Week 5 : Session 1

7. K Means Clustering

So far, we've looked at the techniques such as Linear Regression / Logistic Regression / Naive Bayes / SVM / etc., which are SUPERVISED LEARNING, where there are Labels (ie) Responses.

But, when we simply have the data alone and No Labels associated with the data set (ie) NO Responses, and we have to organize the huge data set, that becomes UNSUPERVISED LEARNING. For instance, Clustering belongs to Unsupervised Learning technique.

K - Means clustering is One of the simplest and efficient clustering technique, where we basically form K clusters via their means of the Centroid.

Clustering is useful to Detect patterns.

Example : Group emails or Search result,
Customer Shopping patterns,
Regions of images, ...

The Basic idea behind clustering is, Grouping together similar instances. Example : 2D point patterns

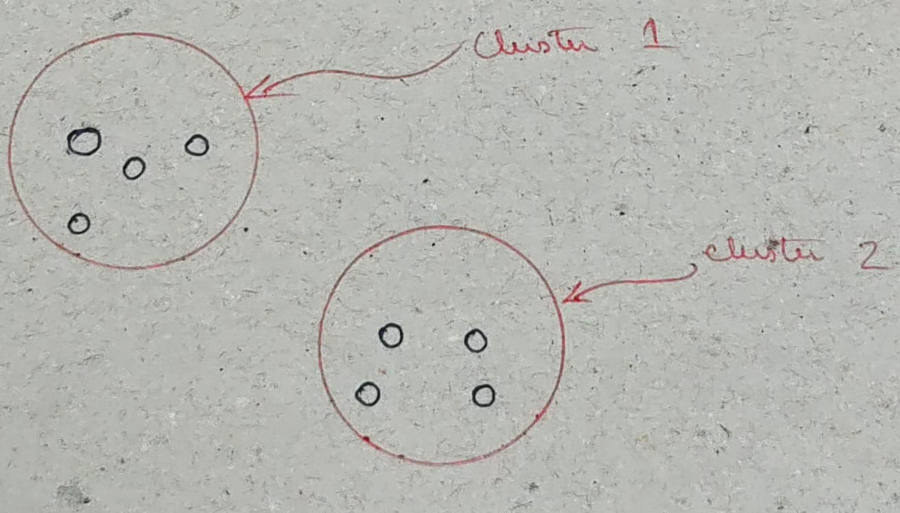


Image Segmentation — The Goal is to Partition an image into perceptually similar regions.

K-Means Algorithm

K-Means is an algorithm for clustering, which divides the data set into K clusters.

organizes

K-Means is an iterative algorithm. Let us consider an Unlabeled data set of n-dimensional vectors.

j^{th} vector → $\bar{x}(j)$

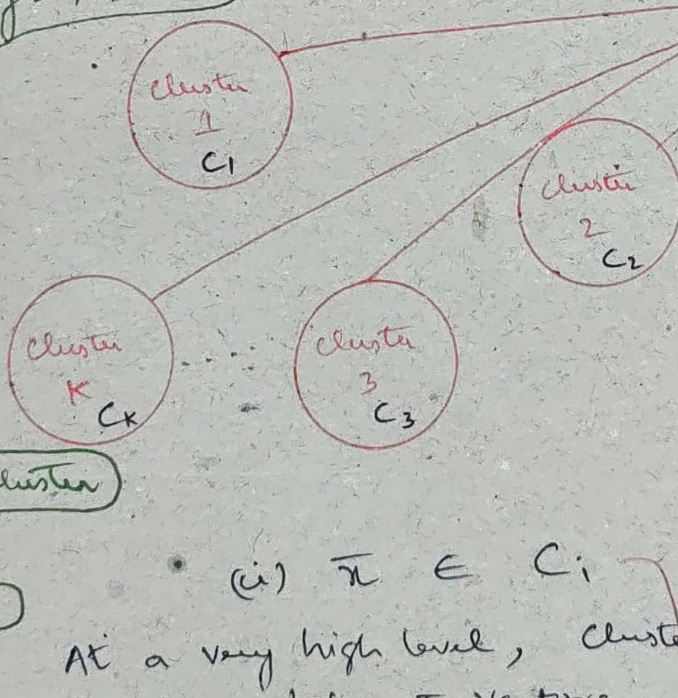
$\bar{x}(1), \bar{x}(2), \ldots, \bar{x}(M)$ → Total of M vectors $j = 1, 2, \ldots M$

And we want to organize the data into K clusters.

where $j = 1, 2, \ldots, M$

K Clusters

$i = 1, 2, \ldots, K$



Let C_i denotes ith cluster. And $\bar{x}$ belongs to a single cluster.

