Clustering

Agenda

Key Takeaways-

- Clustering and its types
- K-Means
- Evaluation measures for K
- K-Means limitations
- Overcoming K-Means limitations

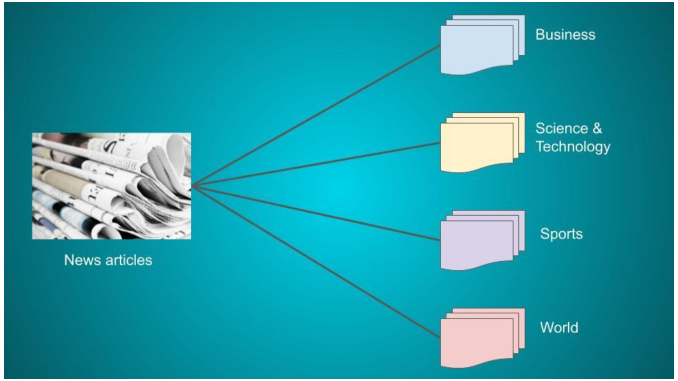
Unsupervised Learning

Unsupervised Learning has no explicit output/target variable, i.e. works without supervision.

So it discovers knowledge, hidden structures or relationship in the unlabelled data. For e.g. it can learn to group or organize data in such a way that similar objects are in the same group.

Similarly, news articles can be put together based on the topics like sports, business, technology, politics etc. This approach is known as Clustering.

Clustering is a technique to group a set of objects in such a way that objects in the same group are much more similar to each other than to those in other groups.



Clustering Use-cases

1. **Clustering based on topics**

   Text documents (such as news articles, white papers, research papers, reports etc.), images are published/available in good quantities so clustering can be used to group or visualize them together based on topics.

2. **Text Summarization**

   Summarizing the text documents to ensure good coverage and avoid redundancy (reducing the size of large documents).

3. **Anomaly Detection**

   Anomaly detection aims to find out the objects that are significantly different from others. E.g. detecting fraudulent transactions in banking and finance.

Types of Clustering Techniques

Most commonly used clustering techniques are -

1. K-Means
2. Hierarchical Clustering
   - Agglomerative
   - Divisive

K-Means

Choose the k (number of clusters).

Step 1. **Initialization :**

Randomly select k data points as initial centroids.

Step 2. **Assignment :**

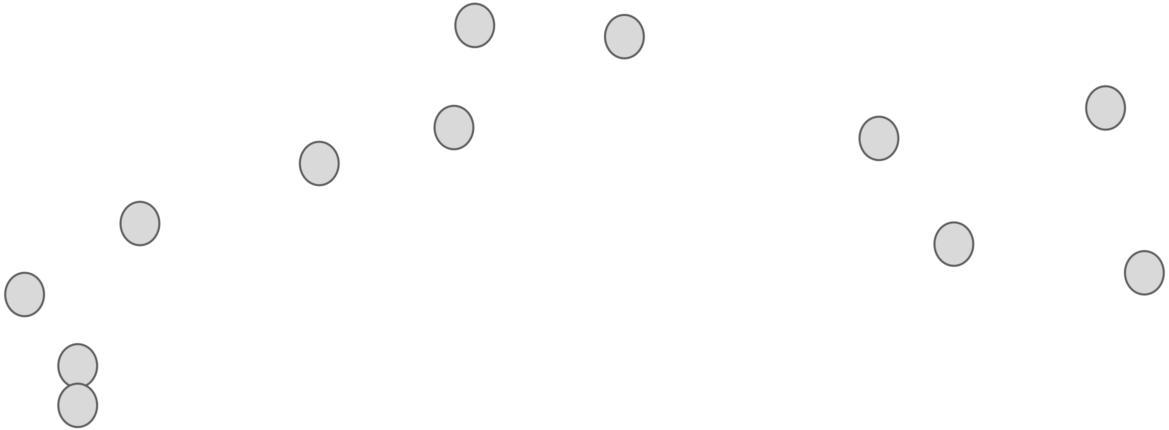
Assign each data point to the closest centroids, that forms k clusters.

Step 3. **Recompute Centroid :**

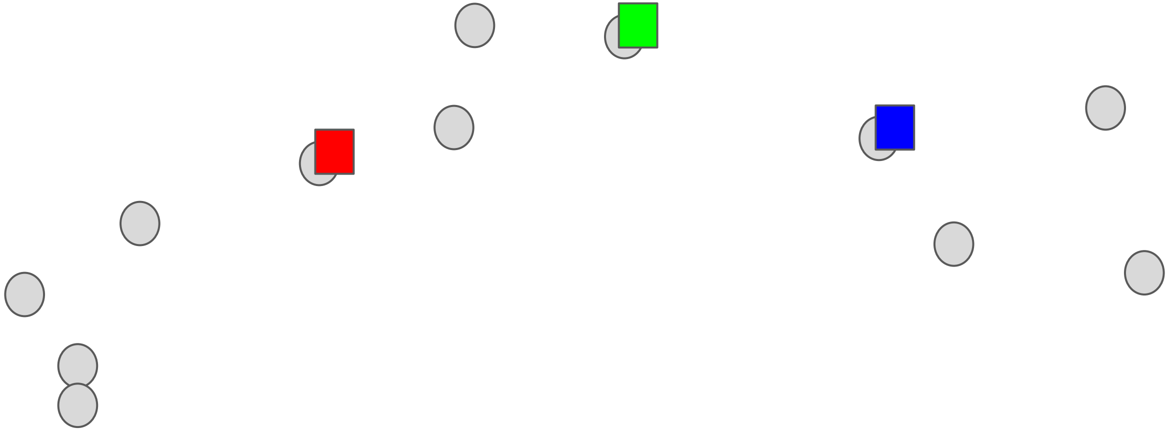
Calculate new cluster centroid for each cluster.

Step 4. Repeat Step 2 and 3 until convergence criterion is met.

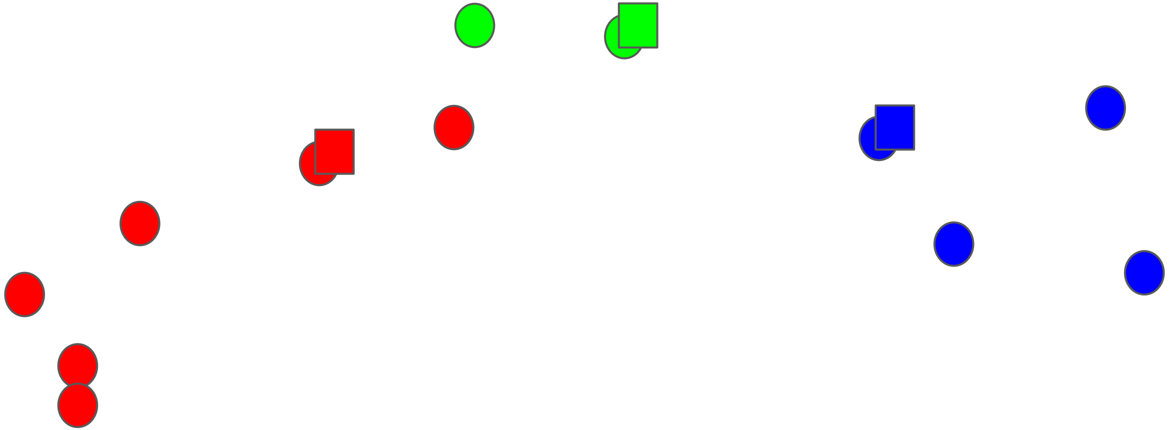
Consider the below data points.



