Clustering Techniques

Machine Learning (CS 306)

Instructor: Dr. Moumita Roy

Teaching Assistants: Indrajit Kalita, Veronica Naosekpam

Email-ids:

moumita@iiitg.ac.in

veronica.naosekpam@iiitg.ac.in

indrajit.kalita@iiitg.ac.in

Mobile No: +91-8420489325 (only for emergency quires)

Reference: slides from Dr. Debasis Samanta (IIT Kharagpur)

- Such problems are already introduced in Artificial Intelligence course (under the topics of local search)

Complete state-formulation (Clustering)

Formulate the following problem for intelligent agent:

Divide the N students in k groups based on their marks in AI such as diversity in each group is minimized.

For demonstration, you may use:

N=6, k=2, set of marks={50, 20,10, 5, 15, 45} (out of 60)

- State representation:
- Initial state:
- Goal state:
- Successor function:
- Objective function:

Complete state-formulation (Version 1) (Clustering)

Set of students/marks $A=\{a_1, a_2, a_3, \ldots\ldots, a_N\}= \{50, 20, 10, 5, 15, 45\}$

- State representation:

  $S=[C_1\ C_2\ C_3\ \ldots\ldots\ C_N]$

  if a_j belongs to i^{th} group, then $C_j=i$, where $i \in \{1, 2, 3, \ldots., k\}$

- Initial state: (example for demonstration)

  $S^0=[1\ 2\ 2\ 2\ 2\ 1]$ → better representation is needed

  $\mathbf{F(S^0)}=(50-45)^2 + [(20-10)^2+(20-5)^2+(20-15)^2+(10-5)^2+(10-15)^2+(5-15)^2]$

- Goal state/test:

  No binary goal state

  report the state when we have the minimum value of objective function

- Successor function: Successor is generated by changing the group of one student at a time
- Objective function: summation of diversity in each cluster

Complete state-formulation (Version 2) (Clustering)

Set of students/marks $A=\{a_1, a_2, a_3,\ldots\ldots,a_N\}= =\{50, 20,10, 5, 15, 45\}$

- State representation:

  $S=[C_1\ C_2\ C_3\ \ldots\ldots\ C_k]$

  if C_j is representative of the j^{th} group, then $C_j \in \{a_1, a_2, a_3,\ldots,a_N\}$

- Initial state: (example for demonstration)

  S^0=[50 20] (assigned in a group with nearest representative; then same as [1 2 2 2 2 1])

  $F(S^0)=[(50-50)^2+(50-45)^2]+[(20-20)^2+(20-10)^2+(20-5)^2+(20-15)^2]$

  $\mathbf{F(S^0)=(50-45)^2 + [(20-10)^2+(20-5)^2+(20-15)^2+(10-5)^2+(10-15)^2+(5-15)^2]}$

- Goal state/test:

  No binary goal state

  report the state when we have the minimum value of objective function
- Successor function: Successor is generated by changing a group representative at a time
- Objective function: summation of diversity in each cluster

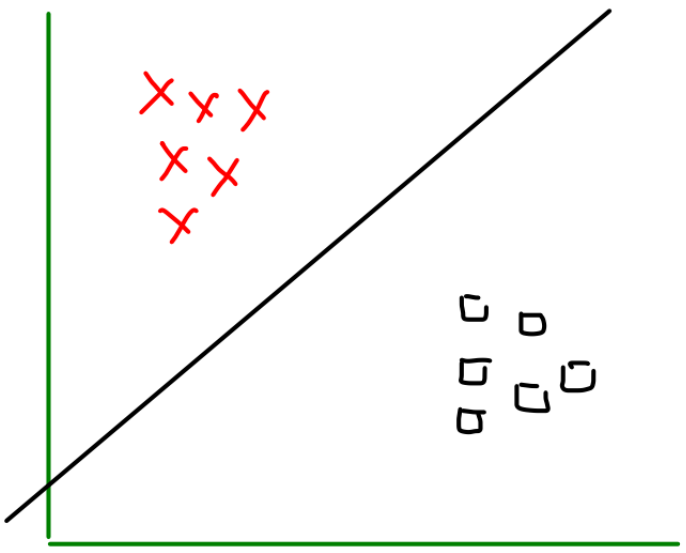
Observation

- We need some strategy to generate successors
- State space may be continuous or discrete
- Solution for such problems are addressed using the following concepts:
  - Genetic algorithms
  - Particle swarm optimization

Clustering techniques

- Clustering is a process of grouping a set of objects (into the groups of similar objects).
- It is most common form of unsupervised learning (learning from raw data without label).
- Groups are called clusters.
- Objects in the same cluster are somehow more similar to each other than the ones in the different clusters.

Classification

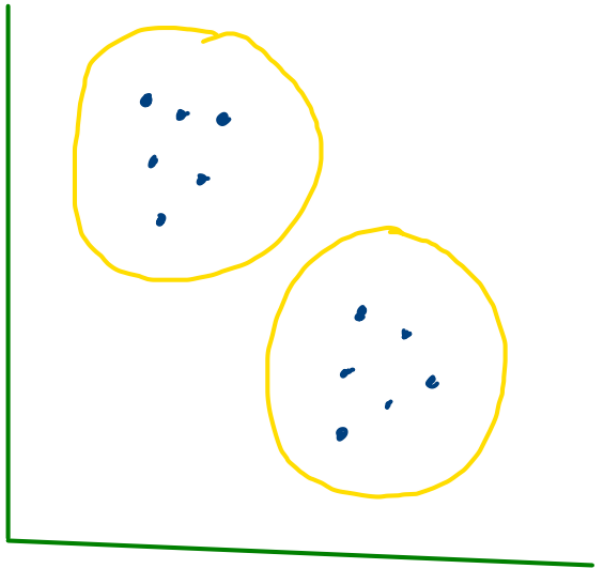


Supervised learning

Training Samples:

