Clustering

Prof. Ankur Sinha

Indian Institute of Management Ahmedabad

Gujarat India

Clustering

- Grouping a set of data objects into different groups based on similarity
- An example of unsupervised learning
- Data objects can be vectors representing different attributes for an object, for example, customer, location, product, etc.

Examples

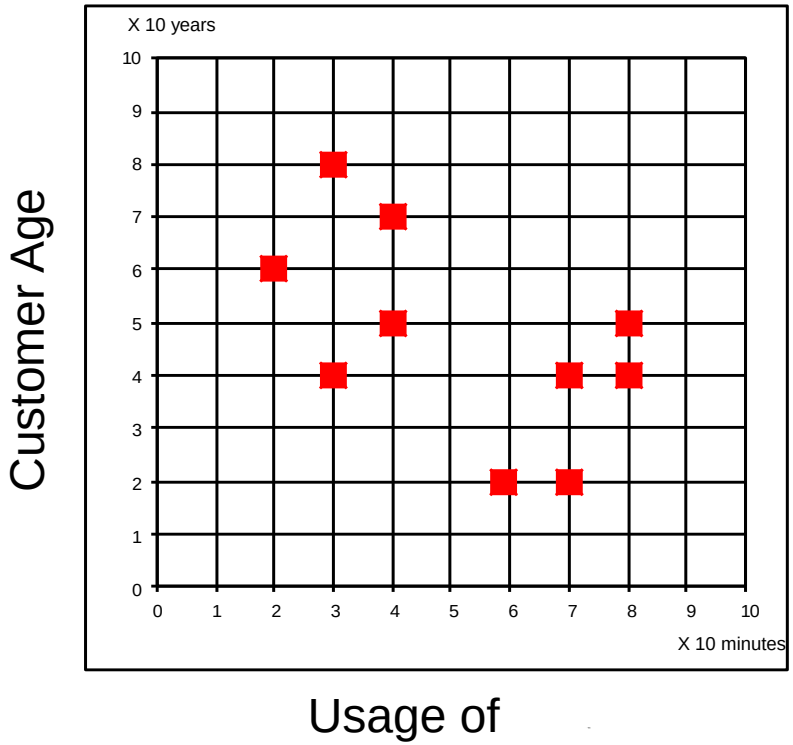
- Used in a variety of areas
  - Marketing
  - Urban planning
  - Customer segmentation
  - Product segmentation
  - Company segmentation
  - Seismology

Similarity Measure

- If two objects i and j are represented by vectors x^i and x^j
  - How do you measure similarity between the two objects
    - Euclidean distance
    - Manhattan distance
    - Mahalanobis distance
  - Similarity can be chosen based on the application

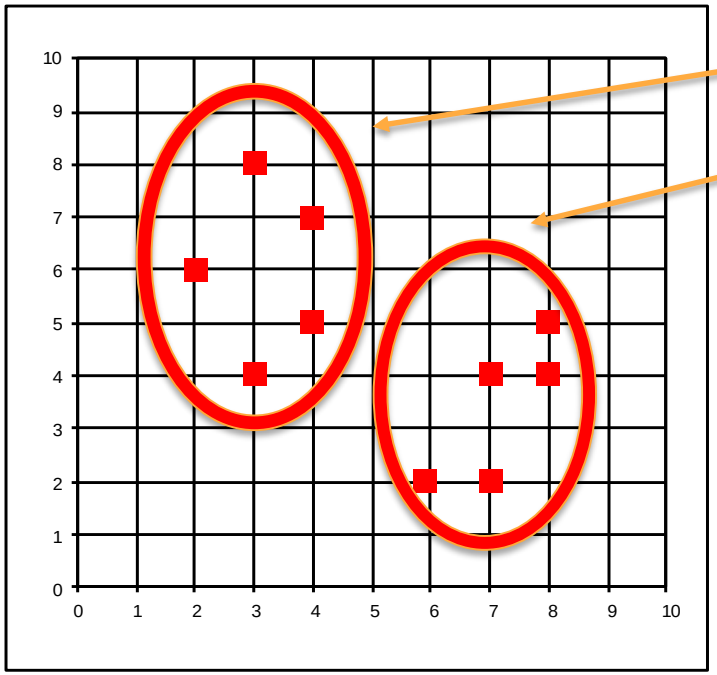
Similarity Measure

- Consider 10 customers with two attributes
  - –Attribute 1: Recent usage of services
  - –Attribute 2: Customer age
- Objective: Cluster the data into two classes and design two marketing campaigns for the two customer segments