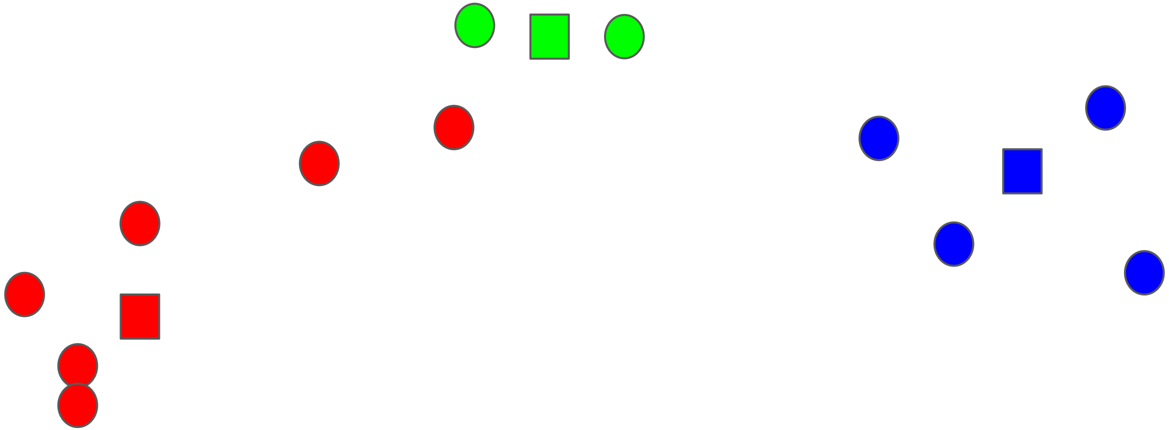
Initialize centroids randomly.



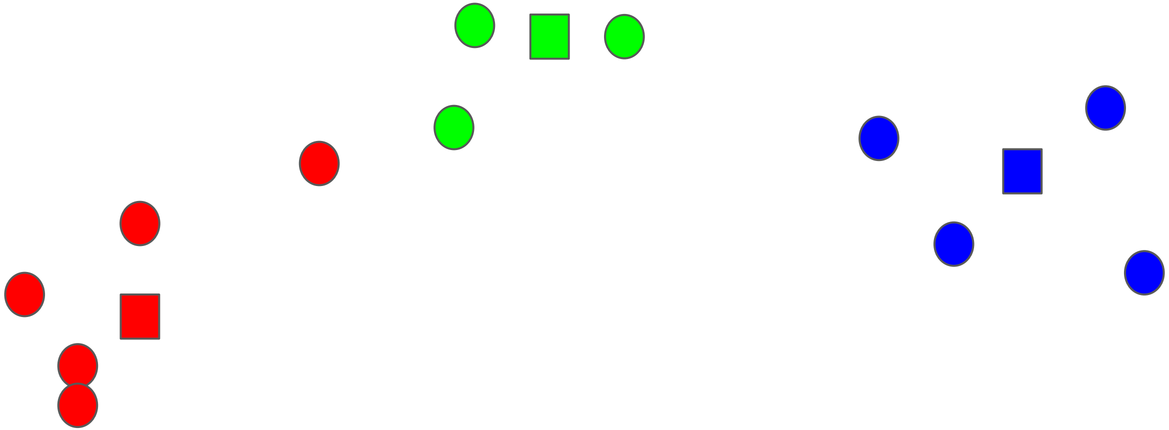
Assign data points to the closest centroid.



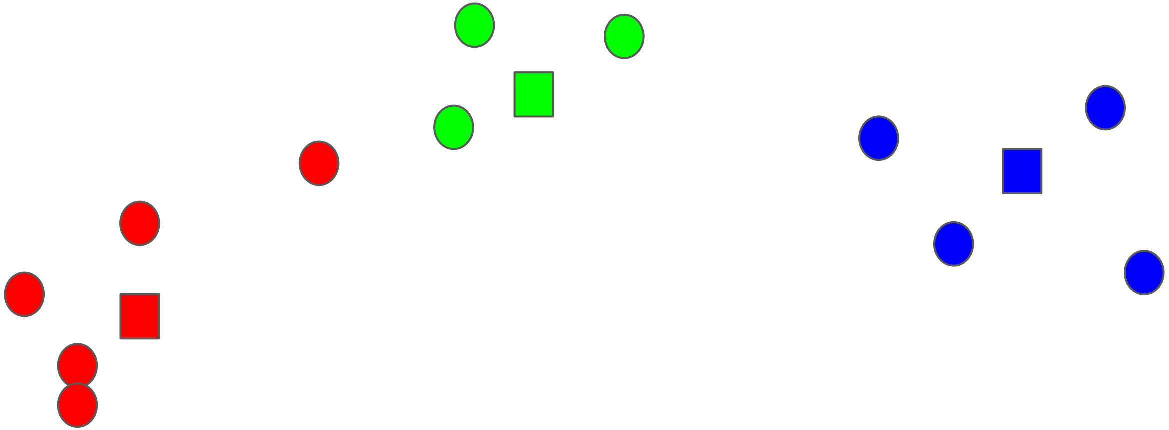
Recompute centroids.



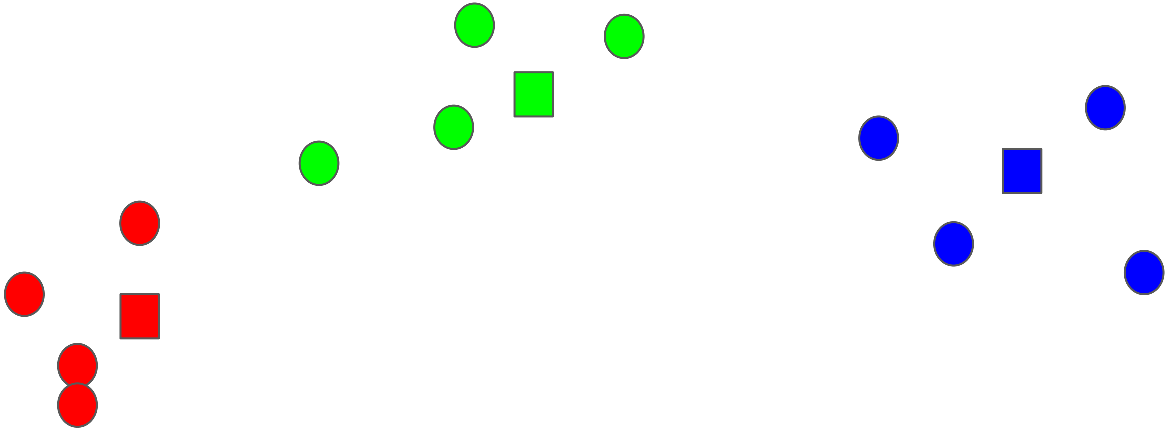
Re-assign data points to the closest centroid.



