

K NEAREST NEIGHBOUR

- MUKUL KADASKAR
(M.TECH RESEARCH)

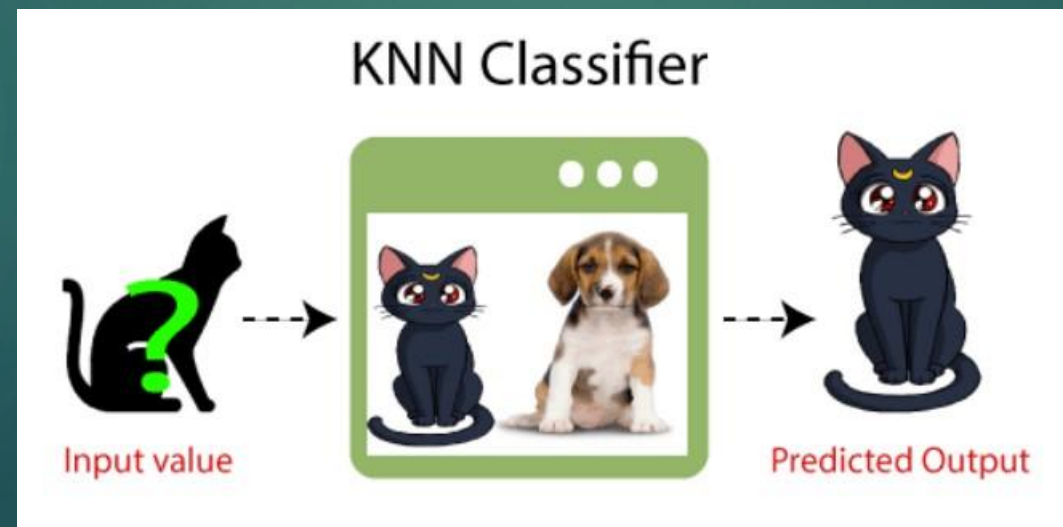
INTRODUCTION :

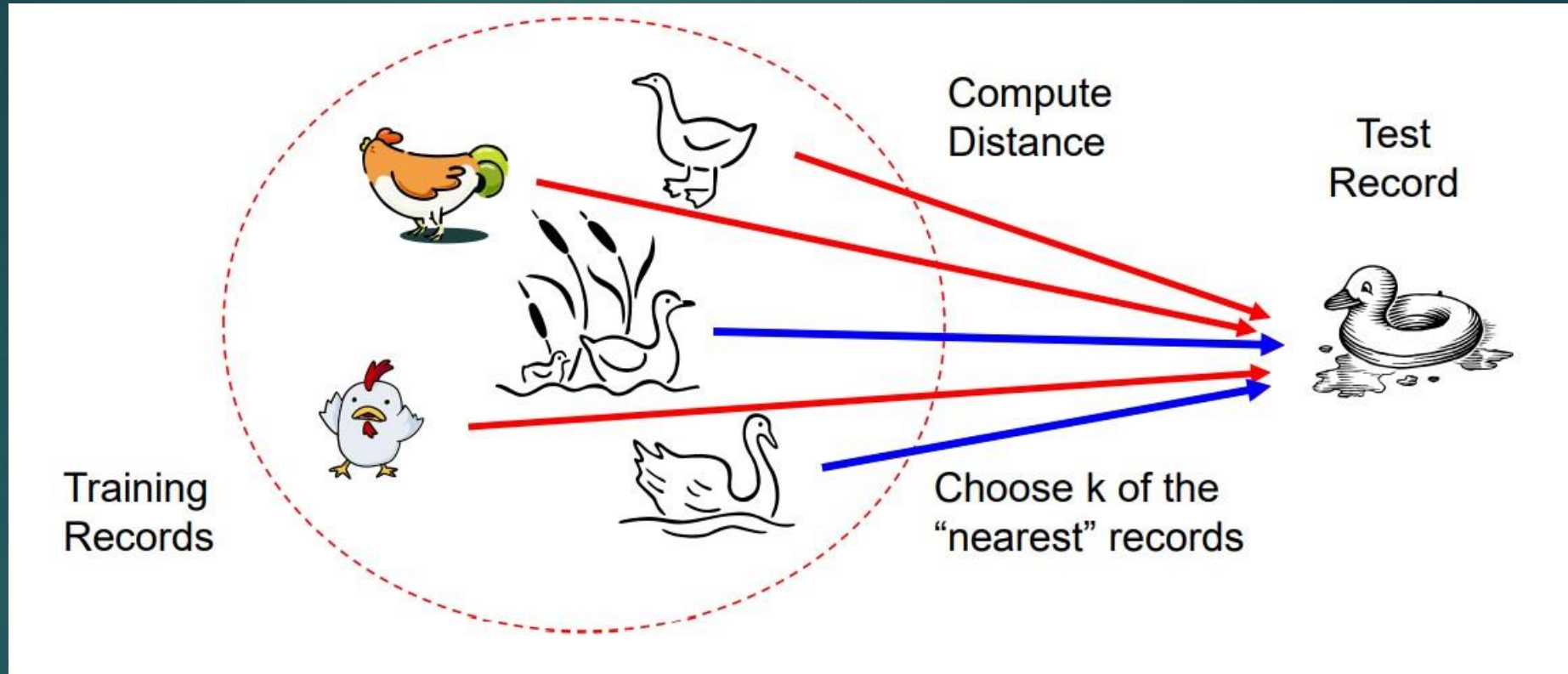
- ▶ K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- ▶ K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- ▶ K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- ▶ K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

