K-Nearest Neighbor

Prepared By: Kimia Aksir

What we will Learn..

- Definition of K-NN
- Properties of K-NN
- K-NN as a Classification Approach
- Measuring Distance: Euclidean Distance
- Example of K-NN
- Choosing the value for “K”
- Binary Class & Multiple Class Classification
- Advantages and Disadvantages of K-NN Algorithms
- Some applications of K-NN

K-NN: K-Nearest Neighbor

K-NN is one of the powerful techniques or, a **Classification** Algorithms that classifies data based on their similarity with the neighbors.

Here, “**K**” stands for the number of Neighbors we are considering to compare for the classification of a new datapoint.

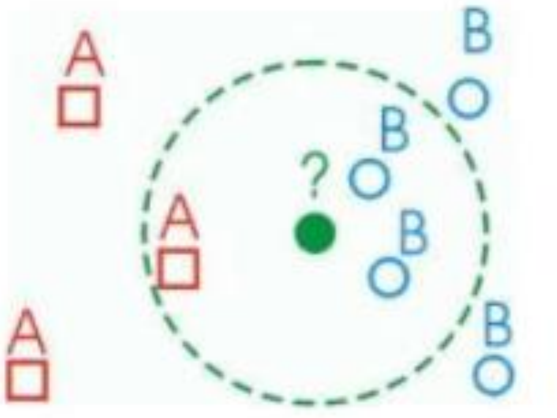
Properties of K-NN

- **Lazy Learning Algorithm:** As, K-NN does not have a training phase and uses all the data while classification

- **Non-Parametric Learning Algorithm:** K-NN is non-parametric learning algorithm as, it doesn’t assume anything about underlying data

Classification Approach of K-NN

- A New Datapoint is classified by a majority votes for its neighbor classes
- The New Datapoint is assigned to the most common “Class” amongst the “K” Nearest Neighbors
- Nearest Neighbors are measured by a Distance Function.



Distance between Neighbors

There are several possible ways of measuring the distance between two Datapoints (Here, one of them is **already existing datapoint** from the existing set and the other one is **the new “Unseen” Datapoint**) in n-dimensional space.

We will use “**Euclidean Distance**” for our **K-NN**.

Euclidean Distance between two Data points:

Considering, the datapoint(s) from the existing Data set are denoted by (a_1 , a_2) and the Unseen Datapoint is denoted by, (b_1 , b_2).