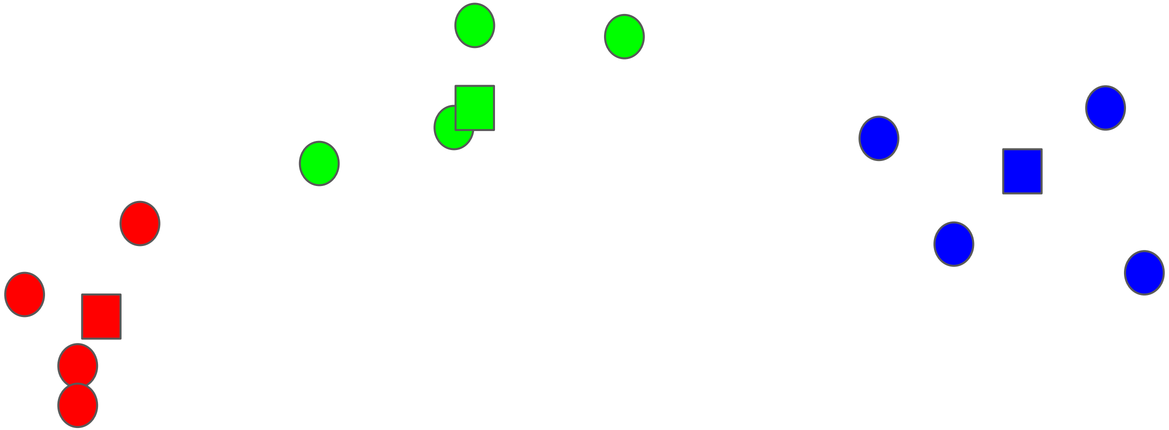
Recompute centroids.



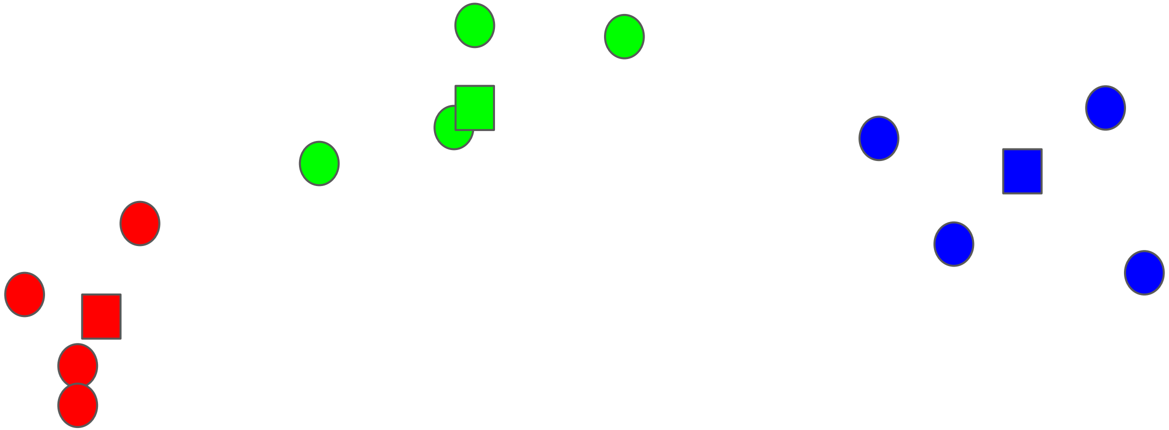
Re-assign data points to the closest centroid.



Recompute centroids.



Re-assign data points to the closest centroid.



Convergence criterion is met.

Quiz 1

Choose the correct statement(s) w.r.t K-Means clustering.

- It is often used for unlabelled data.
- It can be used to segment customers based on their past behaviour/characteristic.
- It puts two dissimilar points in same cluster.
- All of the above

Objective function

The objective of K-Means clustering is to minimize the total intra-cluster distance (squared error).

number of clusters number of cases

centroid for cluster *j*

case *i*

$$\text{objective function} \leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

Where, $$c_j = \frac{\sum_{x_i \in S_j} x_i}{|S_j|}$$

$|S_j|$ = Number of instances in cluster j.

Quiz 2

K in K-Means clustering stands for -

- Number of nearest neighbors
- Number of samples in each cluster
- Minimum distance between the clusters
- Number of clusters

Quiz 3

Which of the following can act as a termination criterion in K-Means?

- Fixed number of iterations
- Stationary centroids appear between successive iterations.
- The distance between the clusters is minimum.
- None of the above

Finding optimal value of k

- Data points clustering is a subjective decision as there is no ground truth available. Domain knowledge or better business understanding may help in getting intuition behind right number of clusters.

- Additionally, there are few methods that help in selecting optimal value of k. Most commonly used are -

  1. Elbow method
  2. Average silhouette method

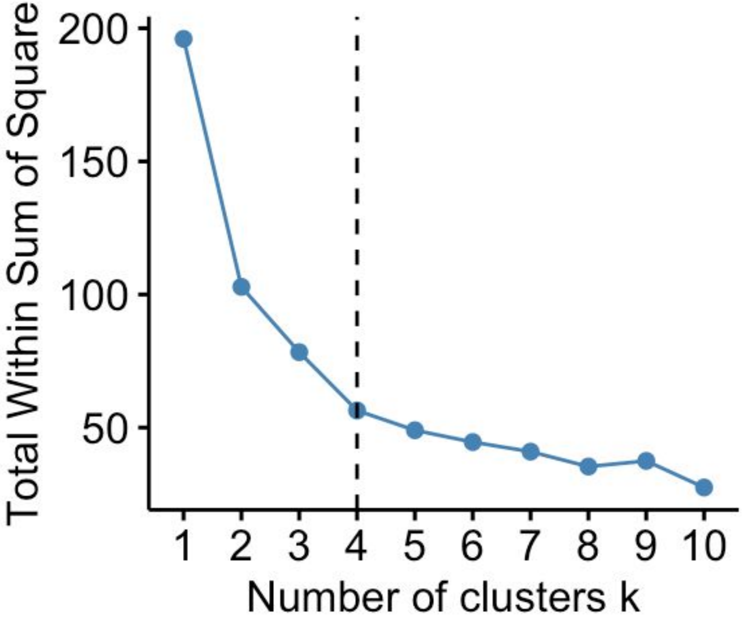
Elbow Method

- The **Elbow Method** is one of the most popular methods to determine this optimal value of k.

- It looks at the inertia for different values of k.

Inertia(or **within cluster sum of squared distance** or **intra cluster distance**) - It is the sum of squared distances of samples to their closest cluster centroid.

Step 1. Perform k-means clustering for different values of k.
Step 2. For each k, calculate the inertia.
Step 3. Plot the curve of inertia according to the number of clusters k.
Step 4. Choose the k where interia stops decreasing abruptly.



NOTE - As k increases, the inertia tends towards zero.

Average Silhouette Method

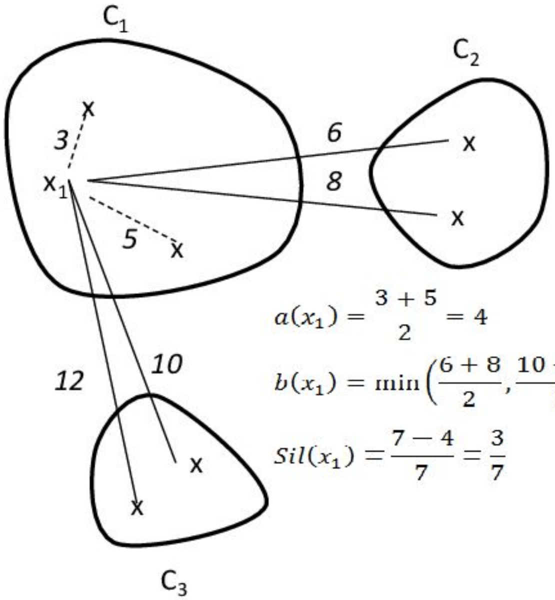
- Silhouette method is used to determine the degree of separation between clusters.