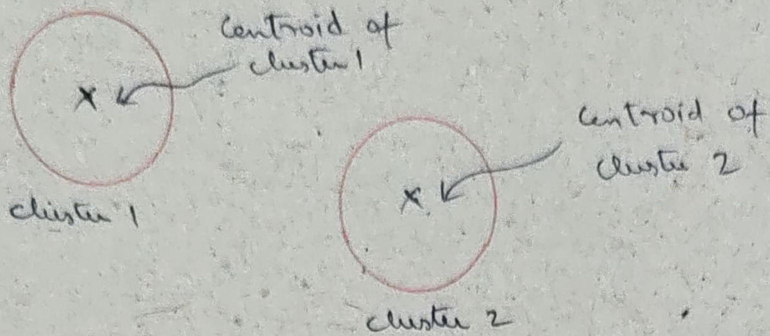
i^{th} cluster → C_i → where $i = 1, 2, \ldots, K$

(i) $\bar{x} \in C_i$

At a very high level, cluster is nothing but Set $\{\ldots\}$ which has different vectors.

The Centroids of the Clusters are

$$\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_K$$



Keep in Mind

M Data Points
K Clusters

j - Data Point Index
i - Cluster index

Cluster assignment

Given a large data set, we have to determine the cluster. Let $\alpha_i(j)$ denote the cluster assignment indicator. We have to assign each point in $\bar{x}$ to a cluster.

$$\alpha_i(j) = \begin{cases} 1 & \text{if } \bar{x}(j) \in C_i \\ 0 & \text{if } \bar{x}(j) \notin C_i \end{cases}$$

(ie) $\alpha_i(j) = \begin{cases} 1 & \text{if } j^{th} \text{ point belong to } i^{th} \text{ cluster} \\ 0 & \text{Else} \end{cases}$

(ii) $\alpha_2(1) = 1 \Rightarrow \bar{x}(1) \in C_2$

(First point belongs to 2^{nd} cluster)

So, for any given j, $\alpha_i(j) = 1$ only for a particular value of i because, any point can only be assigned to a single cluster. So, $\alpha_i(j) \in \{0, 1\}$

Thus, $\sum_{i=1}^{K} \alpha_i(j) = 1$ → Constraint on α

where $\alpha_i(j)$ is a discrete variable

Cost function (How to cluster these points?)

To cluster these points, we construct a cost function.
The K-Means cost function to minimize is given as

$$\min \underbrace{\sum_{j=1}^{M} \underbrace{\sum_{i=1}^{K} \alpha_i(j) \, \|\bar{x}(j) - \bar{\mu}_i\|^2}_{\text{square of distance of } j^{th} \text{ point to its centroid } \because \alpha_i(j)=1 \text{ only for a particular value of } i}}^{\text{sum of squares of distances of all points to respective centroids}}$$

K-Means procedure

There is no Algorithm to solve / to obtain a direct solution. So, the K-Means procedure proceeds iteratively, which is why K-means is called an iterative algorithm.

We randomly initialize centroids $\bar{\mu}_1^{(0)}, \bar{\mu}_2^{(0)}, \ldots, \bar{\mu}_K^{(0)}$ (iteration)

where $\bar{\mu}_i^{(0)}$ denotes i^{th} centroid in iteration 0.

(III) $\bar{\mu}_i^{(l-1)}$ denotes centroid in iteration $l-1$.

Now, we have to determine two things.

(i) $\alpha_i^{(l)}(j) \rightarrow$ cluster Assignment Indicator in iteration l

(ii) $\bar{\mu}_i^{(l)} \rightarrow$ Update the centroid in iteration l.