INTRODUCTION CONTD :

3

- ▶ K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- ▶ It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- ▶ KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

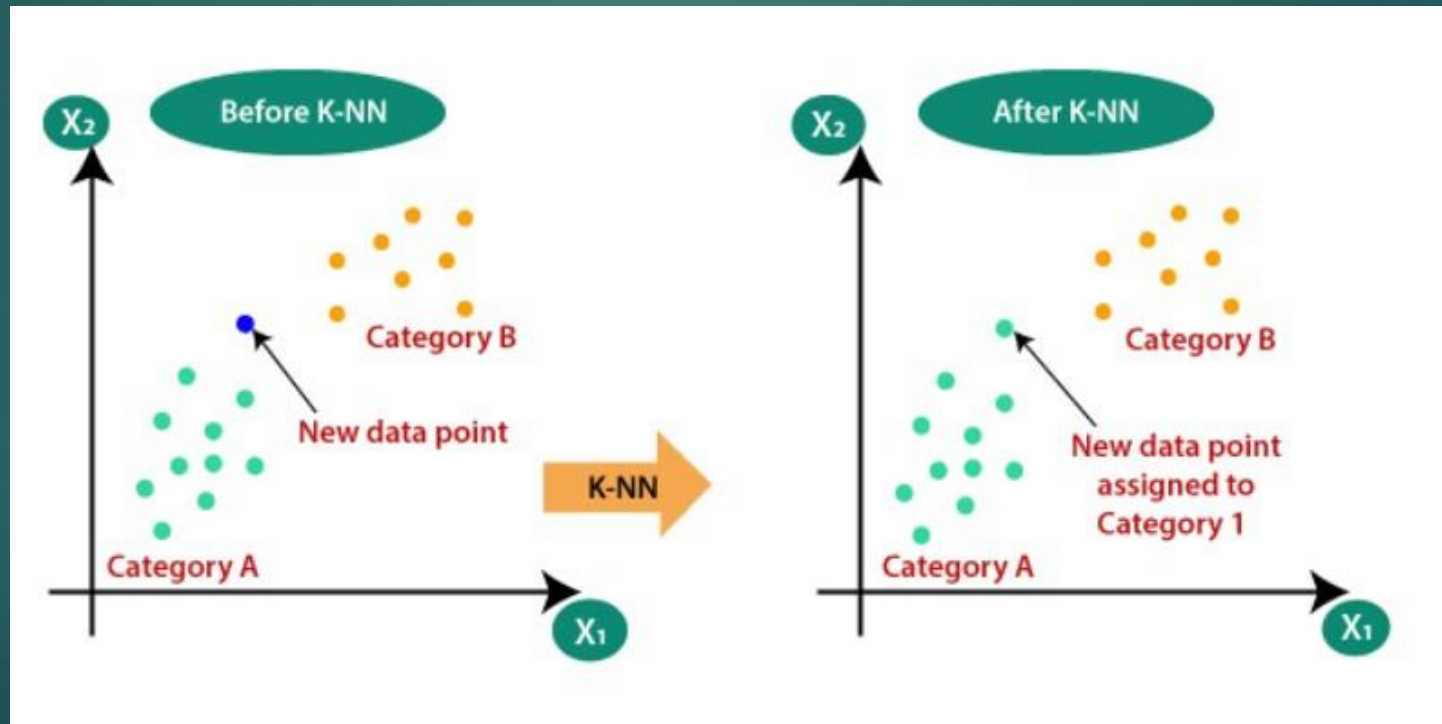




Why do we need a K-NN Algorithm?

5

- ▶ Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x_1 , so this data point will lie in which of these categories.
- ▶ With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



Distance measures :

6

- In order to determine which data points are closest to a given query point, the distance between the query point and the other data points will need to be calculated.

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

$$\text{Manhattan Distance} = d(x,y) = \left(\sum_{i=1}^m |x_i - y_i| \right)$$

$$\text{Minkowski Distance} = \left(\sum_{i=1}^n |x_i - y_i| \right)^{1/p}$$

How does K-NN work ?