For each sample, it computes -

- mean intra-cluster distance (a) : average distance from all data points in the same cluster.
- mean nearest-cluster distance (b) : average distance from all data points in the closest cluster.
- Compute the Silhouette coefficient :

$$\frac{(b^i - a^i)}{max\ (\ b^i, a^i\ )}$$

Like this, mean Silhouette Coefficient over all samples is calculated.



$$a(x_1) = \frac{3+5}{2} = 4$$

$$b(x_1) = \min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$$

$$Sil(x_1) = \frac{7-4}{7} = \frac{3}{7}$$

Average Silhouette Method [Contd.]

- The average Silhouette coefficient ranges from -1 to 1.
- If value = 0 -> the sample is very close to the neighboring clusters.
- If value = 1 -> the sample is far away from the neighboring clusters.
- If value = -1 -> the sample is assigned to the wrong clusters.

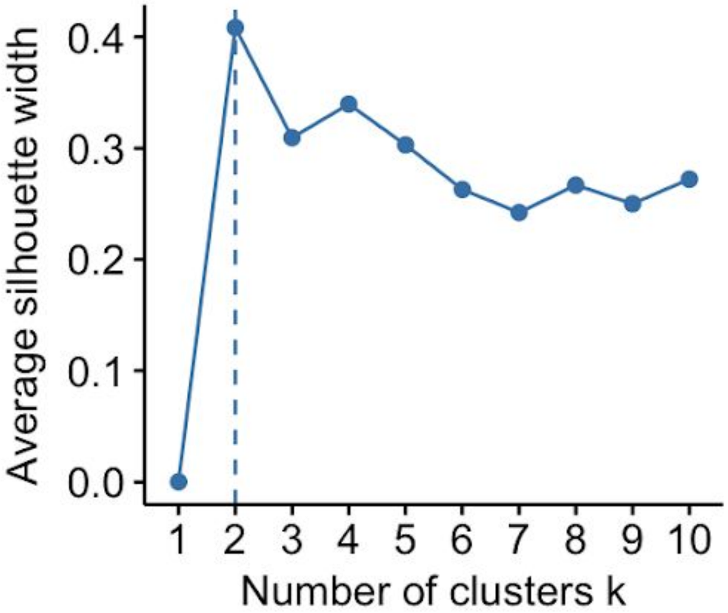
Finding optimal value of k using average Silhouette method

Step 1. Perform k-means clustering for different values of k.

Step 2. For each k, calculate the average Silhouette coefficient.

Step 3. Plot the curve of average Silhouette coefficient according to the number of clusters k.

Step 4. The location of the maximum is considered as the appropriate number of clusters.



K-Means Time and Space Complexity

Time Complexity : O(nkdi)

Where n = number of data points

k = number of clusters

d = dimensionality of data

i = number of iterations

- Generally k ≤ 10 and If d ≤ 300 then O(nd)
- But if n and d are large then it is time consuming.

Space Complexity : O(nd + kd)

Where nd to store **n** d-dimensional data points,

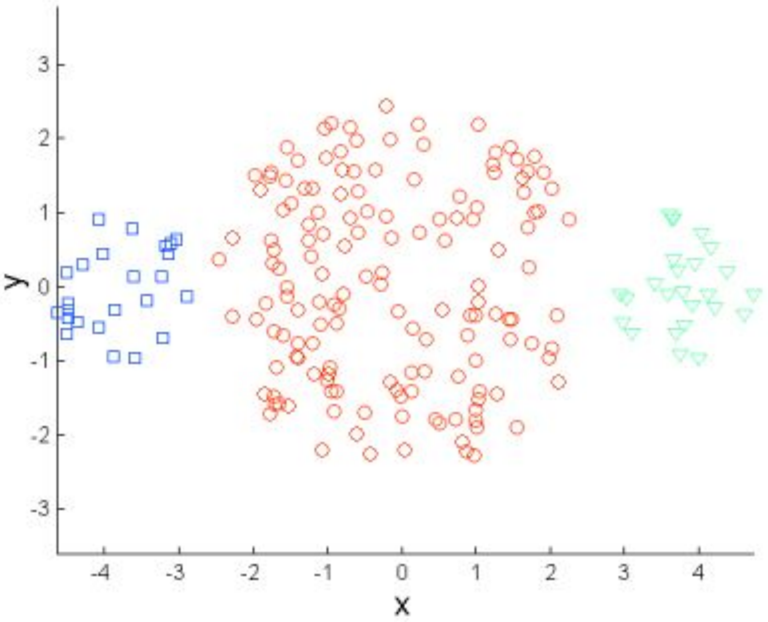
kd to store **k** d-dimensional centroids.

K-Means Limitations

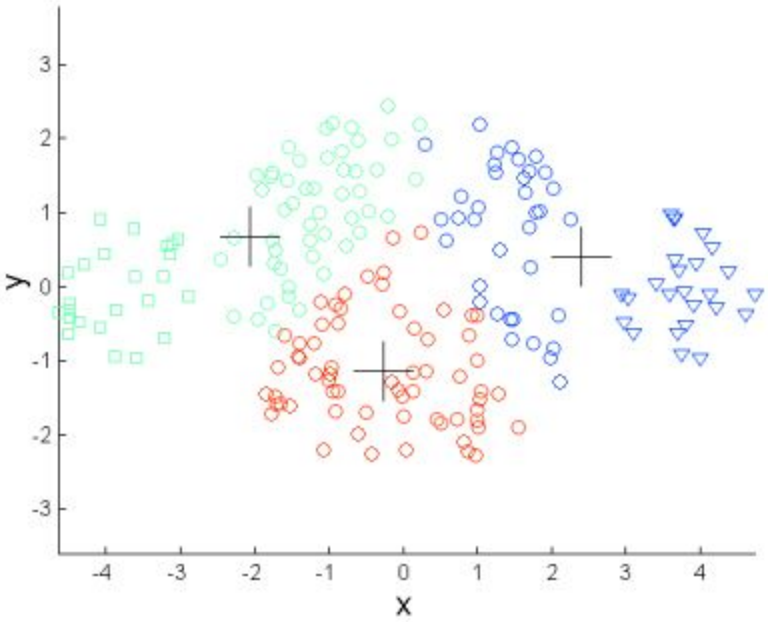
- K-Means has problems when clusters are of different
  - Sizes
  - Densities
  - Non-globular shapes

- K-Means has problems when the data contains outliers.