$\{(x_1, y_1), (x_2, y_2) \cdots (x_i, y_i)\}$

Clustering



Unsupervised learning

Samples:

$\{x_1, x_2, \cdots x_i\}$

Goal of Clustering

- Organizing samples into clusters such that there is
  - **high intra-cluster similarity**
  - **low inter-cluster similarity**

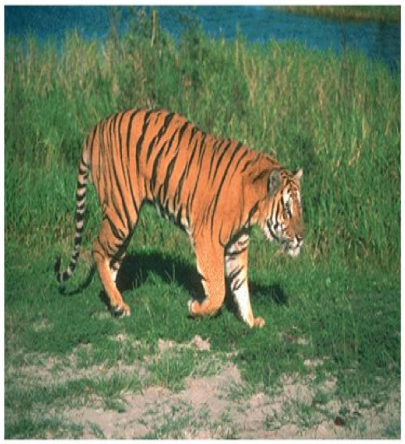
- Requires the definition of a similarity measure
- Try to gain some insight into the structure of the data/samples

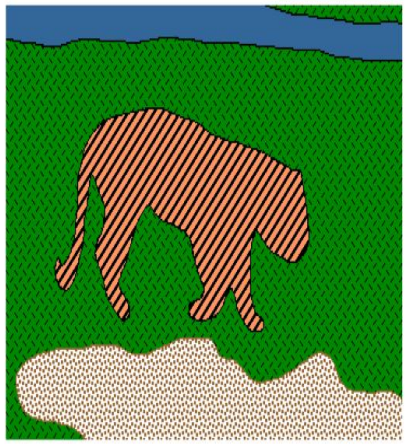
Application of clustering techniques

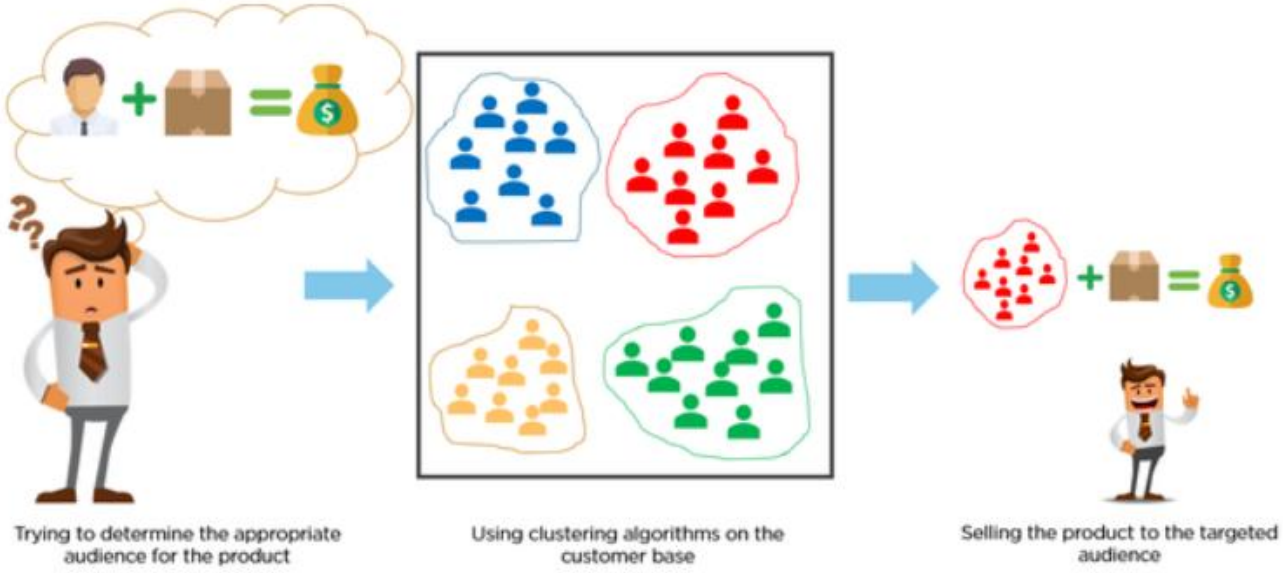
- Marketing policy (customer segmentation): Discover distinct groups of customers by analyzing the customer behavior and develop different marketing strategy for different groups

- Clustering in social network analysis: Discovery of clusters or communities (i.e. a collection of individuals with dense/sparse friendship patterns).