Let's do this step-by-step.

⊙ In iteration l, for each point $\bar{x}(j)$, perform

$$\min \sum_{i=1}^{K} \alpha_i(j) \, \| \bar{x}(j) - \bar{\mu}_i^{(l-1)} \|^2$$

Here, we have the centroids at iteration $l-1$.
We want to come up with $\alpha_i(j)$ such that for each point $\bar{x}(j)$, this cost is minimized. (i.e.) the Square of distance to its respective centroid is minimized. Then automatically, the sum of the square distance of all the points to their respective centroids will be minimized.

Now, which i^{th} cluster should one assign to the point $\bar{x}(j)$? Here it is very obvious that $\alpha_i(j) = 1$ for only one value of i. (i.e) Each point can belong to one cluster. Naturally, the point j should be assigned to the cluster such that the distance from the centroid of the cluster is minimum. If the distance from the centroid is minimum, naturally the square of the distance will be minimum. Then the sum of the squares from its respective centroid will be minimum.

For instance, I have to assign a particular Hostel to a student, such that the distance to the Hostel is minimum. Similarly; assign $\bar{x}(j)$ to cluster $\tilde{i}$ with closest centroid.

This is minimized when $\alpha_{\tilde{i}}^{(l)}(j) = 1$, where

$$\tilde{i} = \arg\min_{i} \underbrace{\| \bar{x}(j) - \bar{\mu}_i^{(l-1)} \|^2}$$

(i.e) The i^{th} cluster for which the square of the distance of the j^{th} point to its centroid is MINIMUM.

centroid for which the square of the distance is minimum.

(ii) assign $\bar{x}(j)$ to the closest centroid $\bar{\mu}_{\tilde{i}}^{(l-1)}$

(iii) $\alpha_i^{(l)}(j) = \begin{cases} 1 & \text{if } i = \tilde{i} \\ 0 & \text{if } i \neq \tilde{i} \end{cases}$

Do this for each Data point j, $1 \leq j \leq M$

Cluster assignment indicator in iteration l

Remember, we have a total of M points. So, we do this assignment of $\bar{x}(j)$ to the closest centroid for each j, where $j = 1, 2, \ldots, M$.

Week 5 : Session 2

We saw that, the K-Means Clustering algorithm proceeds iteratively. In each iteration, we first determine the cluster assignments. (i) given the centroids from the previous iteration (i) $\bar{\mu}_i^{(l-1)}$, we have derived the cluster assignments $\alpha_i^{(l)}(j) = \begin{cases} 1 & i = \tilde{i} \\ 0 & i \neq \tilde{i} \end{cases}$ ($\bar{x}(j)$ is assigned in l^{th} iteration)

Now, let us determine the centroids in iteration l for the given clusters. For this, in each cluster i, minimize the cost function.

$\|\bar{a}\|^2 = \bar{a}^T\bar{a}$

$\bar{a}^T\bar{b} = \bar{b}^T\bar{a} = \sum_k a_k b_k$

$$\min \sum_{j=1}^{M} \alpha_i^{(l)}(j) \, \|\bar{x}(j) - \bar{\mu}_i\|^2$$

We've to determine $\bar{\mu}_i$ such that this cost function is minimized.

$$\Rightarrow \min \sum_{j=1}^{M} \alpha_i^{(l)}(j) \left(\bar{x}(j) - \bar{\mu}_i\right)^T \left(\bar{x}(j) - \bar{\mu}_i\right)$$

$$\Rightarrow \min \sum_{j=1}^{M} \alpha_i^{(l)}(j) \left(\bar{x}(j)^T\bar{x}(j) - \bar{\mu}_i^T\bar{x}(j) - \bar{x}(j)^T\bar{\mu}_i + \bar{\mu}_i^T\bar{\mu}_i\right)$$

$$\Rightarrow \min \sum_{j=1}^{M} \underbrace{\alpha_i^{(l)}(j)} \left(\bar{x}(j)^T\bar{x}(j) - 2\bar{\mu}_i^T\bar{x}(j) + \bar{\mu}_i^T\bar{\mu}_i\right)$$

Cluster Assignments are known.