K-Means Limitations - Differing Sizes

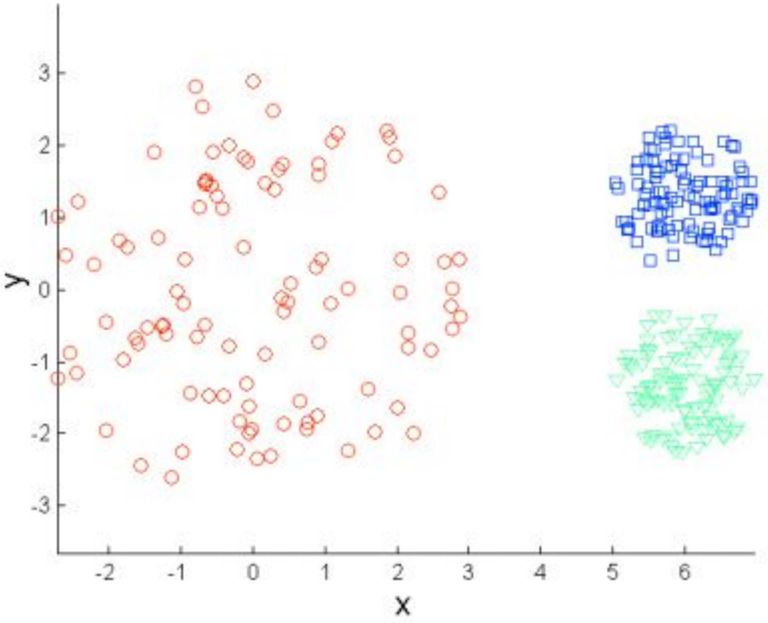


Original Points

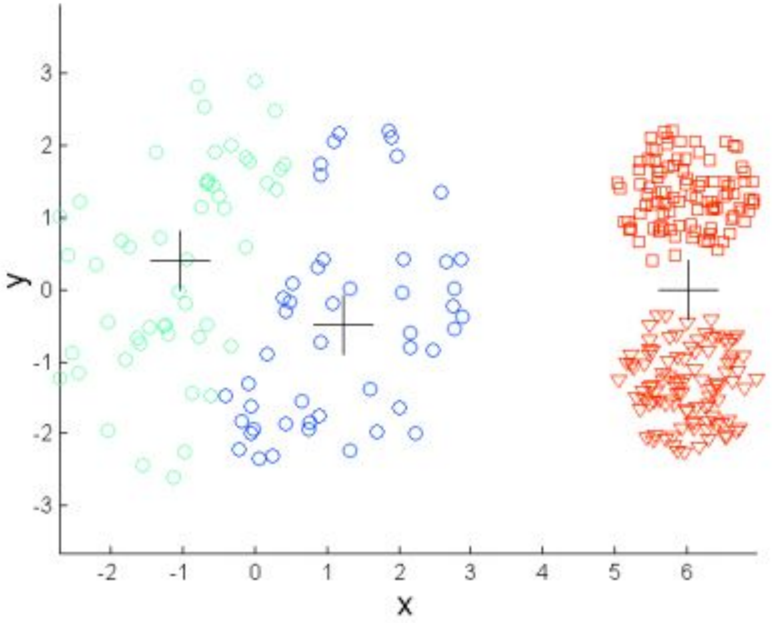


K-Means Clusters

K-Means Limitations - Differing Densities

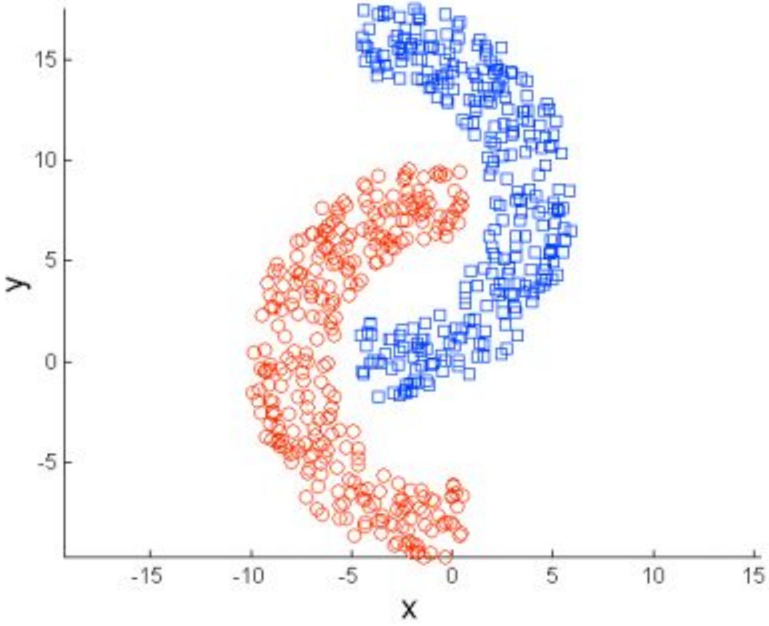


Original Points

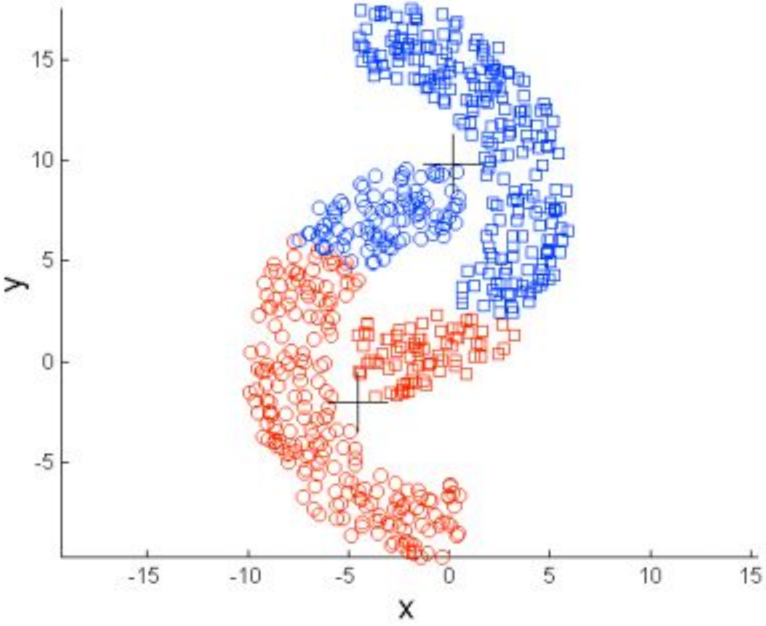


K-Means Clusters

K-Means Limitations - Non-globular shapes

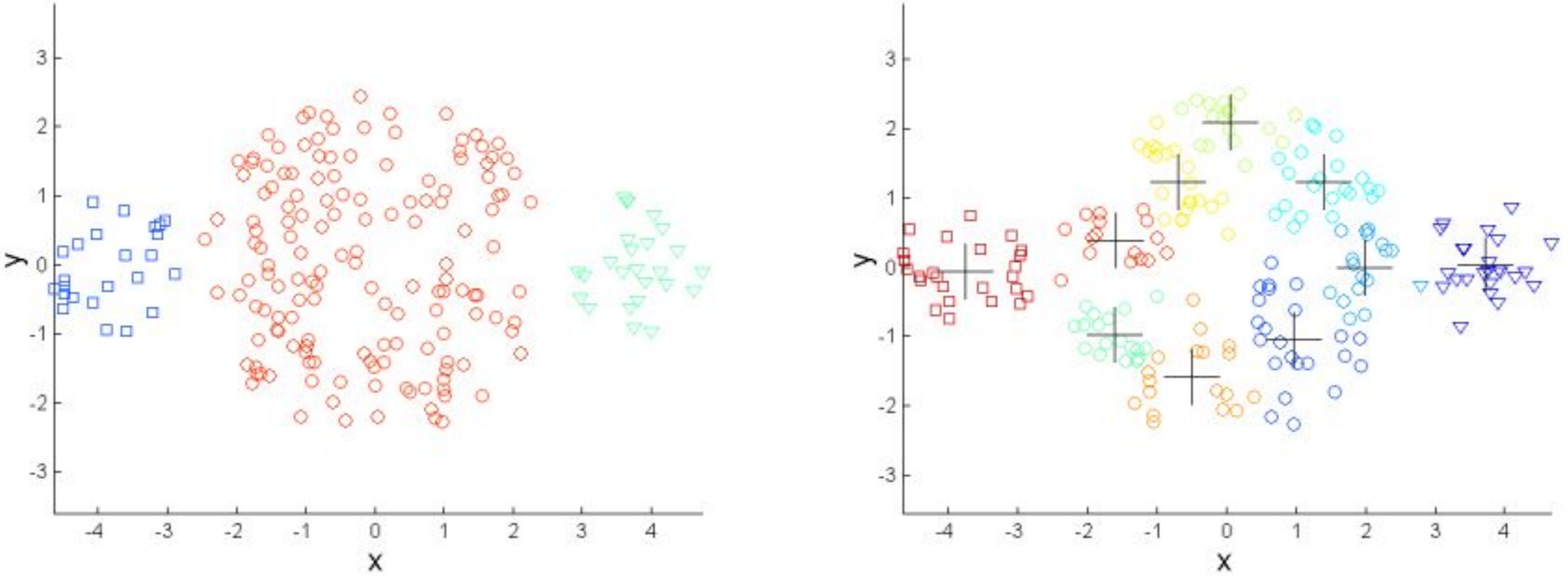


Original Points



K-Means Clusters

Overcoming K-Means Limitation - Differing Sizes



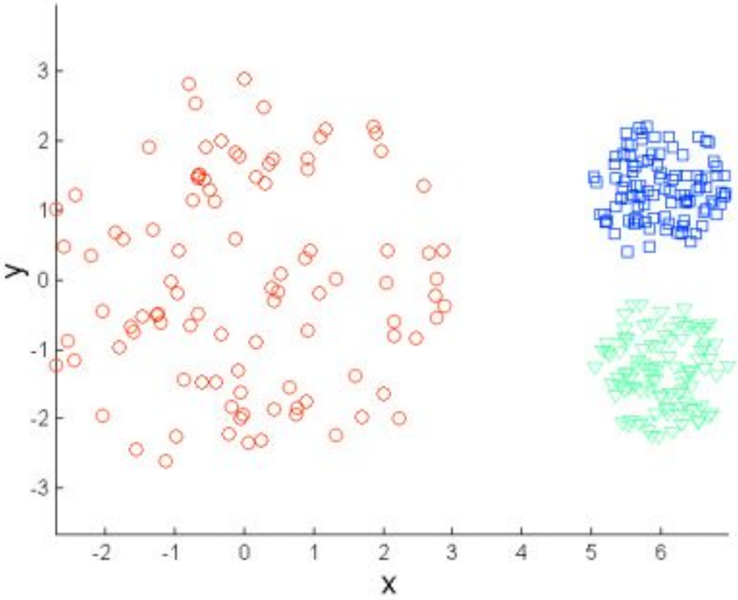
