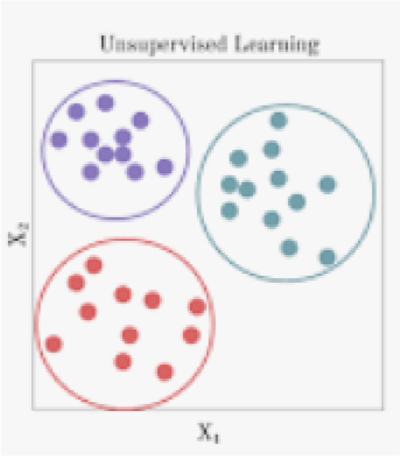
Lecture 18
Artificial Intelligence

Khola Naseem
khola.naseem@uet.edu.pk



Unsupervised learning

Credit: Khola Naseem

Machine learning

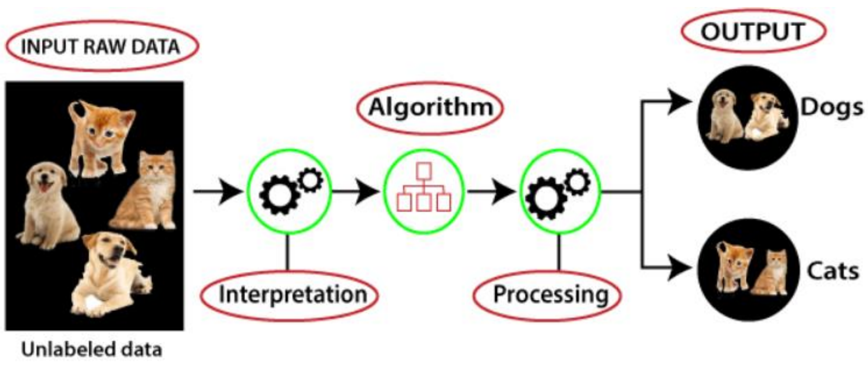
➢Unsupervised learning

- ➢ Unsupervised learning refers to the use of artificial intelligence (AI) algorithms to identify patterns in data sets containing data points that are neither classified nor labeled.
- ➢ The algorithms are thus allowed to classify, label and/or group the data points contained within the data sets without having any external guidance in performing that task.
- ➢ The algorithms analyze the underlying structure of the data sets by extracting useful information or features from them.

Machine learning

➢Unsupervised learning

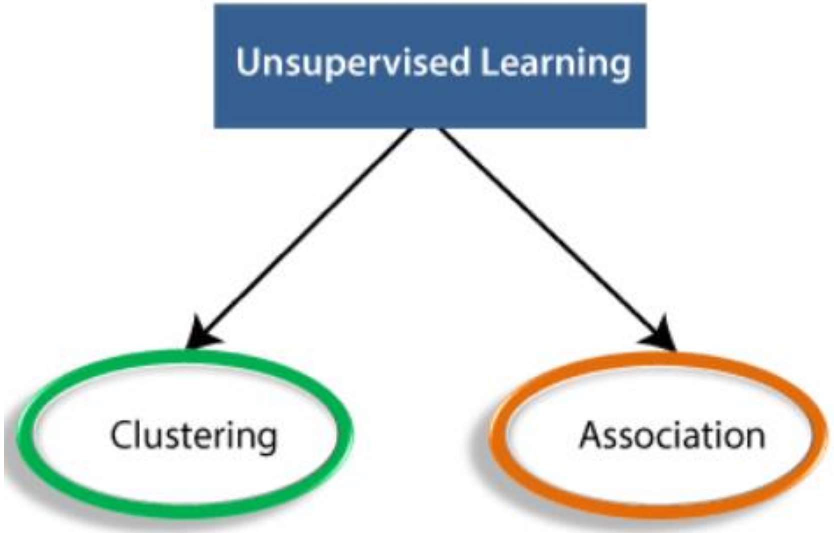
➢ **Example:** Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm has no idea about the features of the dataset. The task of the algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images



Credit: Khola Naseem

Machine learning

- Unsupervised learning
-



Machine learning

- Unsupervised learning
  - **Clustering**: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group.
  - **Association**: is used for finding the relationships between variables in the large database.
    - It determines the set of items that occurs together in the dataset. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

Machine learning

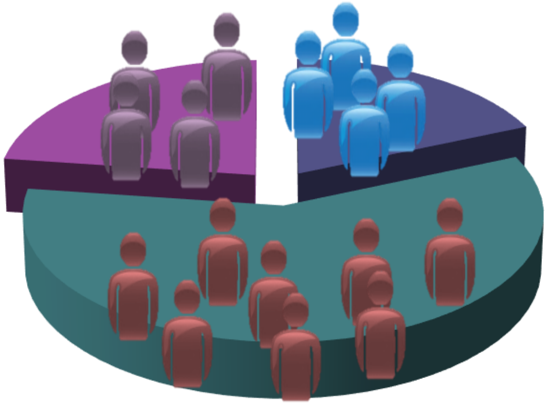
Unsupervised learning

Clustering Applications

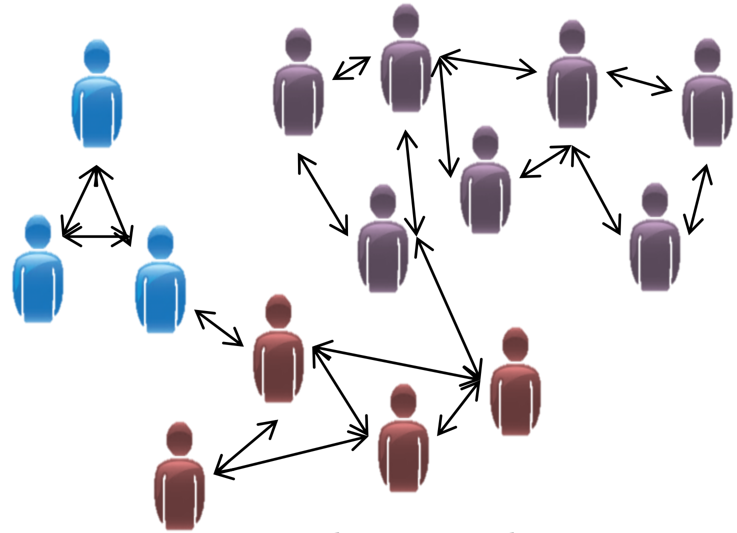
- Document clustering = (webpages, news, blogs, tweets, sentiments of products, etc)
- Cluster the documents, sentiments or short text segments(Helpful for analyzing what are big topics in the collection)
- Image Clustering = Cluster the images into groups that have similar visual contents
- Multimedia Clustering = Voice, Videos, Music

