8. K-Means Clustering

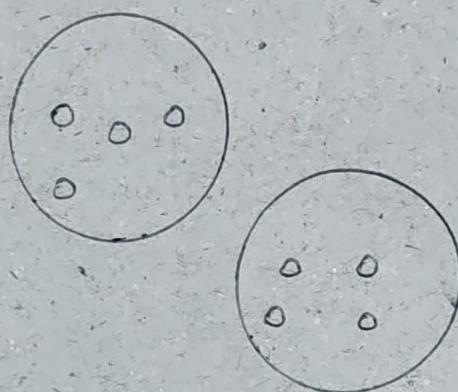
Clustering

- Unsupervised Learning (Requires data, but NO Labels)
- Detect patterns.

  Example:
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- Basic idea : Group together similar instances.

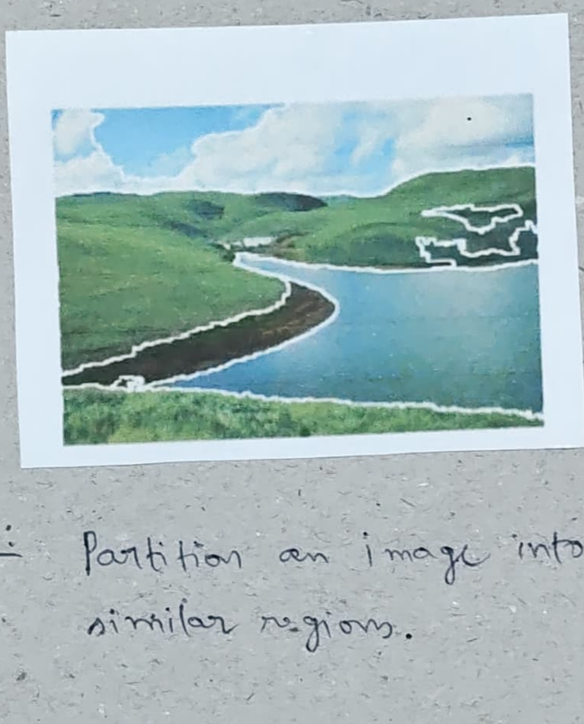
  Example:

  2D point patterns.



- Clustering Example

  Image Segmentation :



Goal : Partition an image into perceptually similar regions.

K-Means formulation

- K-Means is an iterative algorithm
- Consider the dataset of n-dimensional vectors.

  $\bar{x}(1), \bar{x}(2), \ldots, \bar{x}(M)$
- organize the data into K clusters

  $C_1, C_2, \ldots, C_K$
- The centroids for the clusters are

  $\bar{\mu}_1, \bar{\mu}_2, \ldots, \bar{\mu}_K$
- Let $\alpha_i(j)$ denote the cluster assignment indicator

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

- $\alpha_i(j)$ should satisfy the below property.

  $\alpha_i(j) = 1$, only for one i

  Since $\sum_i \alpha_i(j) = 1$.
- The K-mean cost-function to minimize is given as

$$\min \sum_{i=1}^{K} \sum_{j=1}^{M} \underbrace{\alpha_i(j)}_{\text{Cluster Assignment Indicator}} \underbrace{\left\| \bar{x}(j) - \bar{\mu}_i \right\|^2}_{\text{Square of the distance b/w 2 points}}$$

$K \rightarrow$ No. of clusters

$M \rightarrow$ No. of vector points.

K - Means procedure.

- $\bar{\mu}_i^{(\ell)}$ denotes centroids in iteration ℓ
- Initialize centroids randomly

$$\bar{\mu}_1^{(0)}, \bar{\mu}_2^{(0)}, \ldots, \mu_K^{(0)}$$

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

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

- Cluster determination

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

$$\tilde{i} = \arg\min \left\| \bar{x}(j) - \bar{\mu}_i^{(\ell-1)} \right\|^2$$

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

Problem 1: Given the data $\bar{x}^{(1)} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$ and

centroids $\bar{\mu}_0 = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$, $\bar{\mu}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Determine the cluster assignments.

- Distance to centroid 0 $= \sqrt{1^2 + 1^2} = \sqrt{2}$
- Distance to centroid 1 $= \sqrt{2^2 + 4^2} = \sqrt{20}$

Hence, assigned to cluster 0, as distance is minimum.

- $\alpha_0(1) = 1$, $\alpha_1(1) = 0$.

- Centroid determination

Next, determine the centroids for the given clusters.

For this, in each cluster i, minimize

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

This can be expanded as

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

Taking the gradient and setting to zero yields

$$\bar{\mu}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\,\bar{x}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}$$

$$= \frac{\sum_{j:\bar{x}(j) \in C_i} \bar{x}(j)}{\sum_{j:\bar{x}(j)\in C_i} 1}$$

(ie) average of all points assigned to cluster i in iteration l.

Problem 2 : Given the data below, determine the Centroids

$$\bar{x}^{(1)} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}, \quad \bar{x}^{(2)} = \begin{bmatrix} -3 \\ -1 \end{bmatrix} \in C_0$$

$$\bar{x}^{(3)} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \bar{x}^{(4)} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \in C_1$$

$$\bar{\mu}_0 = \frac{\bar{x}^{(1)} + \bar{x}^{(2)}}{2} = \begin{bmatrix} -2 \\ -3/2 \end{bmatrix}$$

$$\bar{\mu}_1 = \frac{\bar{x}^{(3)} + \bar{x}^{(4)}}{2} = \begin{bmatrix} 5/2 \\ 3 \end{bmatrix}$$

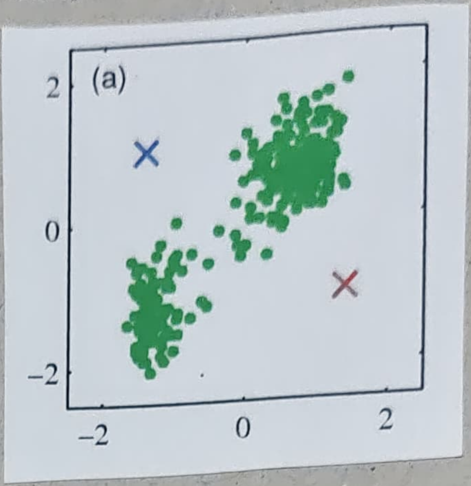
Stopping Criterion

- Stop when clusters are stable.
  (ii) When cluster assignments do NOT change.

