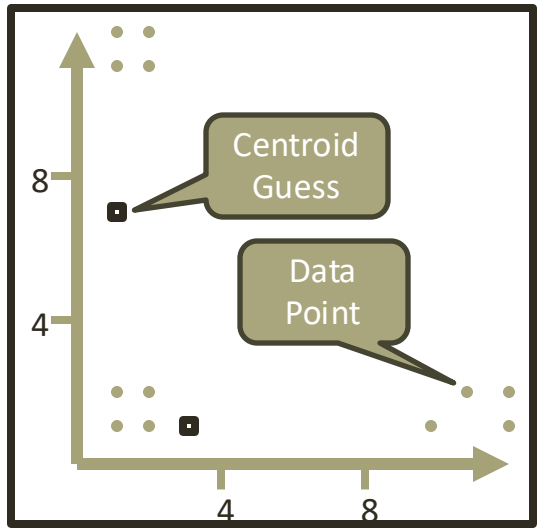
K-Means Clustering (16)



K-means

- A few remarks:
  - The number of clusters is determined before the K-means algorithm by choosing the number of initial guesses for centroids
  - Initial centroid (cluster) number and placement is an art
  - Categorical Data must be one-hot encoded
  - K-means is unsupervised because we do not tell the algorithm what outcome was observed or what outcome is desired

Lesson 08 In-class Quiz#1



How may clusters will be detected given this distribution of points and initial centroid guesses?

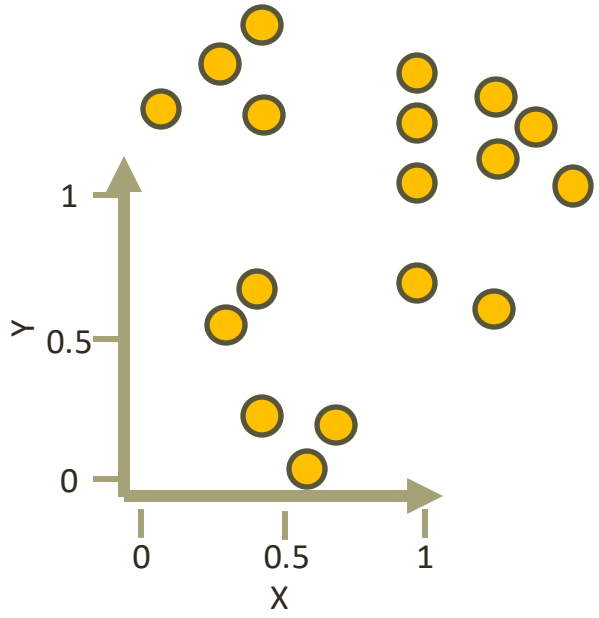
```
data = {'x': [1, 1, 2, 2, 9, 10, 11, 11, 1, 1, 2, 2],
        'y': [1, 2, 1, 2, 1, 2, 1, 2, 11, 12, 11, 12]}
```

initial guesses for centroids: (3,1), (1,7)

