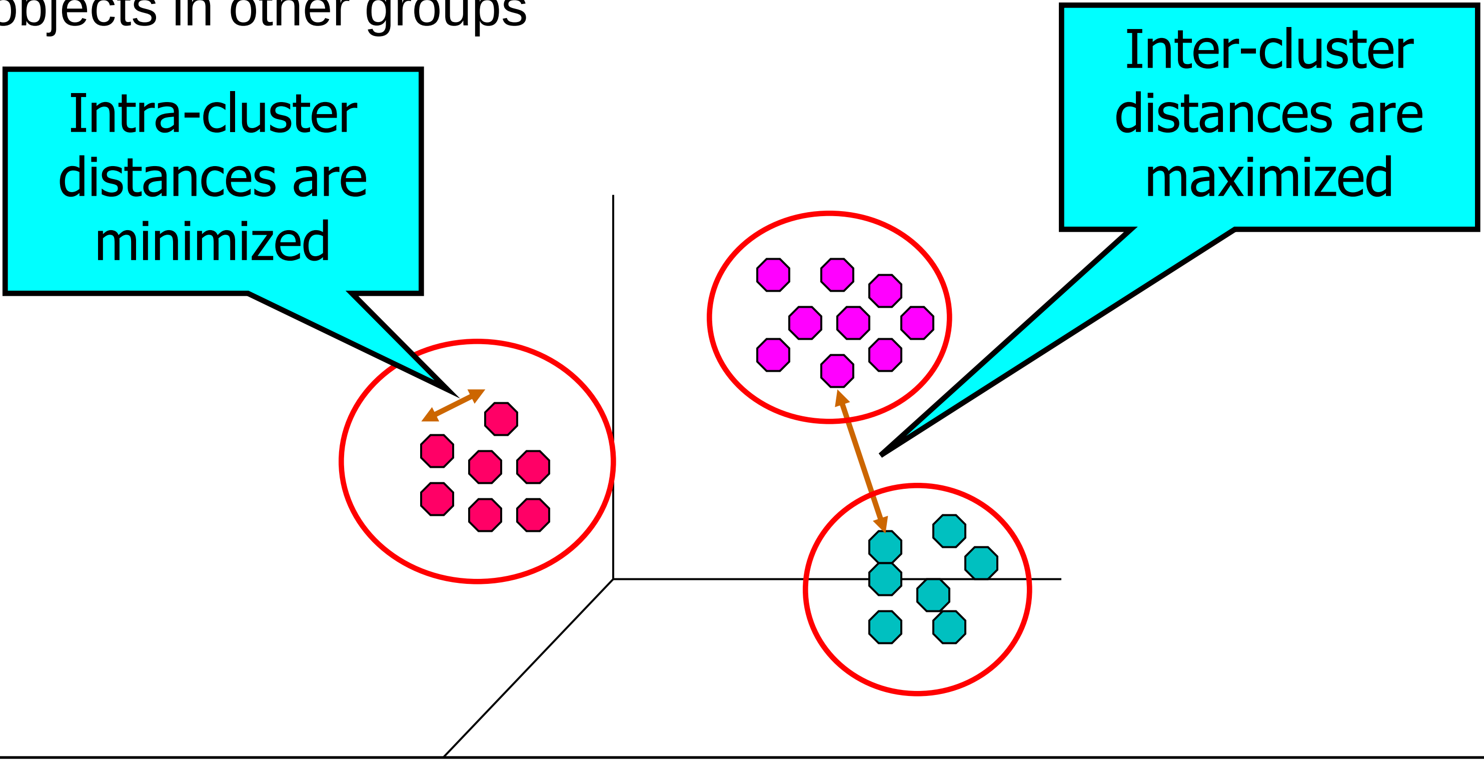
UNIT-5
CLUSTERING

What is Cluster Analysis?

- Cluster: a collection of data objects
  - **Similar to one another within the same cluster**
  - **Dissimilar to the objects in other clusters**
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Intra-cluster distances are minimized

Inter-cluster distances are maximized

General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

Types of Clustering

- Partitional Clustering
  - This method divides a given database of n objects or data tuples into m partitions such that m<=n, where each partition is a cluster.

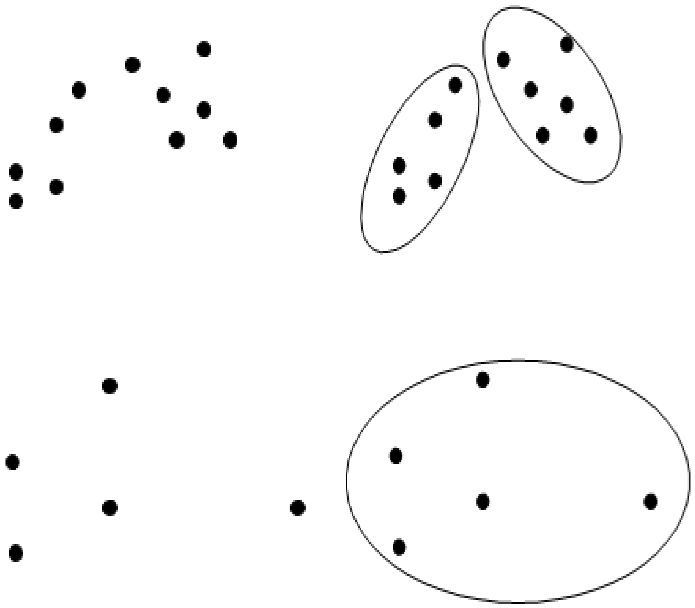
Methods:

- **K-means**:

  Each cluster is represented by the center of the cluster

- **K-mediods:**

  Each cluster is represented by one of the objects in the cluster



Original Points

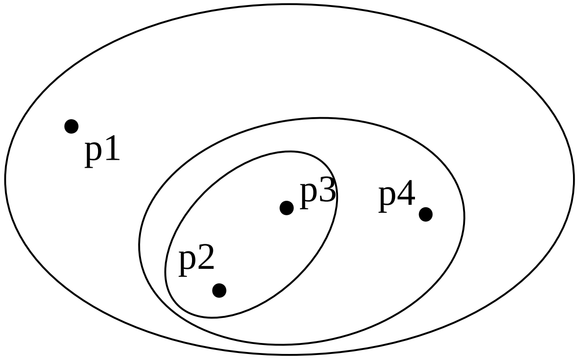
A Partitional Clustering

Types of Clustering

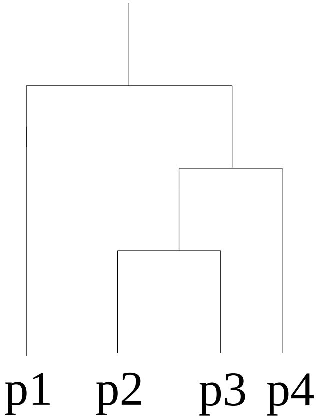
- Hierarchical Clustering
  - A Hierarchical clustering method works by grouping data objects into a tree of clusters.