K-means : Example

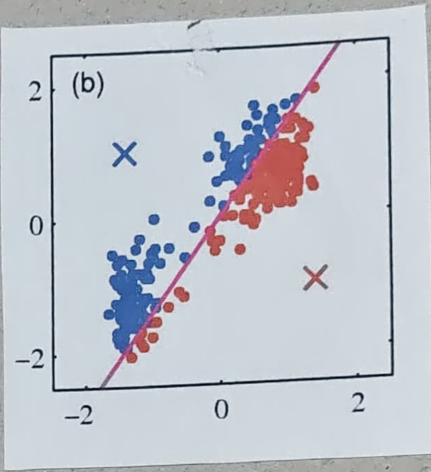
⊙ Pick K random points as cluster centroids.

Shown here for K = 2.



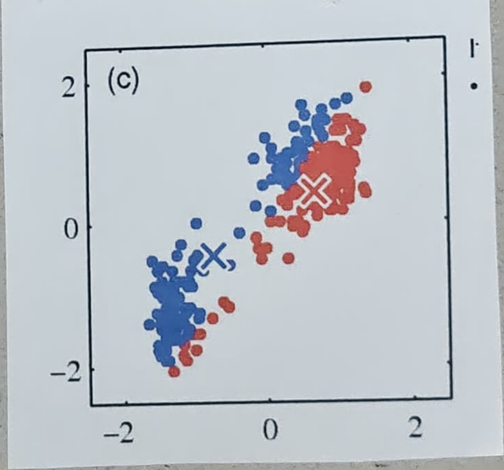
⊙ Iteration # 1

Assign data points to closest centroid.



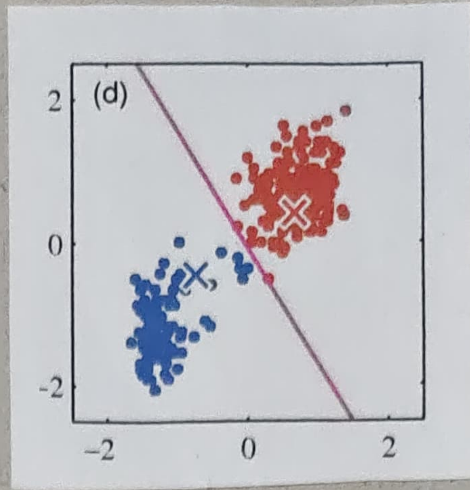
⊙ Iteration # 1

Change each centroid to the average of the assigned points



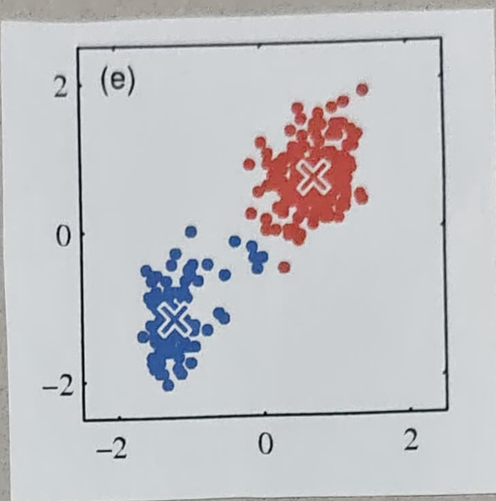
- Iteration # 2

  Assign data points to closest centroid



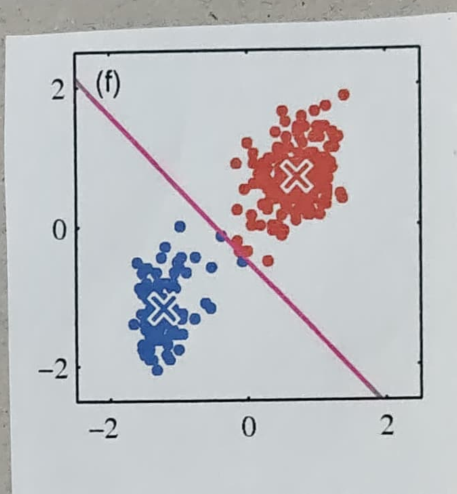
- Iteration # 2'

  Change each centroid to the average of the assigned points.



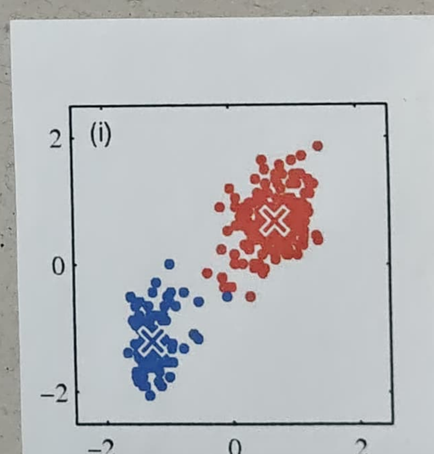
- Iteration # 3

  Assign data points to closest centroid



- Iteration # 3'

  Change each centroid to the average of the assigned points.



- CONVERGENCE ACHIEVED !