K-means Clustering Algorithm

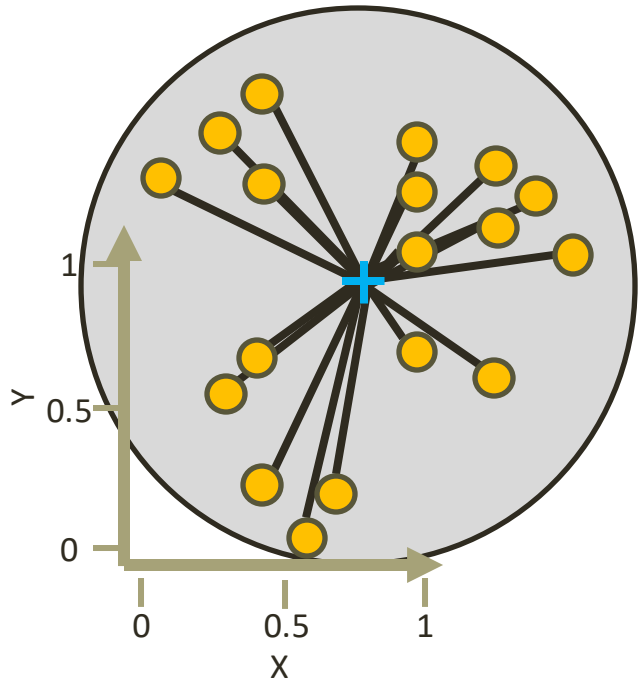
K-Means Number of Clusters

K-Means: K = 0



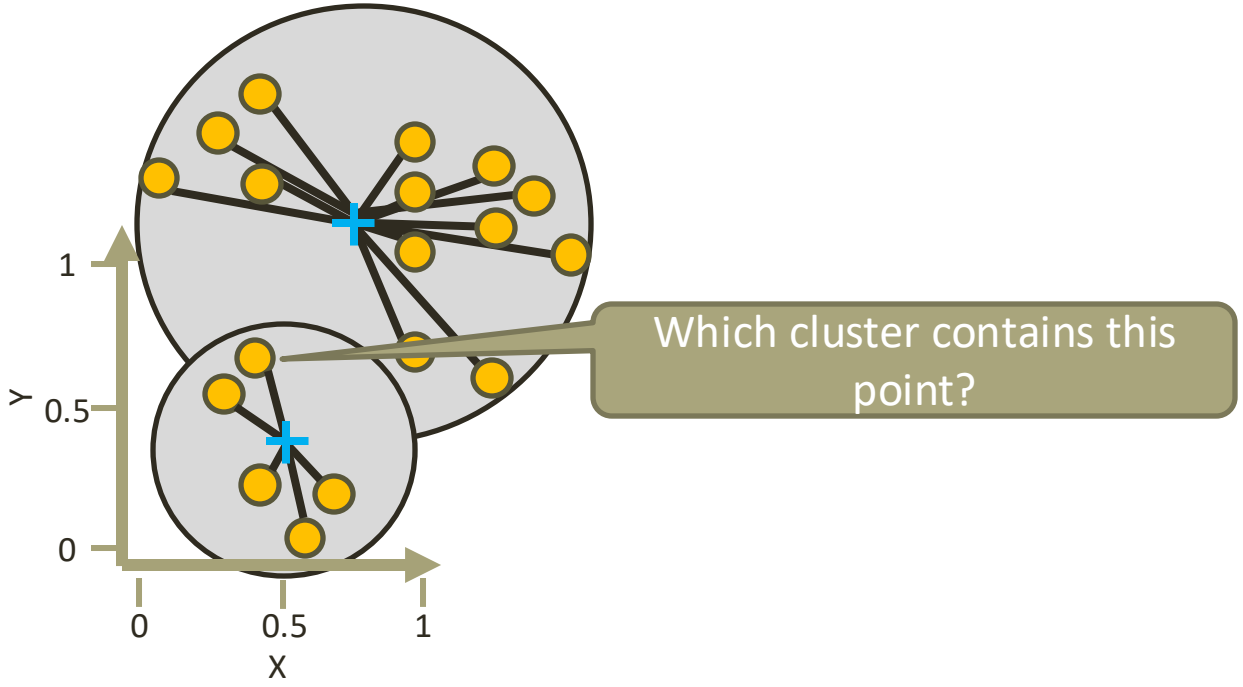
K is the number of clusters. If K = 0, then there is no clustering.