Unsupervised learning

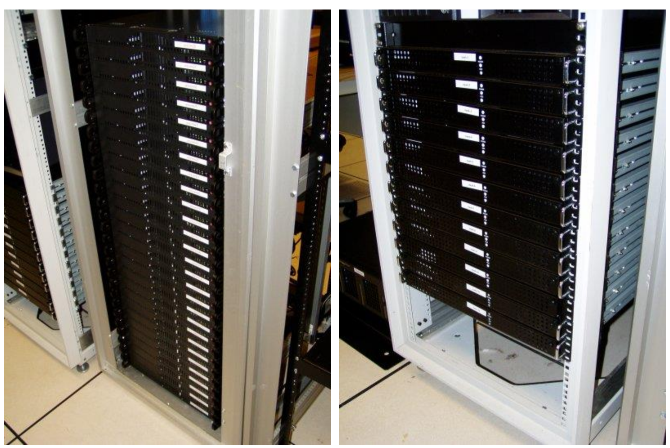
Clustering Applications



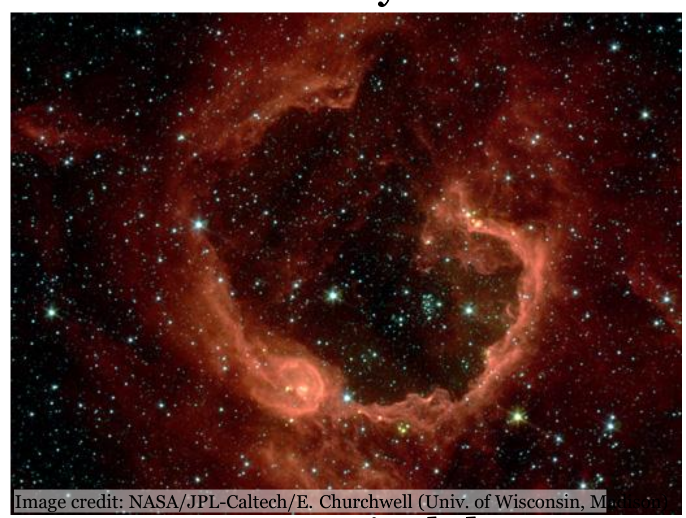
Market segmentation



Social network analysis



Organize computing clusters



Astronomical data analysis

Credit: Khola Naseem

➢Unsupervised learning

Clustering examples

Image segmentation

Goal: Break up the image into meaningful or perceptually similar regions

Machine learning

Unsupervised learning

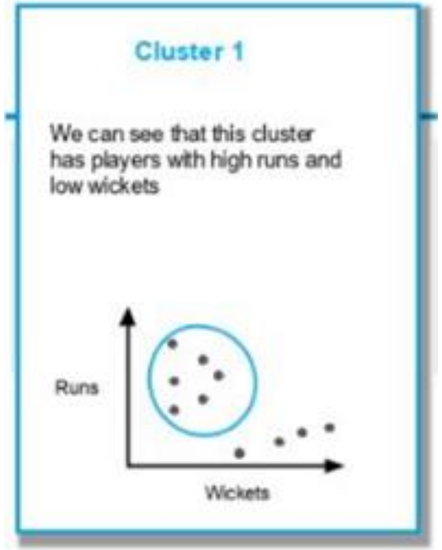
Clustering Applications

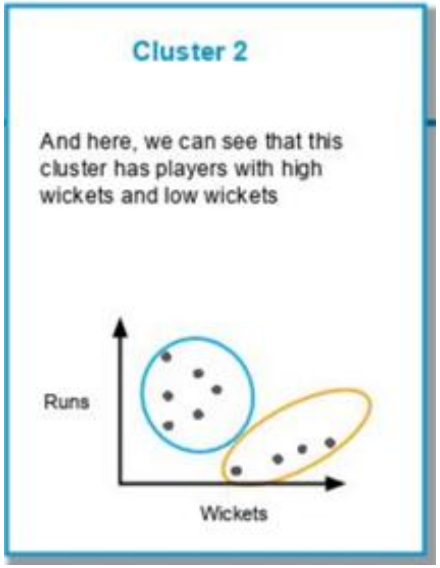
- Document clustering = (webpages, news, blogs, tweets, sentiments of products, etc)
- Cluster the documents, sentiments or short text segments(Helpful for analyzing what are big topics in the collection)
- Image Clustering = Cluster the images into groups that have similar visual contents
- Multimedia Clustering = Voice, Videos, Music

➢Unsupervised learning

➢ K-means clustering:

- ➢ Example:
- ➢ Imagine you received data on a lot of cricket players from all over the world, which gives information on the runs scored by the player and the wickets taken by them in the last ten matches. Based on this information, we need to group the data into two clusters, namely batsman and bowlers.



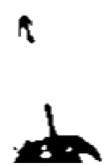


Credit: Khola Naseem

➢Unsupervised learning

- **K-means clustering:**
  - Example:

$$\underbrace{\left.\begin{matrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & & & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN}\end{matrix}\right\} M \text{ data items}}_{N \text{ features}}$$



- We just have a data matrix of data items of N features each, with M records
- Task of unsupervised learning is to find structure in data of this type

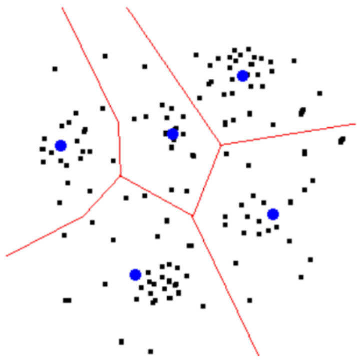
Unsupervised learning

- **K-means clustering:**
  - Example:

- Is there any structure in these data items?
- Yes. Data does not seem random.
- How many groups?
- 2

Unsupervised learning

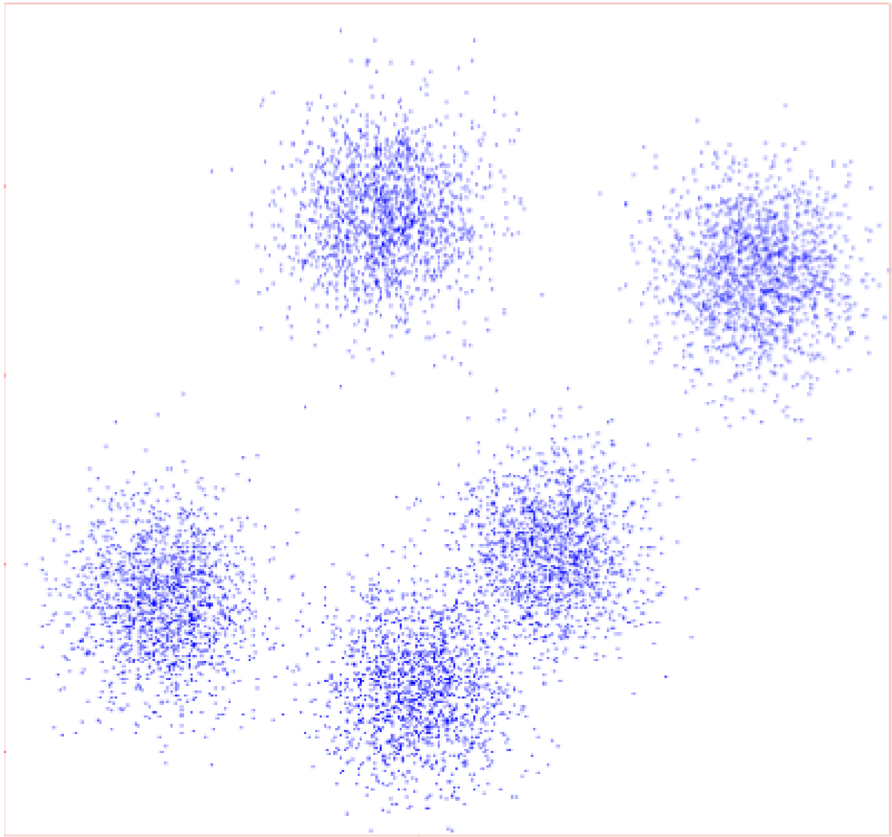
- **K-means clustering:** Working:



  - Step-1: Choose the value of K to decide the number of clusters.
  - Step-2: Select random K points or centroids.
  - Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
  - Step-4: Calculate the variance and place a new centroid of each cluster.
  - Step-5: Repeat the third steps, which means reassigning each data point to the new closest centroid of each cluster.
  - Step-6: If any reassignment occurs, then go to step-4 else you can end it.

