# K-Means Clustering

Contents:

1. Implementation details

Algorithm:

1. Take the number of clusters and the cluster centroids from input or select the centroids randomly based on the number of clusters from the dataset.
2. For each object xi
   1. Calculate the distance between xi and all the K centroids and (re) assign object xi to the cluster whose centroid it is closest to.
   2. Update the centroid value of each cluster after new assignments by computing the mean of the cluster.
3. Iterate over step b until maximum number of iterations are reached or the clusters do not change anymore.
4. Comparison with external index:
   1. Rand coefficient = number of matches / total number of attributes

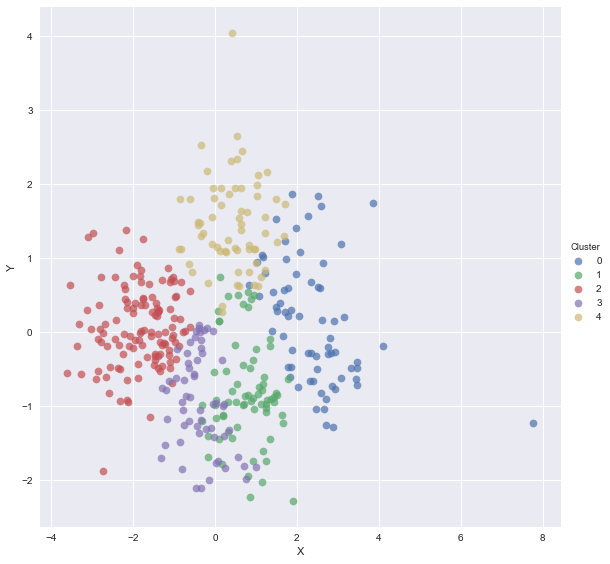
Rand coefficient = M11 + M00 / M11 + M00 + M10 + M01

* + 1. Dataset: cho.txt

Number of clusters: 5

Initial centroids: 5,25,32,100, 132

Rand Coefficient: 0.8111627.

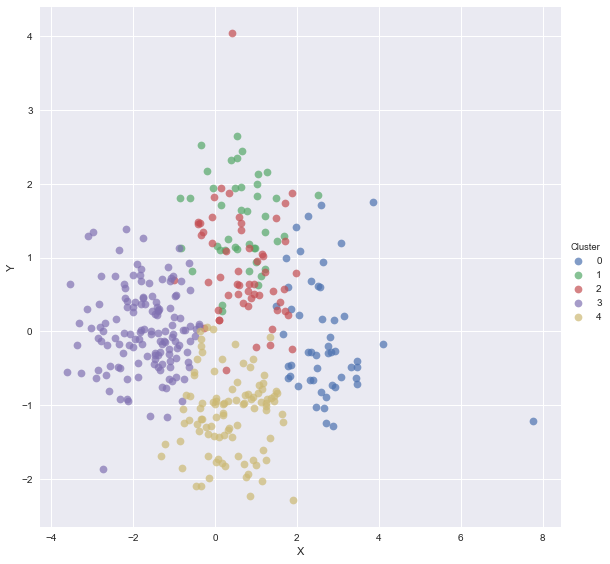


* + 1. Dataset: cho.txt

Number of clusters: 5

Initial centroids: 10, 24, 57, 99, 266

Rand Coefficient: 0.8014577.

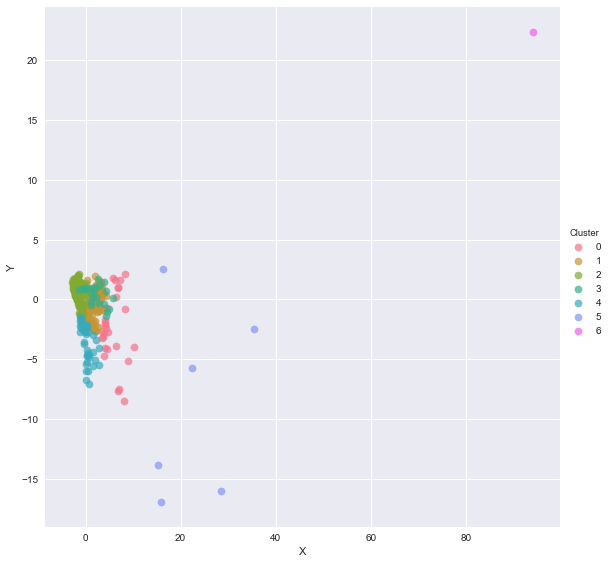


* + 1. Dataset: iyer.txt

Number of clusters: 7

Initial centroids: 8, 19, 46, 88, 146, 201, 299

Rand Coefficient: 0.6743375



* 1. Jaccard coefficient = number of matches / number of not-both-zero attribute value