Similarity Measure

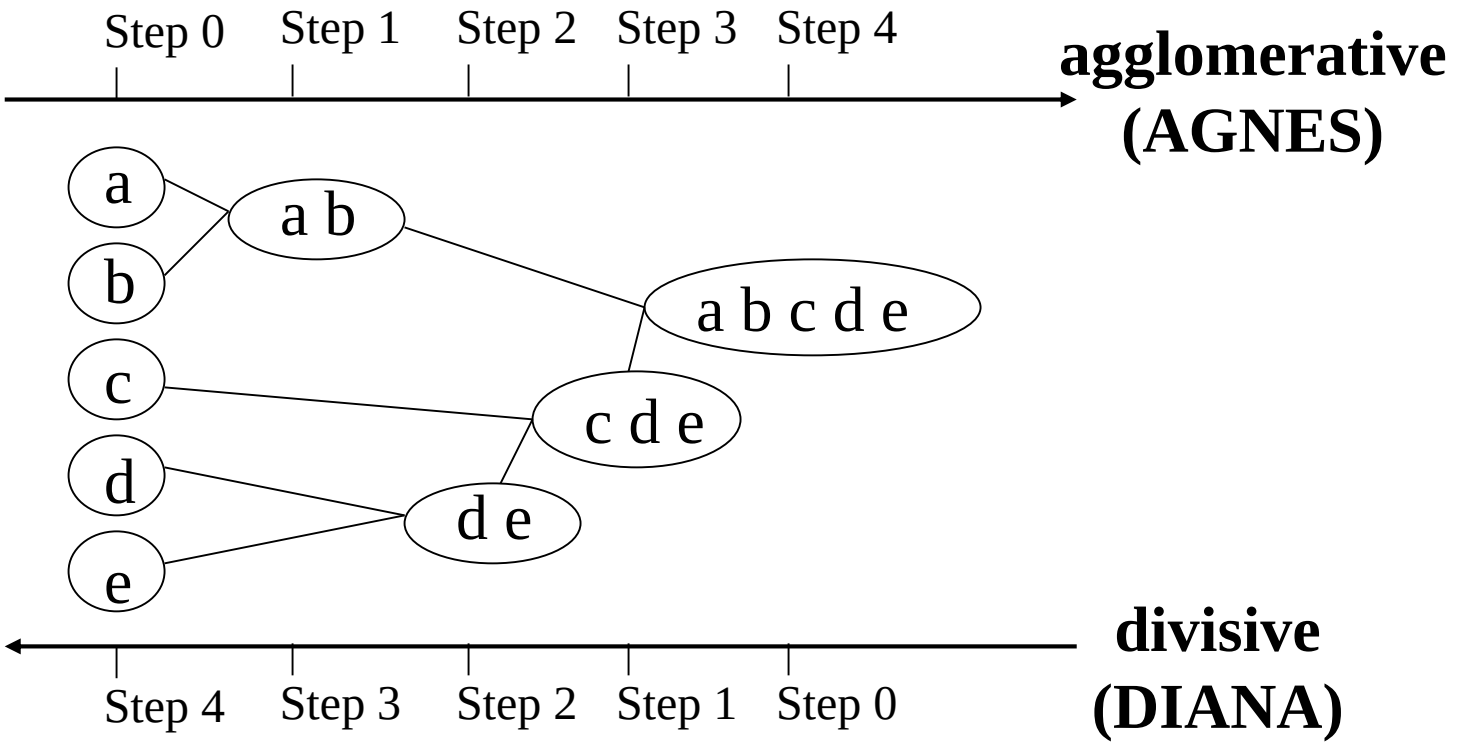
- Consider 10 customers with two attributes
  - Attribute 1: Usage of services
  - Attribute 2: Customer age



Cluster 1	Cluster2
(3,4)	(6,2)
(2,6)	(7,2)
(4,5)	(7,4)
(4,7)	(8,4)
(3,8)	(8,5)

Clustering approaches

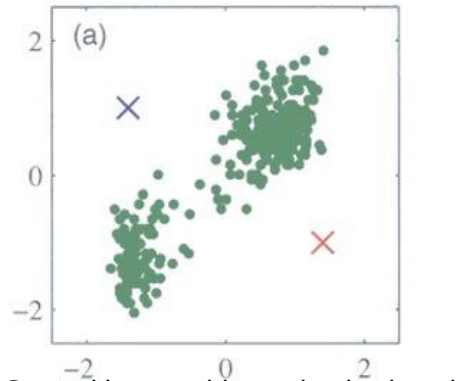
- Hierarchical clustering
  - Agglomerative
  - Divisive



