**Unsupervised learning**

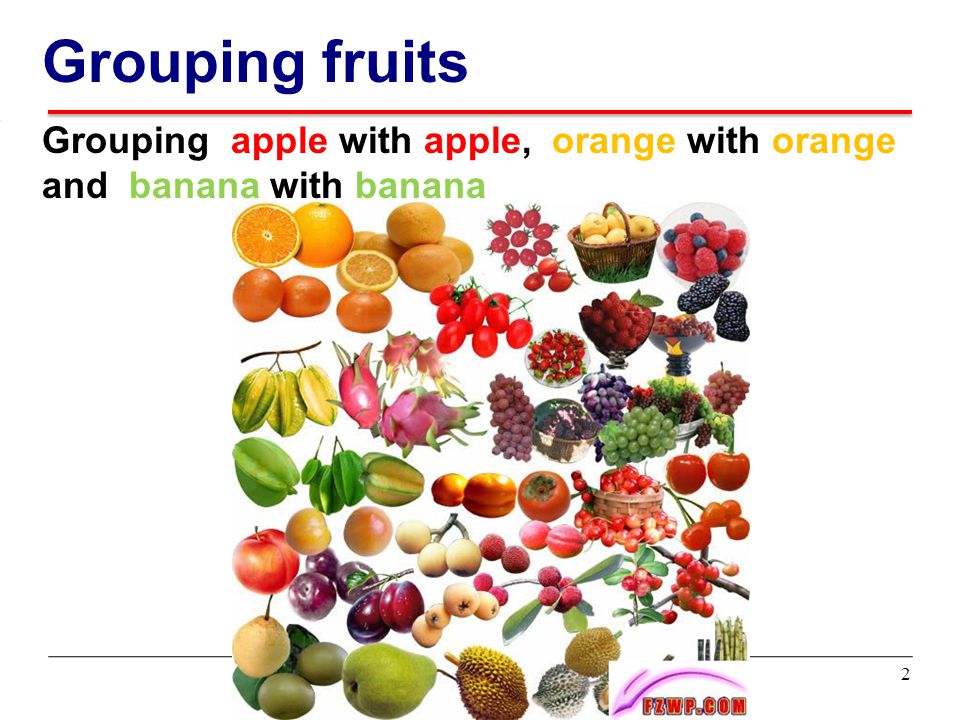
**Unsupervised learning:**

Unsupervised learning is you only have input data (X) and no corresponding output data.

* Suppose you had a basket and it is fuelled with some different types fruits, your task is to arrange them as groups.
* This time you don’t know any thing about that fruits, honestly saying this is the first time you have seen them.
* So how will you arrange them?
* What will you do first???
* You will take a fruit and you will arrange them by considering physical character of that particular fruit. Suppose you have considered colour.
* Then you will arrange them on considering base condition as **colour.**
* Then the groups will be something like this.
* RED COLOR GROUP: apples & cherry fruits.
* GREEN COLOR GROUP: bananas & grapes.
* So now you will take another physical character such as **size.**
* RED COLOR AND BIG SIZE: apple.
* RED COLOR AND SMALL SIZE: cherry fruits.
* GREEN COLOR AND BIG SIZE: bananas.
* GREEN COLOR AND SMALL SIZE: grapes.
* job done happy ending.
* Here you didn’t know learn any thing before, means no train data and no response variable.
* This type of learning is knowing unsupervised learning.
* clustering comes under



**Unsupervised learning:**



**Unsupervised Learning Real Life Examples:**

1. A friend invites you to his party where you meet totally strangers. Now you will classify them using unsupervised learning (no prior knowledge) and this classification can be on the basis of gender, age group, dressing, educational qualification or whatever way you would like. **Why this learning is different from Supervised Learning? Since you didn't use any past/prior knowledge about people and classified them "on-the-go".**
2. NASA discovers new heavenly bodies and finds them different from previously known astronomical objects - stars, planets, asteroids, black holes etc. (i.e. it has no knowledge about these new bodies) and classifies them the way it would like to (distance from Milky way, intensity, gravitational force, red/blue shift or whatever)
3. Let's suppose you have never seen a Cricket match before and by chance watch a video on internet, now you can classify players on the basis of different criterion: Players wearing same sort of kits are in one class, Players of one style are in one class (batsmen, bowler, fielders), or on the basis of playing hand (RH vs LH) or whatever way you would observe [and classify] it.
4. We are conducting a survey of 500 questions about predicting the IQ level of students in a college. Since this questionnaire is too big, so after 100 students, administration decides to trim the questionnaire down to fewer questions and for it we use some statistical procedure like PCA to trim it down.

**Types of unsupervised learning algorithms:**

1. K – means clustering
2. Hierarchical clustering

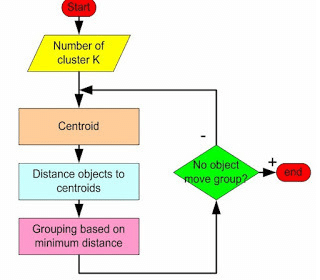
**K- MEANS CLUSTERING**

**K-MEANS CLUSTERING:**

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data.  The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided.