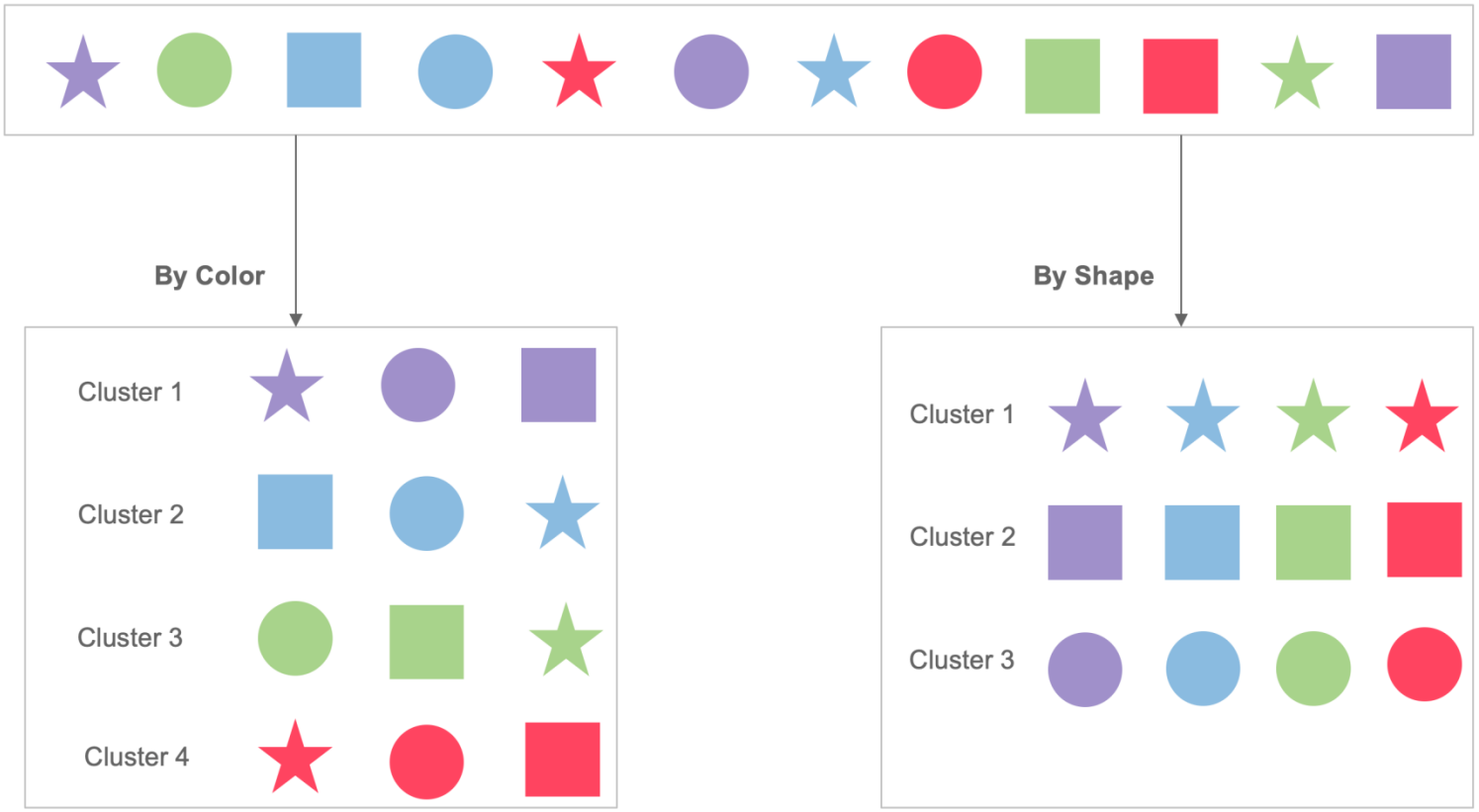
- Land-use monitoring (image segmentation): Identify areas of similar land-use






Trying to determine the appropriate audience for the product
Using clustering algorithms on the customer base
Selling the product to the targeted audience

Clustering (subjective)



Clustering algorithms

- Hard clustering: Each pattern exclusively belongs to a single cluster.
  - Partitioning algorithm (K-means, K-medoids)
  - Density based methods (DBSCAN, CLIQUE)
  - Graph theoretical based methods (MST, OPOSSUM)
  - Model-based methods (SOFM, EM algorithm)
  - Genetic clustering algorithm
  - PSO based clustering
  - Hierarchical clustering (Agglomerative and divisive)
- Soft clustering: Each pattern may belongs to more than one clusters
  - Fuzzy clustering
  - Probabilistic clustering

What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.
Webster's Dictionary

Similarity is hard to define, but…
"*We know it when we see it*"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Similarity and Dissimilarity Measures

- In clustering techniques, similarity (or dissimilarity) is an important measurement.
- Informally, similarity between two objects (e.g., two images, two documents, two records, etc.) is a numerical measure of the degree to which two objects are alike.
- The dissimilarity on the other hand, is another alternative (or opposite) measure of the degree to which two objects are different.
- Both similarity and dissimilarity also termed as proximity.

Note:

- Frequently, the term distance (for example, Euclidian distance) is used as a synonym for dissimilarity
- In fact, it is used to refer as a special case of dissimilarity.

Partitioning algorithm: Basic concepts

- Partition of n objects into k-clusters by optimizing some partitioning criteria.

- One approach: examine all possible partition (too expensive)

- Heuristic based: K-means algorithm and K-medoids algorithm

k-Means Algorithm

- k-Means clustering algorithm proposed by J. Hartigan and M. A. Wong [1979].
- Given a set of *n* distinct objects/patterns, the k-Means clustering algorithm partitions the objects into *k* number of clusters such that intracluster similarity is high but the intercluster similarity is low.
- In this algorithm, we need to specify *k*, the number of clusters and consider the objects are defined with numeric attributes and thus using any one of the distance metric to demarcate the clusters.

k-Means Algorithm

The algorithm can be stated as follows.

- First it selects *k* number of objects at random from the set of n objects. These *k* objects are treated as the centroids of *k* clusters.
- For each of the remaining objects, it is assigned to one of the closest centroid. Thus, it forms a collection of objects assigned to each centroid and is called a cluster.
- Next, the centroid of each cluster is then updated (by calculating the mean values of attributes of each object).
- The assignment and update procedure is until it reaches some stopping criteria (such as, number of iteration, centroids remain unchanged or no assignment, etc.)

k-Means Algorithm

Algorithm: k-Means clustering

Input: D is a dataset containing *n* objects/patterns, *k* is the number of cluster

Output: A set of *k* clusters

Steps:

1. Randomly choose *k* objects from D as the initial cluster centroids.
2. **For** each of the objects in D **do**
   - Compute distance between the current objects and *k* cluster centroids
   - Assign the current object to that cluster to which it is closest.
3. Compute the "cluster centers" of each cluster. These become the new cluster centroids.
4. Repeat step 2-3 until the convergence criterion is satisfied
5. Stop

Important Points

- Distance measure: Usually Euclidian distance is used

- Convergence criteria:

  - maximum number of iteration
  - no change of centroid values in any cluster
  - no change in cluster assignment of the patterns
  - cluster quality reaches a certain level of aceptance

Cluster quality

- There is some objective function to be met as a goal of clustering. One of such objective function is sum-square-error (denoted by SSE).

$$SSE = \sum_{i=1}^{K} \sum_{P \in C_i} |d(P, m_i)|^2$$

$C_i \rightarrow$ denotes i^{th} cluster

$P \rightarrow$ denotes the pattern belongs to any cluster

$m_i \rightarrow$ denotes the centroid of i^{th} cluster

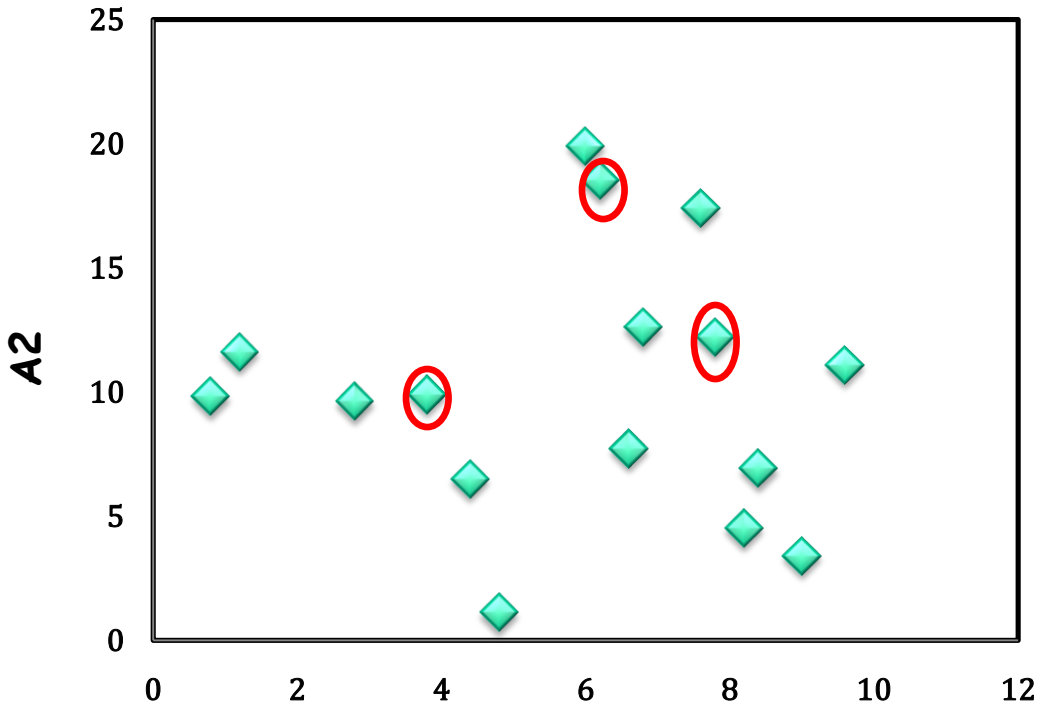
$d(P, m_i) \rightarrow$ denotes distance measure between any pattern and the corresponding cluster

Demonstration of k-Means clustering algorithms

Dataset

A_1	A_2
6.8	12.6
0.8	9.8
1.2	11.6
2.8	9.6
3.8	9.9
4.4	6.5
4.8	1.1
6.0	19.9
6.2	18.5
7.6	17.4
7.8	12.2
6.6	7.7
8.2	4.5
8.4	6.9
9.0	3.4
9.6	11.1

Plotting data (n=?,k=?, Feature=?)



Centroid	Objects	
	A1	A2
c_1	3.8	9.9
c_2	7.8	12.2
c_3	6.2	18.5

Initial Centroids chosen randomly

Demonstration of k-means clustering algorithms

- Suppose, k=3. Three objects are chosen at random shown as red circled. These three centroids are shown below.

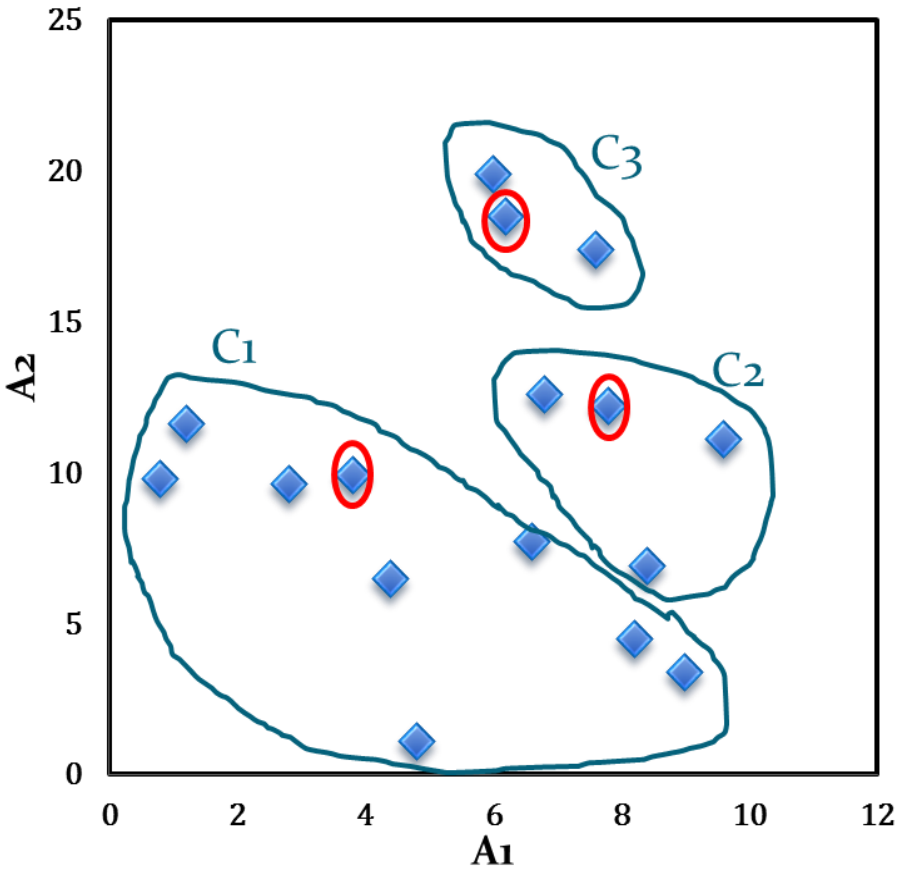
- Let us consider the Euclidean distance measure (L_2 Norm) as the distance measurement in our illustration.
- Let d_1, d_2 and d_3 denote the distance from an object to c_1, c_2 and c_3 respectively.
- Assignment of each object to the respective centroid is shown in the right-most column (in the next table).