Methods:

- Agglomerative Clustering(Bottom-up)
- Divisive Clustering(Top-Down)



Hierarchical Clustering



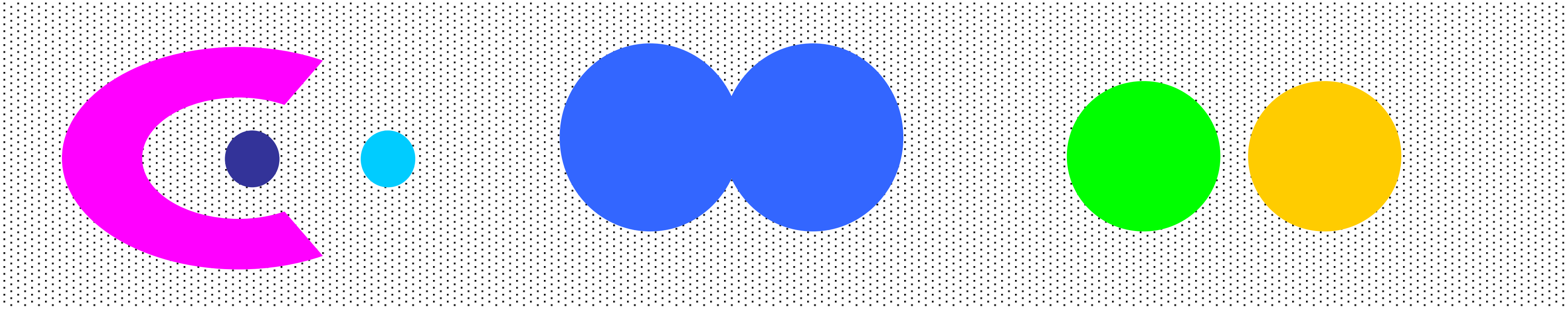
Dendrogram

Types of Clustering

- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.

  Methods:

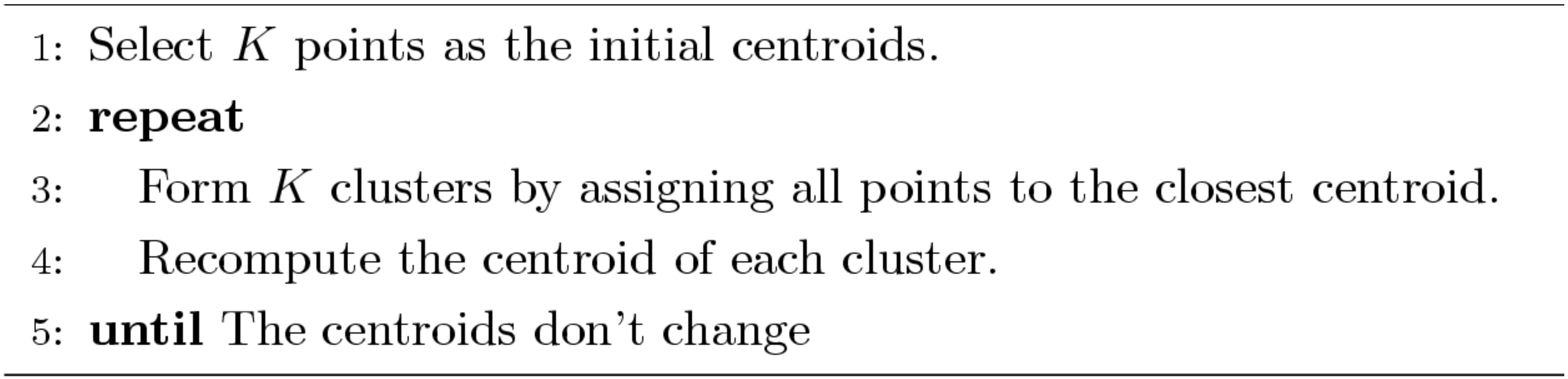
  - DBSCAN
  - DENCLUE
  - OPTICS



Density-based clusters

K-means Clustering(Centroid based technique)