* The Cancroids’ of the *K* clusters, which can be used to label new data.

**Flow chart for k-means clustering:**



**Algorithm:**

* assuming we have inputs x1,x2,x3,…,xn  and value of **K**
* **Step 1** - Pick K random points as cluster centers called centroids.
* **Step 2** - Assign each xi to nearest cluster by calculating its distance to each centroid.
* **Step 3** - Find new cluster center by taking the average of the assigned points.
* **Step 4** - Repeat Step 2 and 3 until none of the cluster assignments change.

**step1:**

We randomly pick **K** cluster centres (centroids). Let’s assume these are c1, c2… ck and we can say that;

C= c1, c2, ck

C is the set of all centroids.

**Step 2:**

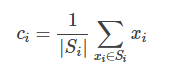
In this step we assign each input value to closest centre. This is done by calculating Euclidean (L2) distance between the point and the each Centroid.

****

Where dist (.) is the Euclidean distance.

**Step 3:**

In this step, we find the new centroid by taking the average of all the points assigned to that cluster.

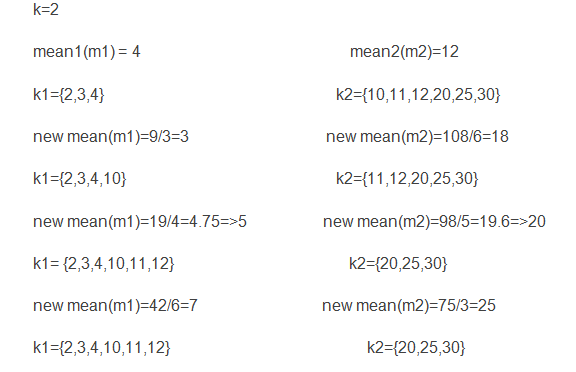


Si is the set of all points assigned to the ith cluster.

Step 4:

#### In this step, we repeat step 2 and 3 until none of the cluster assignments change. That means until our clusters remain stable, we repeat the algorithm.

**ex: K =** {2,3,4,10,11,12,20,25,30}



**K-Means is widely used for many applications:**

* Image Segmentation
* Clustering Gene Segmentation Data
* News Article Clustering
* Clustering Languages
* Species Clustering
* Anomaly Detection

Suppose we have a set of the following two dimensional data instances named D.

D = { (5,3), (10,15), (15,12), (24,10), (30,45), (85,70), (71,80), (60,78), (55,52), (80,91) }

We want to divide this data into two clusters, C1 and C2 based on the similarity between the data points.

The first step is to randomly initialize values for the centroids of both clusters. Let's name centroids of clusters C1 and C2 as c1 and c2 and initialize them with the values of the first two data points i.e. (5, 3) and (10, 15).

Now we have to start the iterations.

| S.No | Data Points | Euclidean Distance from Cluster Centroid c1 = (5,3) | Euclidean Distance from Cluster Centroid c2 = (10,15) | Assigned Cluster |
| --- | --- | --- | --- | --- |
| 1 | (5,3) | 0 | 13 | C1 |
| 2 | (10,15) | 13 | 0 | C2 |
| 3 | (15,12) | 13.45 | 5.83 | C2 |
| 4 | (24,10) | 20.24 | 14.86 | C2 |
| 5 | (30,45) | 48.87 | 36 | C2 |
| 6 | (85,70) | 104.35 | 93 | C2 |
| 7 | (71,80) | 101.41 | 89 | C2 |
| 8 | (60,78) | 93 | 80 | C2 |
| 9 | (55,52) | 70 | 58 | C2 |
| 10 | (80,91) | 115.52 | 103.32 | C2 |

**Iteration 1**

In the table above, the second column contains all the data points. The third column contains the [Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance) between all the data points and centroid c1. Similarly the fourth column contains distance between the c2 centroid and the data points. Finally, in the fifth column we show which cluster the data point is assigned to based on the Euclidean distance between the two cluster centroids. For instance, look at the third data point (15, 12). It has a distance of 13.45 units from c1while a distance of 5.83 units from c2; therefore it has been clustered in C2.

After assigning data points to the corresponding clusters, the next step is to calculate the new centroid values. These values are calculated by finding the means of the coordinates of the data points that belong to a particular cluster.

For cluster C1, there is currently only one point i.e. (5,3), therefore the mean of the coordinates remain same and the new centroid value for c1 will also be (5,3).

For C2, there are currently 9 data points. We name the coordinates of data points as x and y. The new value for x coordinate of centroid c2 can be calculated by determining the mean of xcoordinates of all 9 points that belong to cluster C2 as given below:

c2(x) = (10 + 15 + 24 + 30 + 85 + 71 + 60 + 55 + 80) / 9 = 47.77

The new value for y coordinate of centroid c2 can be calculated by determining the mean of all y coordinates of all 9 points that belong to cluster C2.

c2(y) = (15 + 12 + 10 + 45 + 70 + 80 + 78 + 52 + 91) / 9 = 50.33

The updated centroid value for c2 will now be {47.77, 50.33}.

