K-Means Clustering

Tuesday, 29 April 2025 07:01

Unsupervised Learning

↓

Data with input cols only

Amazon Customers Data

↓

	Name	Age	City	Gender	Purchases	Amount spent
0						
1						
2						
3						
4						
5						

↓

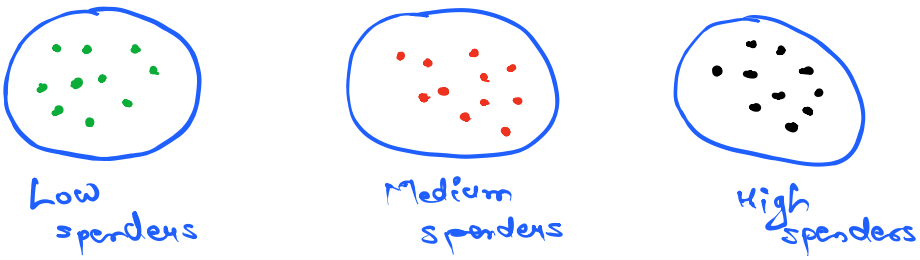
Unsupervised Algorithm

↓

They will learn two things :

→ How many groups / clusters exist in our data ?

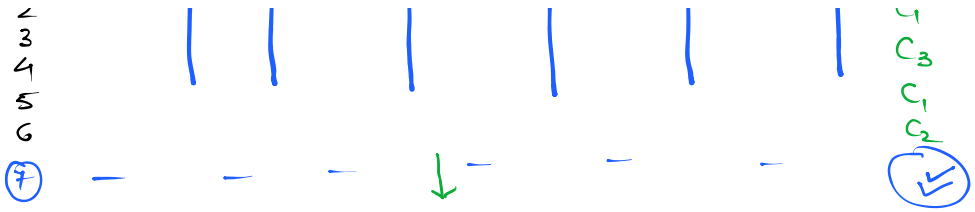
→ Which customer / datapoint belongs to which cluster.

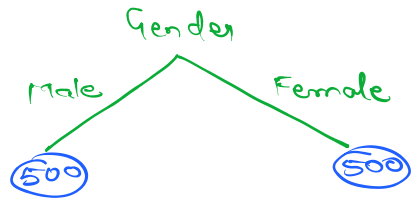


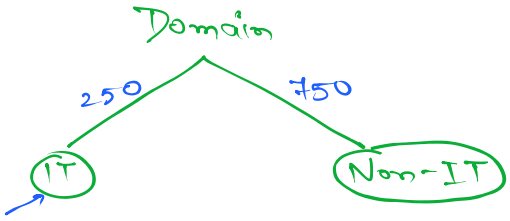
Intellipaat (1000)

↓

	Name	Age	Gender	Educat qual.	Domain	Exp. Level	Cluster Number (output)
0							C_2
1							C_4
2							

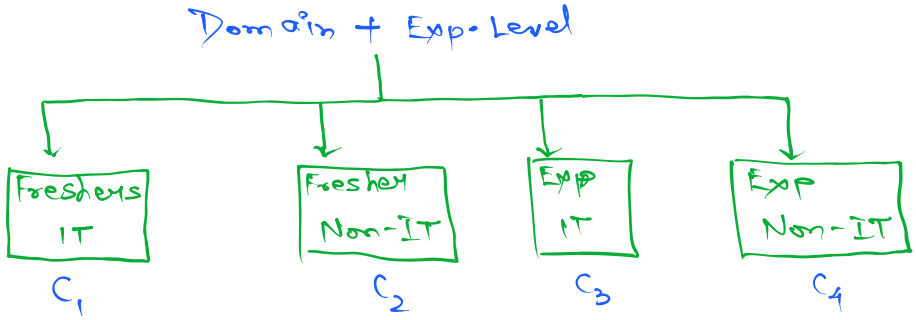






Fresher
< 1

Exp
>= 1 year



1. Targeted marketing
2. Recommendations
3. offers, coupons, discount

Unsupervised Learning Algorithms →

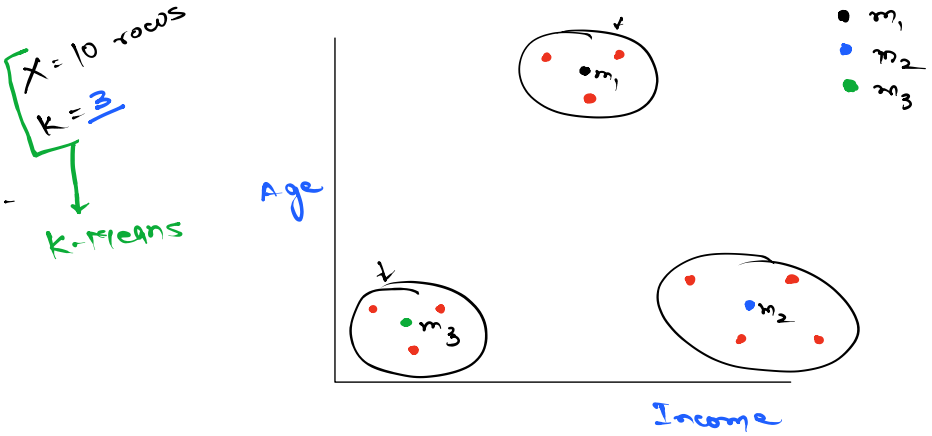
1. K-Means clustering.
2. Hierarchical Clustering.
3. DBScan clustering.
4. Gaussian mixture models (GMMs).