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple



```
Select $K$ points as the initial centroids.
repeat
    Form $K$ clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

K-means Clustering(Example)

Consider a triangular space containing a group of numbered objects. Let, k=2 be the numbers of clusters desired by the user. The algorithm proceeds as follows,

We randomly choose two objects as initial cluster centers marked as ✡ in the figures with mean values as $m_1=2$ and $m_2=4$ and find the Euclidean distance between the mean and the objects to classify objects into two clusters as shown in Fig. 7.4.1.

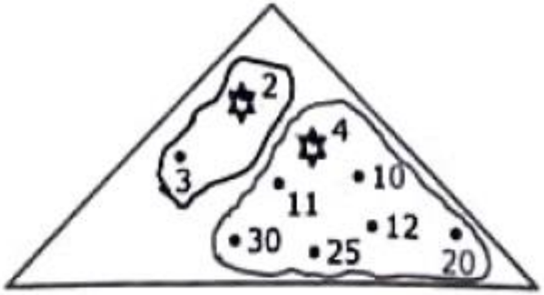


Fig. 7.4.1 Initial Partitioning with $m_1 = 2$ and $m_2 = 4$

K-means Clustering(Example)

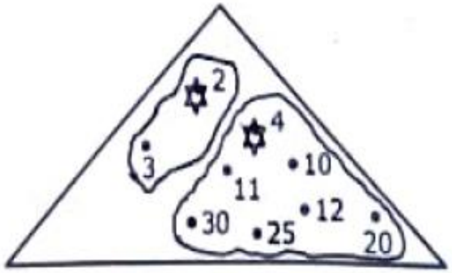
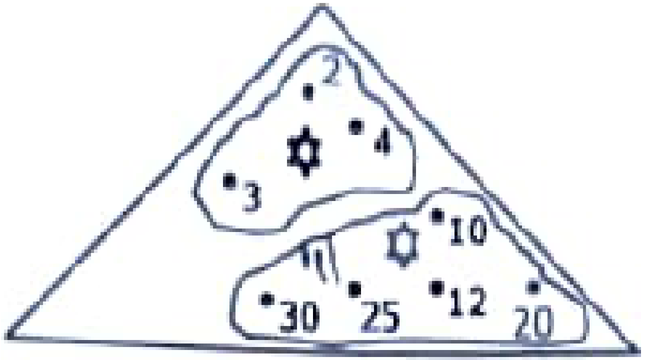


Fig. 7.4.1 Initial Partitioning with $m_1 = 2$ and $m_2 = 4$

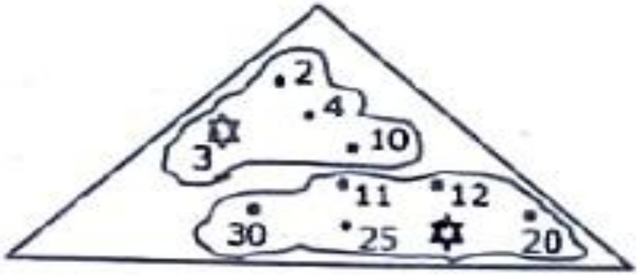
$$m_1 = \frac{2 + 3}{2} = 2.5$$

$$m_2 = \frac{4 + 10 + 11 + 12 + 20 + 25 + 30}{7} = 16$$

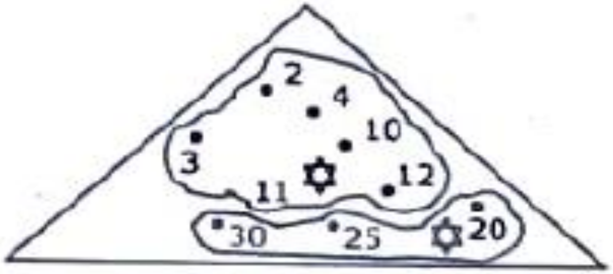
New Clusters Formed with $m_1 = 2.5$ and $m_2 = 16$



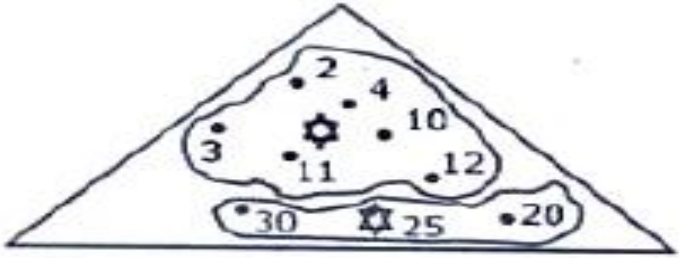
Clusters with $m_1 = 3$ and $m_2 = 18$



Clusters with $m_1 = 4.75$ and $m_2 = 19.6$



Final Clusters are Returned by the Clustering Process with $m_1 = 7$ and $m_2 = 25$



Evaluation of *K-means*

- Strength
  - **Relatively efficient**: O(tkn), where

    n is # objects,

    k is # clusters, and

    t is # iterations.

