ML - 7

Machine Learning

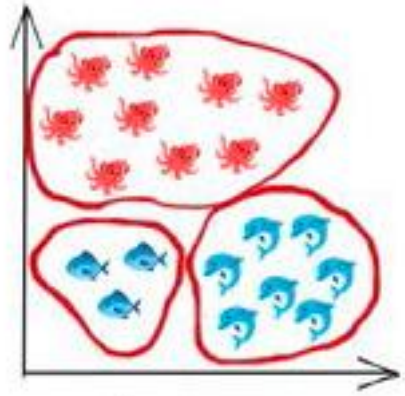
- Unsupervised Learning
  - Distance based grouping
    - K-Means Clustering

Prof. Umut ORHAN

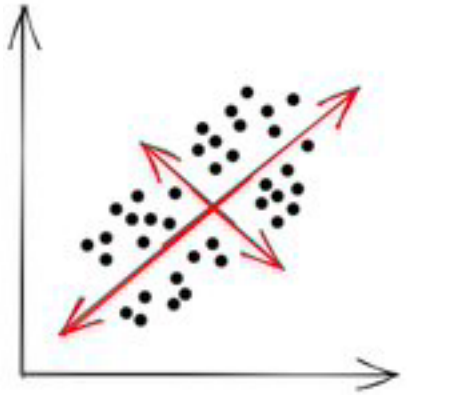
Unsupervised Learning

Unsupervised learning is a branch of machine learning that focuses on discovering the structure of the problem that generates the data, including hidden patterns, groups, or features, using unlabeled data.

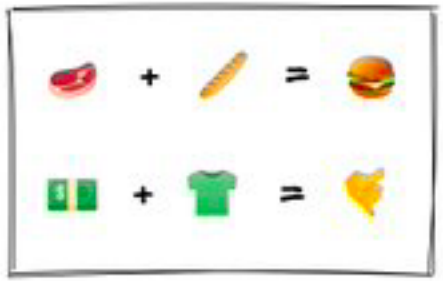
1. Clustering
2. Dimensionality Reduction
3. Association Rules



Clustering



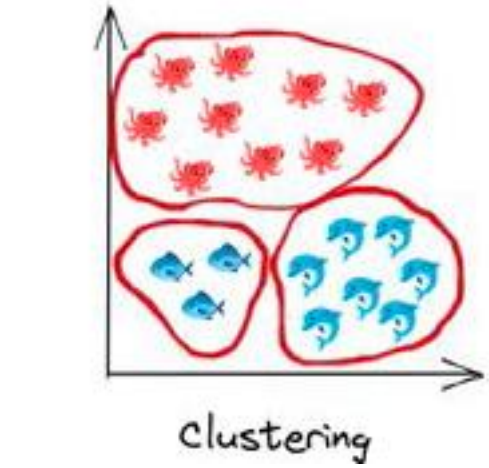
Dimensionality Reduction

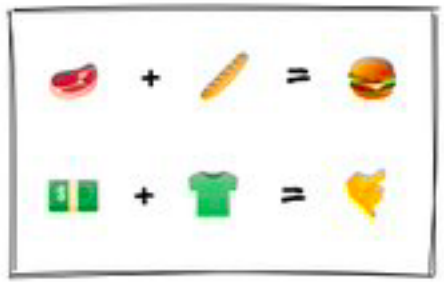


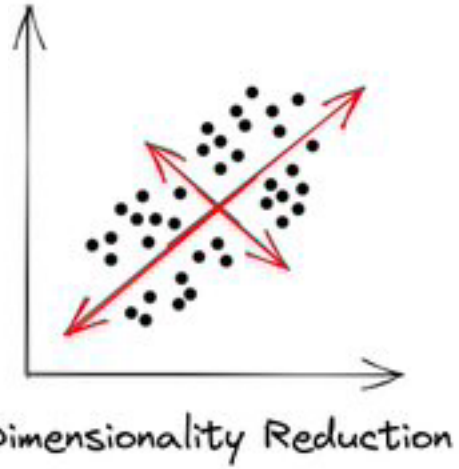
Association Rule

Unsupervised Learning

1. Clustering
   + K-means
2. Dimensionality Reduction
   + Principal Component Analysis (PCA)
3. Association Rules
   + Apriori