K-means Clustering

Two requirements of k-means :

① Data (input cols) → 'X'

② k-value : No of clusters to be identified.



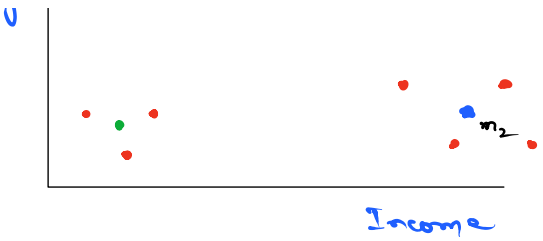
① Create centroids / cluster means on the data.

② Calculate the distance b/w each data point to each centroid / cluster mean.

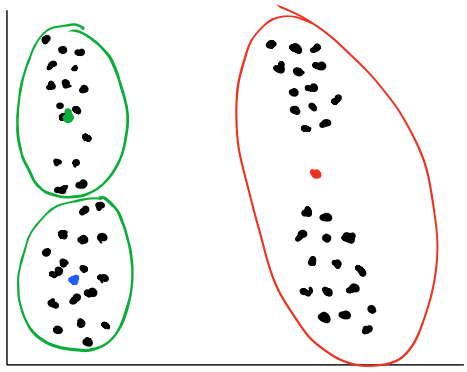
③ Put together the datapoints with their closest centroids in clusters.

④ Update the centroid values.

⑤ Repeat steps 2-4 until no further changes observed.

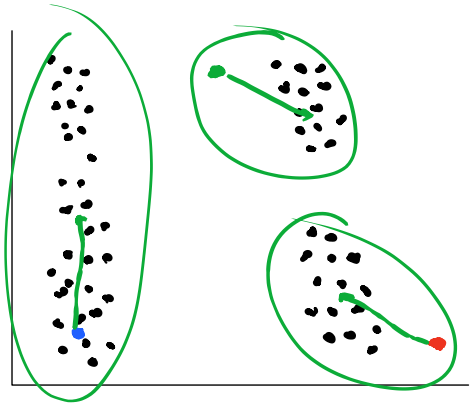


Initialization Trap :



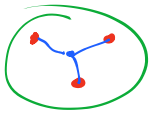
K=3

To overcome this, we now have updated version of k-means which is known as 'k-means++'.



Finding optimal k-value using WCSS method →

WCSS : Within Cluster Sum of squared distances



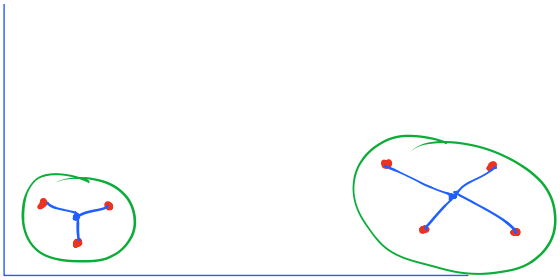
3875

K = ?

K = 1

K = 2

K = 3



$WCSS(K=1) = d_1^2 + d_2^2 + d_3^2 + \dots\dots\dots + d_{10}^2$

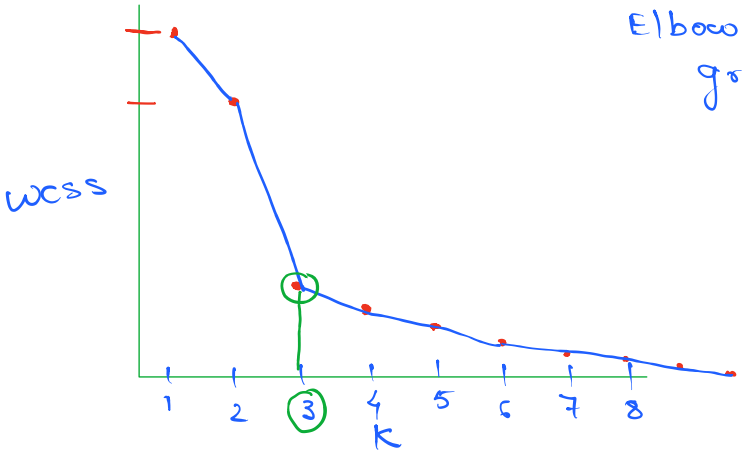
$= 3899$

$WCSS(K=2) = 3125$

$WCSS(K=3) = 1800$

$WCSS(K=4) = 1670$

$WCSS(K=5) = 1450$



Elbow point graph

```
From sklearn.linear_model import LinearRegression
From sklearn.metrics import r2_score
From sklearn.model_selection import train_test_split
```