Cluster Assignment (after Iteration 1)

Distance calculation

A_1	A_2	d_1	d_2	d_3	cluster
6.8	12.6	4.0	1.1	5.9	2
0.8	9.8	3.0	7.4	10.2	1
1.2	11.6	3.1	6.6	8.5	1
2.8	9.6	1.0	5.6	9.5	1
3.8	9.9	0.0	4.6	8.9	1
4.4	6.5	3.5	6.6	12.1	1
4.8	1.1	8.9	11.5	17.5	1
6.0	19.9	10.2	7.9	1.4	3
6.2	18.5	8.9	6.5	0.0	3
7.6	17.4	8.4	5.2	1.8	3
7.8	12.2	4.6	0.0	6.5	2
6.6	7.7	3.6	4.7	10.8	1
8.2	4.5	7.0	7.7	14.1	1
8.4	6.9	5.5	5.3	11.8	2
9.0	3.4	8.3	8.9	15.4	1
9.6	11.1	5.9	2.1	8.1	2

Cluster formation (iteration 1)



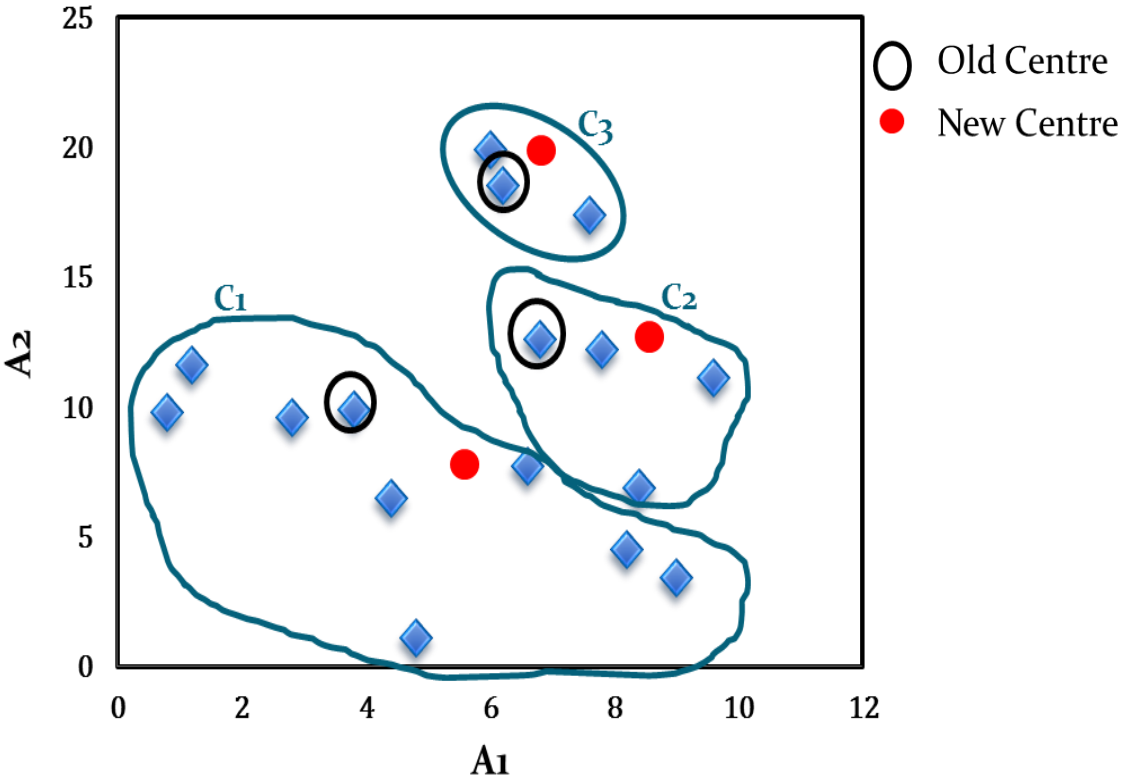
Updating centroids (iteration 1)

The calculation new centroids of the three cluster using the mean of attribute values of A_1 and A_2 is shown in the Table below. The cluster with new centroids are shown in Fig 16.3.

Calculation of new centroids

New Centroid	Objects	
	A1	A2
c_1	4.6	7.1
c_2	8.2	10.7
c_3	6.6	18.6

Repeat the steps until convergence



Initial cluster with new centroids

Evaluating K-means clustering

- Most common measure is sum-square-error (SSE) (choose the one with the smallest error).