K-Means: K = 1

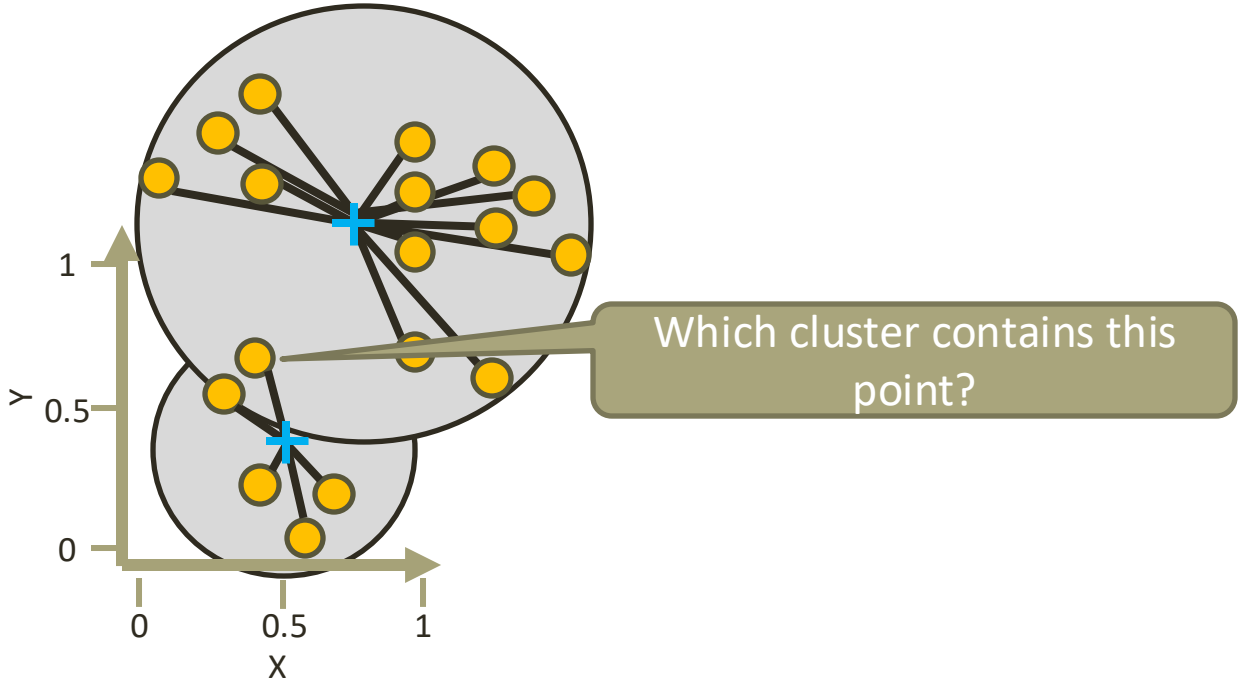


If K = 1, then clustering is trivial. The centroid of all points is the cluster center. The measure of clustering is called inertia and is the sum of the square distances.