For the next iteration, the new centroid values for c1 and c2 will be used and the whole process will be repeated. The iterations continue until the centroid values stop updating. The next iterations are as follows:

**Iteration 2**

| **S.No** | **Data Points** | **Euclidean Distance from Cluster Centroid c1 = (5,3)** | **Euclidean Distance from Cluster Centroid c2 = (47.77,50.33)** | **Assigned Cluster** |
| --- | --- | --- | --- | --- |
| 1 | (5,3) | 0 | 63.79 | C1 |
| 2 | (10,15) | 13 | 51.71 | C1 |
| 3 | (15,12) | 13.45 | 50.42 | C1 |
| 4 | (24,10) | 20.24 | 46.81 | C1 |
| 5 | (30,45) | 48.87 | 18.55 | C2 |
| 6 | (85,70) | 104.35 | 42.10 | C2 |
| 7 | (71,80) | 101.41 | 37.68 | C2 |
| 8 | (60,78) | 93 | 30.25 | C2 |
| 9 | (55,52) | 70 | 7.42 | C2 |
| 10 | (80,91) | 115.52 | 51.89 | C2 |

c1(x) = (5, 10, 15, 24) / 4 = 13.5 c1(y) = (3, 15, 12, 10) / 4 = 10.0

Updated c1 to be (13.5, 10.0).

c2(x) = (30 + 85 + 71 + 60 + 55 + 80) / 6 = 63.5

c2(y) = (45 + 70 + 80 + 78 + 52 +91) / 6 = 69.33

Updated c2 to be (63.5, 69.33).

Updated c2 to be (63.5, 69.33).

**Iteration 3**

| **S.No** | **Data Points** | **Euclidean Distance from Cluster Centroid c1= (13.5,10)** | **Euclidean Distance from Cluster Centroid c2= (63.5,69.33)** | **Assigned Cluster** |
| --- | --- | --- | --- | --- |
| 1 | (5,3) | 11.01 | 88.44 | C1 |
| 2 | (10,15) | 6.10 | 76.24 | C1 |
| 3 | (15,12) | 2.5 | 75.09 | C1 |
| 4 | (24,10) | 10.5 | 71.27 | C1 |
| 5 | (30,45) | 38.69 | 41.40 | C1 |
| 6 | (85,70) | 93.33 | 21.51 | C2 |
| 7 | (71,80) | 90.58 | 13.04 | C2 |
| 8 | (60,78) | 82.37 | 9.34 | C2 |
| 9 | (55,52) | 59.04 | 19.30 | C2 |
| 10 | (80,91) | 104.80 | 27.23 | C2 |

c1(x) = (5, 10, 15, 24, 30) / 5 = 16.8

c1(y) = (3, 15, 12, 10, 45) / 5 = 17.0

Updated c1 to be (16.8, 17.0).

c2(x) = (85 + 71 + 60 + 55 + 80) / 5 = 70.2

c2(y) = (70 + 80 + 78 + 52 + 91) / 5 = 74.2

Updated c2 to be (70.2, 74.2).

**Iteration 4**

| **S.No** | **Data Points** | **Euclidean Distance from Cluster Centroid c1 = (16.8,17)** | **Euclidean Distance from Cluster Centroid c2 = (70.2,74.2)** | **Assigned Cluster** |
| --- | --- | --- | --- | --- |
| 1 | (5,3) | 18.30 | 96.54 | C1 |
| 2 | (10,15) | 7.08 | 84.43 | C1 |
| 3 | (15,12) | 5.31 | 83.16 | C1 |
| 4 | (24,10) | 10.04 | 79.09 | C1 |
| 5 | (30,45) | 30.95 | 49.68 | C1 |
| 6 | (85,70) | 86.37 | 15.38 | C2 |
| 7 | (71,80) | 83.10 | 5.85 | C2 |
| 8 | (60,78) | 74.74 | 10.88 | C2 |
| 9 | (55,52) | 51.80 | 26.90 | C2 |
| 10 | (80,91) | 97.31 | 19.44 | C2 |

c1(x) = (5, 10, 15, 24, 30) / 5 = 16.8

c1(y) = (3, 15, 12, 10, 45) / 5 = 17.0

Updated c1 to be (16.8, 17.0).

c2(x) = (85 + 71 + 60 + 55 + 80) / 5 = 70.2

c2(y) = (70 + 80 + 78 + 52 + 91) / 5 = 74.2

Updated c2 to be (70.2, 74.2).

At the end of fourth iteration, the updated values of C1 and C2 are same as they were at the end of the third iteration. This means that data cannot be clustered any further. c1 and c2 are the centroids for C1 and C2. To classify a new data point, the distance between the data point and the centroids of the clusters is calculated. Data point is assigned to the cluster whose centroid is closest to the data point.