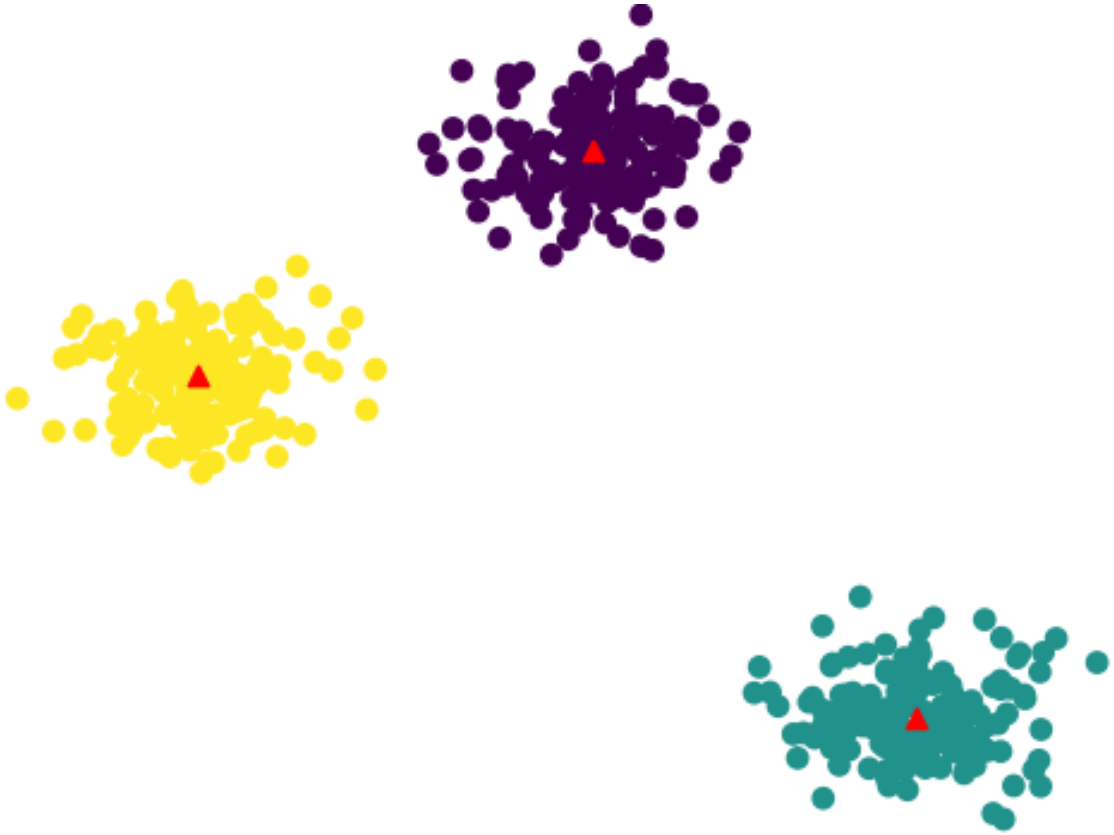


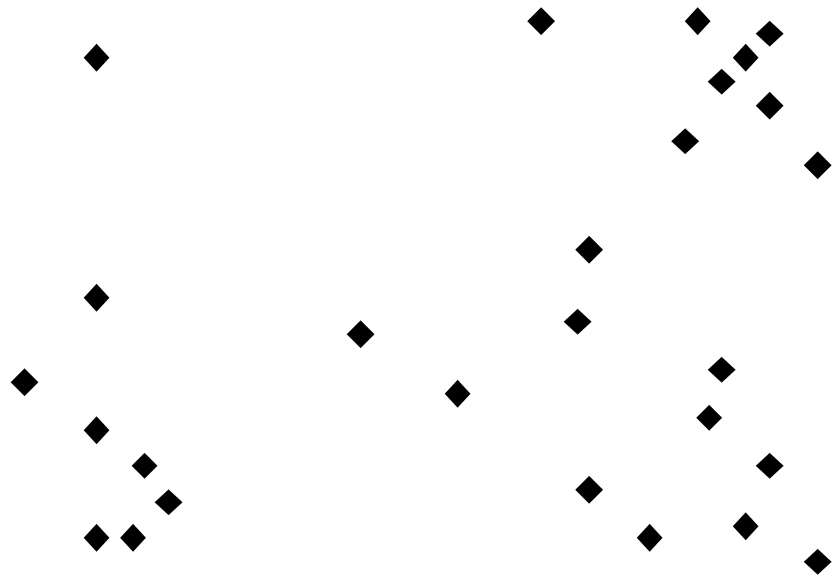




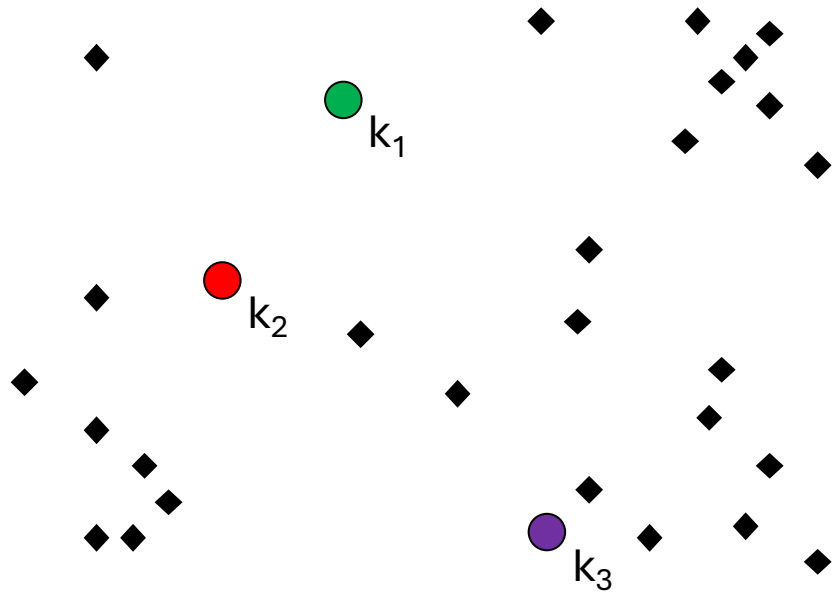
K-means Algorithm

1. Select K random center points
2. Assign each data point to the nearest center
3. Update each center to the mean of its assigned data points
4. Repeat steps 2 and 3 until convergence is achieved