K-Means

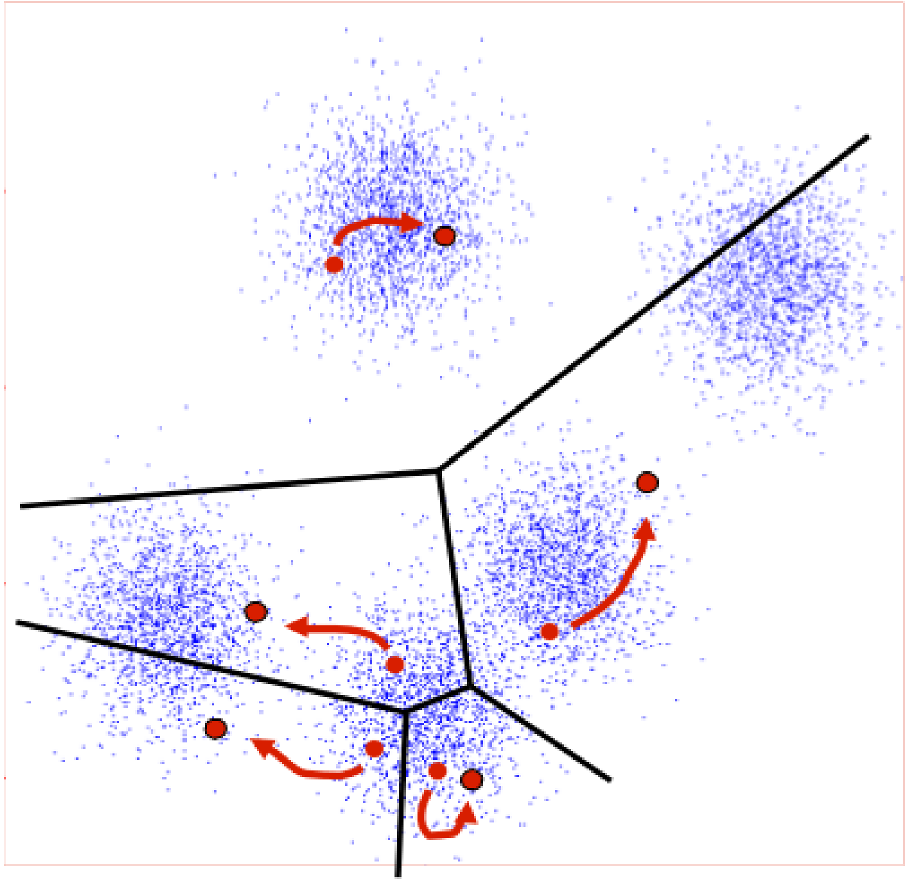
- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



Credit: Khola Naseem

K-Means

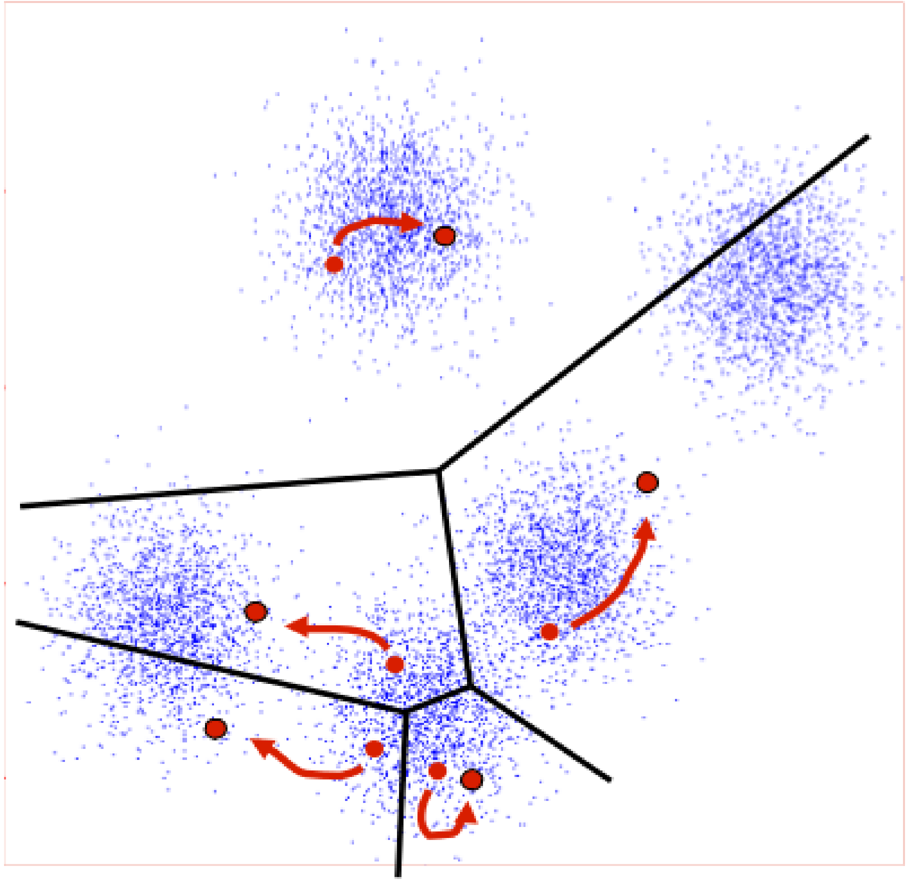
- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