Original Points

K-Means Clusters

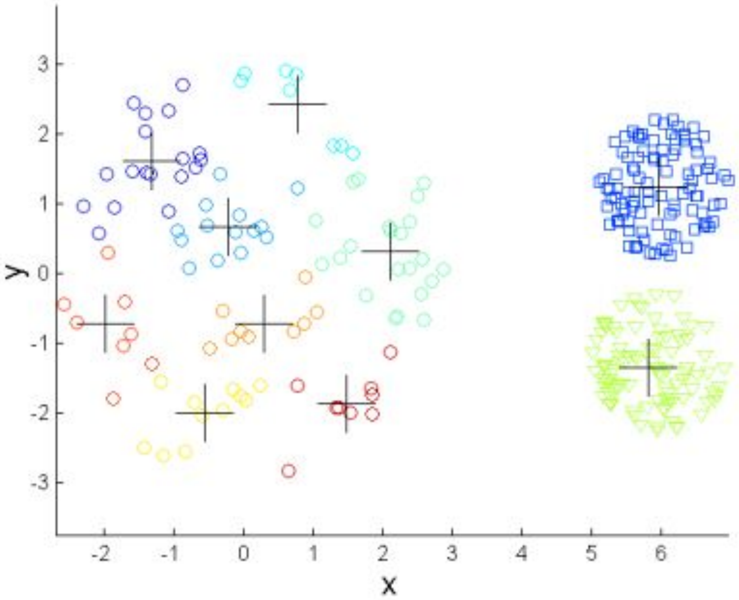
- One solution is trying larger value of k (say 10) and then group smaller clusters into larger.

NOTE - It’s not a perfect solution but works in some cases.

Overcoming K-Means Limitation - Differing Densities



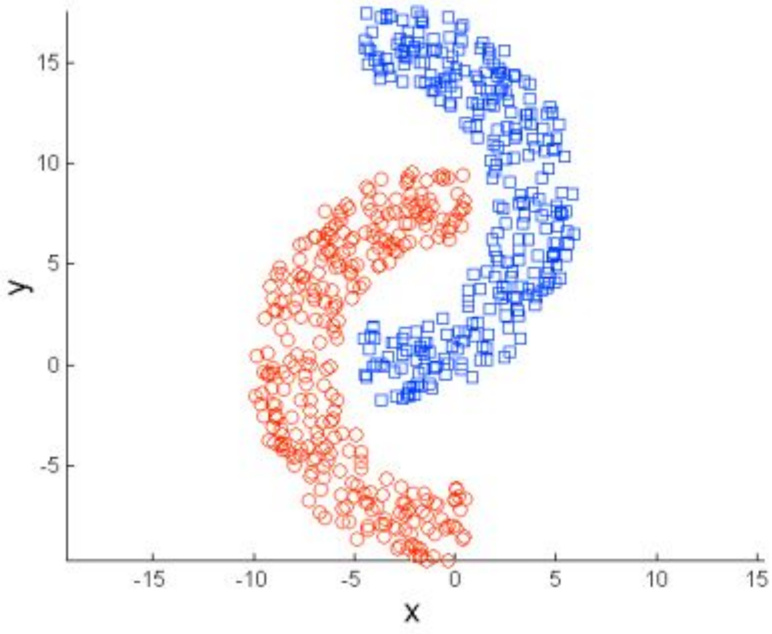
Original Points



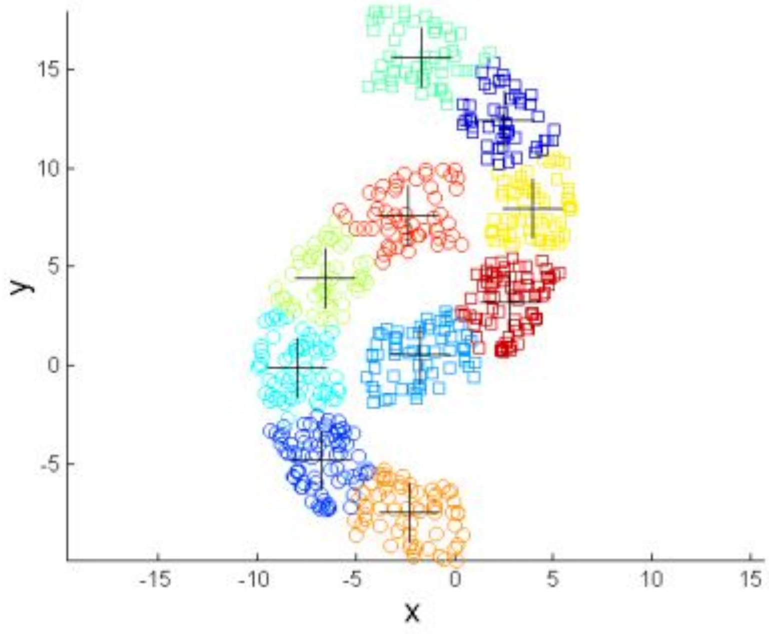
K-Means Clusters

- One solution is trying larger value of k (say 10) and then group smaller clusters into larger.