    Normally, k, t << n.
- Weakness
  - Applicable only when mean is defined, then what about categorical data?
  - Need to specify k, the number of clusters, in advance
  - Unable to handle noisy data and outliers
  - Not suitable for clusters with non-convex shapes

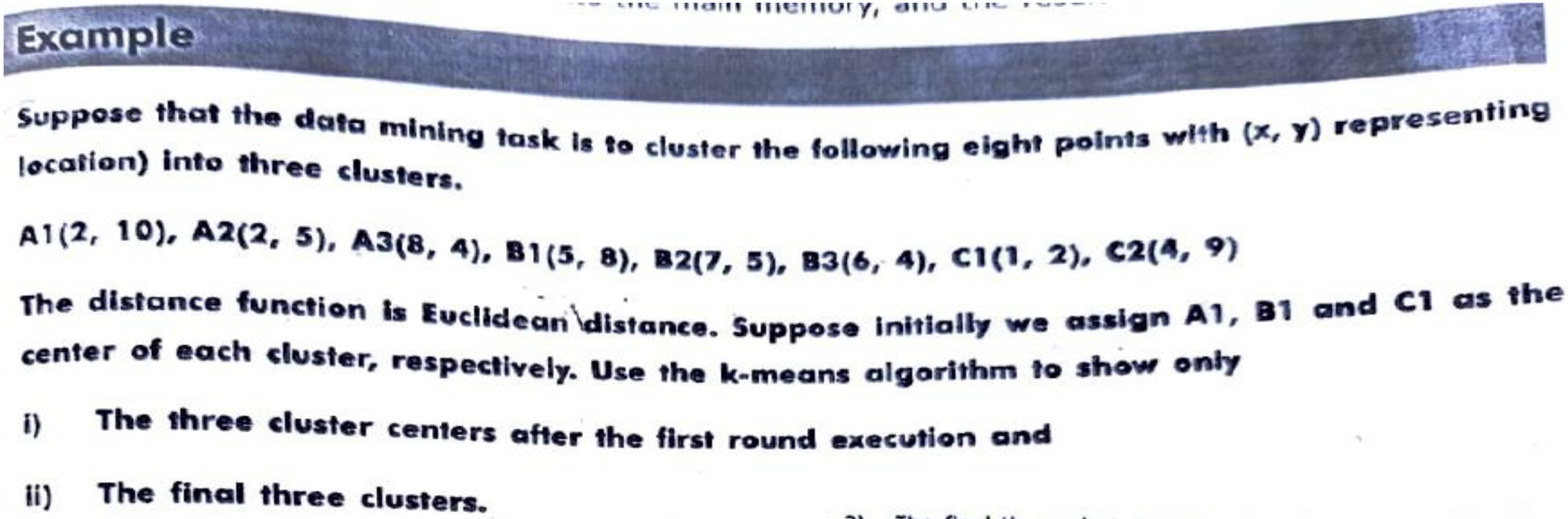
Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - *x* is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - ◆ can show that m_i corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
  - One easy way to reduce SSE is to increase K, the number of clusters
    - ◆ A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

K-means Clustering(Example)



the main memory, and the

Example

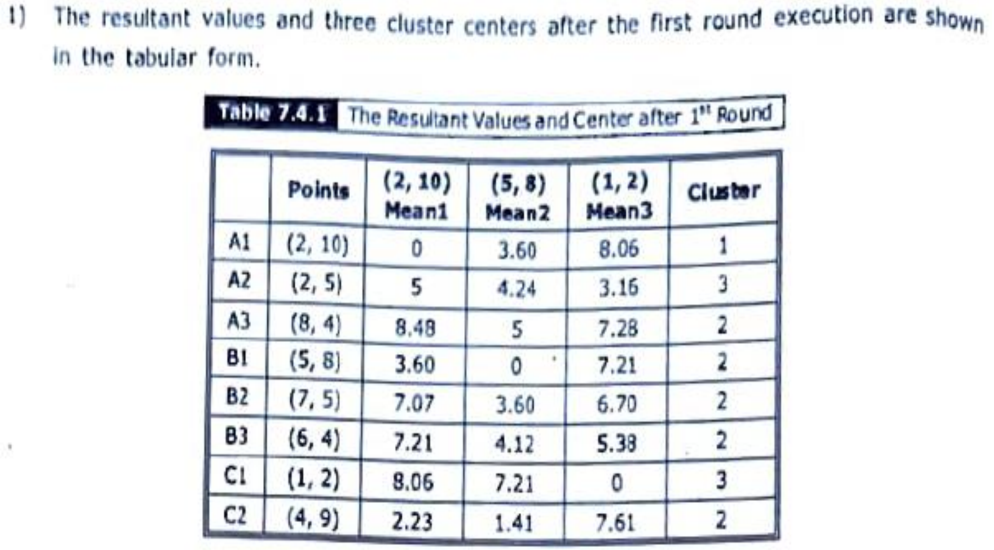
Suppose that the data mining task is to cluster the following eight points with (x, y) representing location) into three clusters.

A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9)

The distance function is Euclidean distance. Suppose initially we assign A1, B1 and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only

i) The three cluster centers after the first round execution and

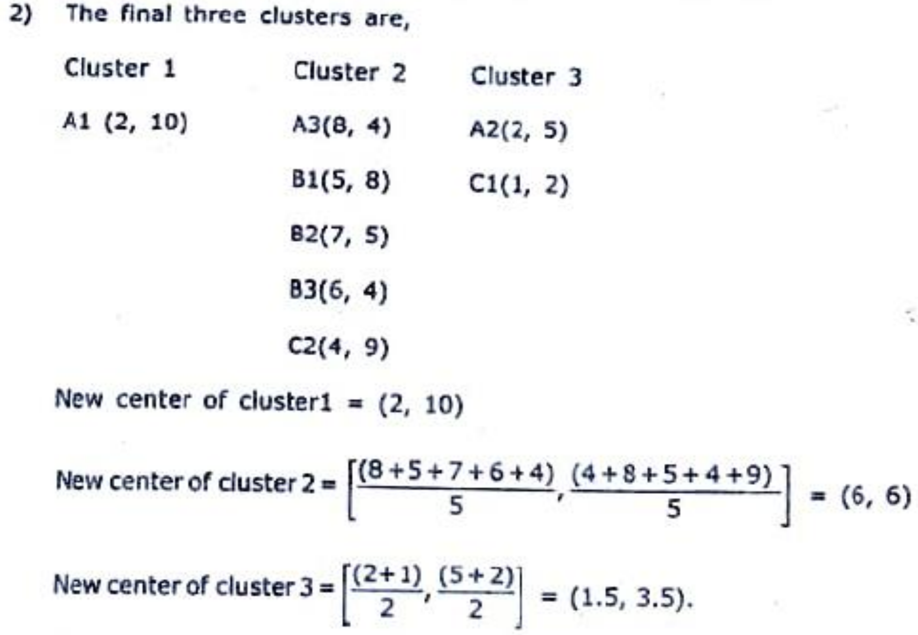
ii) The final three clusters.



1) The resultant values and three cluster centers after the first round execution are shown in the tabular form.

Table 7.4.1 The Resultant Values and Center after 1st Round

	Points	(2, 10) Mean1	(5, 8) Mean2	(1, 2) Mean3	Cluster
A1	(2, 10)	0	3.60	8.06	1
A2	(2, 5)	5	4.24	3.16	3