Credit: Khola Naseem

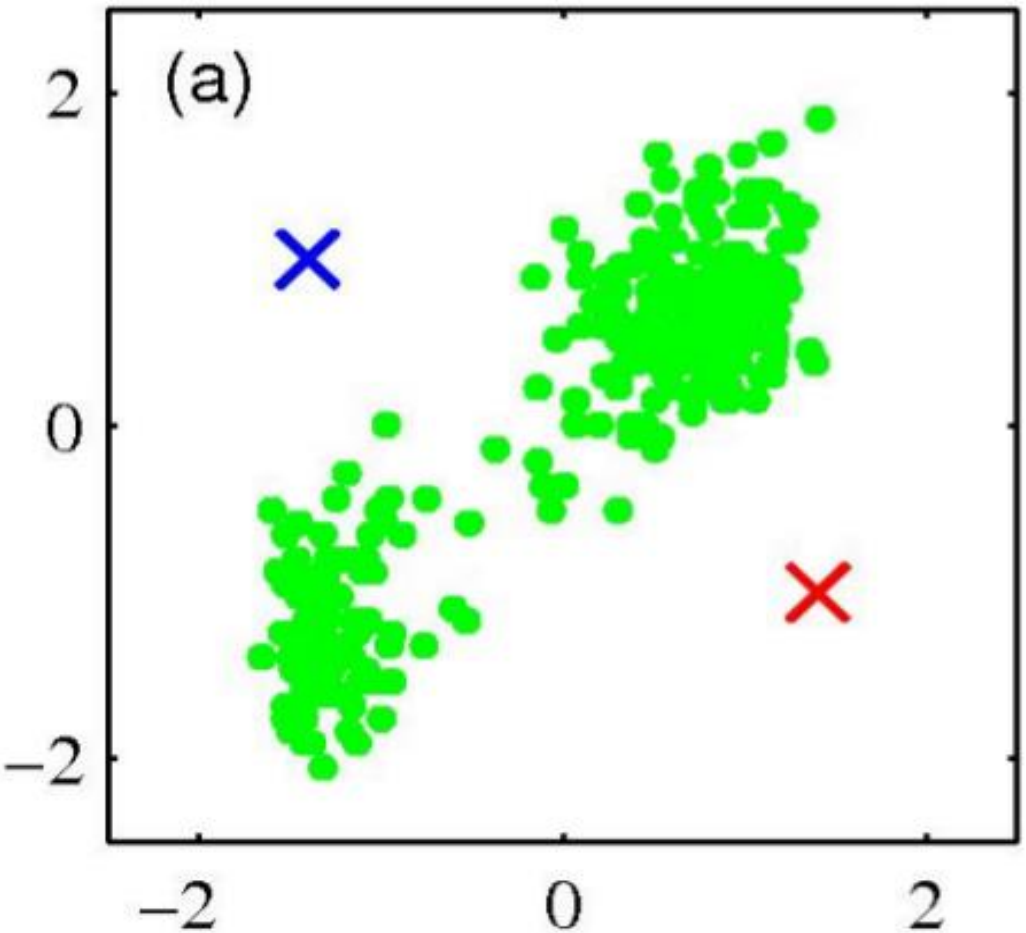
K-Means

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



Credit: Khola Naseem

K-means clustering: Example

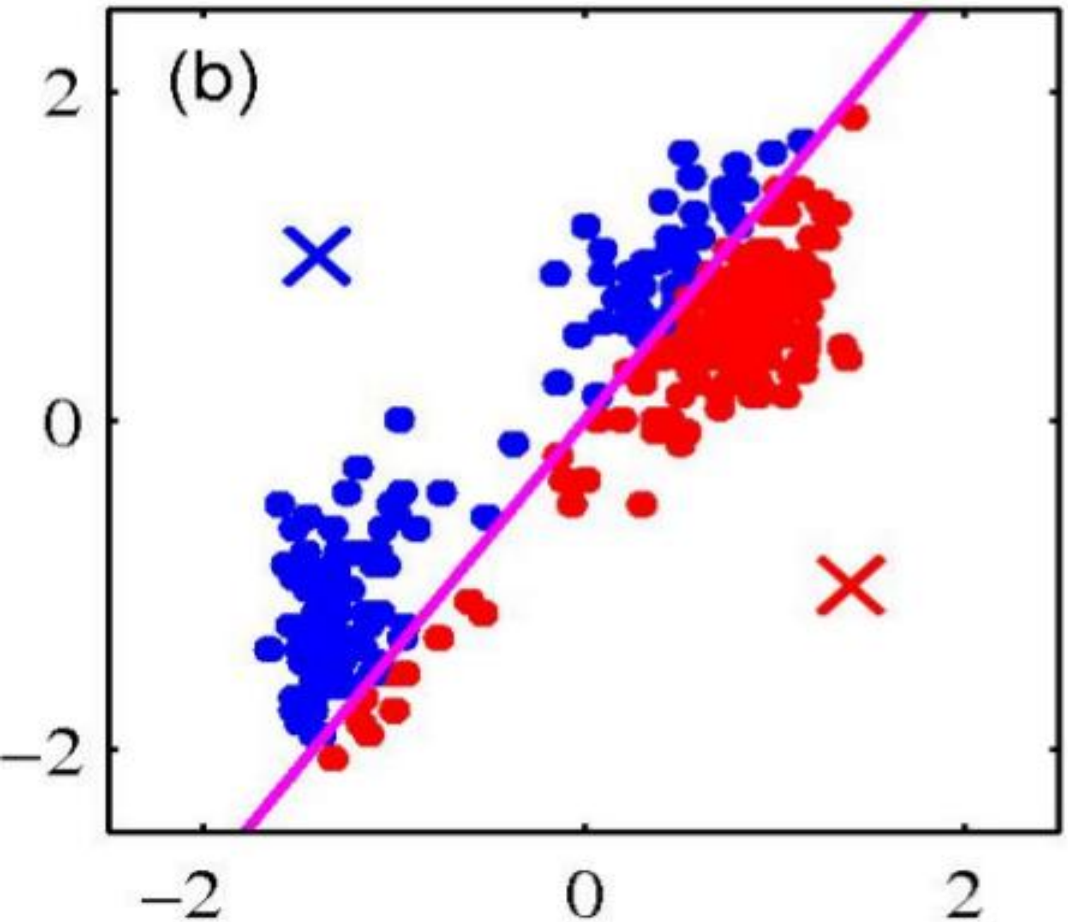


- Pick *K* random points as cluster centers (means)

Shown here for *K*=2

➢Unsupervised learning

K-means clustering: Example



Iterative Step 1

- Assign data points to closest cluster center

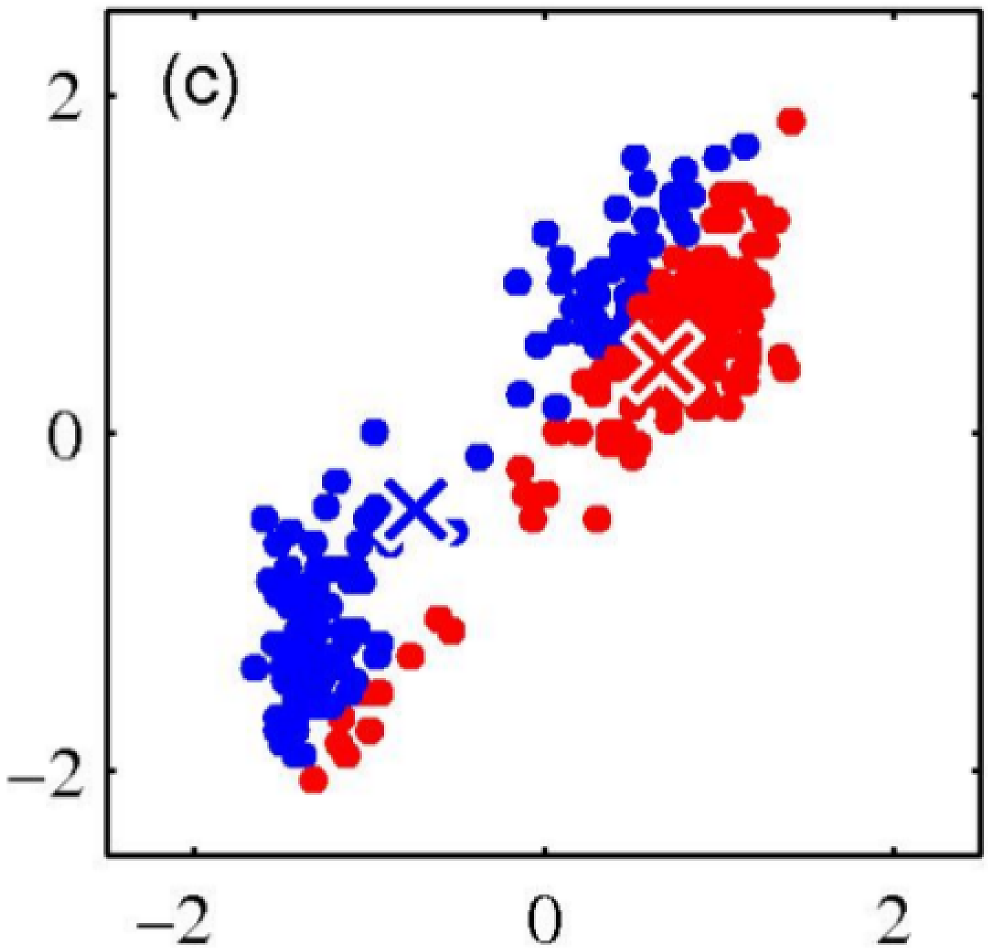
Two Steps:

1. Cluster Assignment
2. Move centroid Step

Credit: Khola Naseem

➢Unsupervised learning

K-means clustering: Example



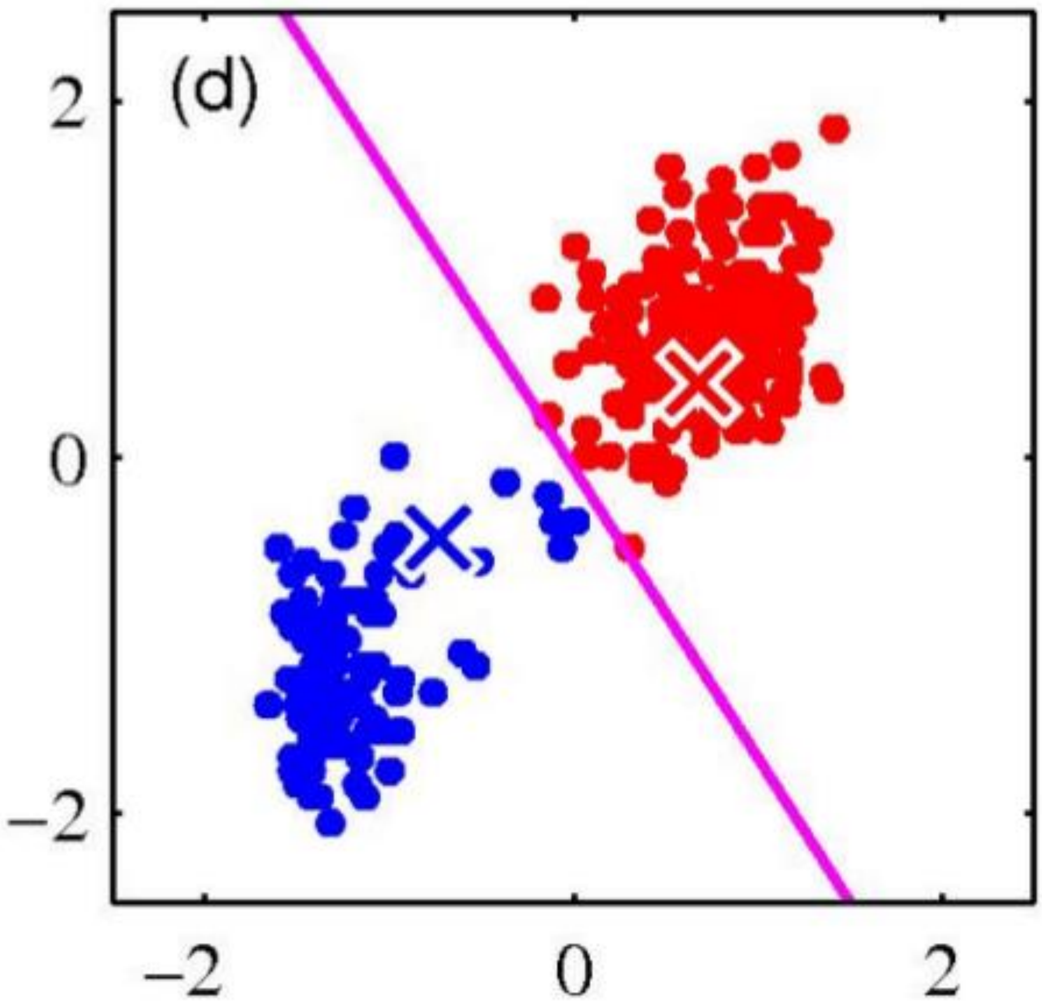
Iterative Step 2

- Change the cluster center to the average of the assigned points

Two Steps:

1. Cluster Assignment
2. Move centroid Step

K-means clustering: Example

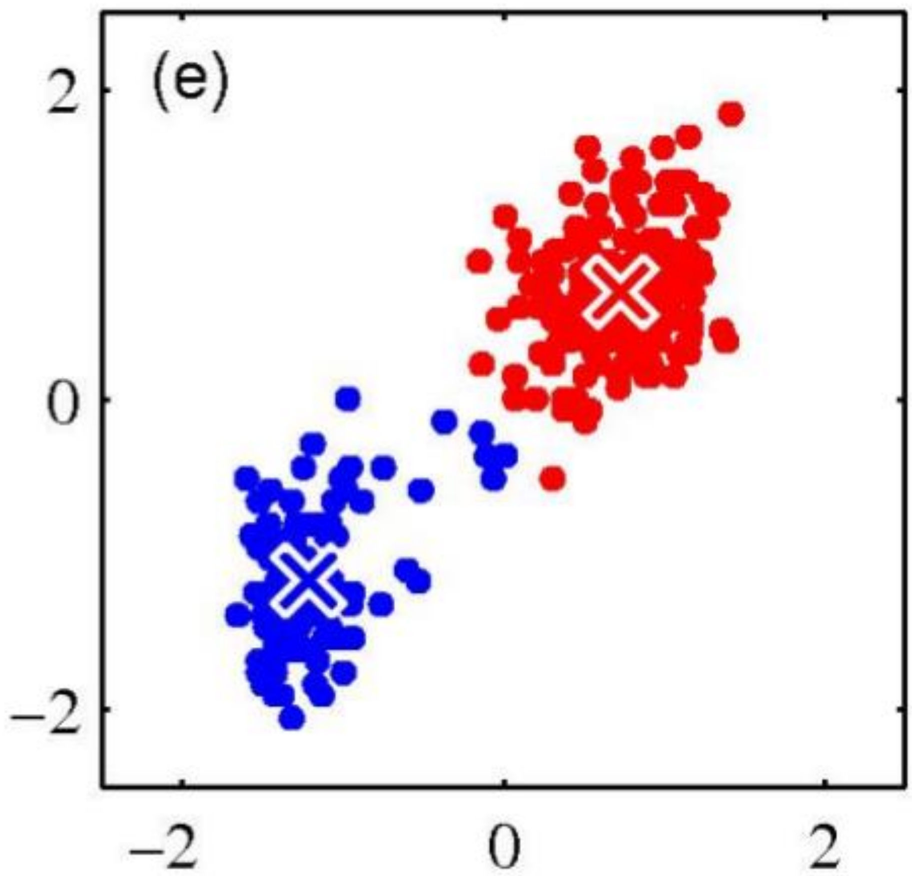


- Repeat until convergence

Two Steps:
1. Cluster Assignment
2. Move centroid Step

➢Unsupervised learning

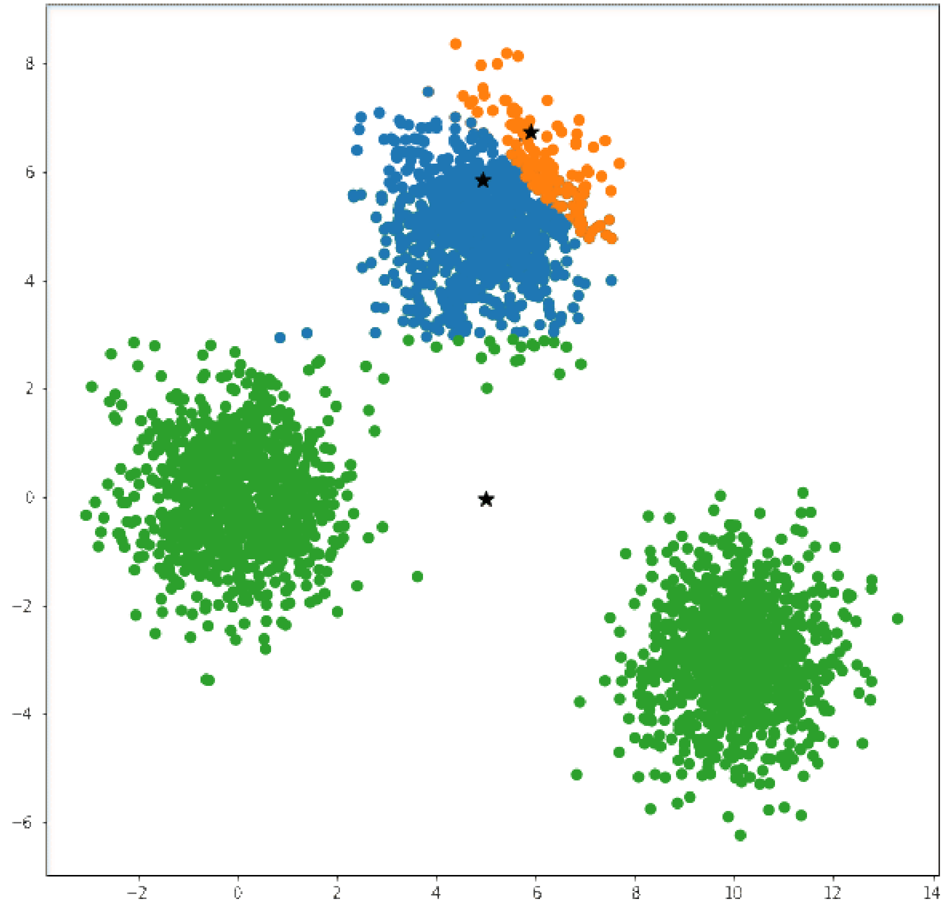
K-means clustering: Example



Two Steps:
1. Cluster Assignment
2. Move centroid Step

➢Unsupervised learning

➢ **K-means clustering:** Working:



Credit: Khola Naseem

Unsupervised learning

K-means clustering:

- Distance:
  - Euclidian Distance $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- Mean point:
  - (Mean of xs , mean of ys)
  - $\left(\frac{x_1+x_2+\cdots+x_n}{n}, \frac{(y_1+y_2+\cdots+y_n)}{n}\right)$

Unsupervised learning

- **K-means clustering: Example:**
- Cluster the following eight points (with (x, y) representing locations) into three clusters:
- A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)
- Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2)

➢Unsupervised learning

- **K-means clustering: Example:**
- The distance may be calculated either by:
- The Manhattan distance is given by:

$$d(x,y)=\sum_{i=1}^{p}|x_i-y_i|$$

- 2. The Euclidean distance (also called taxicab norm or 1-norm) is given by:

$$d(x,y)=\sqrt[2]{\sum_{i=1}^{p}|x_i-y_i|^2}$$

➢Unsupervised learning

- **K-means clustering: Example:**
- Cluster the following eight points (with (x, y) representing locations) into three clusters:
- A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)
- Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2)
- The Manhattan function between two points a = (x1, y1) and b = (x2, y2) is defined as
- $P(a, b) = |x2 - x1| + |y2 - y1|$

Unsupervised learning

- **K-means clustering: Example:**
- Calculating Distance Between A1(2, 10) and C1(2, 10)-
  - P(A1, C1)
  - $= |x2 - x1| + |y2 - y1|$
  - $= |2 - 2| + |10 - 10|$
  - $= 0$
- Calculating Distance Between A1(2, 10) and C2(5, 8)-
  - P(A1, C2)
  - $= |x2 - x1| + |y2 - y1|$
  - $= |5 - 2| + |8 - 10|$
  - $= 3 + 2$
  - $= 5$

Unsupervised learning

- **K-means clustering: Example:**
- First cluster contains points-
  - A1(2, 10)
- Cluster-02:
  - A3(8, 4)
  - A4(5, 8)
  - A5(7, 5)
  - A6(6, 4)
  - A8(4, 9)
- Third cluster contains points-
  - A2(2, 5)
  - A7(1, 2)

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (5, 8) of Cluster-02	Distance from center (1, 2) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

Unsupervised learning

- **K-means clustering: Example:**
- Now,
- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.
- For Cluster-01, We have only one point A1(2, 10) in Cluster-01.
- So, cluster center remains the same.
- Center of Cluster-02
  - $= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)$
  - $= (6, 6)$
- Center of Cluster-03
  - $= ((2 + 1)/2, (5 + 2)/2)$
  - $= (1.5, 3.5)$

Unsupervised learning

- **K-means clustering: Example:**
- This is completion of Iteration-01.
- Iteration-02:
  - We calculate the distance of each point from each of the center of the three clusters.
  - The distance is calculated by using the given distance function.
  - The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Unsupervised learning

K-means clustering: Example:

Given Points	Distance from center (2, 10) of Cluster-01	Distance from center (6, 6) of Cluster-02	Distance from center (1.5, 3.5) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	8	7	C1
A2(2, 5)	5	5	2	C3
A3(8, 4)	12	4	7	C2
A4(5, 8)	5	3	8	C2
A5(7, 5)	10	2	7	C2
A6(6, 4)	10	2	5	C2
A7(1, 2)	9	9	2	C3
A8(4, 9)	3	5	8	C1

- Now,
- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

Unsupervised learning

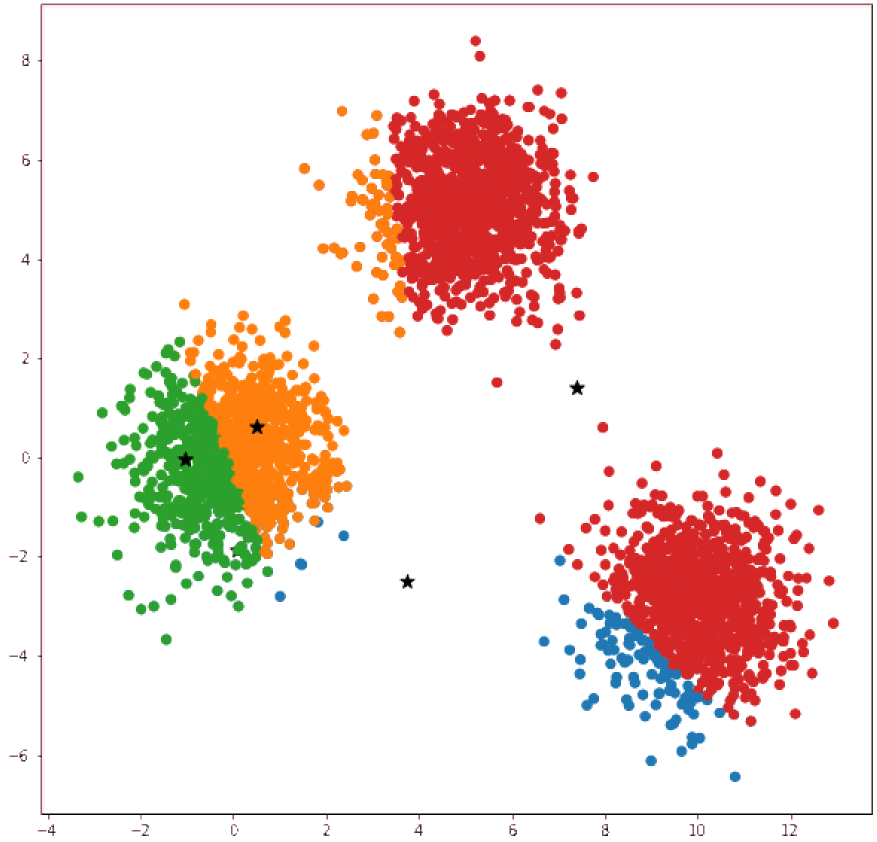
- **K-means clustering: Example:**
- Center of Cluster-01
  - $= ((2 + 4)/2, (10 + 9)/2)$
  - $= (3, 9.5)$
- Center of Cluster-02
  - $= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)$
  - $= (6.5, 5.25)$
- Center of Cluster-03
  - $= ((2 + 1)/2, (5 + 2)/2)$
  - $= (1.5, 3.5)$
- This is completion of Iteration-02.
- After second iteration, the center of the three clusters are-
  - C1(3, 9.5)
  - C2(6.5, 5.25)
  - C3(1.5, 3.5)
- Iteration -03: Repeat the process

