Redesigning Common Mind & Business Towards Excellence

3P urpose rocess eople Culture

sns INSTITUTIONS

www.snsgroups.com

1st GenAI Powered Design Thinking FrameWork

Build an Entrepreneurial Mindset Through Our Design Thinking FrameWork

SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

23AMB201 - MACHINE LEARNING

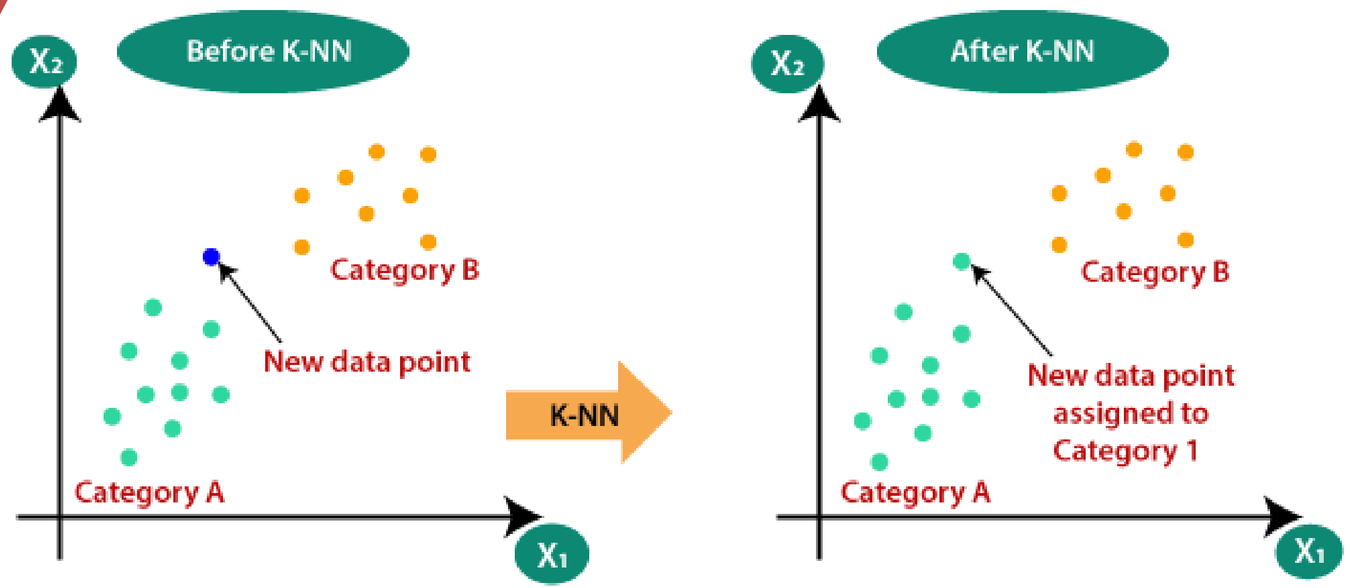
II YEAR IV SEM

UNIT IV – UNSUPERVISED LEARNING ALGORITHM

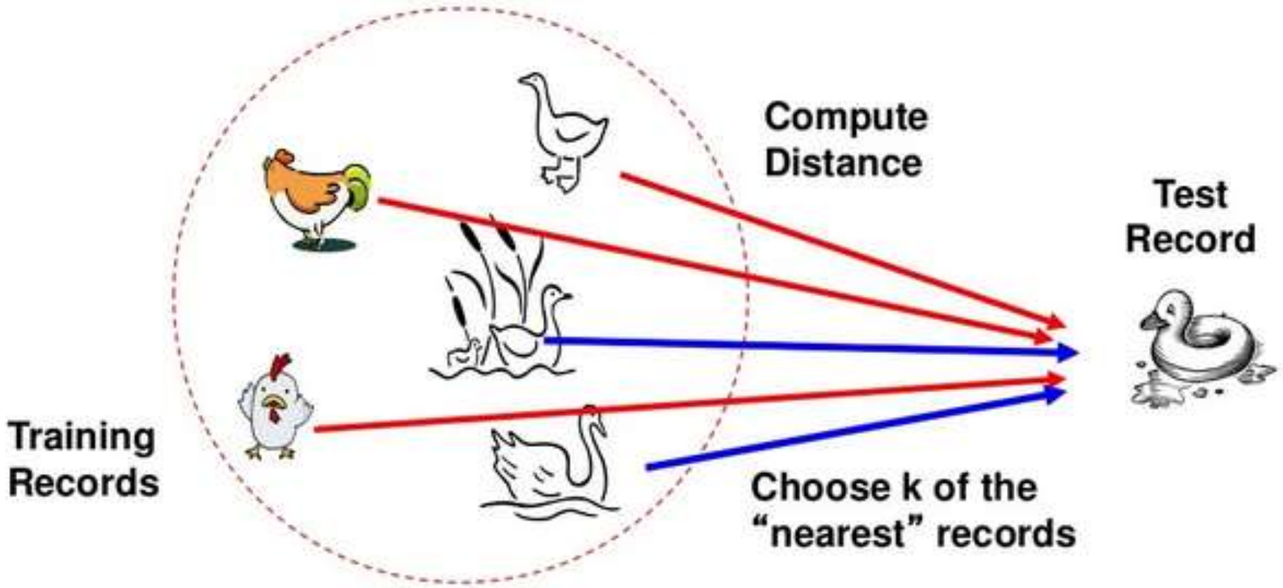
TOPIC 25 – kNN Algorithm

kNN - Definition

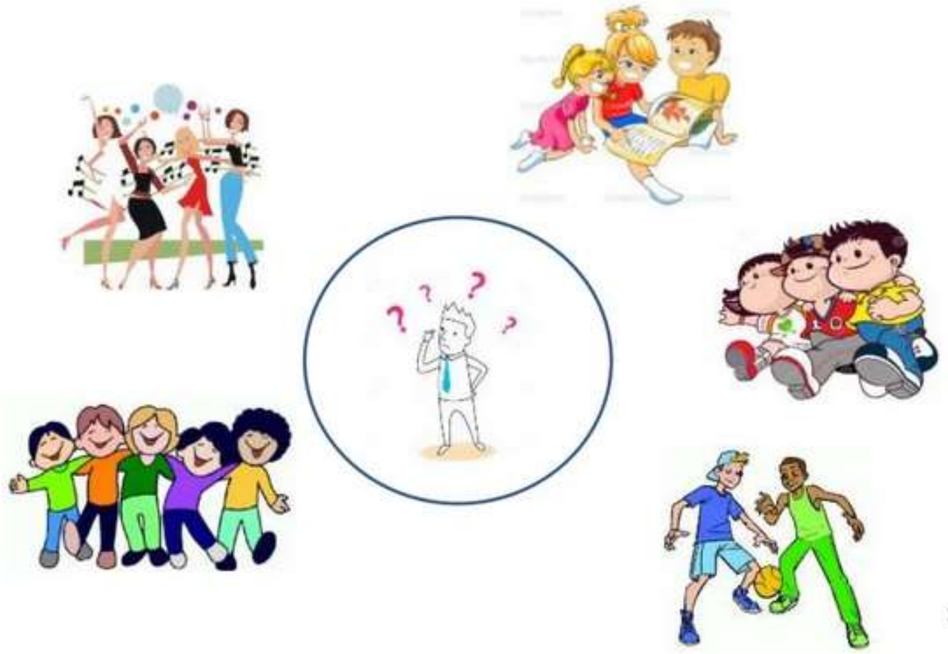
k – Nearest Neighbors is one of the simplest Supervised Learning algorithm mostly used for classification and also regression - Classify the data point based on how its neighbors are classified.



Real case



- Tell me about your friends(*who your neighbors are*) and *I will tell you who you are*.



How do choose k value

Choosing the right value of k is a process called **parameter tuning** and its important for accuracy

If k=1 then overfitting

If k=too large underfitting

Choose the k value using the formula **sqrt(n)**

N: number of dataset

When do we use kNN?

We can use KNN when

Data is labeled

Dog

Dataset is small

Because KNN is a 'lazy learner' i.e. doesn't learn a discriminative function from the training set

Data is noise free

Weight(x2)	Height(y2)	Class
51	167	Underweight
62	182	one-fourty
69	176	23
64	173	hello kitty
65	172	Normal

Noise

How does kNN Algorithm work?

Consider a dataset having two variables: height (cm) & weight (kg) and each point is classified as Normal or Underweight

Weight(x2)	Height(y2)	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

How does kNN Algorithm work?

On the basis of the given data we have to classify the below set as Normal or Underweight using KNN

57 kg	170 cm	?

How to calculate the distance?