A3	(8, 4)	8.48	5	7.28	2
B1	(5, 8)	3.60	0	7.21	2
B2	(7, 5)	7.07	3.60	6.70	2
B3	(6, 4)	7.21	4.12	5.38	2
C1	(1, 2)	8.06	7.21	0	3
C2	(4, 9)	2.23	1.41	7.61	2



2) The final three clusters are,

Cluster 1	Cluster 2	Cluster 3
A1 (2, 10)	A3(8, 4)	A2(2, 5)
	B1(5, 8)	C1(1, 2)
	B2(7, 5)	
	B3(6, 4)	
	C2(4, 9)	

New center of cluster1 = (2, 10)

New center of cluster 2 = $\left[\frac{(8+5+7+6+4)}{5}, \frac{(4+8+5+4+9)}{5}\right]$ = (6, 6)

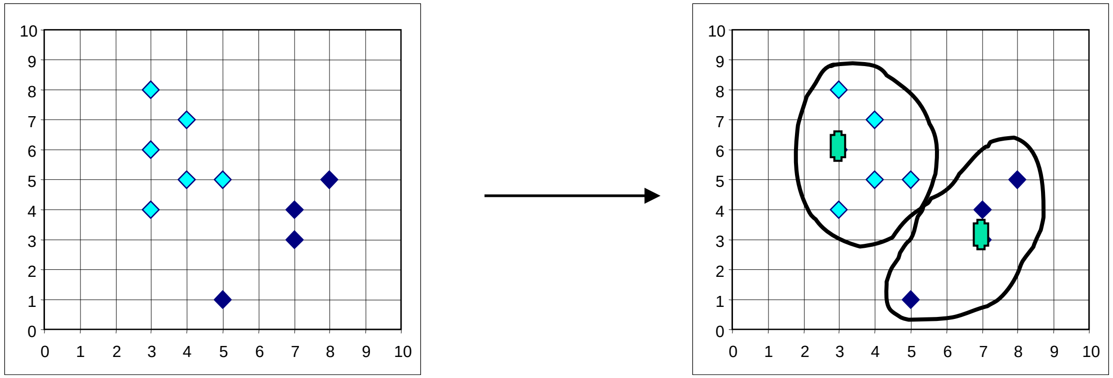
New center of cluster 3 = $\left[\frac{(2+1)}{2}, \frac{(5+2)}{2}\right]$ = (1.5, 3.5).

Variations of the *K-Means* Method

- A few variants of the *k-means* which differ in
  - Selection of the initial *k* means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- Handling categorical data: *k-modes*
  - Replacing means of clusters with modes
  - Using new dissimilarity measures to deal with categorical objects
  - Using a frequency-based method to update modes of clusters
  - A mixture of categorical and numerical data: *k-prototype* method

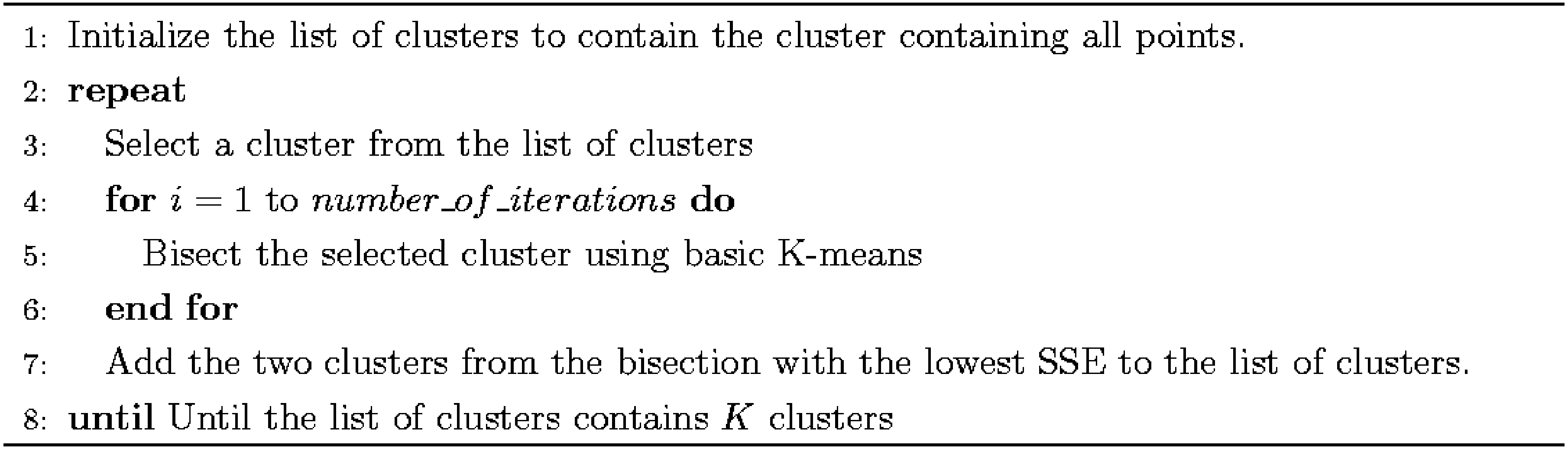
What is the problem of k-Means Method?

- The k-means algorithm is sensitive to outliers !
  - Since an object with an extremely large value may substantially distort the distribution of the data.
- K-Medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster.



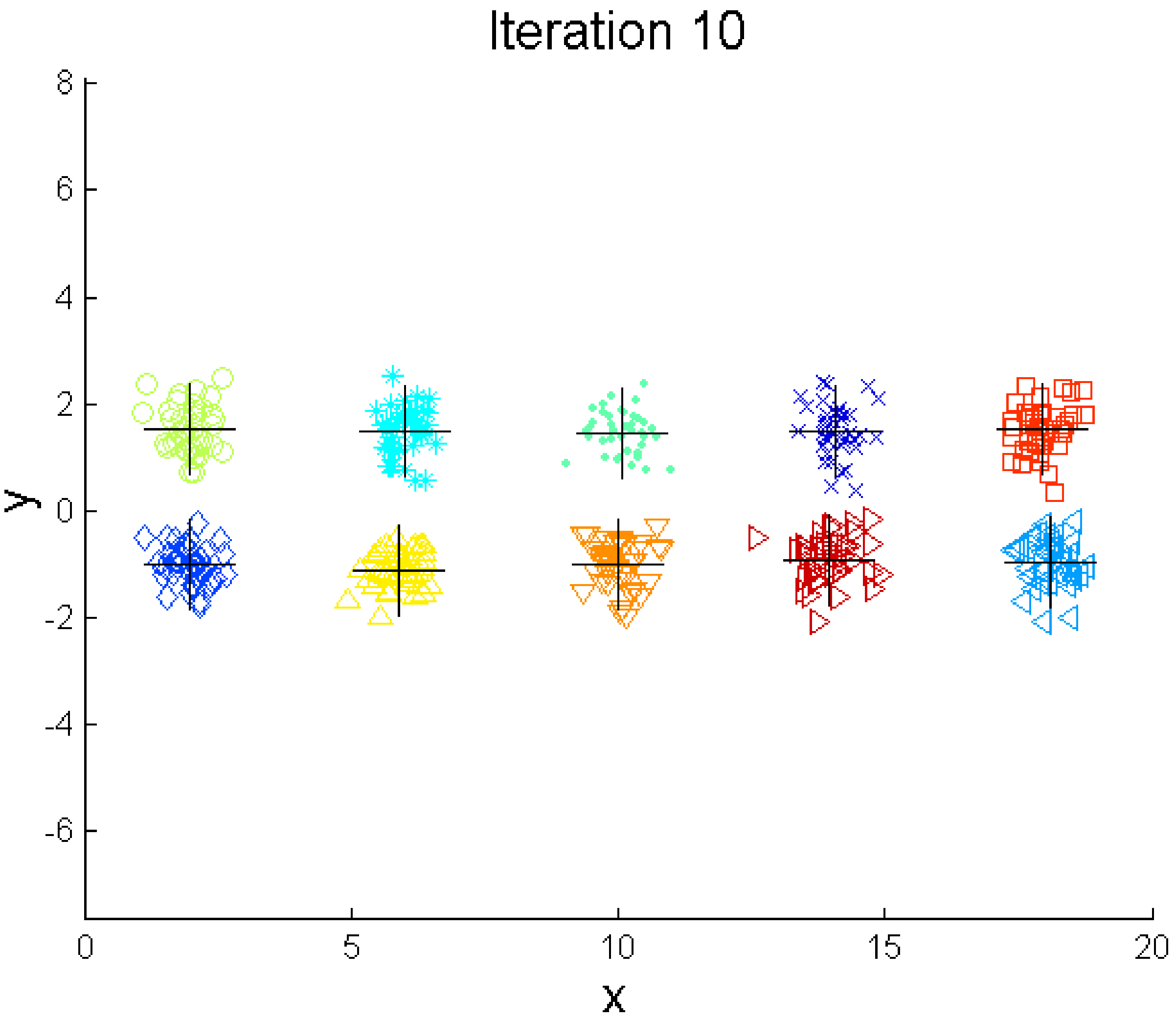
Bisecting K-means

- Bisecting K-means algorithm
  - Variant of K-means that can produce a partitional or a hierarchical clustering