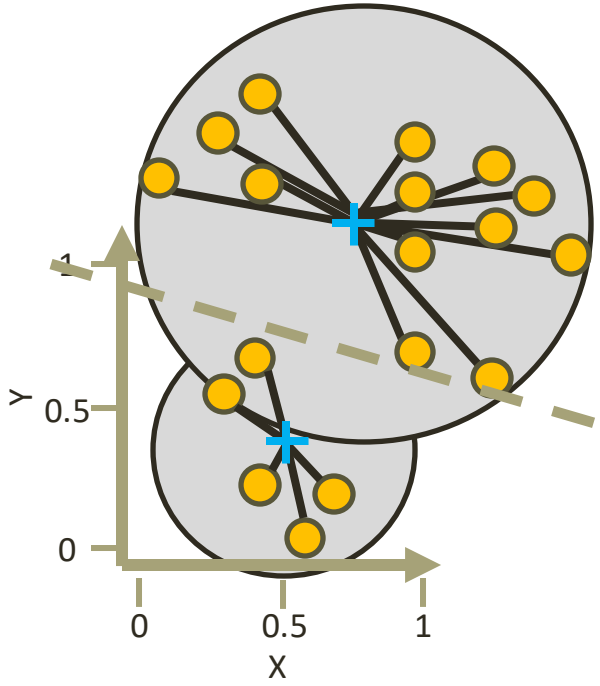
K-Means: K = 2



K-Means: K = 2

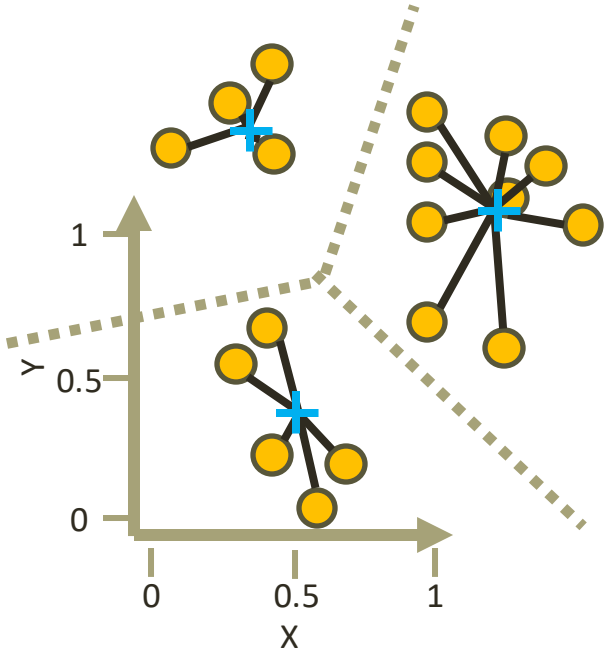


K-Means: K = 2

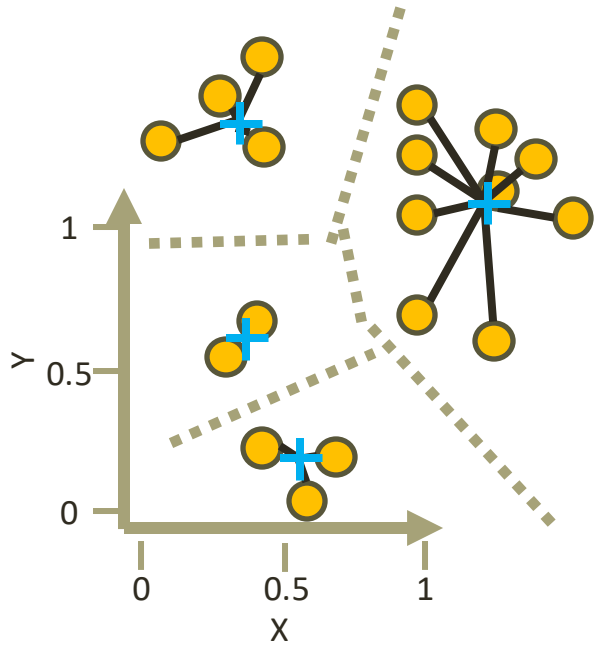


Only the distance to the cluster center counts. The circles are just for illustration.

K-Means: K = 3

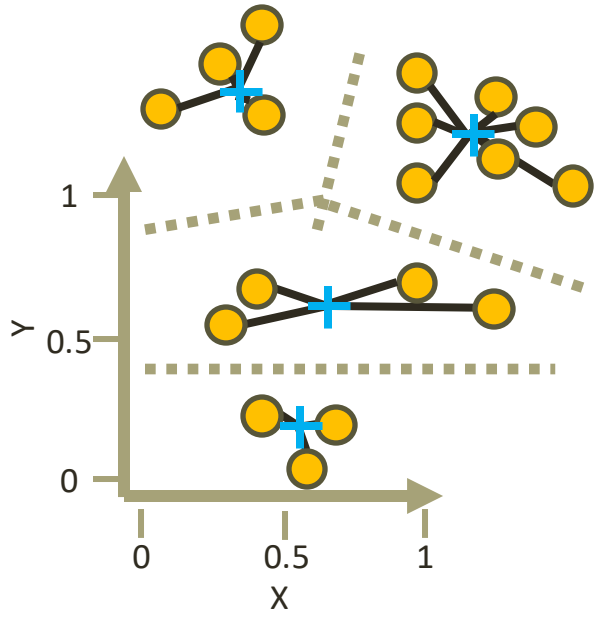


K-Means: K = 4 (1st Attempt)



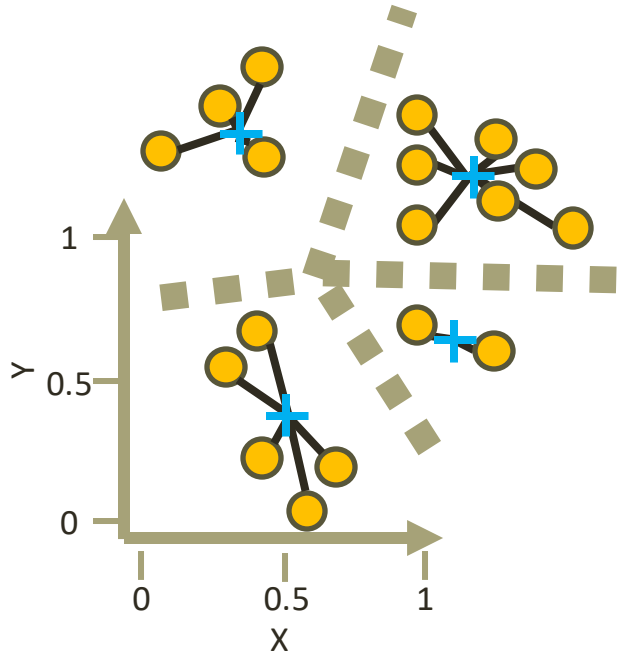
Clustering may have multiple outcomes. For a given number of clusters, clustering should be repeated multiple times to determine the best outcome. The best outcome has the smallest inertia

