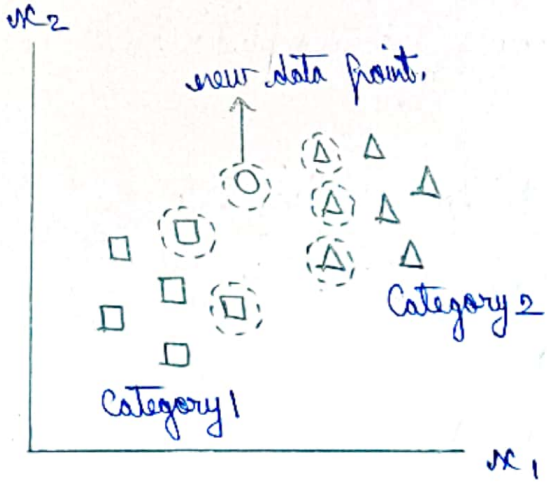
SUBHAPREET PATRO
2211CS010547
G-3

RS ASSIGNMENT-1

K- Nearest Neighbor (KNN) Algorithm :-

K-Nearest Neighbors (KNN) is a simple way to classify things by looking at what's nearby. Imagine a streaming service wants to predict if a new user is likely to cancel their subscription (churn) based on their age. They check the ages of its existing users & whether they churned or stayed. If most of the 'K' closest users in age of new user cancelled their subscription KNN will predict the new user might churn too. The key idea is that users with similar ages tend to have similar behaviors & KNN uses this closeness to make decisions.

K-Nearest Neighbors is also called as a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset & at the time of classification, it performs an action on the dataset.



The new data point is classified as Category 2 because most of its closest neighbors are Triangles. KNN assigns the category based on the majority of nearby points.

The image shows how KNN predicts the category of a new data point based on its closest neighbors.

→ The squares represent Category 1 & the triangles represent Category 2.
→ The new data point checks its closest neighbors (circled points)
→ Since the majority of its closest neighbors are triangles (Category 2) KNN predicts the new data point belongs to Category 2.

KNN works by using proximity & majority voting to make predictions

How to choose the value of k for KNN Algorithm?

The value of k is critical in KNN as it determines the number of numbers to consider when making predictions. Selecting the optimal value of k depends on the characteristics of the input data. If the dataset has significant outliers or noise a higher k can help smooth out the predictions & reduce the influence of noisy data. However choosing a very high value can lead to underfitting where the model becomes too simplistic.

~~Selecting the~~

Statistical Methods for selecting k:-

1. Cross-Validation
2. Elbow Method
3. Odd values for k

Distance Metrics used in KNN Algorithm:-

KNN uses distance metrics to identify nearest neighbour, these neighbours are used for classification & regression task.

(i) Euclidean distance :-

$$\text{distance}(x, X_i) = \sqrt{\sum_{j=1}^{d} (x_j - X_{i_j})^2}$$

(ii) Manhattan distance :-

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

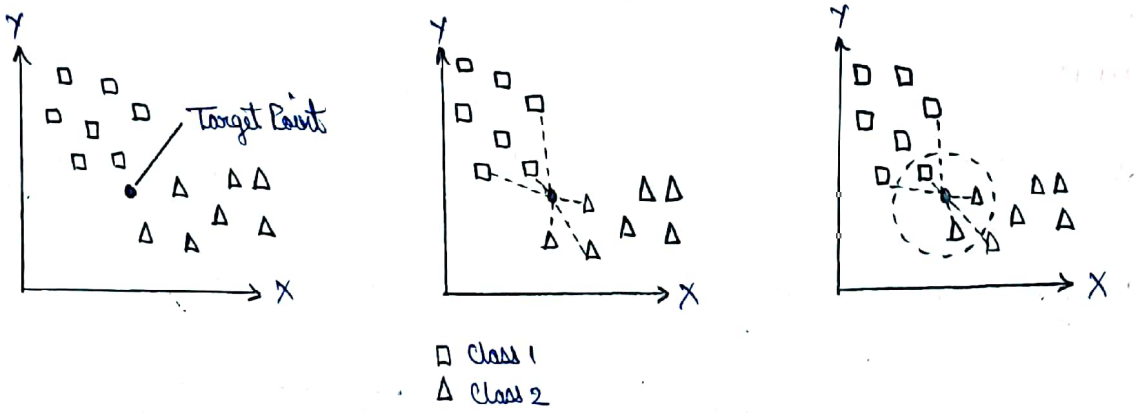
(iii) Minkowski distance :-

$$d(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^p \right)^{\frac{1}{p}}$$

When $p = 2$ then it is the same as Euclidean distance
When $p = 1$ then we obtain the formula of Manhattan distance

Working of KNN Algorithm:-

The K-nearest neighbors algorithm operates on the principle of similarity where it predicts the label or value of a new data point by considering the labels or values of its K nearest neighbors in the training dataset.



Step 1:- Selecting the optimal value of K:-

K represents the number of nearest neighbors that needs to be considered while making predictions

Step 2:- Calculating distance:-

To measure the similarity between target & training data points, Euclidean distance is used.

Step 3:- Finding Nearest Neighbors:-

The k data points with the smallest distances to the target point are nearest neighbors.

Step 4:- Voting for Classification or Taking Average for Regression:-

When classifying data using the KNN algorithm, it identifies the K nearest neighbors and assigns the data point to the most common category among these neighbors, known as majority ~~resp~~ voting.

For regression, it also finds the K closest points but predicts the value by averaging these neighbors' values.