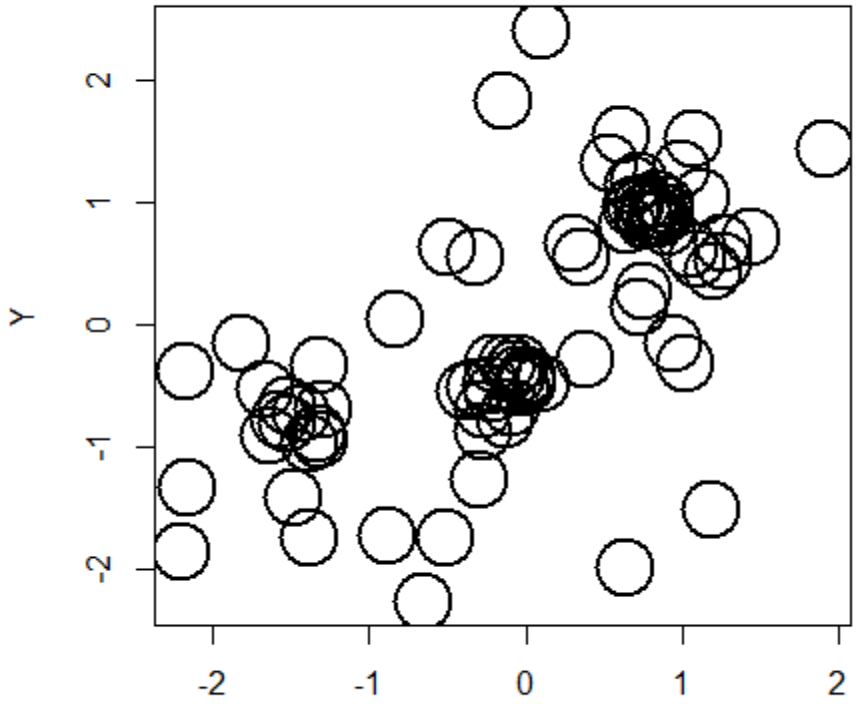
Introduction to K-means Clustering

K-means clustering: Algorithm

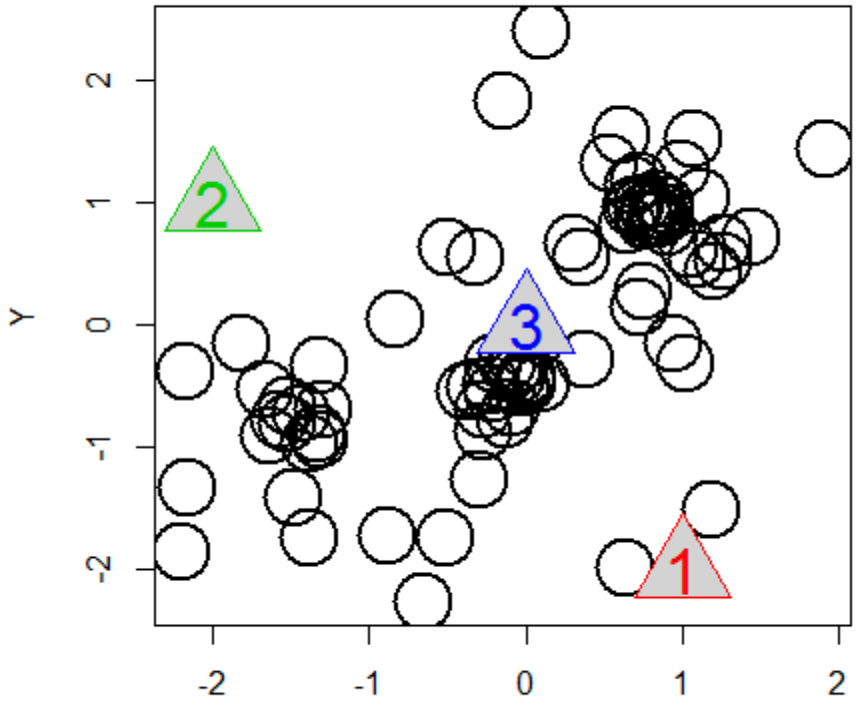
- Pre-requisites
  1. Get points in multi-dimensional space.
     - table, matrix, rectangular dataset
  2. Specify guesses for cluster centers
     - Specify number of clusters: Weakest point in algorithm
     - Choose a center for each cluster: Second Weakest point in algorithm because data does not determine outcome of algorithm.
- Repeat until convergence:
  1. For each point, determine its closest cluster center and assign that point to that cluster
  2. Designate the centroid (mean) for each cluster of points as the cluster center