K-Means: K = 4 (2nd Attempt)



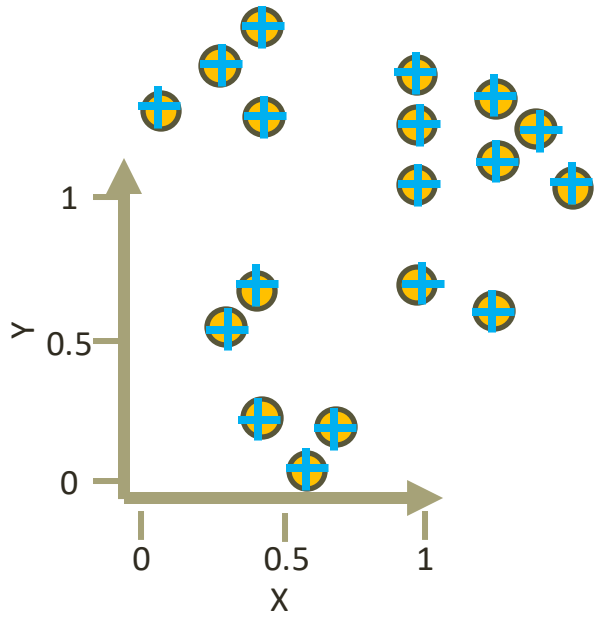
Clustering may have multiple outcomes. For a given number of clusters, clustering should be repeated multiple times to determine the best outcome. The best outcome has the smallest inertia

K-Means: K = 4 (3rd Attempt)



Clustering may have multiple outcomes. For a given number of clusters, clustering should be repeated multiple times to determine the best outcome. The best outcome has the smallest inertia

K-Means: Max K = 18



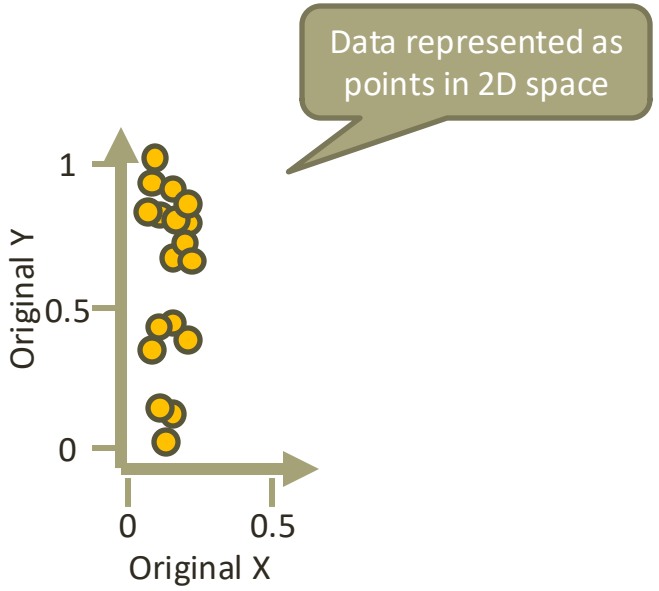
Increasing K will reduce inertia. When K = Number of Points, then inertia is zero.

K-Means: Number of Clusters

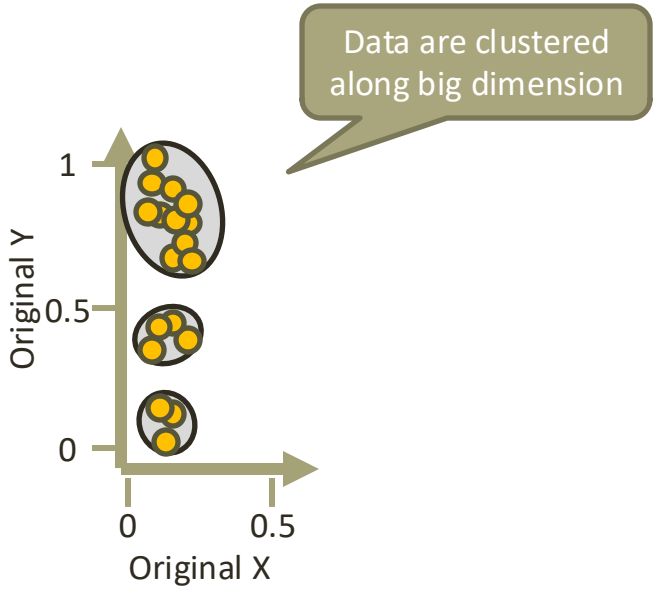
- Initial centroid number and placement is an art
- Inertia in K-Means is the sum of squared distances of points to their closest cluster center
  - It is a measure of cluster density. A smaller value indicates denser clusters
- Varying number of clusters and placements
  - Typically one tries different number of clusters
  - Typically one repeats K-Means multiple times for each number of clusters but with different cluster placements
  - The result with the lowest inertia is chosen