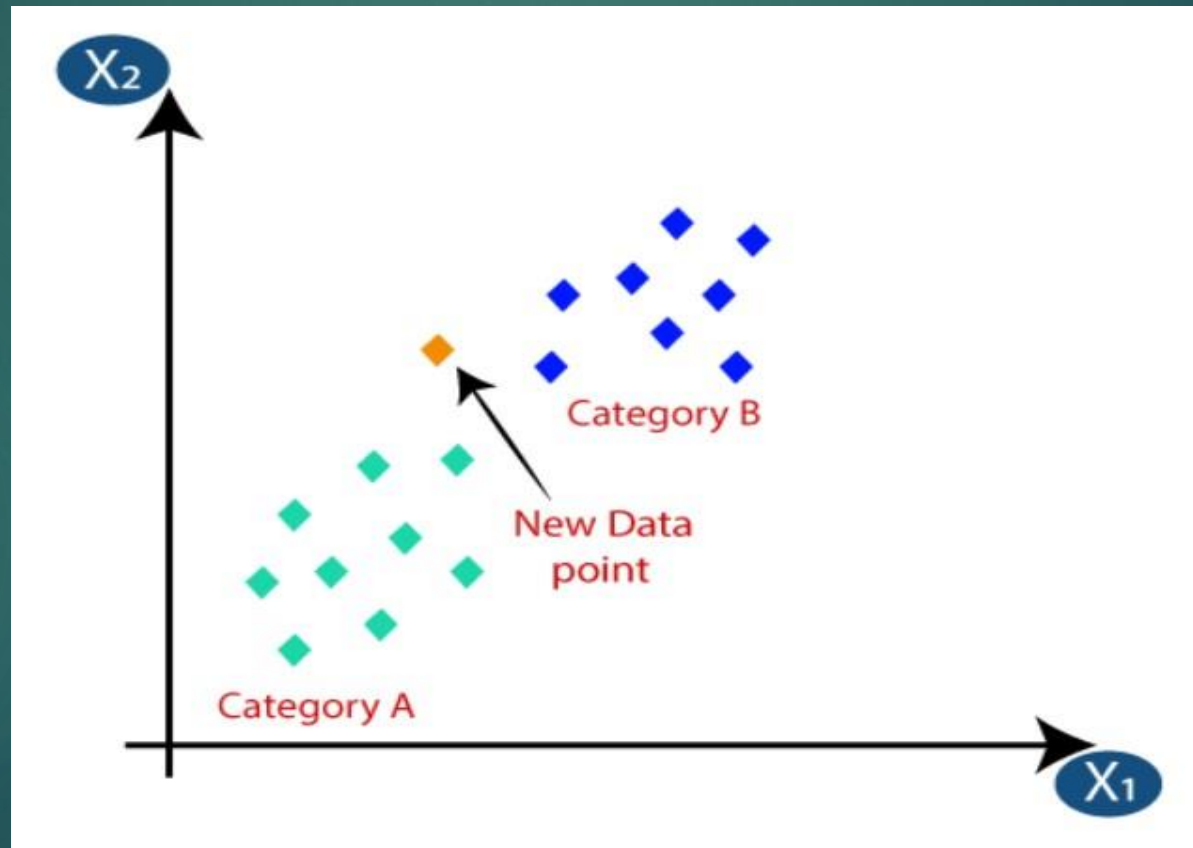
7

- ▶ **Step-1:** Select the number K of the neighbors
- ▶ **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- ▶ **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- ▶ **Step-4:** Among these k neighbors, count the number of the data points in each category.