➢Unsupervised learning

- **K-means clustering Problem:**
- Dependence on Initialization
  - A big problem with k-means clustering is its consistency. k-means isn't very consistent due to its random initialization. The algorithm can produce very different results depending on where you initialize the cluster centers.
- We Don't Always Know K
  - Since this is unsupervised learning we might not know the real number of clusters. This will also result in erroneous cluster centers.

➢Unsupervised learning

➢ K-means clustering Problem:

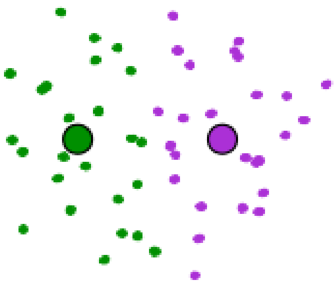


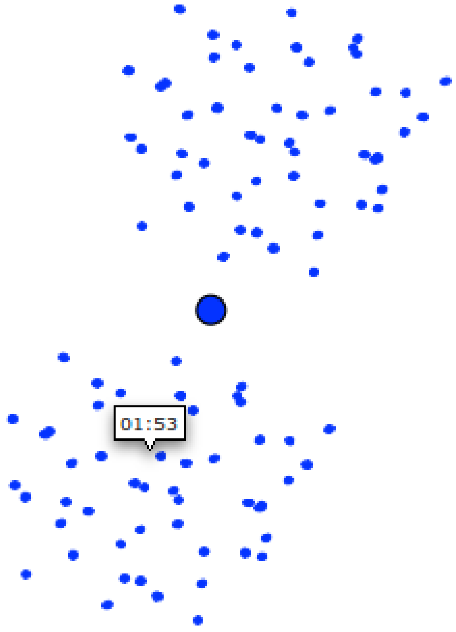
➢Unsupervised learning

➢ **K-means clustering Problem:**

K-Means Getting Stuck

A local optimum:





➢Unsupervised learning

- **K-means clustering Problem:**
- Solution:
- How to choose the value of "K number of clusters" in K-means Clustering
- The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. There are some different ways to find the optimal number of clusters
- The **Elbow method** is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares. The formula to calculate the value of WCSS (for 3 clusters) is given below:

$$\text{WCSS}= \sum P_{i \text{ in Cluster1}} \text{distance}(P_i\ C_1)^2 + \sum P_{i \text{ in Cluster2}} \text{distance}(P_i\ C_2)^2 + \sum P_{i \text{ in CLuster3}} \text{distance}(P_i\ C_3)^2$$

➢Unsupervised learning

- **K-means clustering Problem:**

- Solution:

- To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance

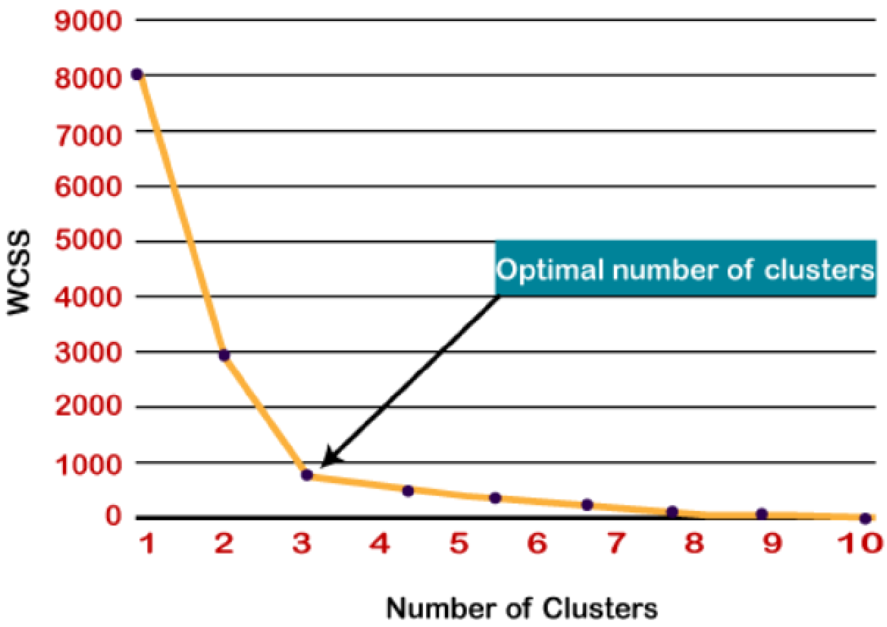
$$\text{WCSS} = \sum_{P_{i \text{ in Cluster1}}} \text{distance}(P_i\, C_1)^2 + \sum_{P_{i \text{ in Cluster2}}} \text{distance}(P_i\, C_2)^2 + \sum_{P_{i \text{ in CLuster3}}} \text{distance}(P_i\, C_3)^2$$

➢Unsupervised learning

Solution:

- To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance

$$WCSS = \sum_{P_i \text{ in Cluster1}} distance(P_i\ C_1)^2 + \sum_{P_i \text{ in Cluster2}} distance(P_i\ C_2)^2 + \sum_{P_i \text{ in CLuster3}} distance(P_i\ C_3)^2$$



Credit: Khola Naseem

➢Unsupervised learning

➢KMean

What properties should a distance measure have?

- Symmetric
  - D(A,B)=D(B,A)
  - Otherwise, we can say A looks like B but B does not look like A
- Positivity, and self-similarity
  - D(A,B)≥0, and D(A,B)=0 iff A=B
  - Otherwise there will different objects that we cannot tell apart
- Triangle inequality
  - D(A,B)+D(B,C) ≥ D(A,C)
  - Otherwise one can say "A is like B, B is like C, but A is not like C at all"