Now, the distance between two different datapoints can be measured by,

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

Distance between Neighbors continued…

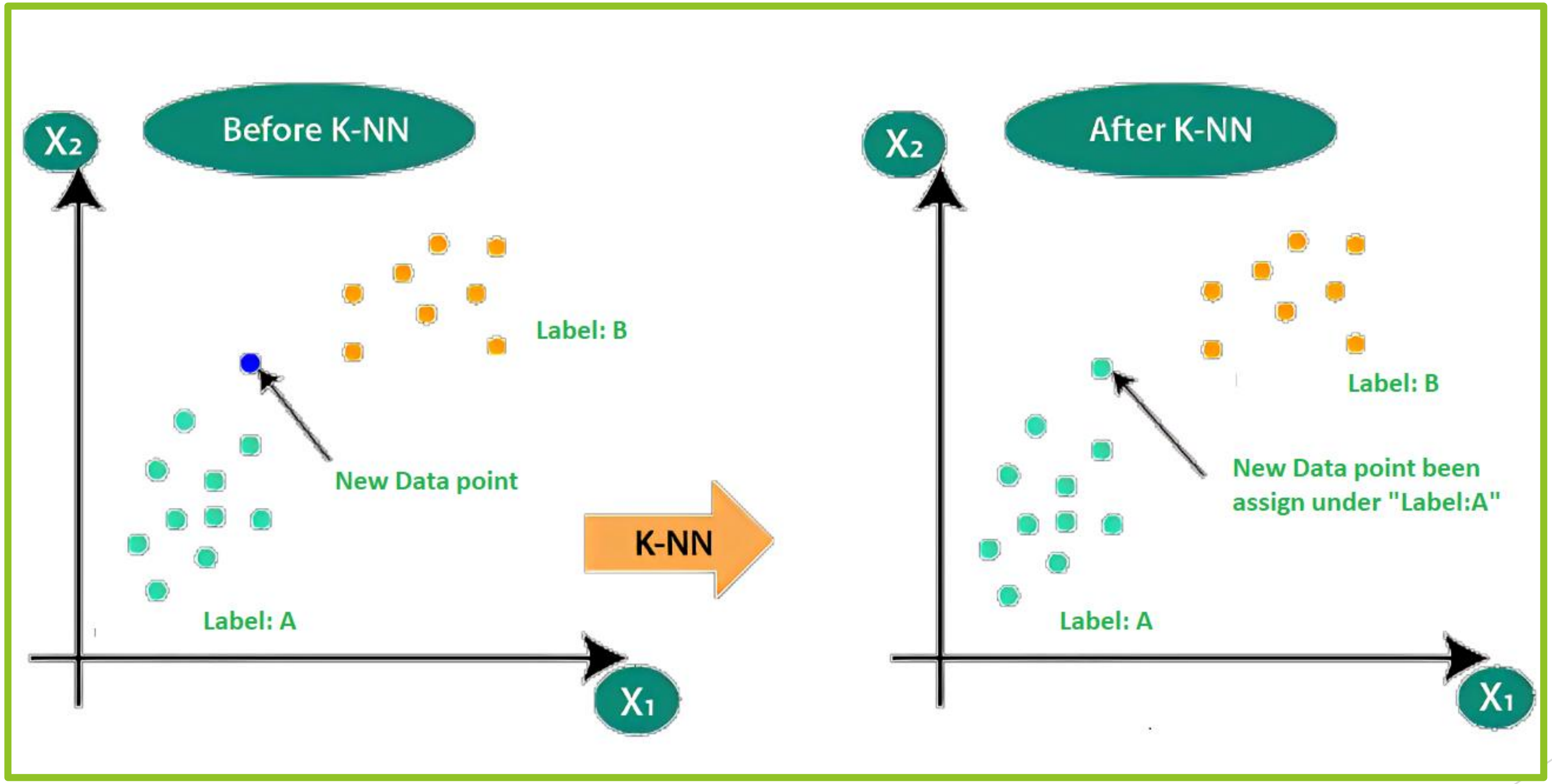
The **Euclidean distance** between two points in a **three-dimensional** space can be correspondingly measured by,

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

So, **Euclidean distance** between points ***$(a_1 , a_2 ,\ldots.. a_n)$*** and ***$(b_1 , b_2 ,\ldots.. b_n)$*** in **n-dimensional** space can be generalized by,

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Why measuring Distance is important?



Some other methods used for measuring distance

- Manhattan Distance
- Minkowski Distance
- Hamming Distance

Example of K-NN

Customer	Age	Income £1000	Number of Credit Cards	Class
George	35	35	3	No
Rachel	22	50	2	Yes
Steve	63	200	1	No
Tom	59	170	1	No
Anne	25	40	4	Yes
John	37	50	2	?????

Assume, you as a bank want to decide whether John can be provided with a Loan, the result can be either "Yes" or, "No" in this case.

Example of K-NN Continued…

Customer	Age	Income £1000	Number of Credit Cards	Class
George	35	35	3	No
Rachel	22	50	2	Yes
Steve	63	200	1	No
Tom	59	170	1	No
Anne	25	40	4	Yes
John	37	50	2	?????

Distance from “John”:

Sqrt $[(35 – 37)^2 + (35 – 50)^2 + (3 -2)^2] = 15.16$

Sqrt $[(22 – 37)^2 + (50 – 50)^2 + (2 -2)^2] = 15$

Sqrt $[(63 – 37)^2 + (200 – 50)^2 + (1 -2)^2] = 152.23$

Sqrt $[(59 – 37)^2 + (170 – 50)^2 + (1 -2)^2] = 122$

Sqrt $[(25 – 37)^2 + (40 – 50)^2 + (4 -2)^2] = 15.74$

Now, we have the distance of the new datapoint with all the existing datapoints exist in the dataset.

Example of K-NN Continued…

Customer	Age	Income £1000	Number of Credit Cards	Class
George	35	35	3	*No*
Rachel	22	50	2	*Yes*
Steve	63	200	1	No
Tom	59	170	1	No
Anne	25	40	4	*Yes*
John	37	50	2	*Yes*

Let's say K=3 and we will compare "3" minimum distances' corresponding "Class" values for our experiment.

Three minimum Distance from "John" are:

Sqrt $[(35 – 37)^2 + (35 – 50)^2 + (3 -2)^2]$ = *15.16*

Sqrt $[(22 – 37)^2 + (50 – 50)^2 + (2 -2)^2]$ = *15*

Sqrt $[(63 – 37)^2 + (200 – 50)^2 + (1 -2)^2]$ = 152.23

Sqrt $[(59 – 37)^2 + (170 – 50)^2 + (1 -2)^2]$ = 122

Sqrt $[(25 – 37)^2 + (40 – 50)^2 + (4 -2)^2]$ = *15.74*

Example of K-NN Continued…

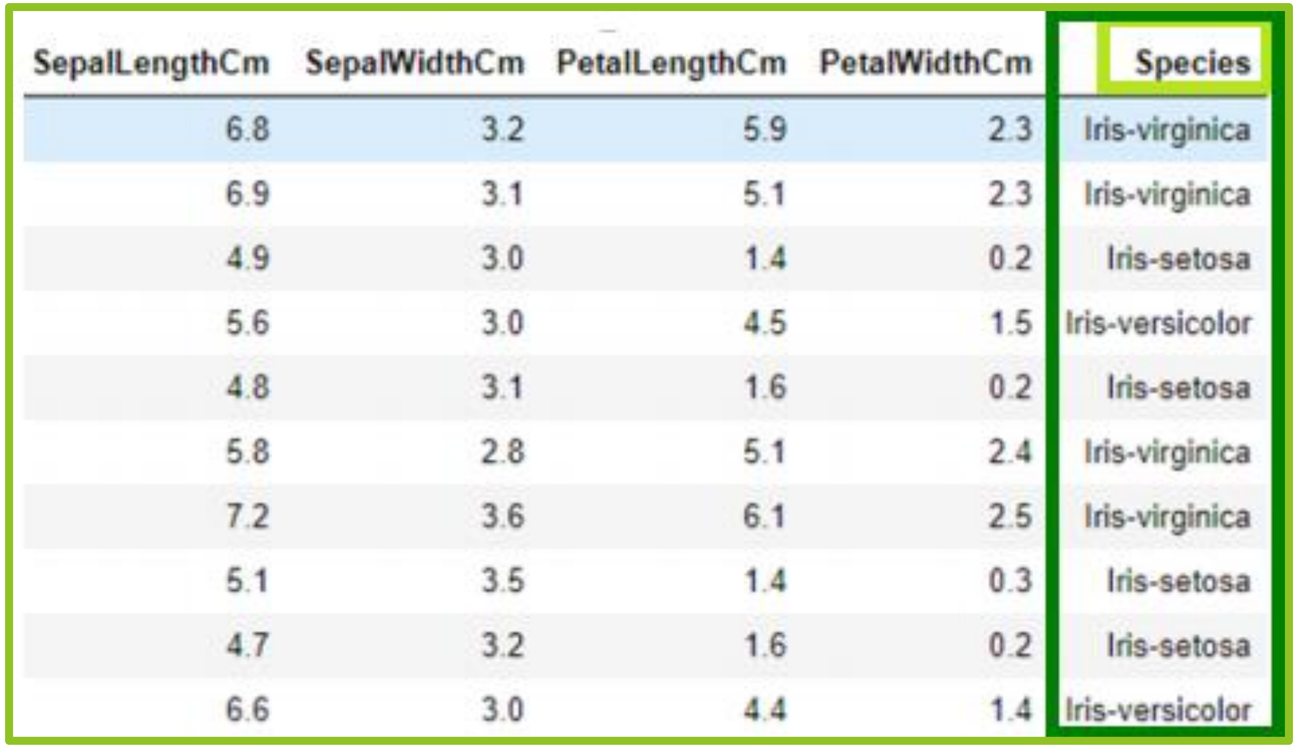
Customer	Age	Income £1000	Number of Credit Cards	Class
George	35	35	3	*No*
Rachel	22	50	2	*Yes*
Steve	63	200	1	No
Tom	59	170	1	No
Anne	25	40	4	*Yes*
John	37	50	2	*Yes*

Now, John has been classified as "Yes" for being provided with the loan.

Binary-Class Classification and Muti-Class Classification Problem(s)

Like Datasets having "Binary" Class values, K-NN can also be applied on Datasets that have more than two Class values or, "Multiple" Class values.

Multi-Class Classification Example



SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
6.8	3.2	5.9	2.3	Iris-virginica
6.9	3.1	5.1	2.3	Iris-virginica
4.9	3.0	1.4	0.2	Iris-setosa
5.6	3.0	4.5	1.5	Iris-versicolor
4.8	3.1	1.6	0.2	Iris-setosa
5.8	2.8	5.1	2.4	Iris-virginica
7.2	3.6	6.1	2.5	Iris-virginica
5.1	3.5	1.4	0.3	Iris-setosa
4.7	3.2	1.6	0.2	Iris-setosa
6.6	3.0	4.4	1.4	Iris-versicolor

Here, in this Dataset, we can see three different Class Values available.

How to determine the value of “K”

Theoretically, choosing “K” value for Binary Classification is concerned with two following points:

- K = sqrt (Total number of Data points)
- Odd Number of K is always selected to avoid any confusion between two classes

For further explanation on why the square root of N is a good estimation of K, you can explore the proofs in Chapters 5, 6, 11 & 26:

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer. https://doi.org/10.1007/978-1-4612-0711-5.

K-NN is applicable on Numeric Data only?

Age	Grade	Rank	Badge	Class
20	Pass	First	Red	Accepted
22.5	Fail	Third	Yellow	Rejected
21	Pass	First	Blue	Accepted
23	Pass	Second	Green	Accepted

22.5	**Pass**	**Second**	**Blue**	**???**

K-NN is applicable on Numeric Data only? Continued…

No!!!

It can work on Continuous, Binary, Ordinal and even Nominal Data as well, however, some prior modification is required for that:

K-NN is applicable on Numeric Data only?

Age	Grade	Rank	Badge	Class
20	Pass	First	Red	Accepted
22.5	Fail	Third	Yellow	Rejected
21	Pass	First	Blue	Accepted
23	Pass	Second	Green	Accepted

Before starting to measure the distance of the new data point following moderations can be done on the example dataset:

1. For Ordinal values, the values can be converted to numeric ones in their order, e.g., 1,2 3…
2. For Binary values, they can be converted into 0s and 1s
3. For Nominal values, it is a bit trickier as it requires the usage of Hamming distance between the values if the attribute

Advantages of K-NN

- As a Classification Algorithm, this is very simple and Intuitive
- It can be applied to the data from any Distribution
- This technique can work good on small dataset

Disadvantages of K-NN

- Takes more time to Classify a new, unseen Datapoint as calculation and comparison both takes time
- Most of the time, choosing “K” value is tricky
- Likely to determine less accurate output when the dataset is too large

Some Applications of K-NN Classification

- Used in Banking System
- Calculating Credit Ratings
- Election or, Voting
- Used to find “Missing” value
- Used for pattern recognition
- Used to measure document similarity

and many more…

Overview: K-NN

- Each instance present are numerical attribute(s)
- Each Datapoint consists of input(s) must associated with a Label
- Classification is done by comparing “K” number of nearest features in the Dataset

Do you have any
QUESTIONS?