K-Means Clustering



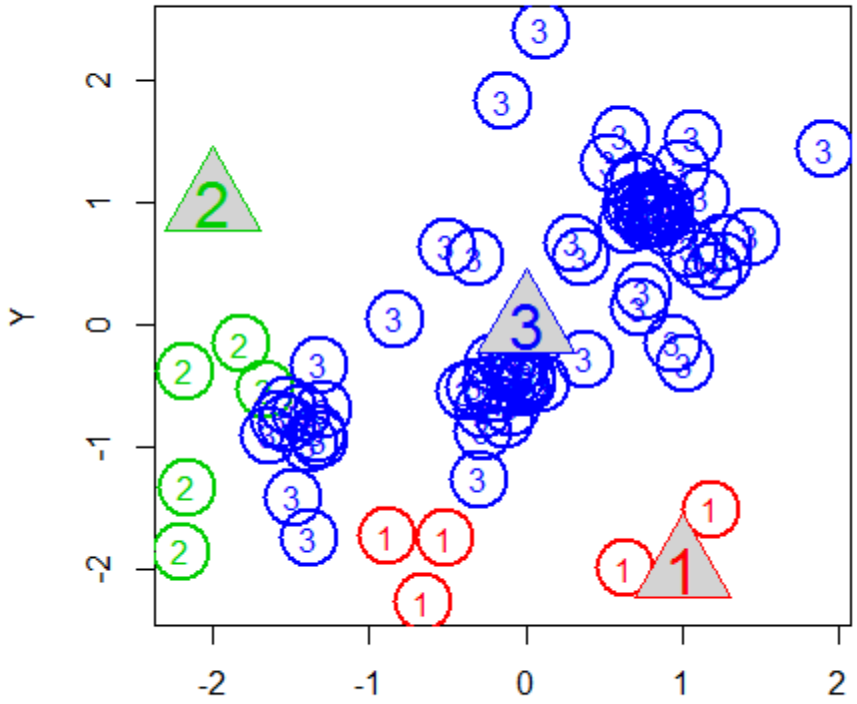
- Clustering starts by getting the data and representing the data as points in space. In this example the space is 2-dimensional.
- Each point describes an observation. An observation is an individual item.
- The dimensions are attributes that describe the item.

K-Means Clustering



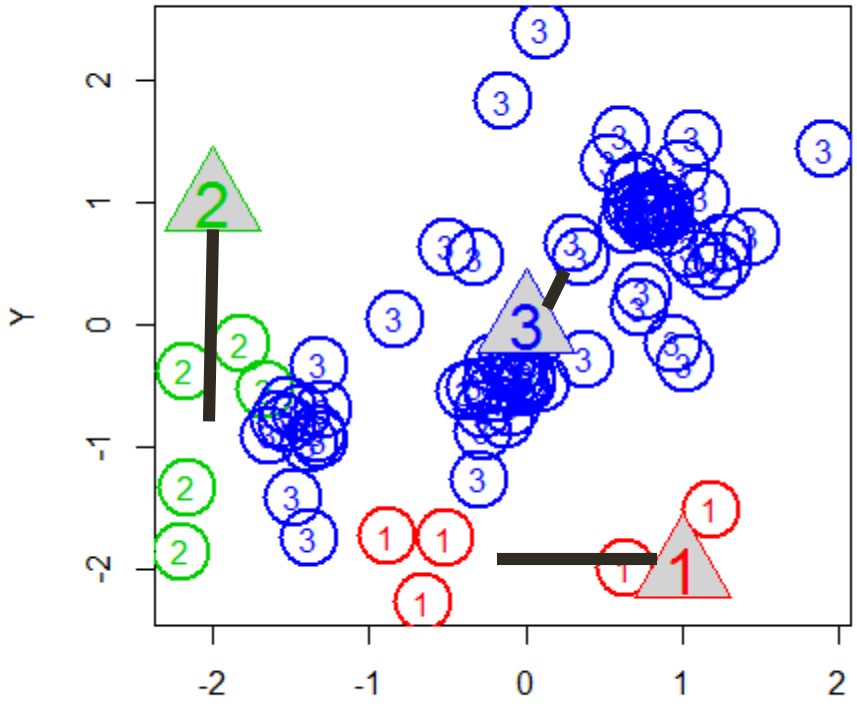
- Clustering continues by guessing, presuming, or specifying a number of clusters.
- Each centroid represents a cluster.
- The centroid positions are determined randomly. The centroids should be within the bounds of the points.