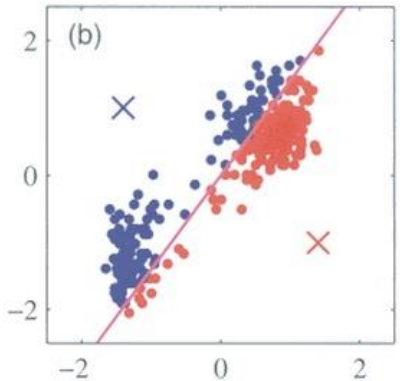
Clustering approaches

- K-means Clustering
  - Select initial centroids randomly
  - Assign objects to centroids based on similarity measure
  - Compute new centroid as mean of each class
  - Repeat the above two steps until there is no change

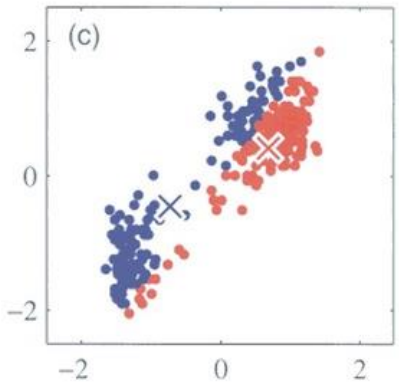
K-Means Clustering



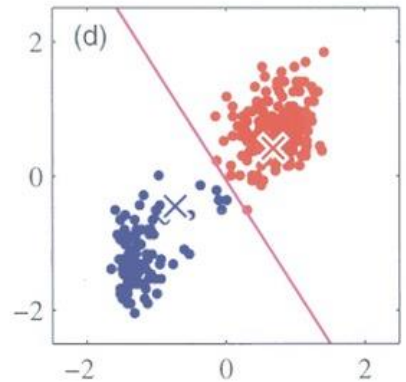
Start with centroids randomly placed



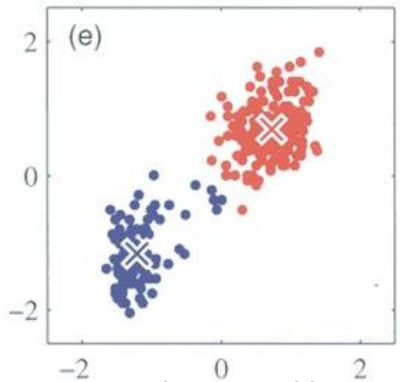
Assign points to the centroids



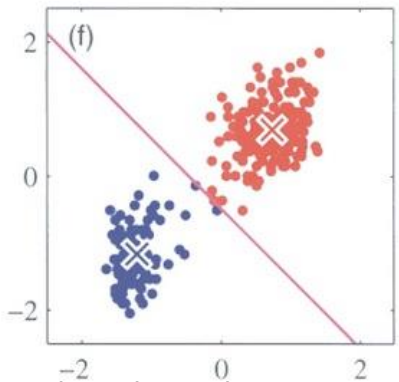
Update centroids



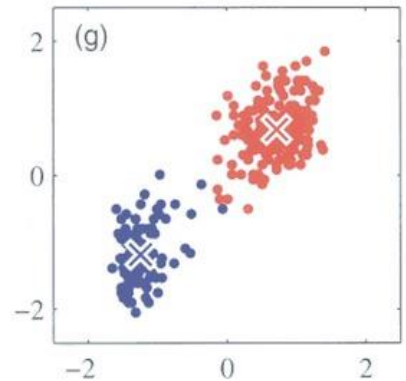
Assign points to the new centroids

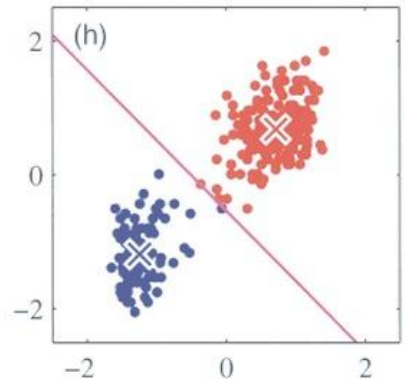


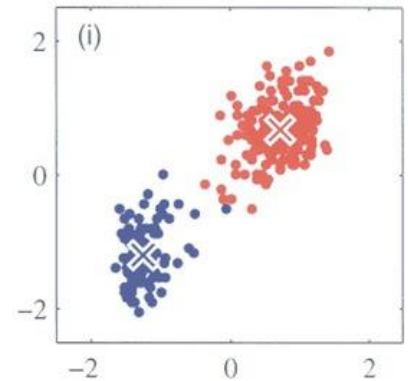