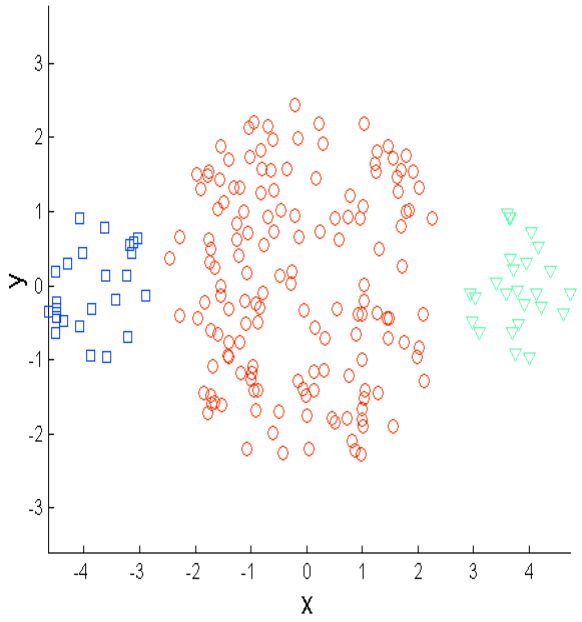
- One easy way to reduce SSE is to increase K, i.e. the number of clusters. A good clustering with smaller K can have a lower SSE than a poor clustering with higher K.

- Silhouette Coefficient or silhouette score is also used to calculate the goodness of a clustering technique (self-study).

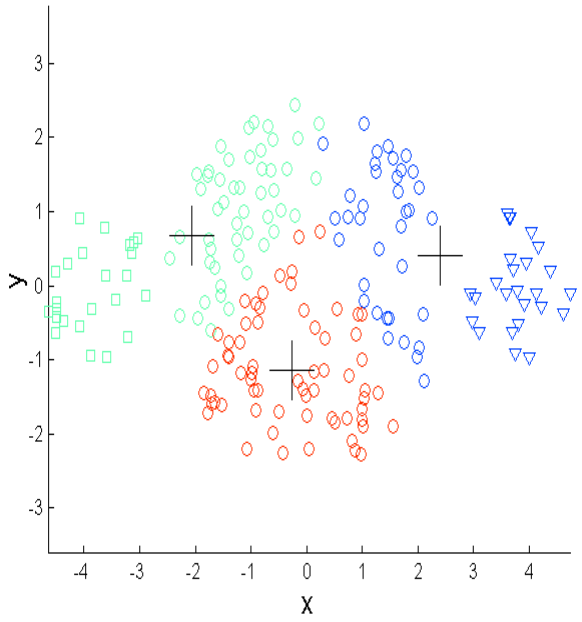
Comments on K-means clustering

- Simple clustering technique
- Specify the value of k: Elbow Method for optimal value of k (self-study)
- Choose initial centroids: usually terminate with any initial choice of the cluster centroids; initial choice influences the ultimate cluster quality (trapped in local optimum); choose initial centroids in multiple runs
- Sensitive to outliers: When the SSE is used as objective function, outliers can unduly influence the cluster that are produced. More precisely, in the presence of outliers, the cluster centroids, in fact, not truly as representative as they would be otherwise. It also influence the SSE measure as well; (one solution may be choose centroids from samples)
- k-Means algorithm cannot handle non-globular clusters, clusters of different sizes and densities

Limitations of K-means: Differing Sizes

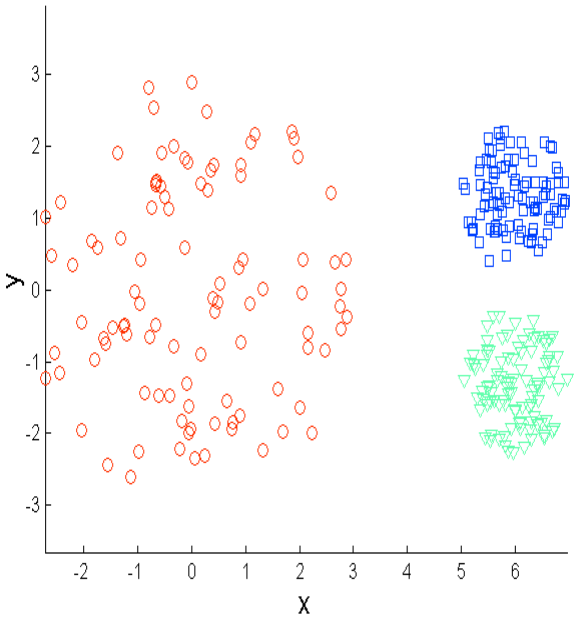


Original Points

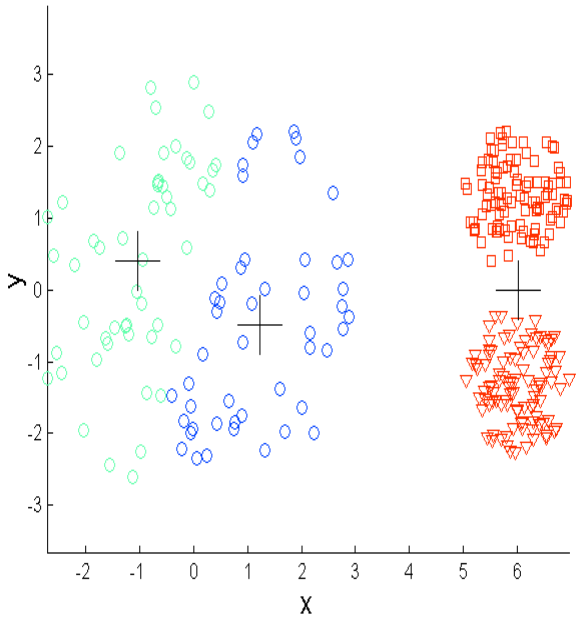


K-means (3 Clusters)