Jaccard coefficient = M11 / M11 + M10 + M01

* + 1. Dataset: cho.tx

Number of clusters: 5

Initial centroids: 5, 25, 32, 100, 132

Jaccard Coefficient: 0.4064635

* + 1. Dataset: cho.txt

Number of clusters: 5

Initial centroids: 10, 24, 57, 99, 266

Jaccard Coefficient: 0.4100590

* + 1. Dataset: iyer.txt

Number of clusters: 7

Initial centroids: 8, 19, 46, 88, 146, 201, 299

Jaccard Coefficient: 0.29449429

1. Pros and Cons of k-means algorithm:
   1. Pros:
      1. Easy to implement.
      2. Effective in terms of time complexity. (O(tkn) where t - number of iterations, k – the number of clusters and n – the number of objects).
   2. Cons:
      1. Need to specify number of clusters, k.
      2. Error might vary with initial centroid assignment.
      3. Empty clusters can appear
      4. Sensitive to outliers
      5. K-means has problems when clusters are of differing sizes, densities and shapes.
2. Findings from the algorithms:
   1. It can be seen that Jaccard coefficient is less when outliers are present in the data as k-means algorithm is sensitive to outliers and assigns them to the closest centroid without treating them separately.

# Hierarchical Agglomerative clustering with Single Link (Min)

# Contents:

1. Implementation details

Algorithm

* 1. Compute the distance matrix
  2. Let each data point be a cluster
  3. Repeat
     1. Merge the two closest clusters
     2. Update the distance matrix

d. Until only a single cluster remains

1. Comparison with external index
2. Pros and Cons of HAC with single link

Pros:

1. Easy to implement and gives best result.
2. No apriori information about the number of clusters required.

Cons:

1. Algorithm can never undo what was done previously.
2. Time complexity of at least O(*n2 log n*) is required, where *‘n’* is the number of data points.
3. Sensitivity to noise and outliers.

# Density based Clustering

Contents:

* + - 1. Implementation details

Algorithm:

a. Consider each unvisited point / object in the dataset and check if the neighboring points (distance less than E) are equal to or more than min points. If so, the point has high density and mark it as a core point. This is the start point of a new cluster.

b. Else the point is an outlier, mark it as noise.

c. Now, consider neighboring points of the core point and check for high density areas using their E-neighborhood points that are more than the min points. Add these points to the cluster.

d. Keep increasing the cluster this way until there are no unvisited elements within E-neighborhood of the neighboring points.

e. Append this cluster to the array of clusters and repeat steps 1-4 for the next core point in the data until all core points are covered.

2. Comparison with external index

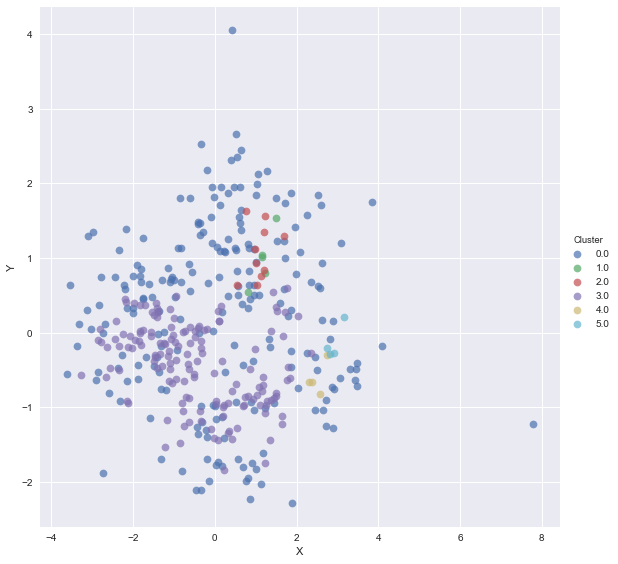
i. Dataset: cho.txt

Eps: 1.03

Min Points: 4

Rand Coefficient = 0.546148

Jaccard Coefficient = 0.20352877



Dataset: iyer.txt

Eps: 1.03

Min Points: 4

Rand Coefficient: 0.652323

Jaccard Coefficient: 0.2840689



1. Pros and cons of Density based clustering:

Pros:

Resistant to noise

Outliers have kth nearest neighbors at farther distance so it is easier to eliminate them.

Can handle clusters of different sizes and shapes

Cons:

Sensitive to parameters, hard to determine the correct set of parameters

Cannot handle varying densities

1. Findings: