DS unit-3

Q1 - Overview of Clustering

Clustering is the use of unsupervised techniques for grouping similar objects. In ML, unsupervised refers to the pattern of finding hidden structure within unlabeled data.

- Clustering is a method often used for exploratory analysis of the data.
- In clustering there are no predictions made ~~paths~~ further.
- Clustering methods find the similarities b/w objects according to the object attributes and group similar objects into cluster.
- cluster techniques are utilized in marketing, economics and various branches of science.
A popular clustering method is K-means.

Imp
Q2) K Means → Given a collection of objects each with n measurable attributes, K-means is an analytical techniques that for a chosen value of K, identifies K clusts of objects based on the objects proximity to the centers of the K groups.

Algorithm:-
1) Selects the no of cluster K=3.
2) Randomly select 3 distinct data points
3) Mean distance b/w the 1st pnt & selected 3 clusters

"K means is a clustering algorithm whose main goal is to group similar elements or data points into a cluster"

Note:- "K" in K-means represent the number of clusters"

Algorithm:-

Steps:-

1. Select the number K to decide the number of clusters.
2. Select random K points as centroids.
3. Assign each data point to their closest Centroid, which will form the predefined K clusters.
4. Calculate the variance & place a new Centroid for each clusters
5. Repeat the step 3.
6. If any reassignment occurs then go to step 4. else to finish
7. Stop.

Example for K-means.

10 marks

2022 Q] Suppose we have 4 different medicines with their weight index & PH as mentioned below, making use of K means algorithm group the below objects into 2 groups. use $c_1 = (1,1)$ & $c_2 = (2,1)$ as initial clusters

Objects	weight (x_1)	PH (y_1)	$c_1(1,1)$	$c_2(2,1)$	clusters
A1	2	1	0	1	C1
B2	2	1	1	0	C2
C4	4	3	3.60	2.82	C2
D5	5	4	5	4.24	C2

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

C1 = (A1) C2 = (B2, C4, D5)

what is Regression Analysis? -

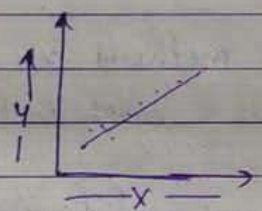
⇒ they are used to determine the relationship blw a dataset's dependent & independent variables. (x, y)

- It is widely used when the dependent & independent variables are linked in a linear or non-linear fashion. & target variable has a set of continuous values.

Example :- the best way to Examine the relationship blw sales & advertising Expenditures.

1. Linear Regression :-

Linear Regression is a analytical technique used to model the relationship blw a dependent variable (y) and an independent variable (x).



it employes a regression line, also known as best-fit line.

The line connection is defined as

$y = c + m * x + e$.

c = intercept, m = slope of line,

e = error term.

2) logistic Regression. List the reasons to choose blw the regression algo & Explain the same :-

⇒ when choosing blw regression algorithms, several factors should be considered based on the specific problem at hand & the charact— of the dataset. Some reasons to choose linear regression. :-

1. <u>Interpretability</u>:- The it provides a clear pse 'interpretation' of the relshp b/w the variables.
Coefficients of the linear regression equation indicate the magnitude & direction.
2. <u>Linearity assumption</u>:- Linear Regr assumes a linear rel b/w the independent variables & the target variable.
3. <u>Simplicity and speed</u>:- it is Computationally efficient and relatively simple to implement compared to other regression algo.
4. <u>Limited training data</u>:- it perform well even when the dataset is small or the number of training instances is limited.
5. <u>Baseline model</u>:- it is a good baseline model to compare against more complex algo.

Q) Explain how least squared methods are used in minimizing errors in linear regression?

→ If the data shows a <u>lean relationship</u> between two variables, it results in a least-squares regression line.
- This <u>minimizes the vertical distance</u> from the data points to the regression line.
- The term least squares is used bcz it is the <u>smallest sum of squares of error</u>, which is also called the variance.

5) In K means algorithm, how to choose the appropriate value of k and discuss the mathematical basis for the same.

→ Elbow Method.

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: within cluster sum of squares, which defines the total variations within a cluster.

Formula is :-

for K=2

$SSE_1 = \Sigma \text{ distance } (x - C_1)^2$

$SSE_2 = \Sigma \text{ distance } (x - C_2)^2$

$SSE = SSE_1 + SSE_2$

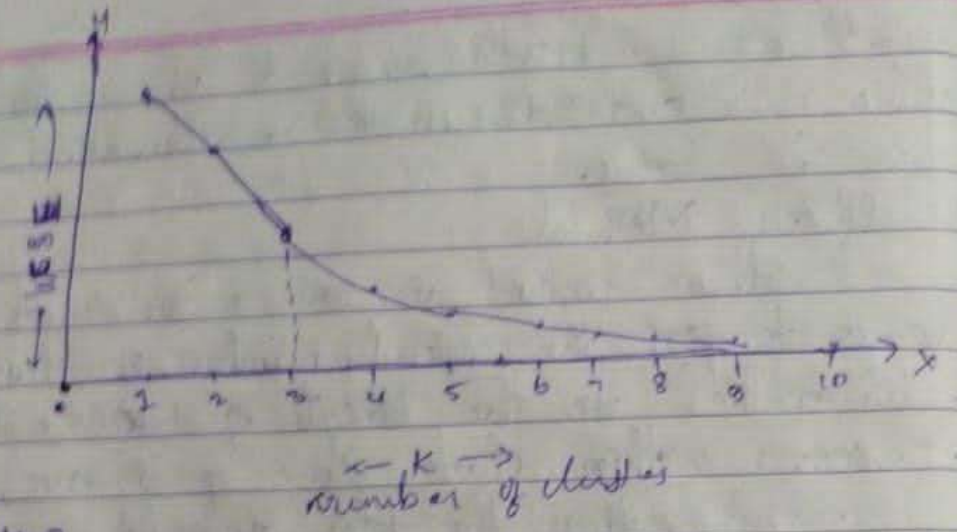
- $WCSS = \Sigma P_{i \text{ in cluster}_1} \text{distance}(P_i C_1)^2 + \Sigma P_{i \text{ in } C_2} \text{distance}(P_i C_2)^2 + \ldots$