K-Means Clustering



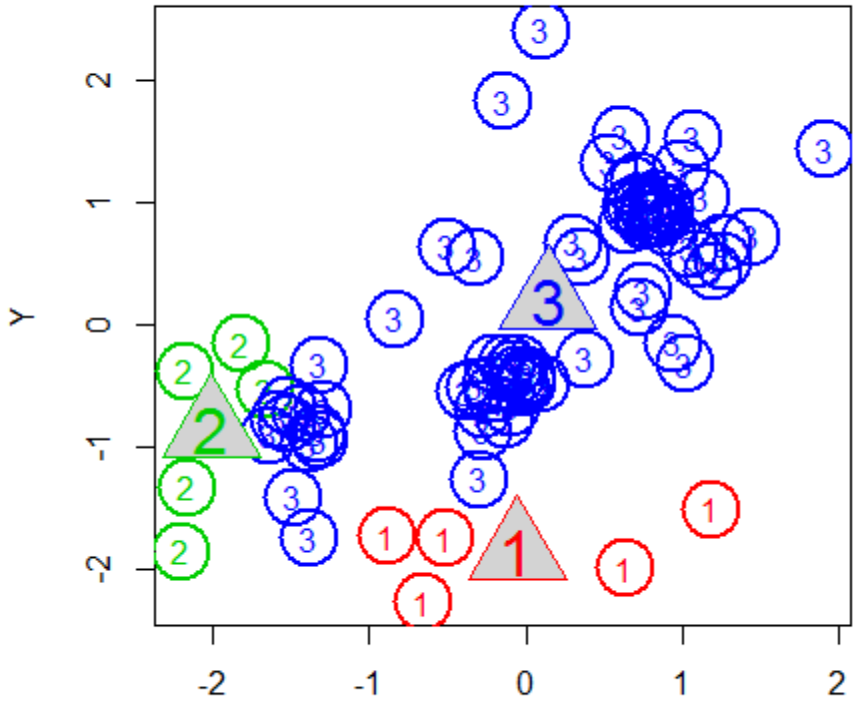
- Clustering continues by assigning each point to a cluster.
- For each point, the algorithm measures the distance to each centroid.
- For each point, the smallest distance to a centroid indicates the assignment.

K-Means Clustering



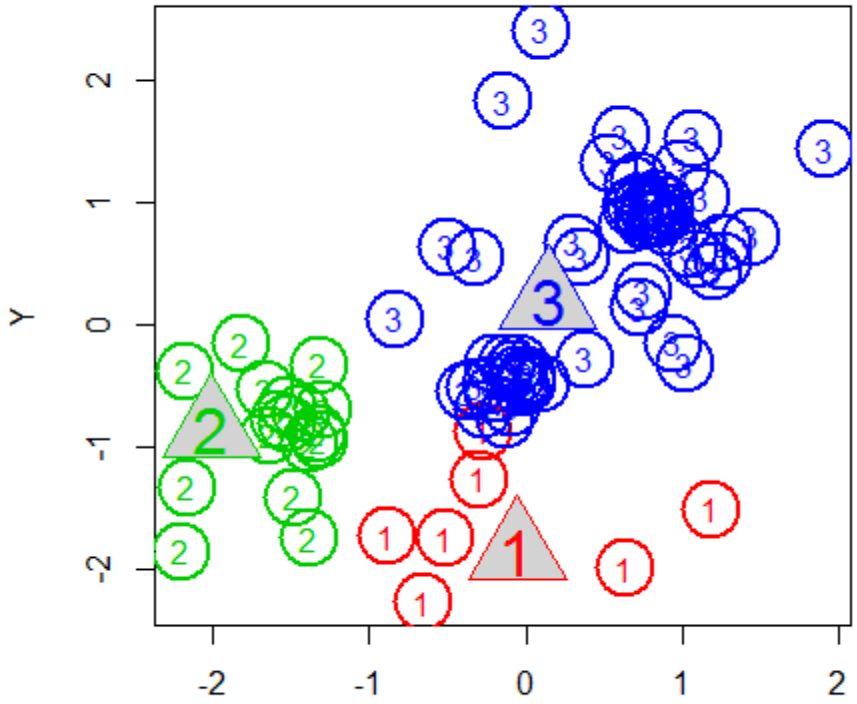
- Clustering continues by moving each centroid to the center of its cluster.