```
Initialize the list of clusters to contain the cluster containing all points.
repeat
    Select a cluster from the list of clusters
    for $i = 1$ to $number\_of\_iterations$ do
        Bisect the selected cluster using basic K-means
    end for
    Add the two clusters from the bisection with the lowest SSE to the list of clusters.
until Until the list of clusters contains $K$ clusters
```

Bisecting K-means Example

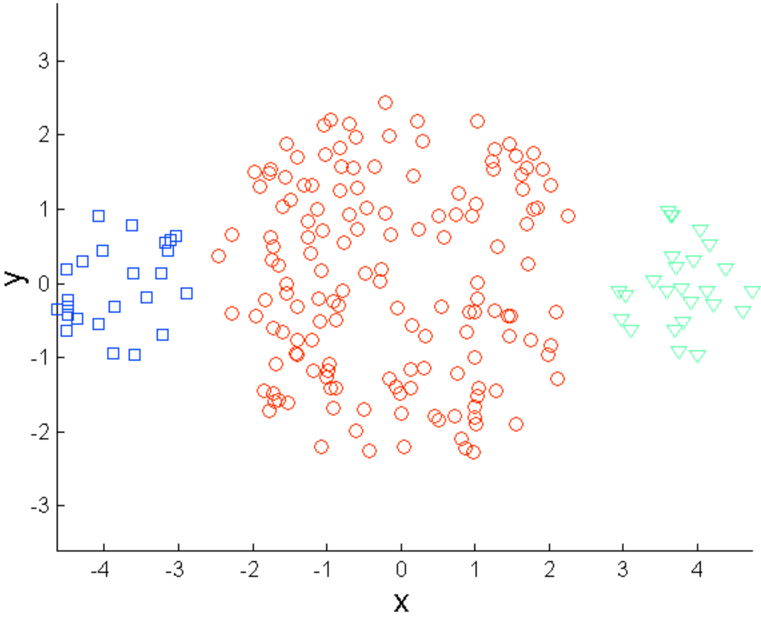


Limitations of K-means

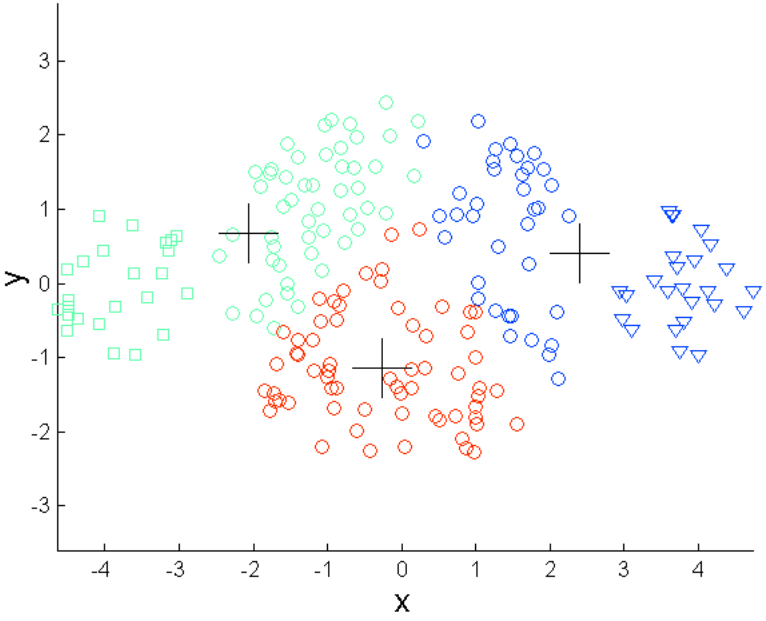
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

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

Limitations of K-means: Differing Sizes

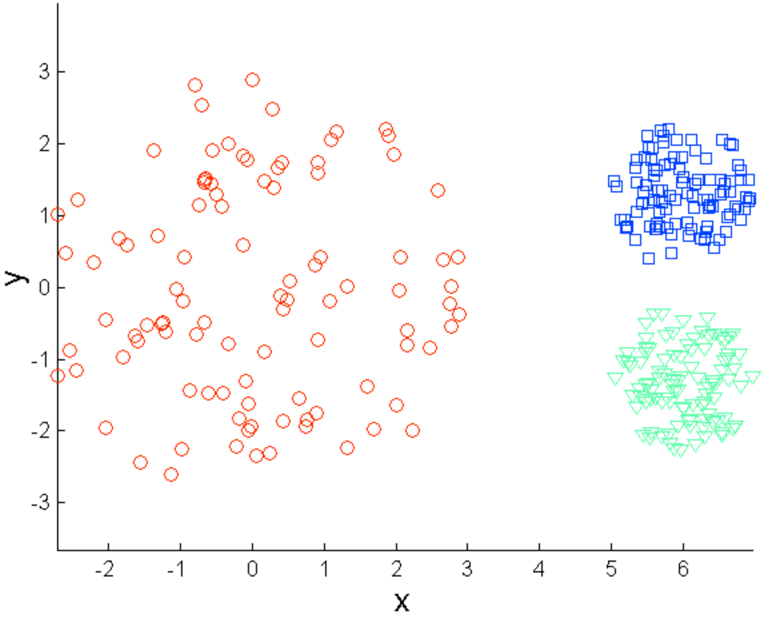