Now, determine $\bar{\mu}_i$ for which cost is minimum.

Since we've to determine $\bar{\mu}_i$, we can differentiate w.r.t $\bar{\mu}_i$.
Since $\bar{\mu}_i$ is vector, we've to do vector differentiation, which is Gradient.

To differentiate the functions of vector, we've to take gradient (partial derivative w.r.t each element of that vector).

$$\Rightarrow \min \sum_{j=1}^{M} \alpha_i^{(l)}(j) \left(\bar{x}^T(j)\,\bar{x}(j) + \bar{\mu}_i^T \mu_i - 2\bar{x}^T(j)\,\bar{\mu}_i \right)$$

To achieve minimum or maximum, we've to take gradient and set equal to zero. (i.e.) $\frac{\nabla f}{\bar{\mu}_i} = \begin{bmatrix} \partial f / \partial \bar{\mu}_1 \\ \partial f / \partial \bar{\mu}_2 \\ \vdots \end{bmatrix}$.

$$\Rightarrow \sum_{j=1}^{M} \alpha_i^{(l)}(j) \left(0 + 2\bar{\mu}_i - 2\bar{x}(j) \right) = 0$$

$$\Rightarrow \sum_{j=1}^{M} \alpha_i^{(l)}(j) \left(\bar{\mu}_i - \bar{x}(j) \right) = 0$$

$$\Rightarrow \bar{\mu}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j) \cdot \bar{x}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)} = \frac{\text{Sum of all points in cluster } i}{\text{Total number of points in cluster } i}$$

$$\Rightarrow \bar{\mu}_i^{(l)} = \frac{\sum_{j : \bar{x}(j) \in C_i} \bar{x}(j)}{\sum_{j : \bar{x}(j) \in C_i} 1} = \text{Average/Mean of all points assigned to cluster } i \text{ in iteration } l$$

Centroid of cluster i in iteration l

Point j lies in cluster i

Stopping Criterion:

Stop when clusters are STABLE.

(i.e) When cluster assignments do NOT change

(ii) Centroids do not change significantly.

K-Means Example.

Iteration 0

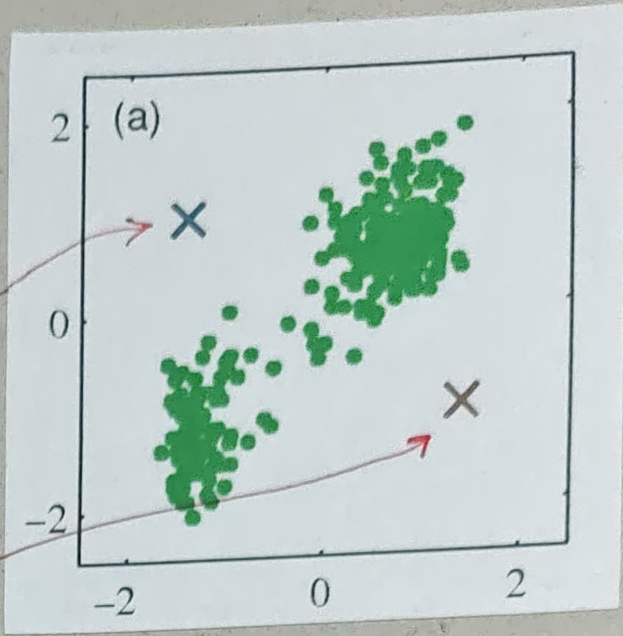
- Pick K random points as cluster Centroids.
- Shown here for $K=2$