Credit: Khola Naseem

➢Unsupervised learning

Example: K-Means for Segmentation

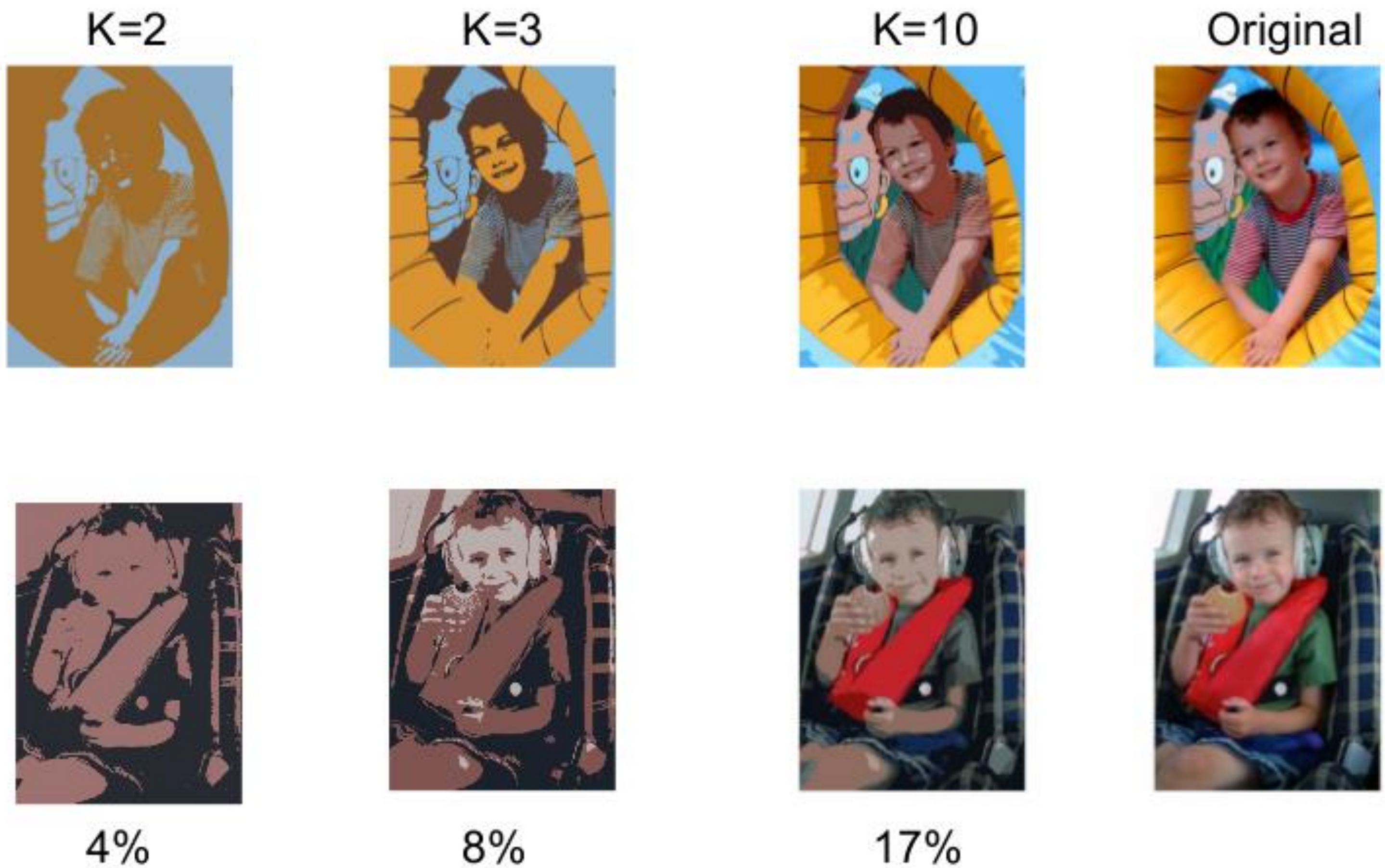
K=2

Goal of Segmentation is to partition an image into regions each of which has reasonably homogenous visual appearance.

Original

➢Unsupervised learning

Example: K-Means for Segmentation



Credit: Khola Naseem

Unsupervised learning

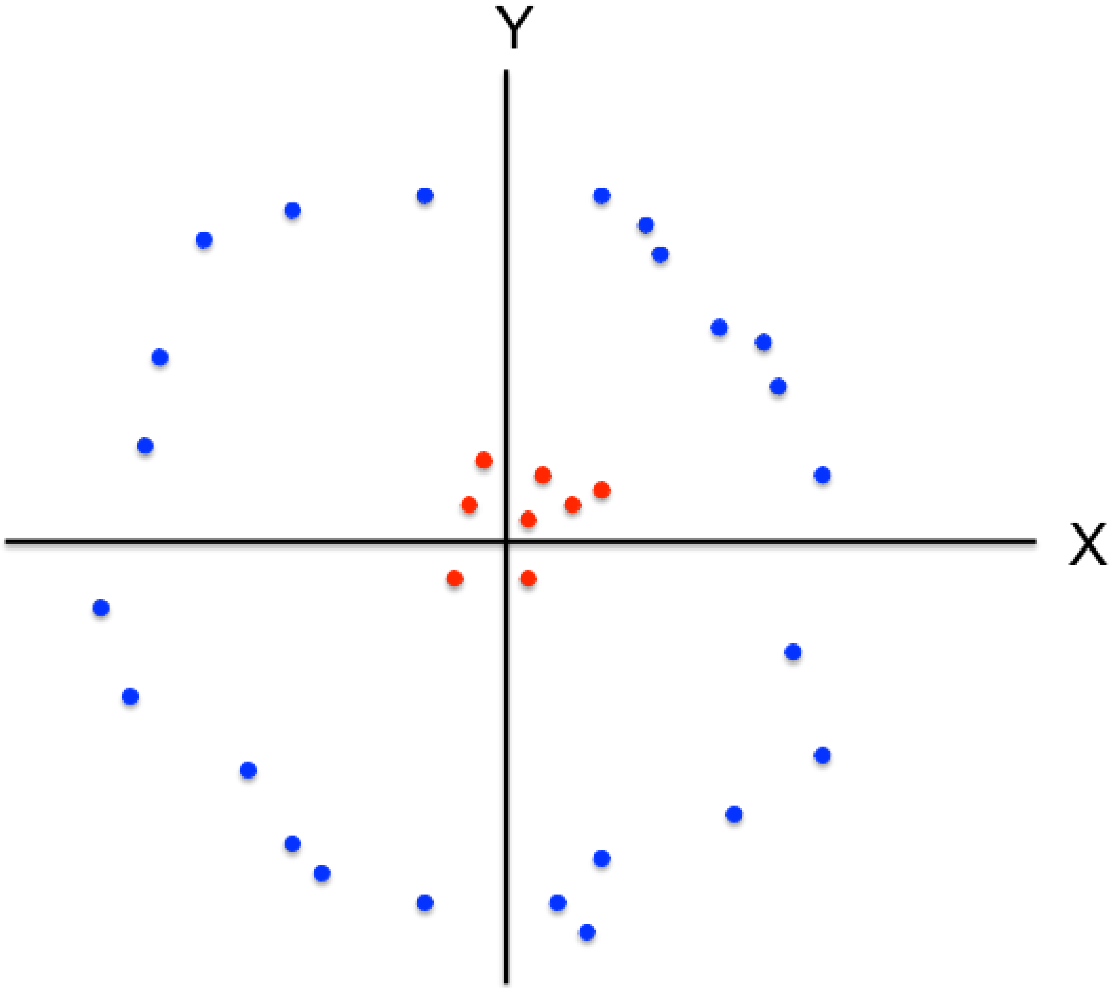
Kmeans algorithm:

```
kmeans(D, k)
    choose K initial means randomly (e.g., pick K points randomly from D)

    while means_are_changing
            %  assign each point to a cluster
            for i = 1: m                          %(m records in D)
                  membership[x(i)] = cluster with mean closest to x(i)
            end
            % update the means
            for k = 1:K
                  mean_k = average of vectors x(i) assigned to cluster k
            end

            % check for convergence
            if (new means are the same as old means) then break
            else means_are_changing = 1
    end
```

Credit: Khola Naseem

➢Unsupervised learning

K-means not able to properly cluster



Credit: Khola Naseem