- ▶ **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- ▶ **Step-6:** Our model is ready.

How does K-NN work ?

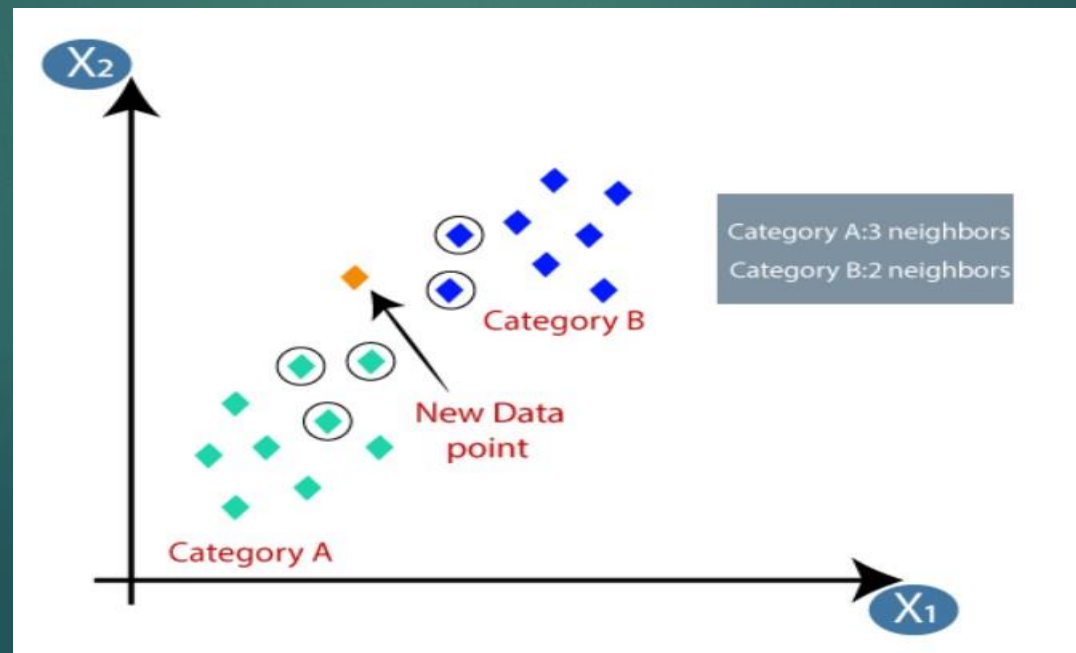
8

- Suppose we have a new data point and we need to put it in the required category. Consider the below image:



How does K-NN work ?