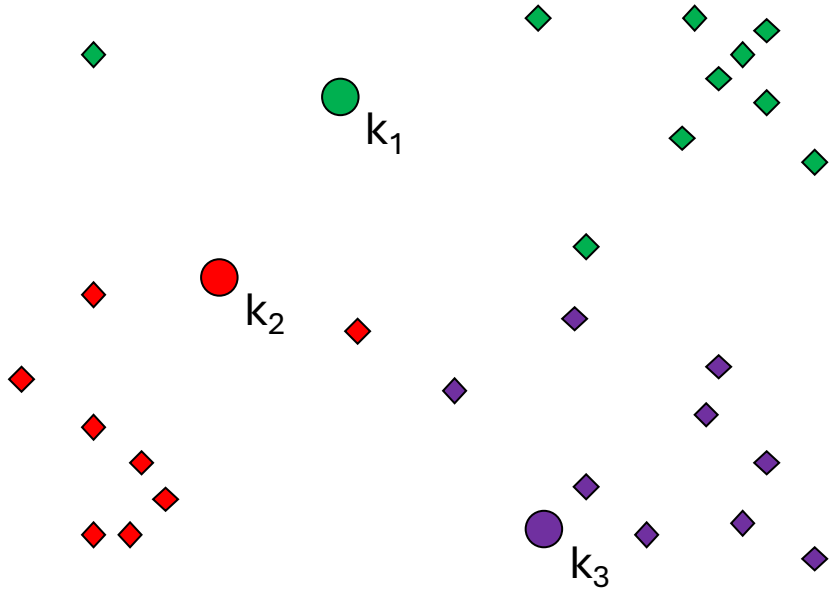




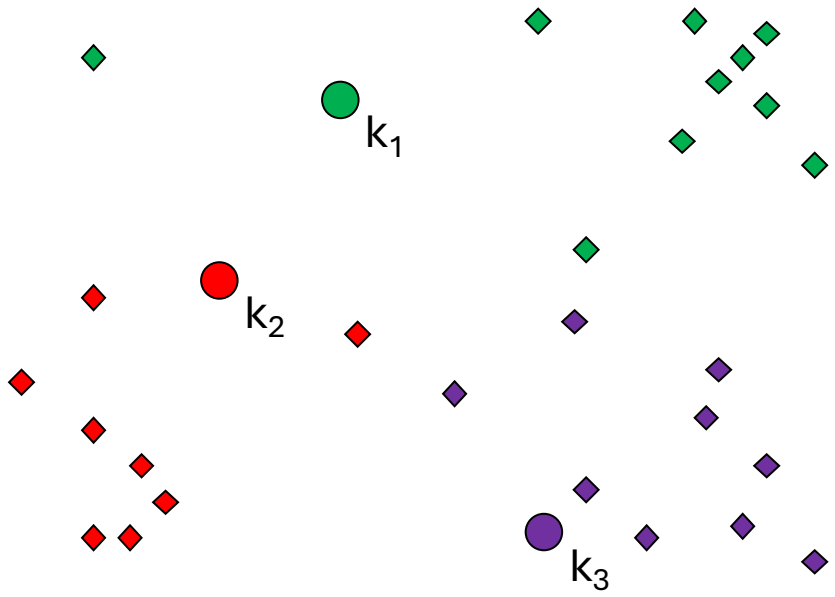
K-means Algorithm



K-means Algorithm