Update centroids



Assign points to the new centroids

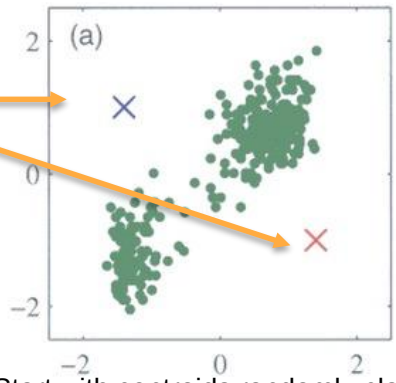




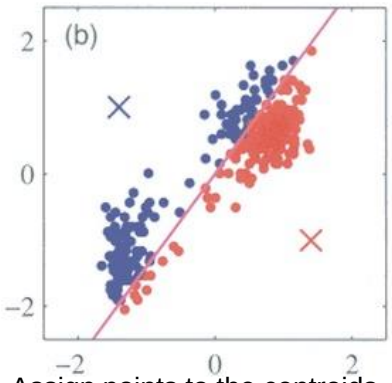


K-Means Clustering

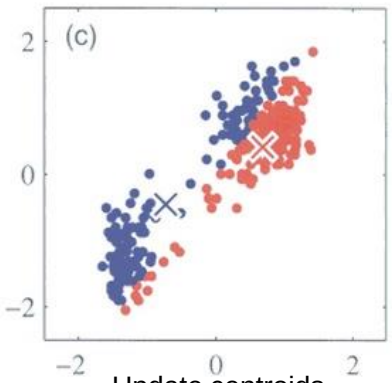
Random centroids



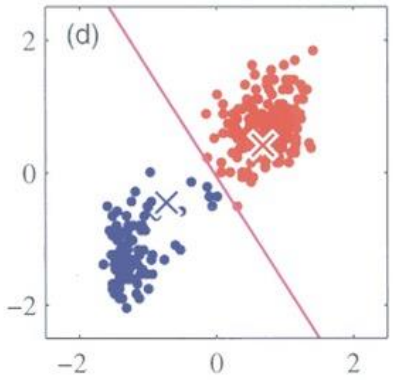
Start with centroids randomly placed



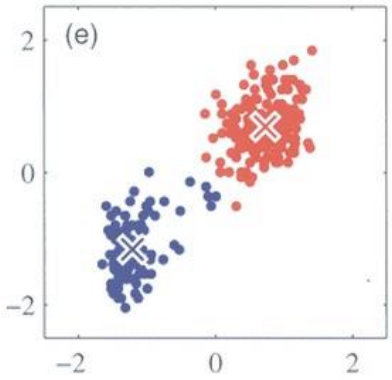
Assign points to the centroids



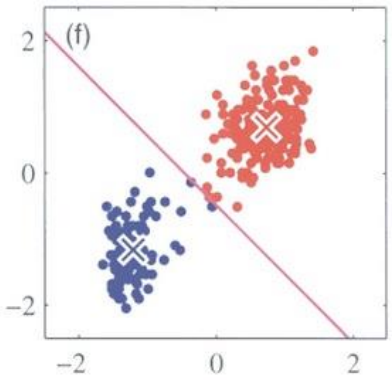
Update centroids



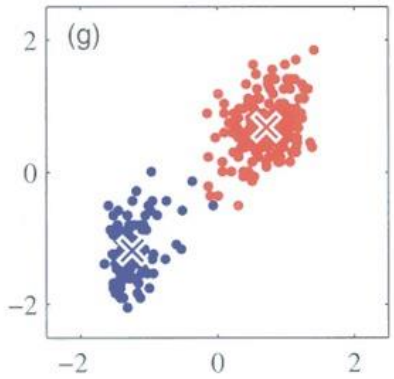
Assign points to the new centroids

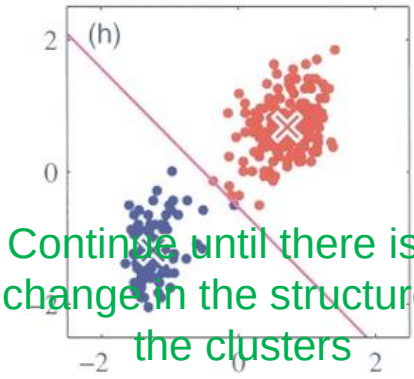


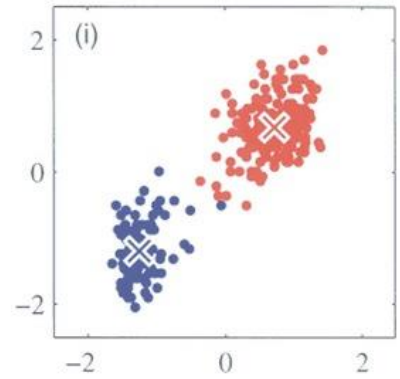
Update centroids



Assign points to the new centroids







Continue until there is no change in the structure of the clusters