$\bar{\mu}_1^{(0)}$

Centroid of cluster 1 in iteration 0

$\mu_2^{(0)}$

centroid of cluster 2 in iteration 0

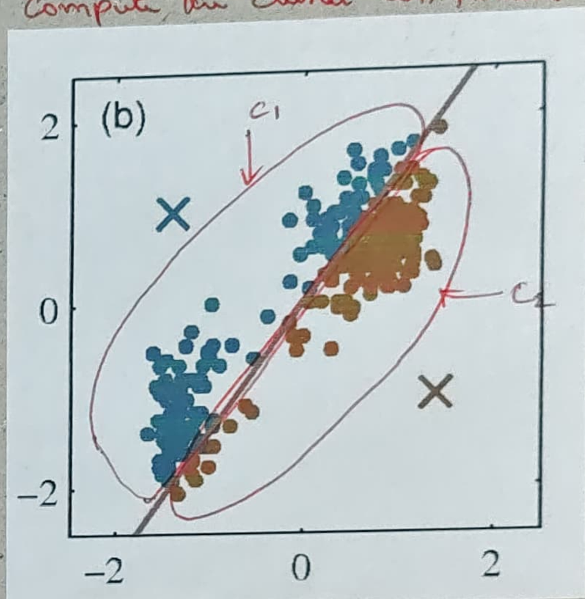


Randomly Pick the Centroids.

Iteration 1

- Assign data points to closest centroid

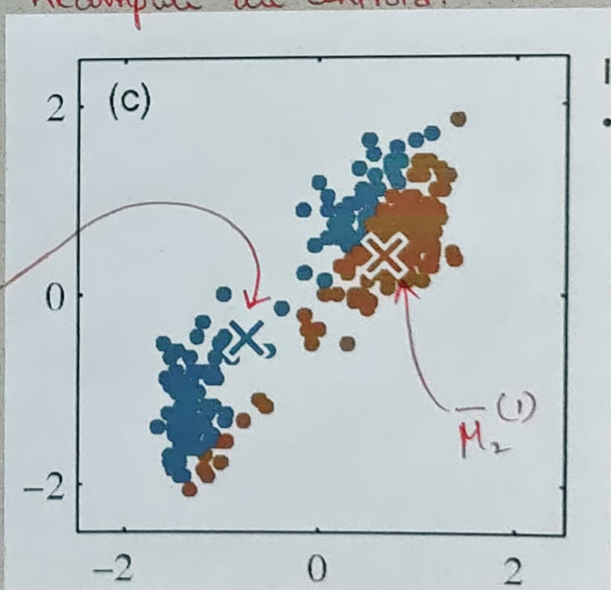
Compute the cluster assignment



Recompute the centroid.