Overcoming K-Means Limitation - Non-globular shapes



Original Points



K-Means Clusters

- One solution is trying larger value of k (say 10) and then group smaller clusters into larger.

K-Means Variations

K-Medoid

Similar to K-Means but the centroid of the cluster is defined to be one of the actual data points in the cluster (the medoid).

Hierarchical Clustering

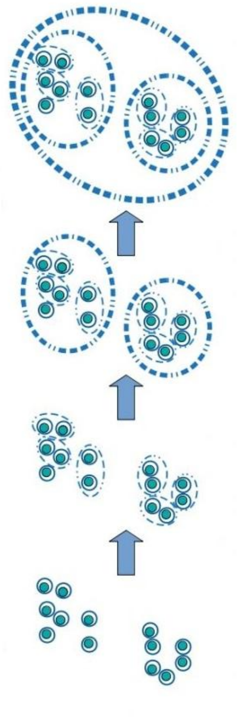
- As the name suggests, hierarchical clustering maintains a well defined hierarchy while forming/dealing with clusters.

- The clustering process relies on a similarity/distance matrix for making decisions.

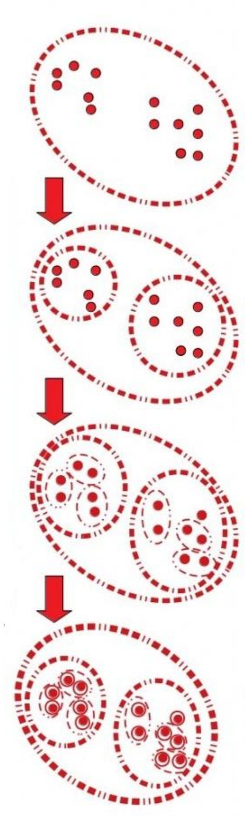
- The algorithm deals with two clusters at a time. Based on the similarity/distance matrix, it decides which two clusters need to be merged or how to divide a cluster into two.

- There are two types of hierarchical clustering based on the way of developing hierarchy.
  1. **Agglomerative** : A bottom-up hierarchical clustering
  2. **Divisive** : A top-down hierarchical clustering

Agglomerative vs. Divisive



Agglomerative



Divisive

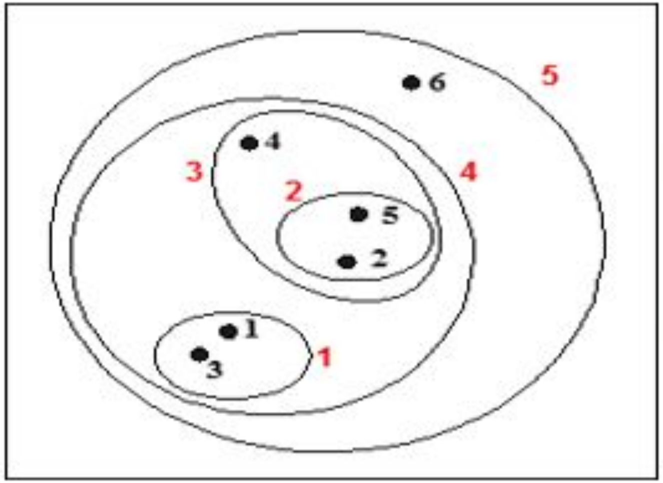
Agglomerative Clustering

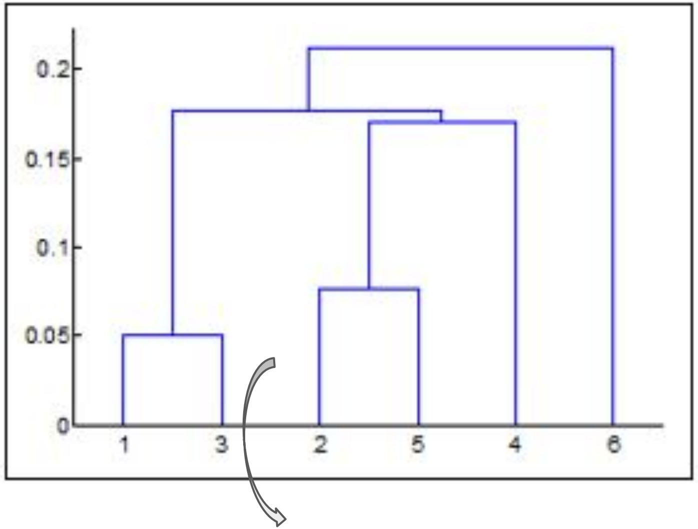
Step 1. Initially treat each data point as one cluster. Therefore initially, # of clusters = n (size of data).

Step 2. Form a cluster by joining two closest data points/clusters.

Step 3. Repeat step 2 until one big cluster is formed.

Agglomerative Clustering [Contd.]





Dendrogram

- Dendrogram = Dendro + gramma. In Greek, dandro means tree and gramma means drawing.

- So dendrogram is a tree diagram/hierarchy which keep track of the order/sequence of merging.

Linkage Mechanisms

- The distance between the clusters (or point to cluster) is also known as linkage.

- There are multiple ways to compute this linkage(distance).

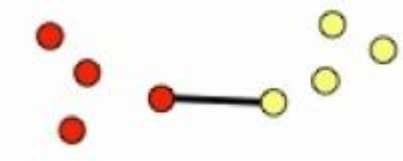
1. Single or MIN Linkage
2. Complete or MAX Linkage
3. Average or Group Linkage
4. Centroid Linkage
5. Ward Linkage

Linkage Mechanisms [Contd.]

1. **Single or MIN Linkage**

   The distance between two clusters is the smallest distance between two points from cluster C1 and C2.

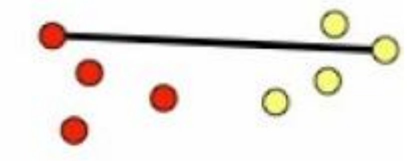
$$d_{12} = min_{i,j}\,(d\,(C1_i,\; C2_j))$$



2. **Complete or MAX Linkage**

   The distance between two clusters is the longest distance between two points from cluster C1 and C2.

$$d_{12} = max_{i,j}\,(d\,(C1_i,\; C2_j))$$

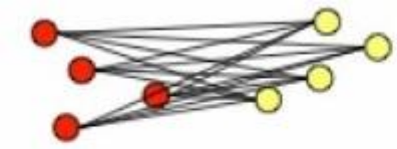


Linkage Mechanisms [Contd.]

3. Average or Group Linkage

The distance between clusters is the average distance between each point in one cluster to every point in other cluster.

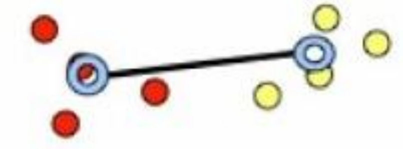
$$d_{12} = mean_{i,j}\,(d\,(C1_i,\ C2_j))$$



4. Centroid Linkage

The distance between two clusters is the distance between centroid of cluster C1 and C2.

$$d_{12} = d\,(\overline{C1},\ \overline{C2})$$



5. Ward’s Linkage

Exactly the same as Group Average except that it calculates the sum of the square of the distances.

Quiz 4

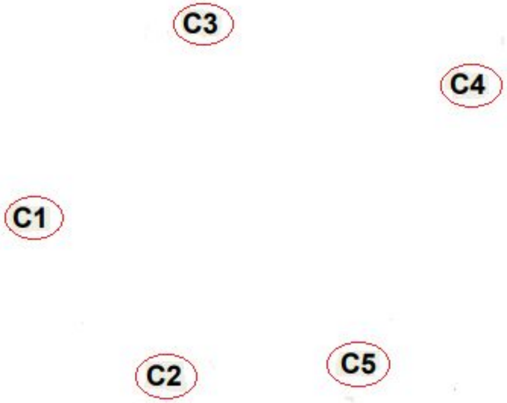
Choose the correct statement(s).