$\Sigma P_{i \text{ in cluster}_1} \text{distance}(P_i C_1)^2$:- it is Sum of the Square of the distance blw each data point and its centroid within a cluster1 & the same for other terms.

- To measure the distance blw data points & Centroid, we can use any method such as Euclidean distance or Manhattan distance.
- To find optimal value, elbow method follows the steps:-

1. It executes the K-means clustering on a given dataset for different K value (ranges from 1-10)
2. For each value of K, calculate the WCSS value.
3. Plots a curve blw calculated WCSS values & the number of cluster k
4. The shape of bend or a point of the plot look like an arm, that point is considered as the best value of K

graph :-



(K) $\frac{K=3}{}$ As K value increase, the SSE value decrea

2) Illustrate the use of Box-Jenkin Methodology for time Series analysis.

→ A time Series consist of an ordered sequen of equally spaced values over time.

Example: Monthly unemployment rates, daily website visits, or stock prices every second.

- A time Series can consist of the following compone
  - Trend
  - Seasonality
  - Cyclic
  - Random.

1. Trend :- refers to the long-term movement in a time S it indicates whether the observation values are incre or decreasing over time.
   eg :- increase in Sales over month.
2. Seasonality :- it describes the fixed, periodic fluctuation in the observation over time.
   - its often related to the calendar.
   eg :- Monthly retail Sales can fluctuate over the year due to the weather and holidays

3) Cyclic: it refers to a periodic fluctuation, but one that is not as fixed as in case of a seasonality component.
Eg:- Retails sales are influenced by the general state of the economy. Thus, retail sales time series can often follow the lengthy boom-bust cycles of the economy.

4) Random :- component is what remains, after accounting for the other 3 components.
Although noise is certainly part of this random component, there is often some underlying structure to this random component that need to be forecast future values.

- Developed by Georg Box - The Box-Jenkins methodology for time series analysis involves the 3 main steps

1. Condition data and select a model.
   - Identify & account for any trends or seasonality in the time series.
   - Examine the remaining time series & determine a suitable model.
2. Estimate the model parameters.
3. Assess the model and return to step 1, if necessary

VSMP Q. Explain the K nearest Neighbour algorithm. List out the modeling assumptions to be made while using KNN algorithm.

⇒ K-nearest Neighbour is one of the simplest ML algo based on supervised learning technique

⇒ KNN assumes the Similarity b/w the new cases and available cases & put the new cases into the category that is most similar to the available categories.

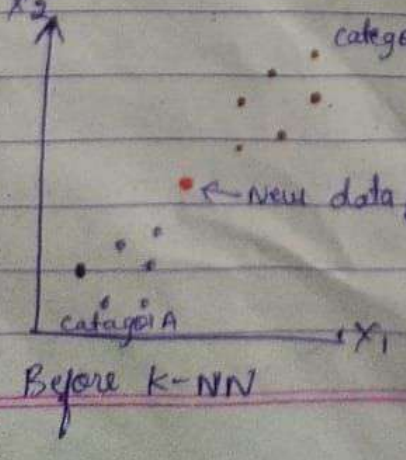
⇒ KNN is used for the classification problems.

⇒ KNN is a non-parametric algo, means it does not make any assumption on underlying data.

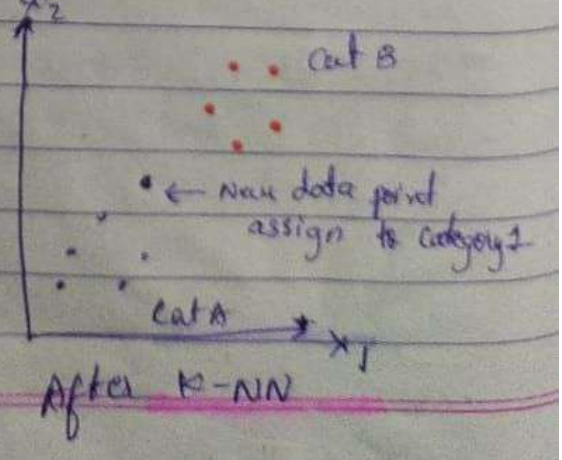
⇒ it is called lazy learner algo, bcz it does not learn from the training set immediately. it performs on action on the dataset.

Eg:- Suppose we have 2 image of animal that looks similar to cat & dog, but we want to know its a cat / Dog. so here we can use KNN algo as it work on similarity measure.

- It will find the similar features of the new data set to the cats & dog images & based on the most similar features it will put it in either cat or dog category.



Before K-NN



After K-NN