- Change each centroid to the Average/Mean of the assigned points,

$\bar{\mu}_1^{(1)}$

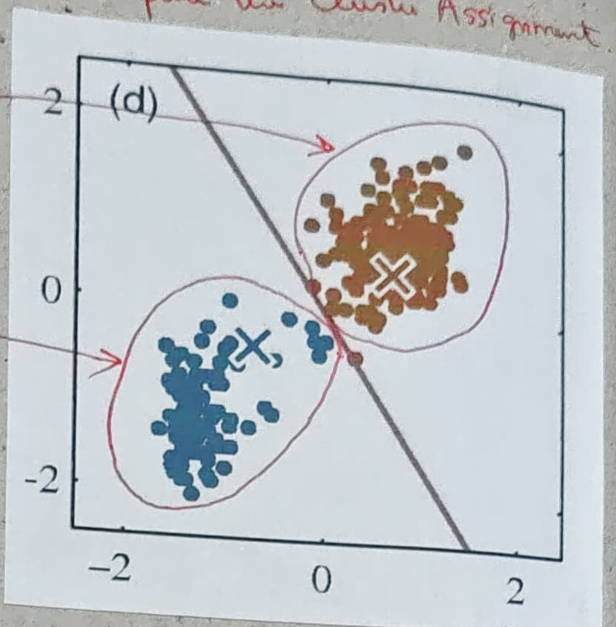


$\bar{\mu}_2^{(1)}$

Iteration 2

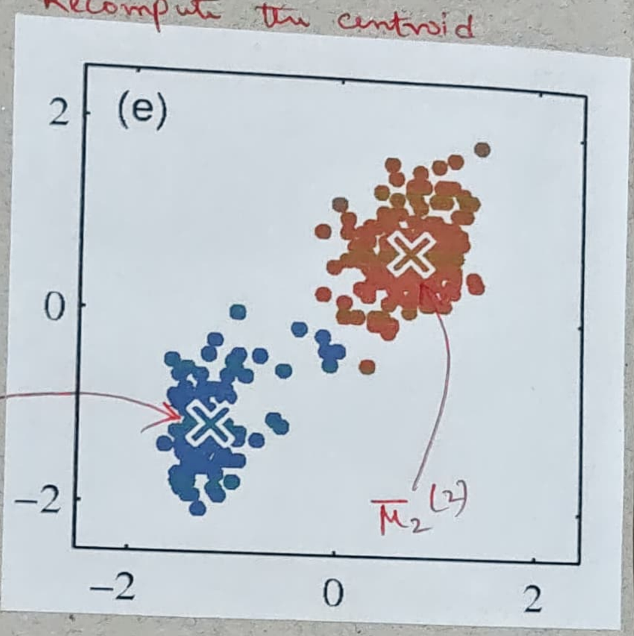
- Assign data points to closest centroid

Recompute the cluster Assignment



- Change each centroid to the Average / Mean of the assigned points.

Recompute the centroid



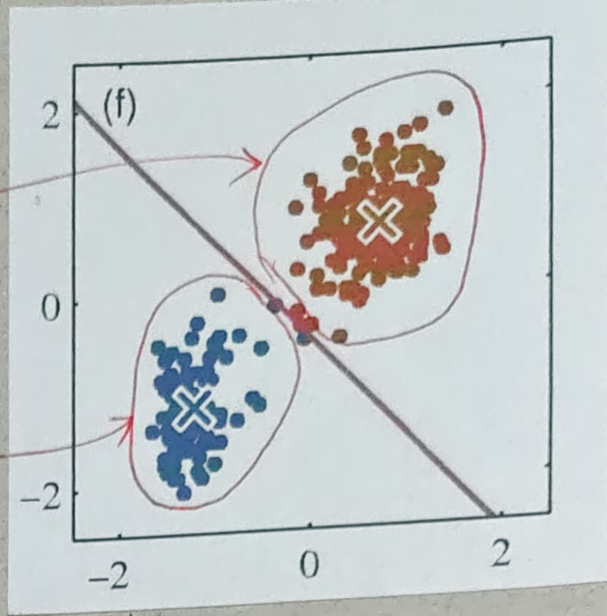
Iteration 3.

(i) Assign data points to closest centroid

Recompute the cluster Assignment

C_2

C_1



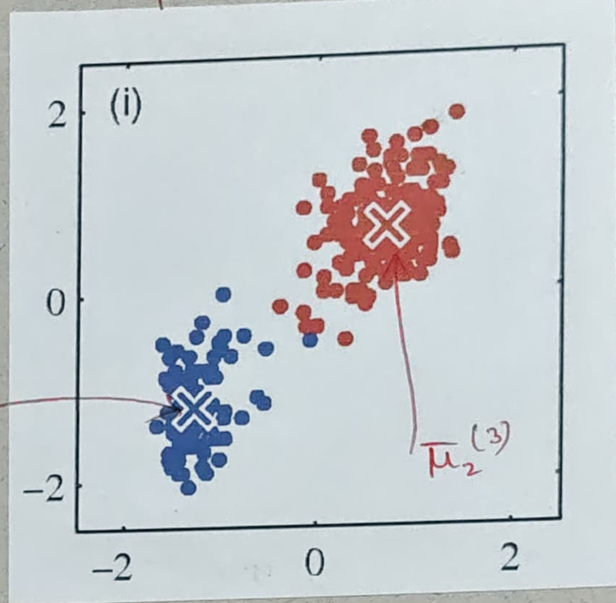
(ii) Change each centroid to the Average / Mean of the assigned points.

(iii) Here, Centroids / clusters have NOT changed significantly from previous iteration.

⇒ Convergence achieved.

Recompute the Centroid

$\bar{\mu}_1^{(3)}$



$\bar{\mu}_2^{(3)}$

(i.e.) We see that the clusters / centroid are Stable (Not changed significantly from the previous location. So, we can say that "Convergence achieved!".

This ML algorithm can be used for practical applications, for instance, "Placing Base Stations in a Wireless Cellular Network".

(i.e.) How to decide where to place the Base Stations, given the density of the subscribers.