K-Means Clustering



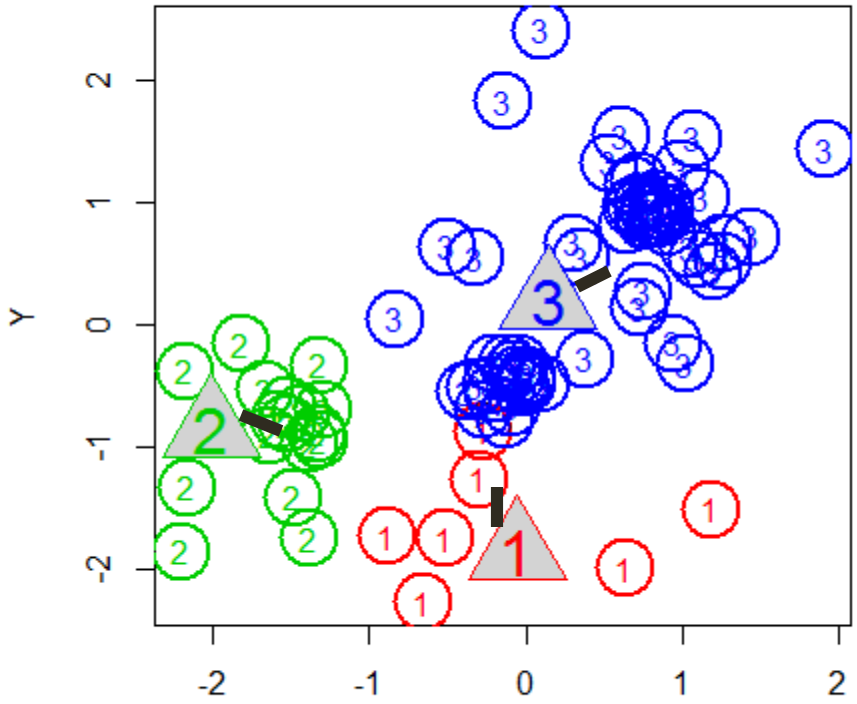
- Clustering continues by moving each centroid to the center of its cluster.

K-Means Clustering



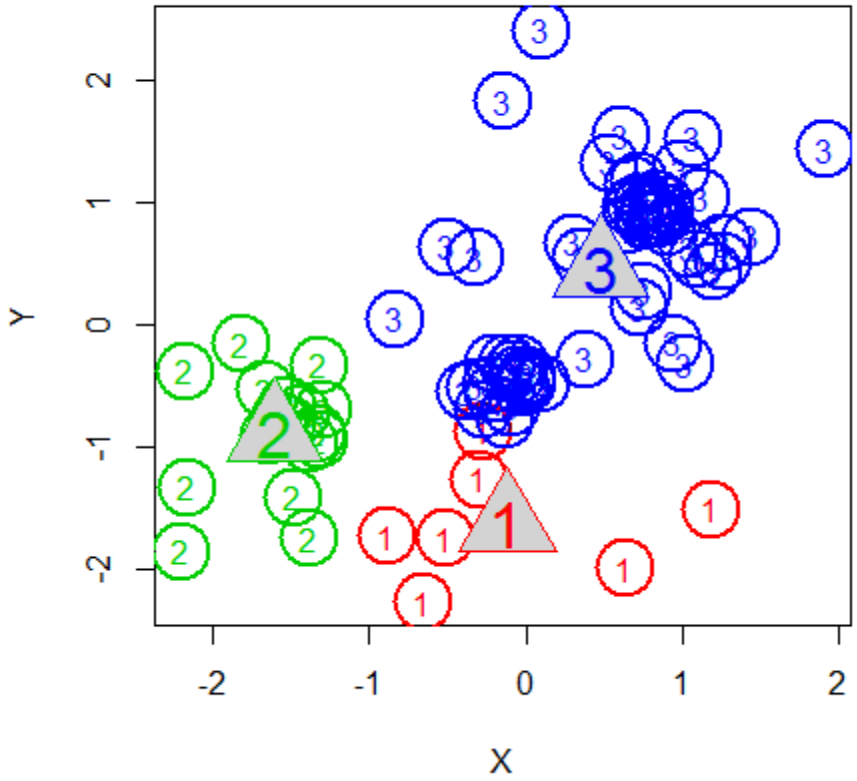
- Clustering continues by assigning each point to a cluster.
- For each point, the algorithm measures the distance to each centroid.
- For each point, the smallest distance to a centroid indicates the assignment.