- ▶ Firstly, we will choose the number of neighbors, so we will choose the $k=5$.
- ▶ Next, we will calculate the **Euclidean distance** between the data points.
- ▶ By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:
- ▶ As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.



How to select the value of K in the K-NN Algorithm?

10

- ▶ There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- ▶ A very low value for K such as $K=1$ or $K=2$, can be noisy and lead to the effects of outliers in the model.
- ▶ Large values for K are good, but it may find some difficulties.

Advantages of KNN Algorithm:

- ▶ Easy to implement.
- ▶ Adapts easily.
- ▶ Few hyper parameters .
- ▶ It is robust to the noisy training data.

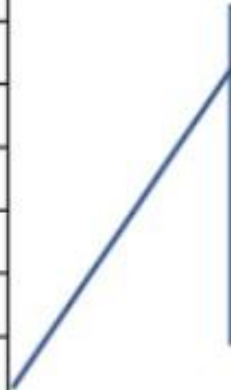
Disadvantages of KNN Algorithm:

- ▶ Value of k .
- ▶ Does not scale well.
- ▶ Curse of dimensionality.
- ▶ Prone to overfitting.
- ▶ The computation cost is high.

Applications of k-NN in machine learning:

- ▶ Data pre-processing .
- ▶ Recommendation Engines .
- ▶ Finance.
- ▶ Healthcare.
- ▶ Pattern Recognition.

Customer	Age	Loan	Default
John	25	40000	N
Smith	35	60000	N
Alex	45	80000	N
Jade	20	20000	N
Kate	35	120000	N
Mark	52	18000	N
Anil	23	95000	Y
Pat	40	62000	Y
George	60	100000	Y
Jim	48	220000	Y
Jack	33	150000	Y
Andrew	48	142000	?



We need to predict
Andrew default status
by using Euclidean
distance

We need to predict Andrew default status (Yes or No).

Customer	Age	Loan	Default	Euclidean distance
John	25	40000	N	1,02,000.00
Smith	35	60000	N	82,000.00
Alex	45	80000	N	62,000.00
Jade	20	20000	N	1,22,000.00
Kate	35	120000	N	22,000.00
Mark	52	18000	N	1,24,000.00
Anil	23	95000	Y	47,000.01
Pat	40	62000	Y	80,000.00
George	60	100000	Y	42,000.00
Jim	48	220000	Y	78,000.00
Jack	33	150000	Y	8,000.01
Andrew	48	142000	?	

First Step calculate the Euclidean distance $\text{dist}(d) = \text{Sq.rt} (x_1 - y_1)^2 + (x_2 - y_2)^2$
 $= \text{Sq.rt}(48-25)^2 + (142000 - 40000)^2$
 $\text{dist}(d_1) = 1,02,000.$

We need to calculate the distance for all the datapoints

Calculate Euclidean distance for all the data points.

Customer	Age	Loan	Default	Euclidean distance	Minimum Euclidean Distance
John	25	40000	N	1,02,000.00	
Smith	35	60000	N	82,000.00	
Alex	45	80000	N	62,000.00	5
Jade	20	20000	N	1,22,000.00	
Kate	35	120000	N	22,000.00	2
Mark	52	18000	N	1,24,000.00	
Anil	23	95000	Y	47,000.01	4
Pat	40	62000	Y	80,000.00	
George	60	100000	Y	42,000.00	3
Jim	48	220000	Y	78,000.00	
Jack	33	150000	Y	8,000.01	1
Andrew	48	142000	?		

Let assume K = 5

Find minimum euclidean distance and rank in order (ascending)

In this case, 5 minimum euclidean distance.
With k=5, there are two Default = N and three Default = Y out of five closest neighbors.

We can say Andrew default status is 'Y' (Yes)

With K=5, there are two Default=N and three Default=Y out of five closest neighbors. We can say default status for Andrew is 'Y' based on the major similarity of 3 points out of 5.