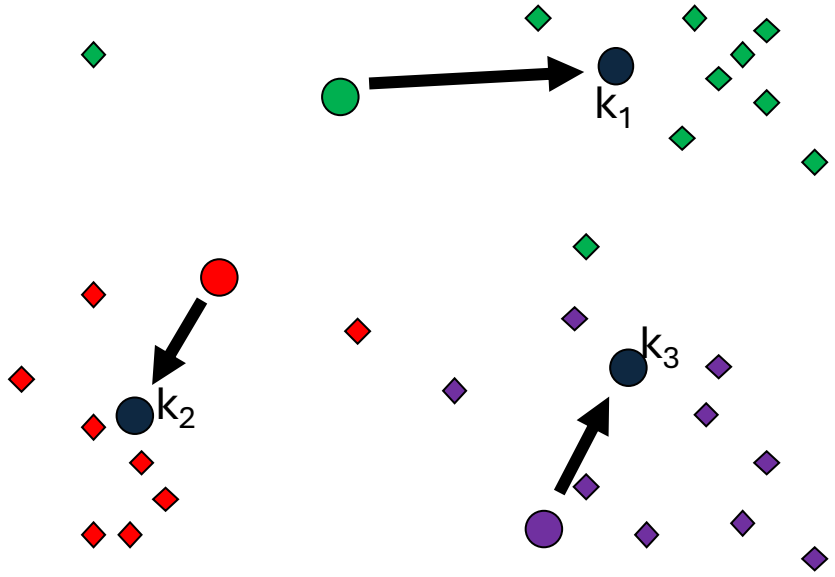
K-means Algorithm



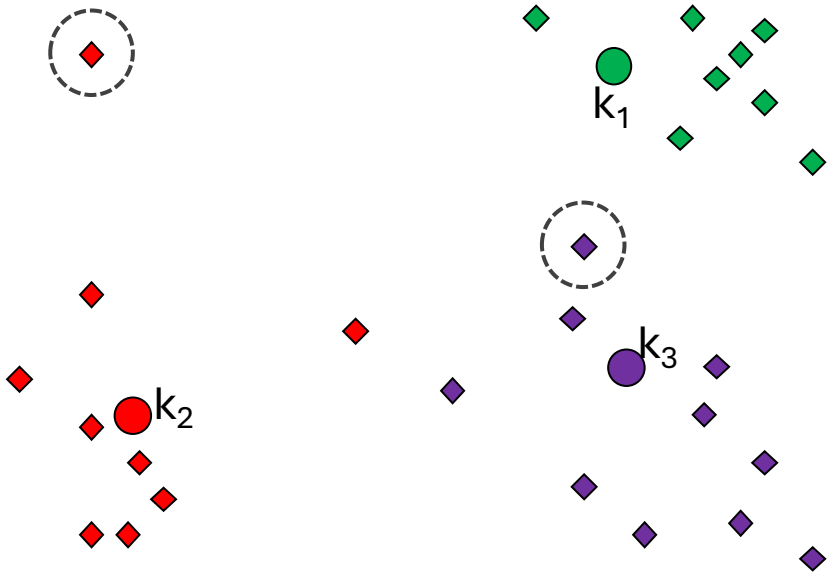
Cost function:

$$J = \sum_{i=1}^{K} \left(\sum_{m:x_m \in C_i} \|x_m - c_i\|^2 \right)$$

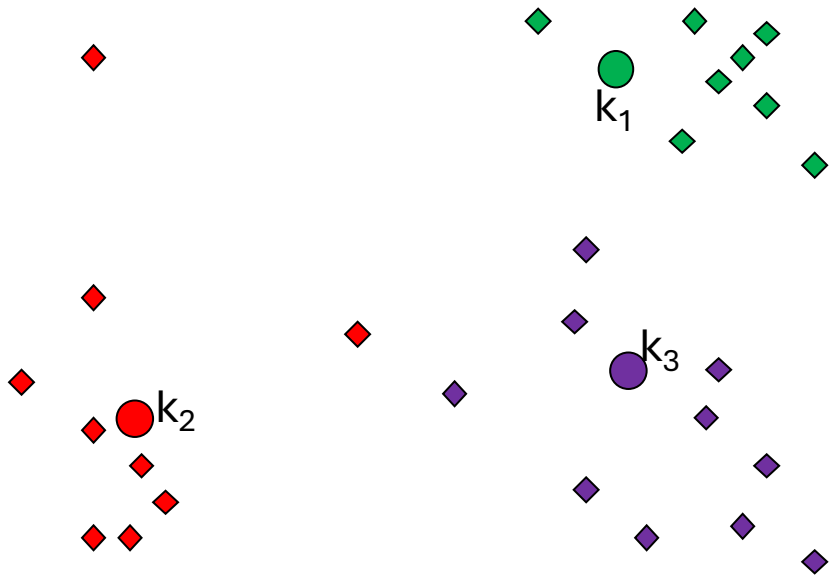
K-means Algorithm



K-means Algorithm



K-means Algorithm



Cost function:

$$J = \sum_{i=1}^{K} \left(\sum_{m:x_m \in C_i} \|x_m - c_i\|^2 \right)$$

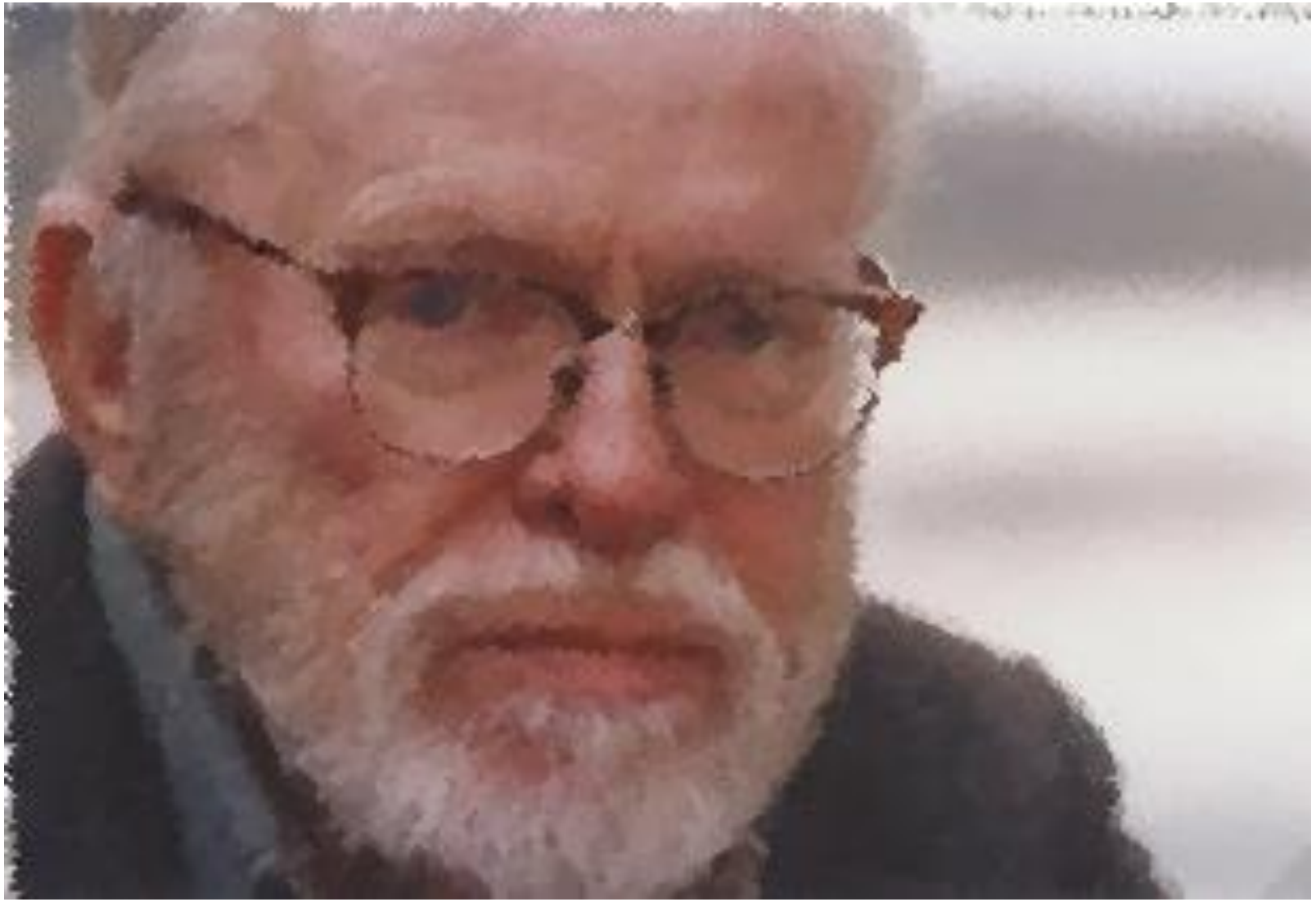
K-means Algorithm

K-means Cost function

Cost function:

$$J = \sum_{i=1}^{K} \left(\sum_{m:x_m \in C_i} \|x_m - c_i\|^2 \right)$$

Here, x_m is each data point belonging to cluster C_i and c_i is the center point of C_i .



James B. MacQueen

(1929-2014)

K-means Advantages

1. Simplicity and Understandability
2. Fast Performance
3. Low Memory Usage
4. Adaptability
5. Ease of Parameter Tuning
6. General Applicability

K-means Disadvantages

1. Pre-Specified K Value
2. Sensitivity to Initial Centroids
3. Cluster Shape Assumptions
4. Sensitivity to Outliers
5. Equal Cluster Distribution Assumption
6. Convergence to Local Minima

Manual Calculations

	X_1	X_2
1st	2	3
2nd	5	4
3rd	0	5
4th	-3	0

Using the Manhattan distance and the initial center points provided below, find the new positions of the cluster centers after the first iteration.

Assume the initial centers are $c_1 = (0, 1)$ and $c_2 = (5, 5)$

Manual Calculations

	X_1	X_2
1st	2	3
2nd	5	4
3rd	0	5
4th	-3	0

Using the Manhattan distance and the initial center points provided below, find the new positions of the cluster centers after the first iteration.

Assume the initial centers are $c_1 = (0, 1)$ and $c_2 = (5, 5)$

At first, calculate all distances:

$d_{11}=|0-2| + |1-3|=4$ $d_{12}=|5-2|+|5-3|=5$ (it belongs to c_1)

$d_{21}=|0-5| + |1-4|=8$ $d_{22}=|5-5|+|5-4|=1$ (it belongs to c_2)

$d_{31}=|0-0| + |1-5|=4$ $d_{32}=|5-0|+|5-5|=5$ (it belongs to c_1)

$d_{41}=|0-(-3)| + |1-0|=4$ $d_{42}=|5-(-3)|+|5-0|=13$ (it belongs to c_1)

Now, we can update the cluster centers

$c_{1\text{-New}}$ = mean or average of x_i values belongs to $c_1 = (-1/3 , 8/3)$

$c_{2\text{-New}}$ = mean or average of x_i values belongs to $c_2 = (5, 4)$

Machine Learning

7. week

Thanks for watching