K-Means Clustering

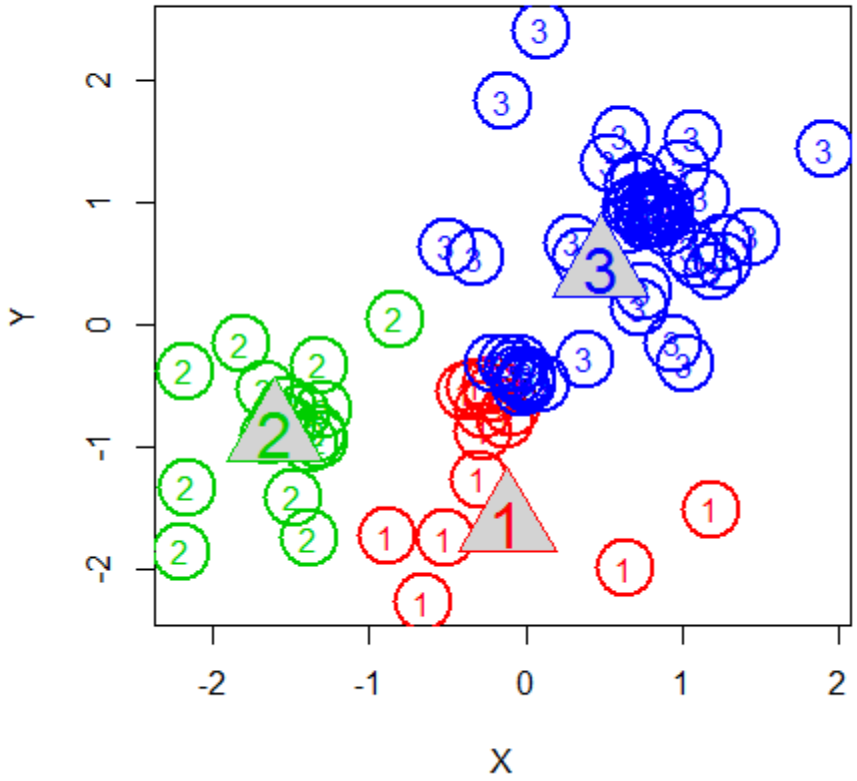


- Clustering continues by assigning each point to a cluster.
- For each point, the algorithm measures the distance to each centroid.
- For each point, the smallest distance to a centroid indicates the assignment.