Original Points

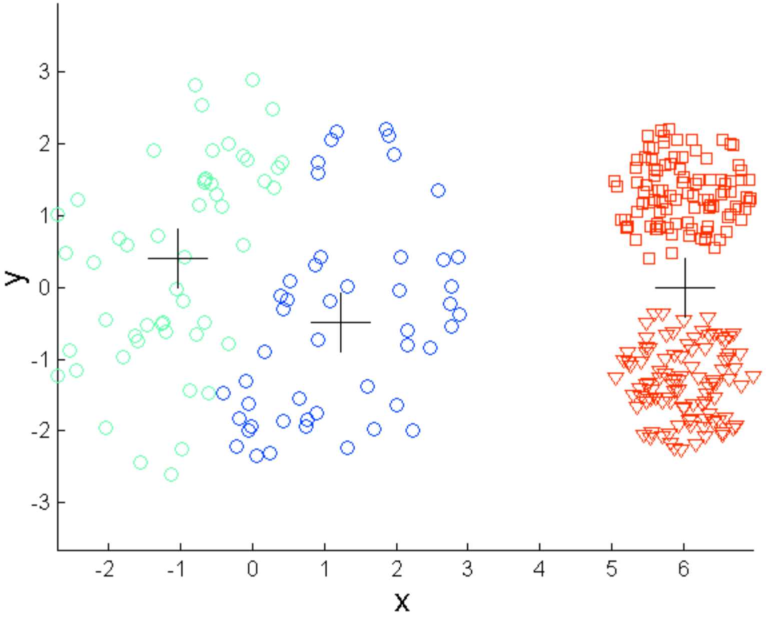


K-means (3 Clusters)

Limitations of K-means: Differing Density

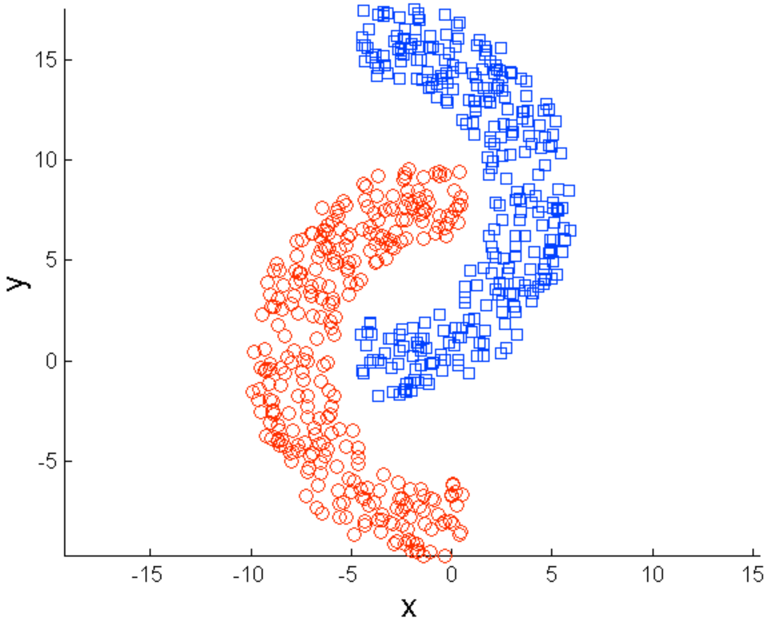


Original Points

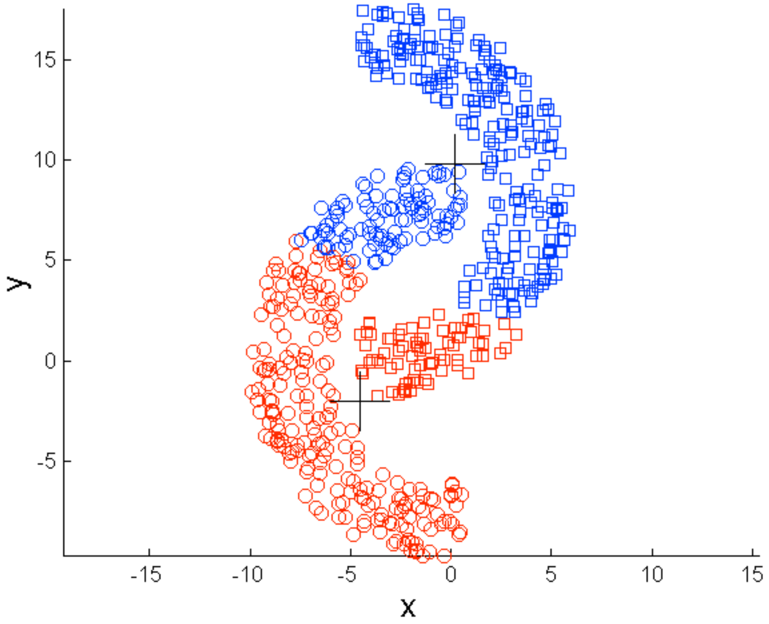


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

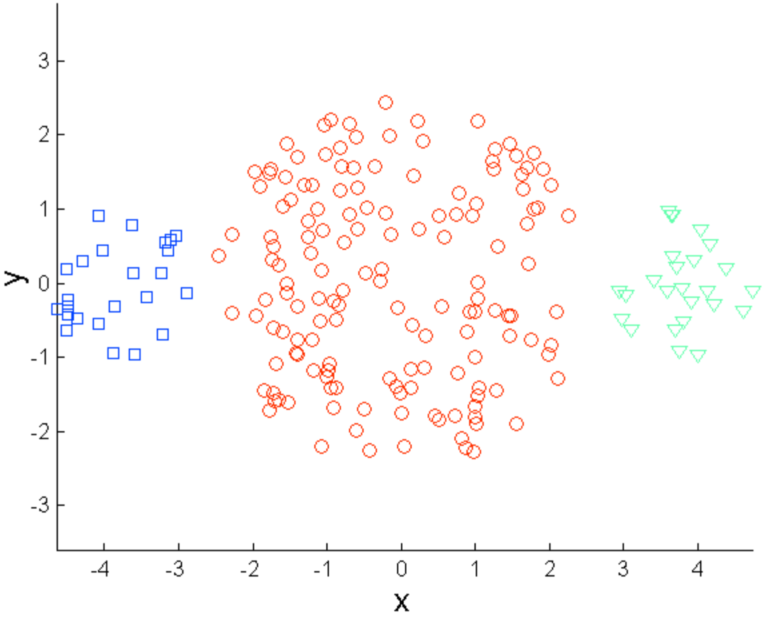


Original Points

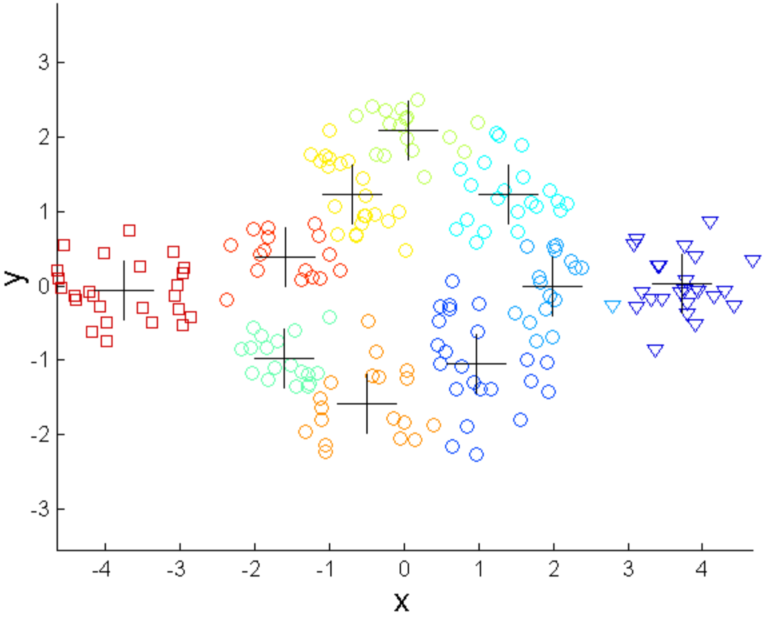


K-means (2 Clusters)

Overcoming K-means Limitations



Original Points

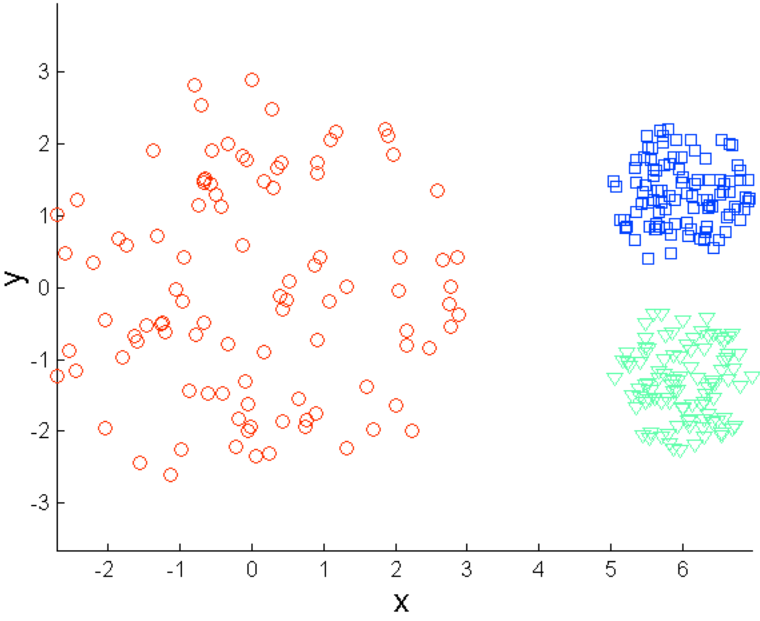


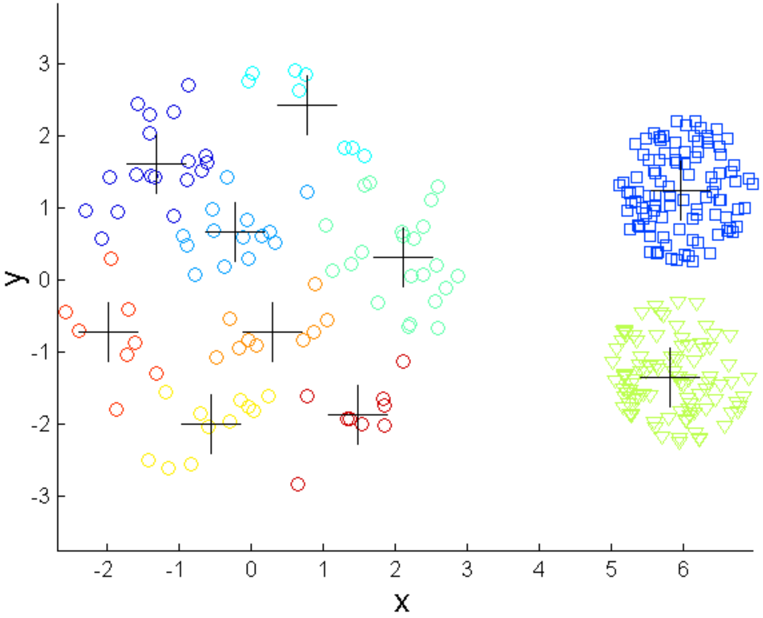
K-means Clusters

One solution is to use many clusters.

Find parts of clusters, but need to put together.

Overcoming K-means Limitations

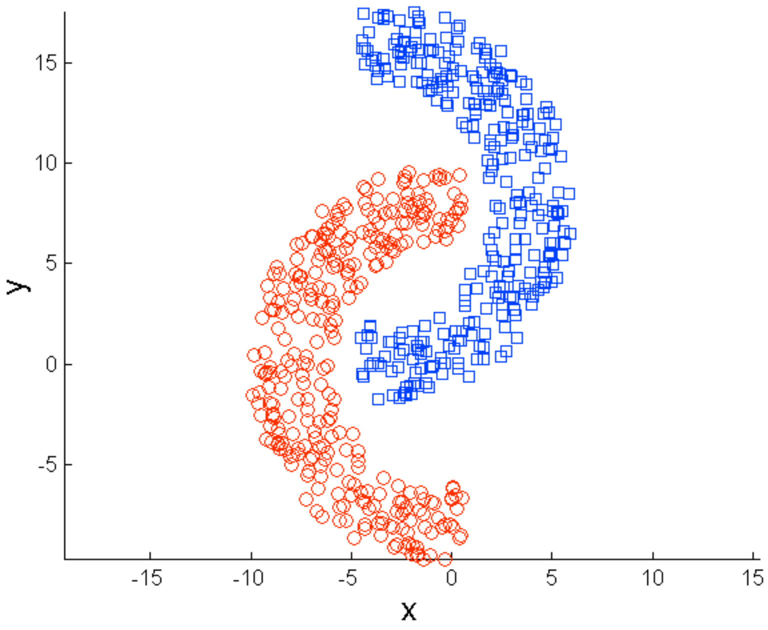