- Complete linkage is the smallest distance between two points from cluster C1 and C2.
- Average linkage is the median distance between each point in one cluster to every point in other cluster.
- Centroid linkage is distance between centroid of cluster C1 and C2.
- All the above

Agglomerative Clustering using Distance/Proximity matrix

Step 1.

- Start with clusters of individual points and compute the proximity/distance matrix of data.

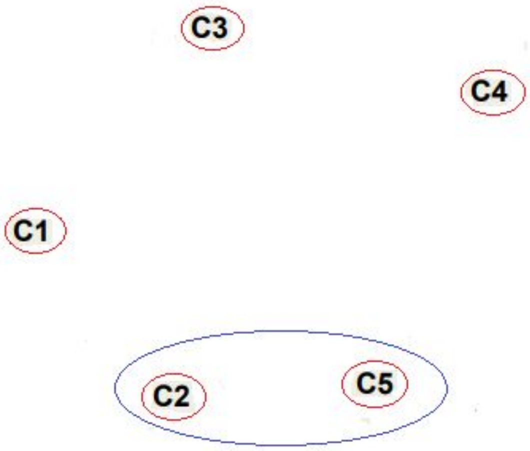


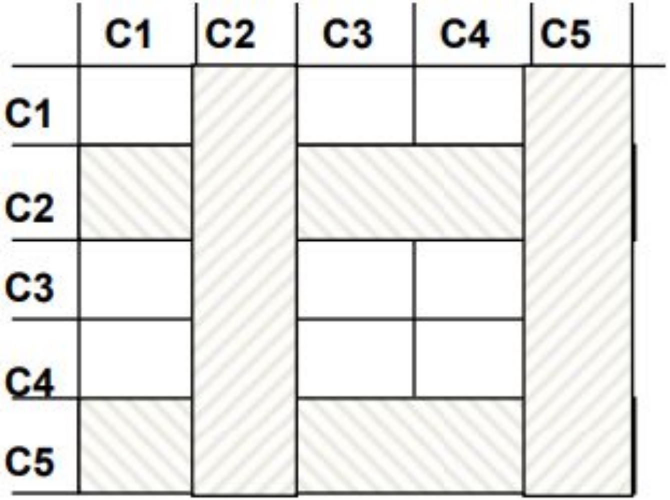
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Agglomerative Clustering using Distance/Proximity matrix[Contd.]

Step 2 (a).

- Merge the two closest clusters. (Consider the off diagonal elements and merge the two clusters with minimum value in the matrix)





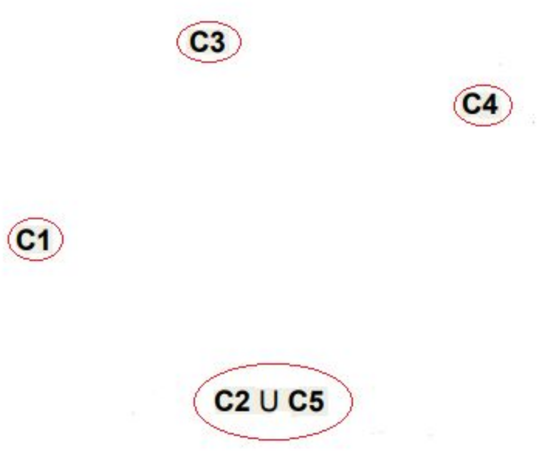
	C1	C2	C3	C4	C5
C1		▨			▨
C2	▨	▨	▨	▨	▨
C3		▨			▨
C4		▨			▨
C5	▨	▨	▨	▨	▨

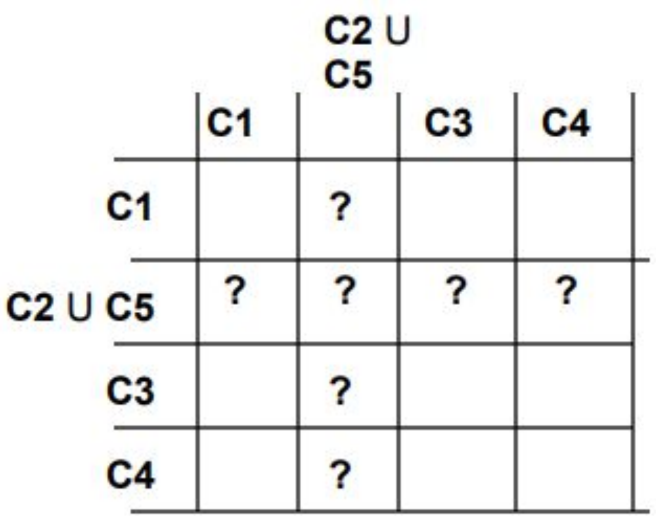
- Assume **C2** and **C5** got merged.

Agglomerative Clustering using Distance/Proximity matrix[Contd.]

Step 2 (b).

- Update the proximity matrix after merging.





	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Step 3. Repeat Step 2 (a) and 2 (b) until a single cluster remains.

Comparison with K-Means

- In K-Means we need to pass the k(number of clusters) upfront whereas there is no need to mention number of clusters in agglomerative clustering.

- Therefore, once the agglomerative clustering is trained we can move from one level to another without any recomputation, whereas in K-Means, for any different value of k, retraining is required.

Quiz 5

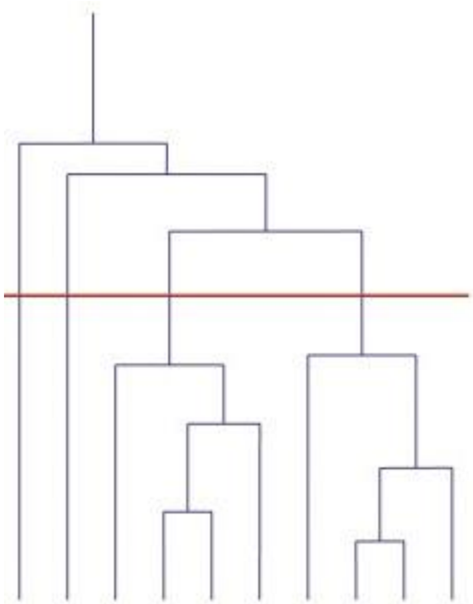
Choose the correct statement(s) about clustering techniques.

- Agglomerative clustering never make use of centroid concept.
- K-Means is sensitive to centroid initialization.
- Agglomerative is bottom-up approach whereas divisive is a top-down approach.
- All the above.

Finding Optimal number of clusters

NOTE - The below approach is not a clear-cut solution as it may not work in many cases.

Step 1. Determine the longest vertical line through which no any extending horizontal line passes.
Step 2. Draw a new horizontal line at both extremities.
Step 3. The optimal number of clusters is equal to number of intersection points on this new horizontal line.



Agglomerative Time and Space Complexity

Time Complexity : O(n^2*n) = O(n^3)

Where n^2 to update proximity(distance) matrix in each iteration

n = almost n iterations

- With better calculation O(n^2 logn) but still it is expensive.

Space Complexity : O(n^2)

Where n^2 to store proximity(distance) matrix

- If n = 1M points, then O(n^2) = 10^12 ≅ 1TB and this need to be stored in RAM.

So if **n** is too large then it is not much helpful because of space and time complexity.