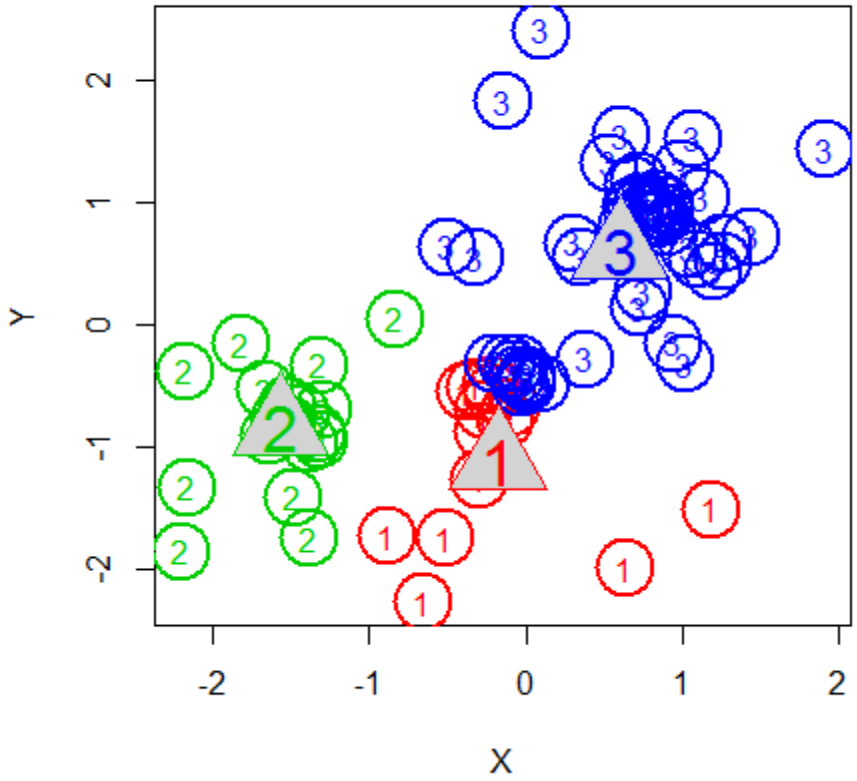
K-Means Clustering



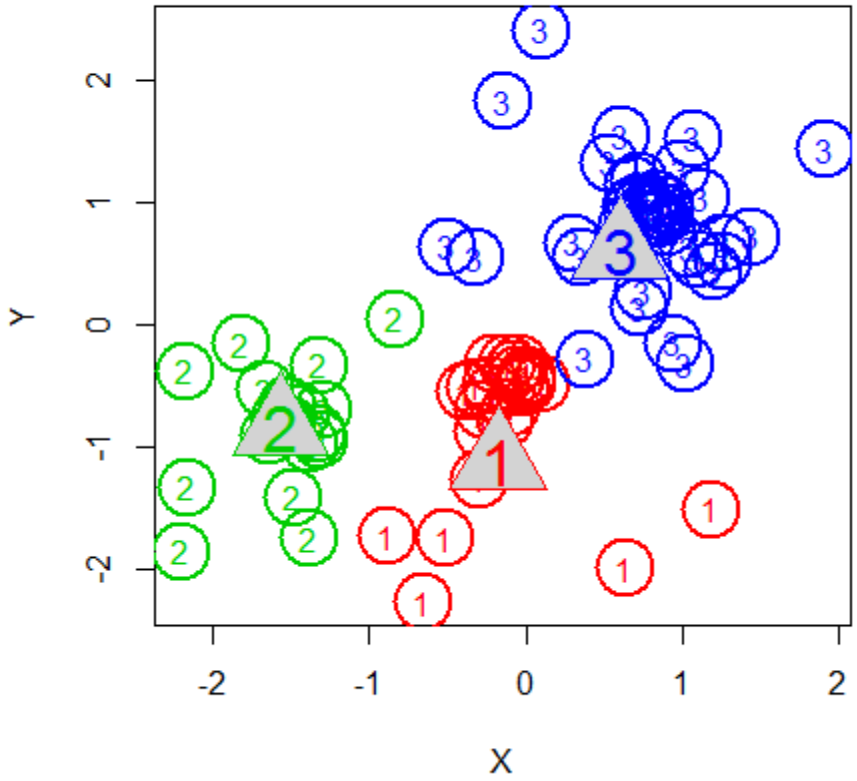
K-Means Clustering



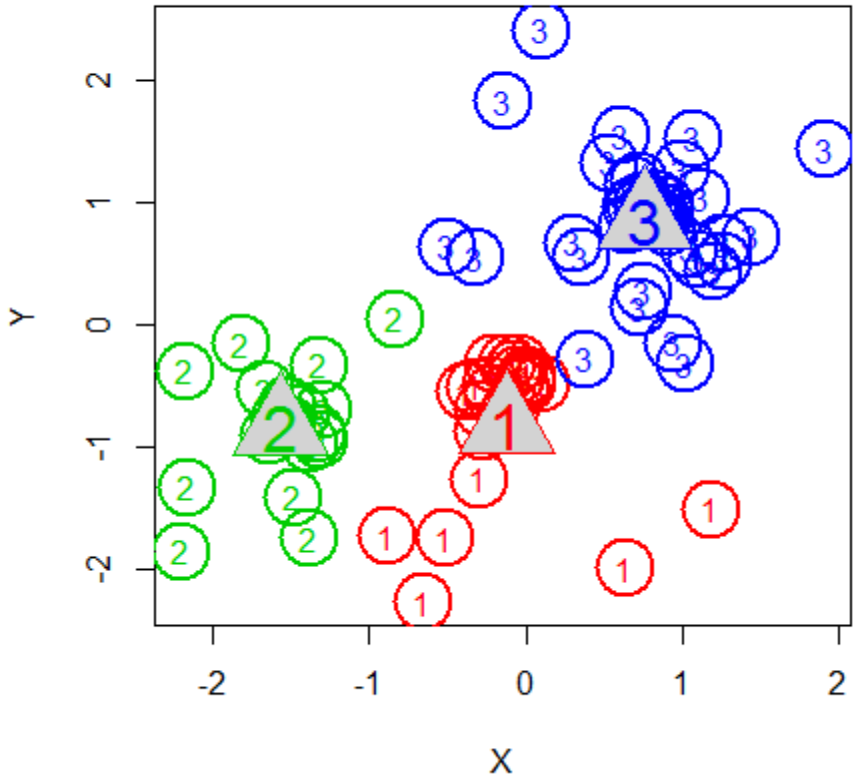
K-Means Clustering



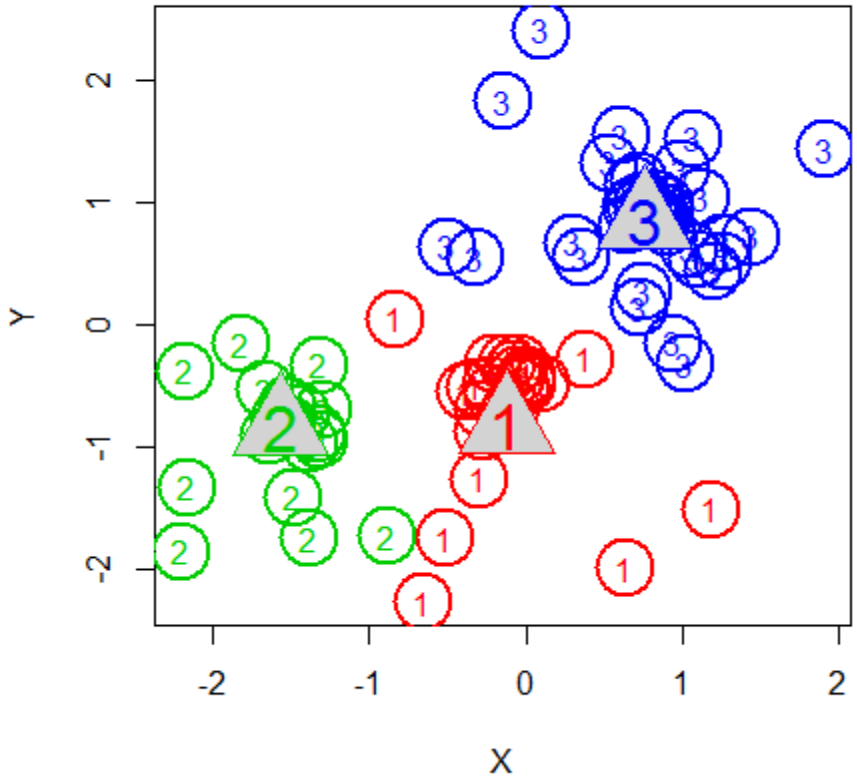
K-Means Clustering



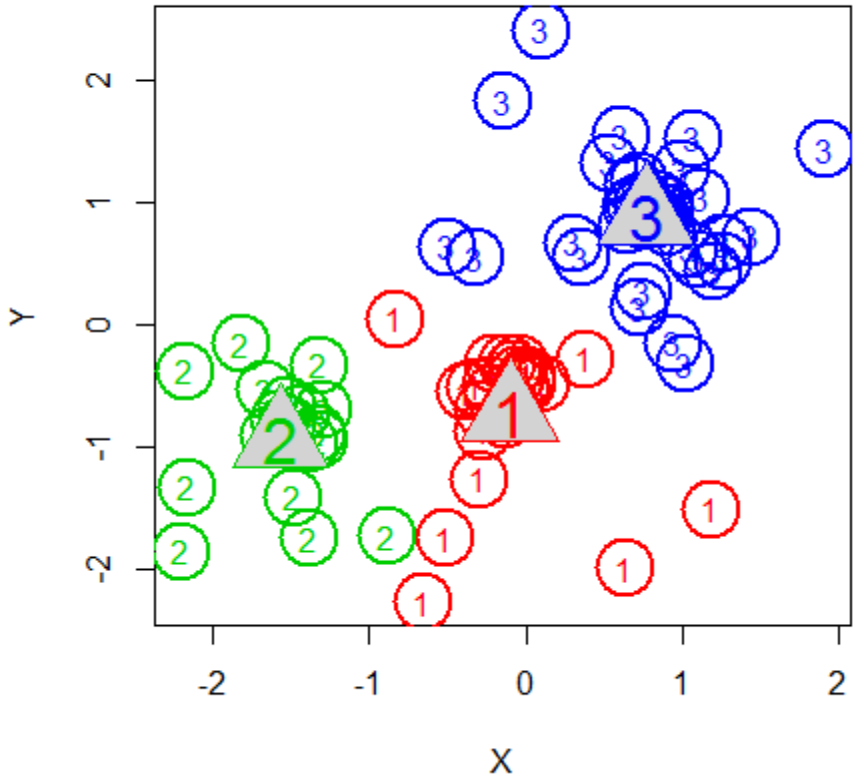
K-Means Clustering