Limitations of K-means: Differing Density

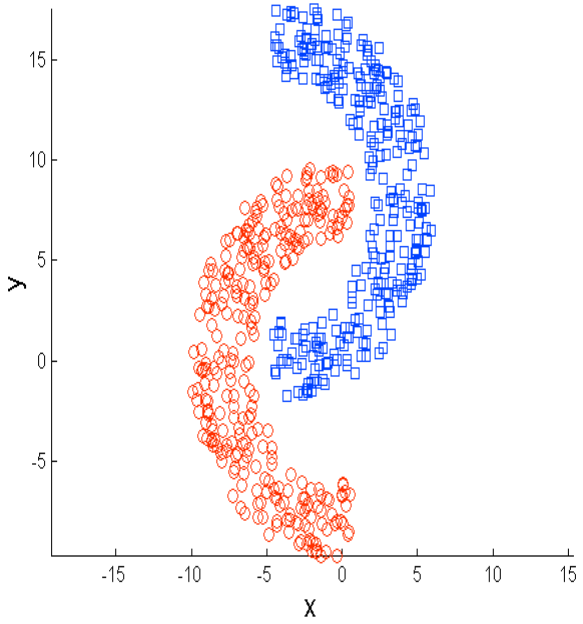


Original Points

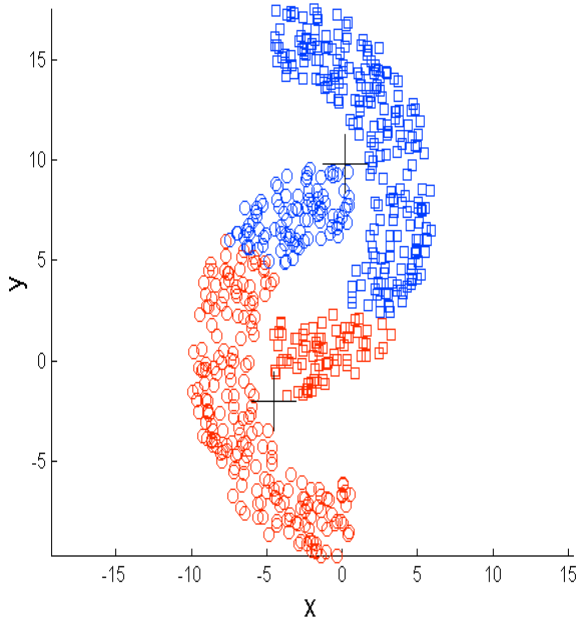


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes



Original Points



K-means (2 Clusters)

K-medoids clustering algorithm

Motivation: The k-Medoids algorithm aims to diminish the effect of outliers. Medoids are similar in concept to means or centroids, but medoids are always restricted to be members of the data set.

Basic concepts:

- The basic concepts of this algorithm is to select an object as a cluster center (one representative object per cluster) instead of taking the mean value of the objects in a cluster (as in k-Means algorithm).
- We call this cluster representative as a cluster medoid or simply medoid.

1. Initially, it selects a random set of *k* objects as the set of medoids.
2. Then at each step, all objects from the set of objects, which are not currently medoids are examined one by one to see if they should be medoids.

This version of K-medoids clustering algorithm is also called partition around medoids (PAM)

Algorithm: Partition around medoids (PAM)

1. Select ***k*** representative objects randomly as medoids and assign each non-selected object to the most similar representative object/medoid (form initial cluster)

2. For each pair of selected (medoid) object (***i)*** and non-selected (non-medoid) object (***h)***, calculate the Total swapping Cost (***SWC_{ih})***

3. For each pair of ***i*** and ***h***,
    1. If ***SWC_{ih}*** < 0, ***i*** is replaced by ***h (new set of medoids that forms more compact cluster)***
    2. Then assign each non-selected object to the most similar representative object

4. Repeat steps 2-3 until there is no change in the medoids**.**

Observation: work for small datasets

Calculation of total swapping cost

- Suppose, present set of medoids $m=[m_1\ m_2\ m_3]$ *and* E_m *is the current* SSE.

- Choose any non-medoid object *h* and swap it with *i* (the medoid object)

  new set of selected objects is $m_{new}=[m_1\ h\ m_3]$ *and* E_{mnew} *is the corresponding* SSE

- $SWC= E_{mnew} - E_m$

Some clustering algorithms for self-study

- Hierarchical clustering
- DBSCAN (density based clustering)

Fuzzy clustering algorithm

- Fuzzy clustering is a powerful unsupervised method for the analysis of data and construction of models
- In many situations, fuzzy clustering is more natural than hard clustering.
- Objects on the boundaries between several clusters are not forced to fully belong to one of the clusters, but rather are assigned membership degrees between 0 and 1 indicating their partial membership.
- The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.

Each pattern belongs to simultaneously more than one cluster with a membership values.

Concepts:

$X_i \rightarrow i^{th}$ pattern

$C_1\ C_2\ C_3 \cdots C_j \rightarrow$ Cluster centers

$\mu_{i1}\ \mu_{i2}\ \mu_{i3} \cdots \mu_{ij}$ → membership value of i^{th} pattern in j^{th} cluster