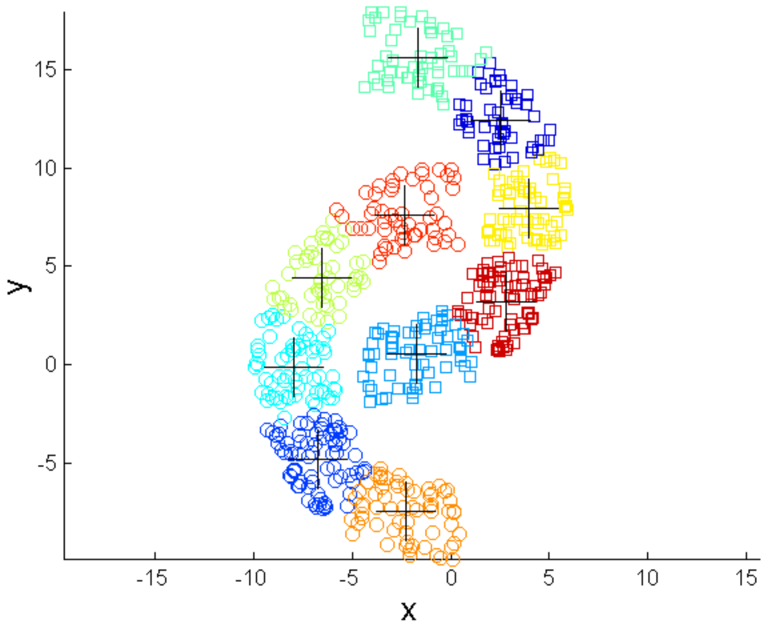
Original Points

K-means Clusters

Overcoming K-means Limitations



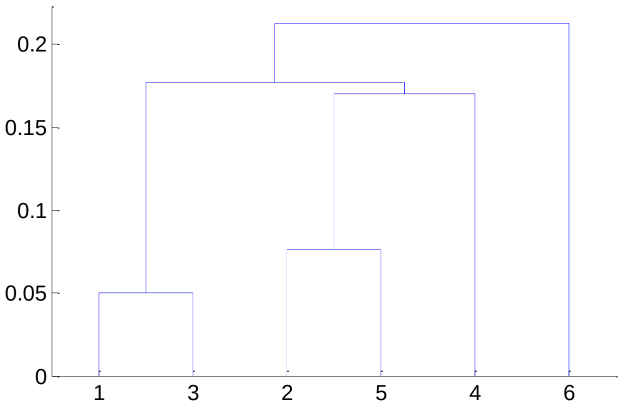
Original Points

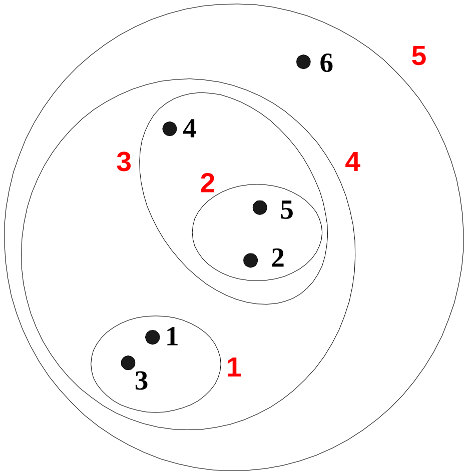


K-means Clusters

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits





Hierarchical Clustering

- Two main types of hierarchical clustering
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)
- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

Agglomerative   Divisive

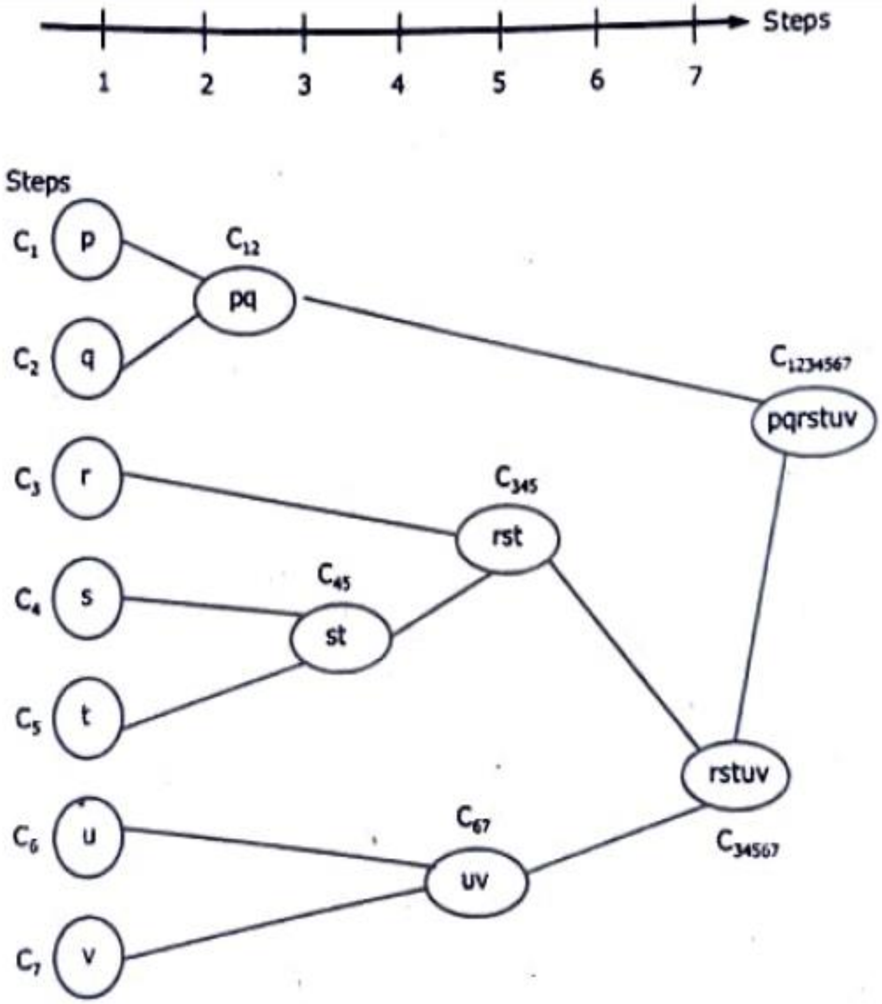


Fig. 7.5.1 First AGNES on Data Objects {p, q, r, s, t, u, v}

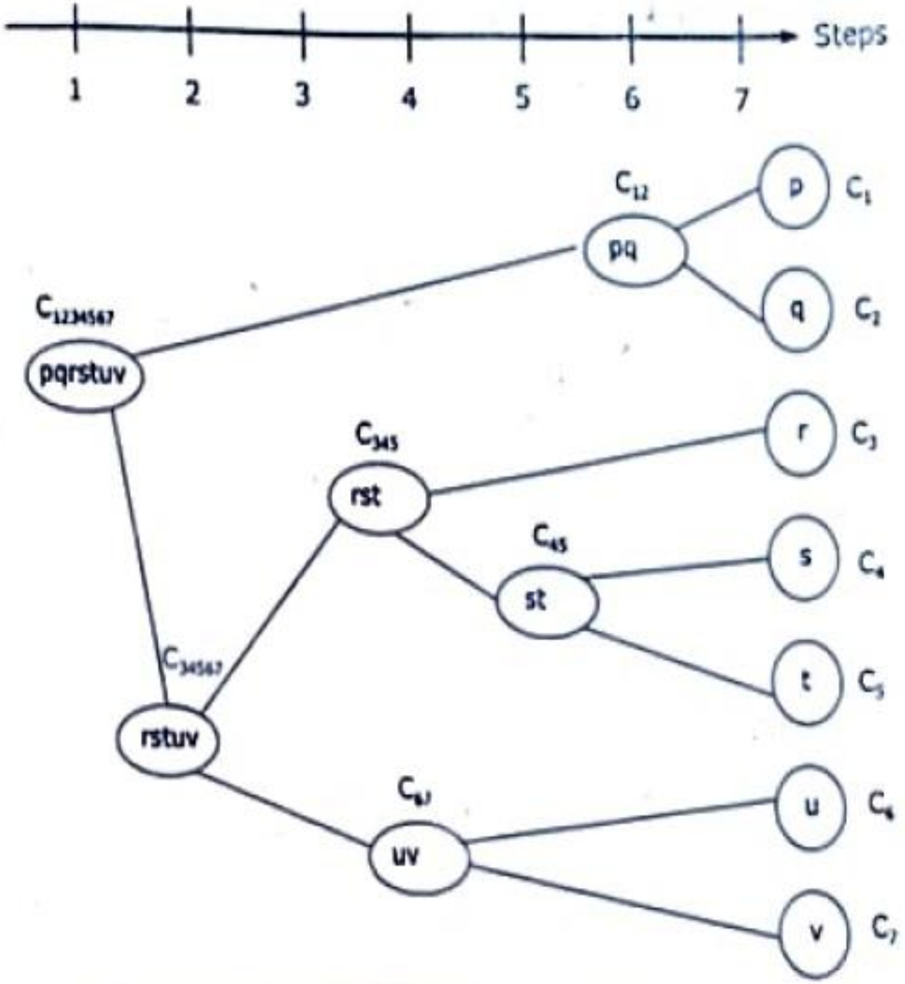


Fig. 7.5.2 DIANA on Data Objects {p, q, r, s, t, u, v}