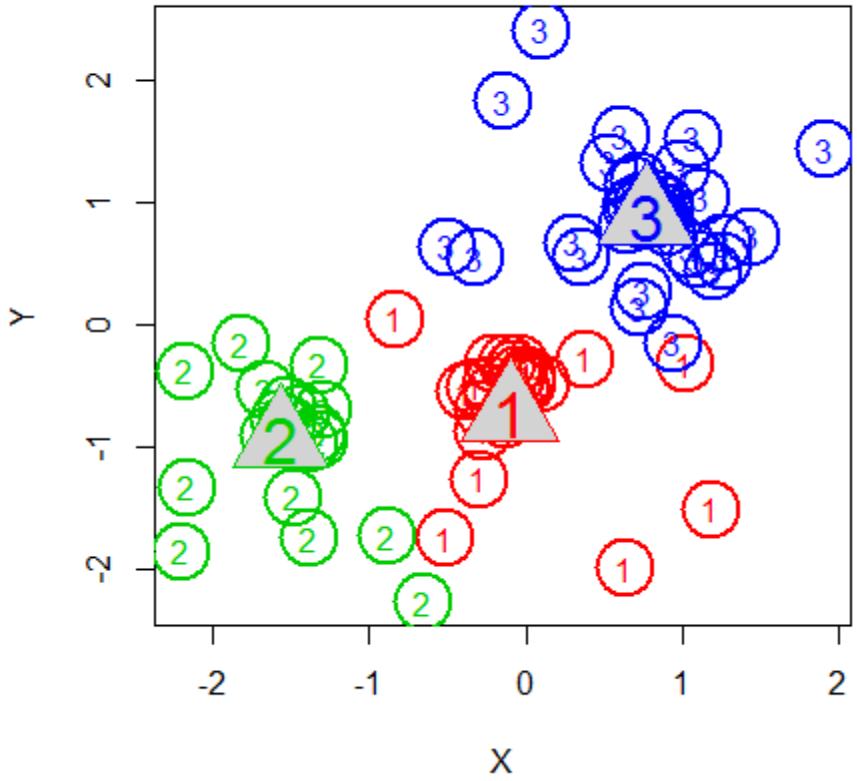
K-Means Clustering



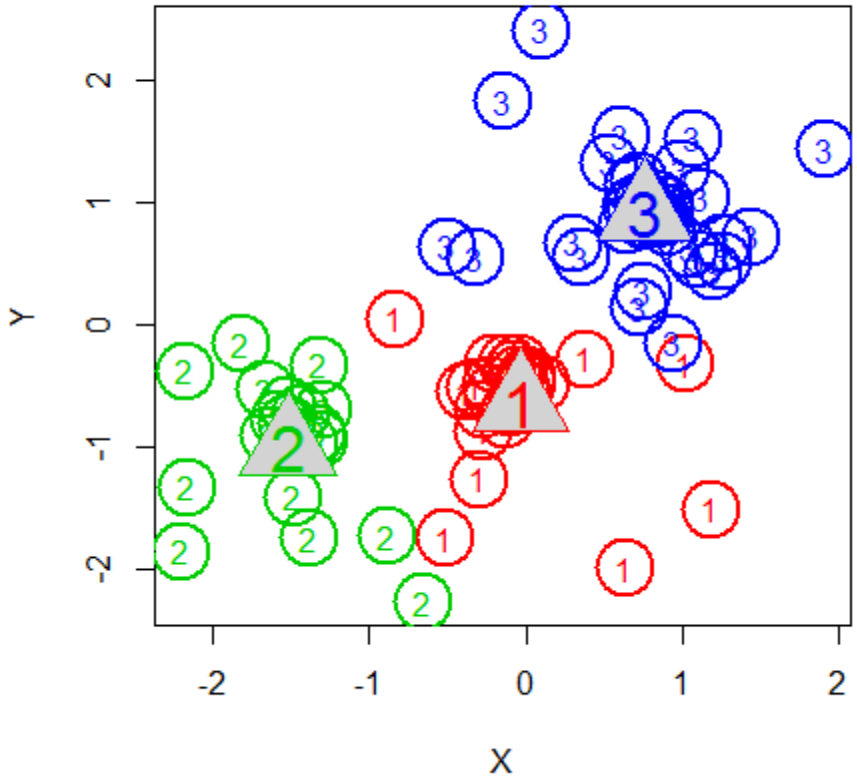
K-Means Clustering



K-Means Clustering



K-Means Clustering



K-means

- Some Points:
  - Initial centroid 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.
  - Normalizations are important to put data on equal terms

In-Class Exercise

Complete K-Means in Python

- Download L07-1-KMeans_Incomplete.py to your working directory.
- Open L07-1-KMeans_Incomplete.py in Spyder
- Run the script (The result will be wrong)
- Complete the function KMeans()
- Specifically, replace all lines that say: "**Replace this line with code**".
- Run the script (The result will be correct)