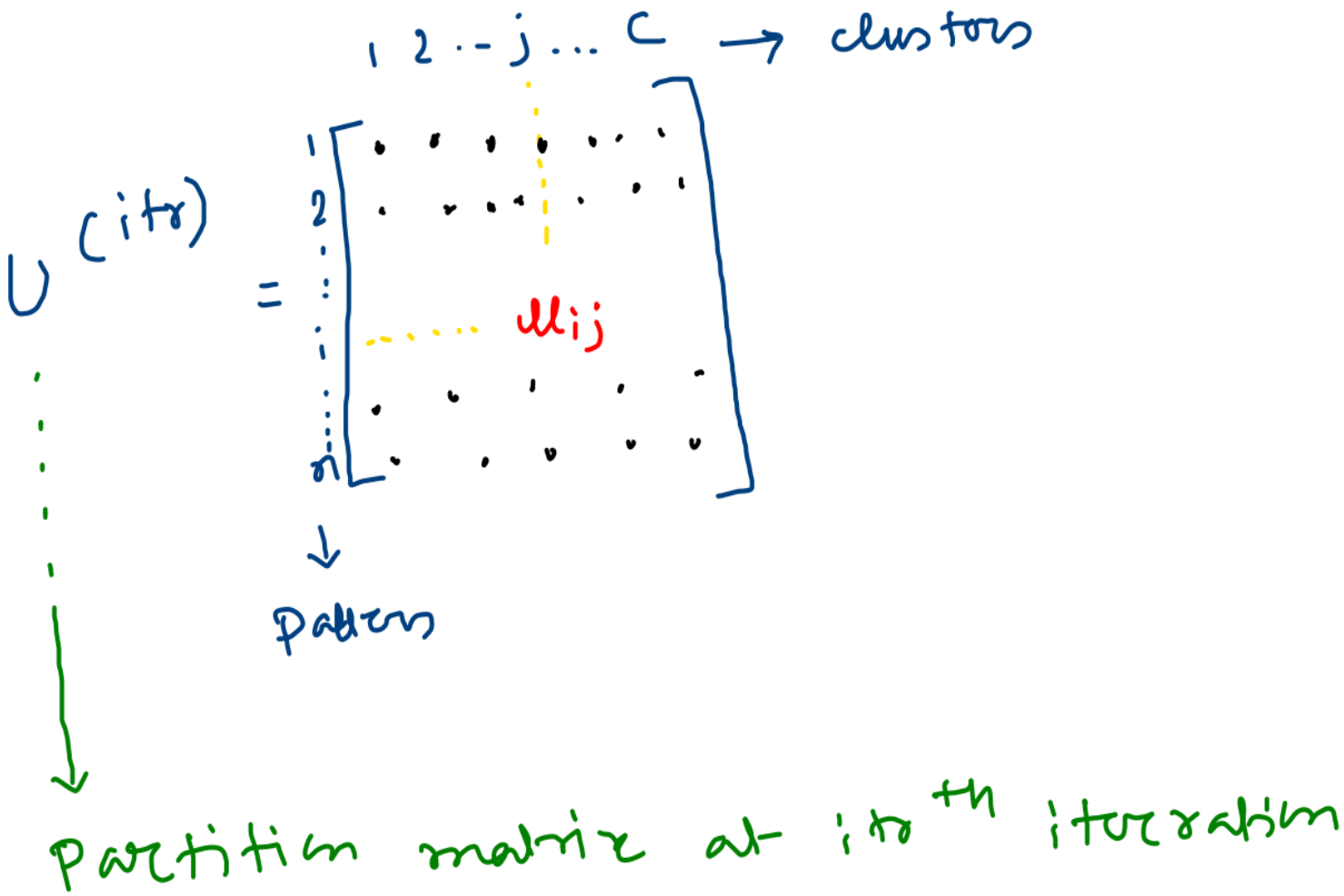
- if it is either 0 or 1 then hard clustering
- Here, the value is between 0 to 1 (including both)

Constraint

$\mu_{i1} + \mu_{i2} + \cdots + \mu_{ic} = 1$

C → number of clusters

Partition matrix



FCM Algorithm

Step 1: Initialize partition matrix randomly $itr = 0$

$U^{(0)}$ (maintain the constraint)

Step 2: Calculate the centers

$$C_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{n} u_{ij}^{m}}$$

degree of fuzzification $m > 1$

(usually assign $m = 2$)

Step 3: Update partition matrix $U^{(itr+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - C_j\|}{\|x_i - C_k\|} \right)^{2/m-1}}$$

FCM Algorithm (continue..)

Step4: Repeat steps 2 and 3 until there is a change in cluster centers

Otherwise, goto to step5

Step5: Report clustering results with patition matrix

Question 1

Perform FCM for the following datasets:

P1: [1 , 2], P2: [1 , 4], P3: [-2 , -3]

Note the following for demonstration:

Number of iterations: 1

Degree of fuzzification:2

Number of clusters: 2

Initial partition matrix=[0.1 0.9; 0.6 0.4; 0.2 0.8]

FCM

Iteration 1

Step 1 : Initialize partition matrix

$$U = \begin{bmatrix} 0.1 & 0.9 \\ 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$$

Step 2 : Update the cluster centers

$$C_1 = [\, C_{11} \quad C_{12} \,]$$

$$C_2 = [\, C_{21} \quad C_{22} \,]$$

$$C_1 = \frac{(0.1)^2 * [1 \quad 2] + 0.6^2 * [1 \quad 4] + 0.2^2 * [-2 \quad -3]}{0.1^2 + 0.6^2 + 0.2^2}$$

$$C_2 = \frac{(0.9)^2 * [1 \quad 2] + 0.4^2 * [1 \quad 4] + 0.3^2 * [-2 \quad -3]}{0.9^2 + 0.4^2 + 0.3^2}$$

Step 3 Updation of partition matrix

$$u_{11} = \frac{1}{\left(\frac{\|[1\ 2] - C_1\|}{\|[1\ 2] - C_1\|}\right)^{2/m-1} + \left(\frac{\|[1\ 2] - C_1\|}{\|[1\ 2] - C_2\|}\right)^{2/m-1}}$$

u_{12}, u_{21}, u_{22}, u_{21}, u_{32}, u_{31} → ?

Repeat step 2 and 3 until convergence

Question 2

- Demonstrate PAM clustering technique for 2 clusters problem using the following dataset:

Pattern no	Feature1	Feature2
1	2	4
2	2	6
3	−1	0
4	−3	4

Random sequence of pattern no:

2 1 3