K-means Clustering Normalization

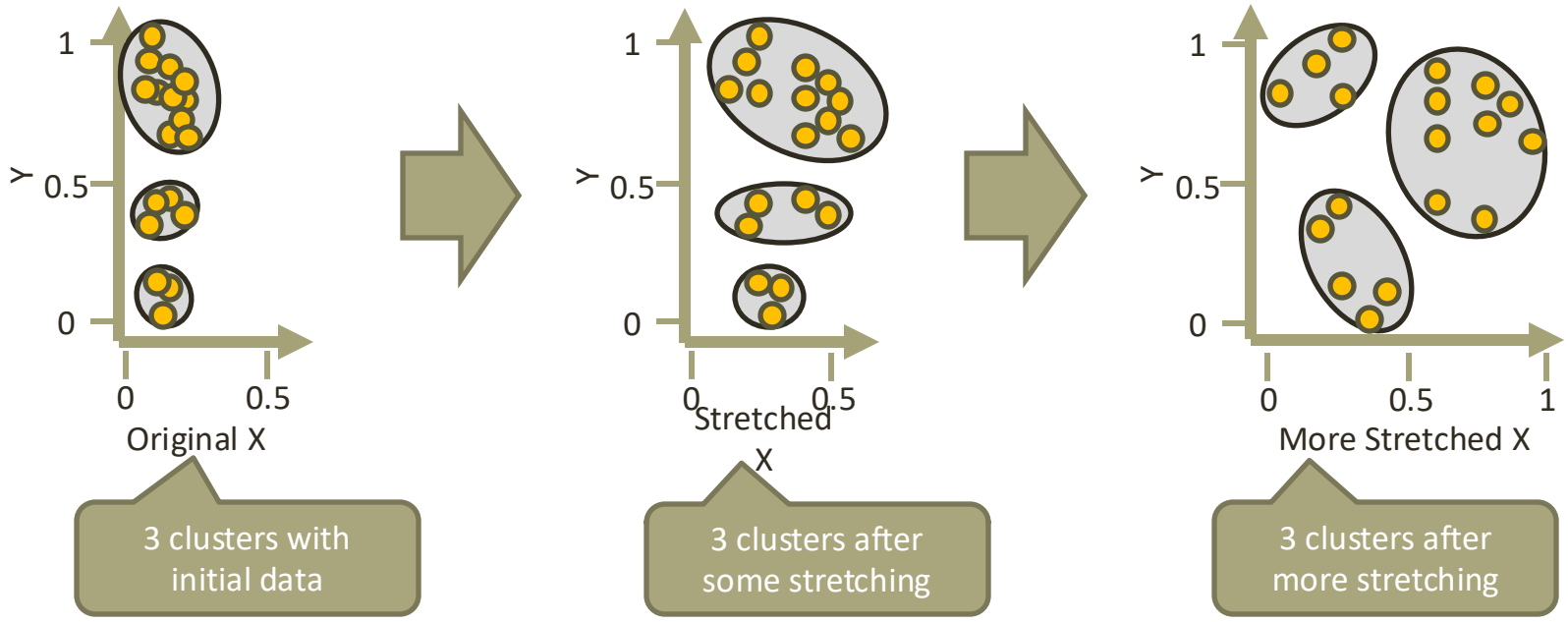
Normalization and Clustering



Normalization and Clustering



Normalization and Clustering



Normalization and Clustering

- Clustering is typically carried out in normalized space
- There is no need to normalize after K-means is completed
- Centroids may need to be de-normalized after clustering.
  - Denormalization needs to be done using the original normalization parameters
- There is no need to de-normalize points as those would be the original points
- Normalizations are important to put data on equal terms