The relative difference between two objects in a problem domain

Euclidean Distance

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Manhattan Distance

$$D_m = \sum_{i=1}^{n} |p_i - q_i|$$

Minkowski Distance

$$D = \left(\sum_{i=1}^{n} |p_i - q_i|^p \right)^{1/p}$$

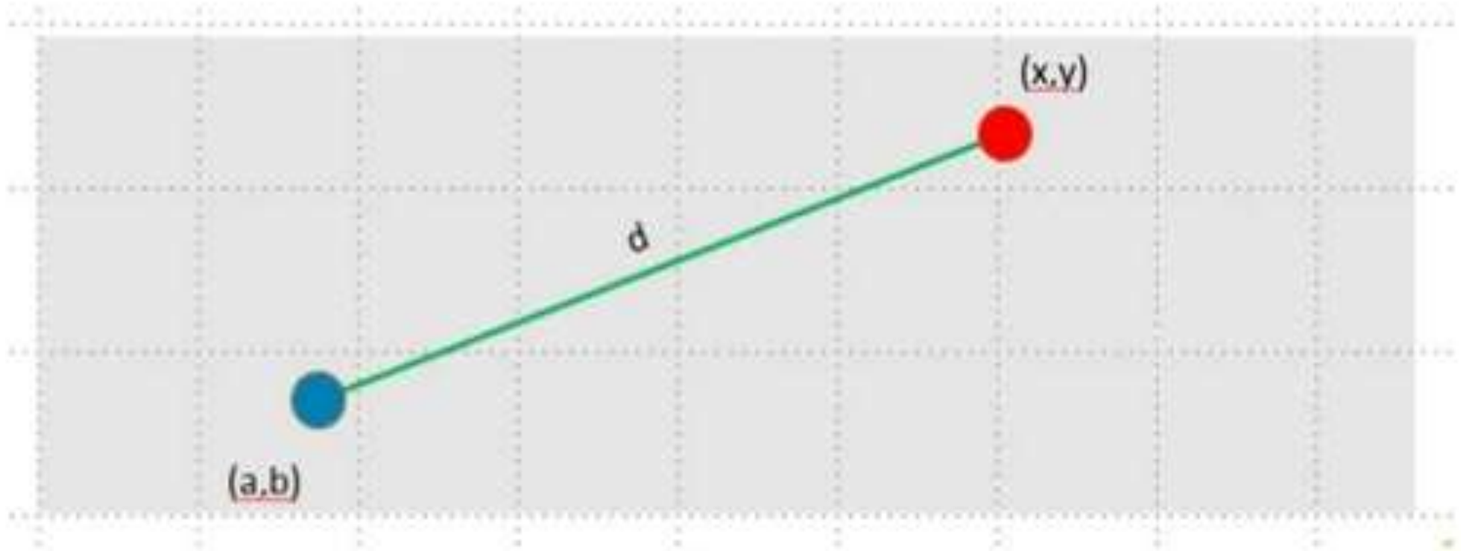
Hamming Distance: (Euclidean and Manhattan)

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

Euclidean distance

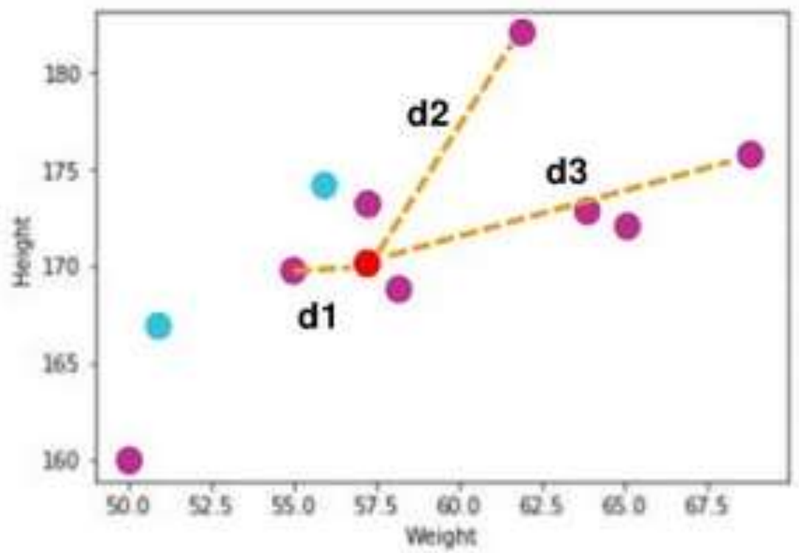
According to the **Euclidean distance** formula, the **distance** between two points in the plane with coordinates (x, y) and (a, b) is given by:

$$dist(d) = \sqrt{(x - a)^2 + (y - b)^2}$$



Calculation

Let's calculate it to understand clearly:



dist(**d1**)= √(170-167)² + (57-51)² ~= 6.7

dist(**d2**)= √(170-182)² + (57-62)² ~= 13

dist(**d3**)= √(170-176)² + (57-69)² ~= 13.4

Similarly, we will calculate Euclidean distance of unknown data point from all the points in the dataset

● Unknown data point

Calculation

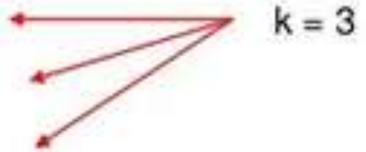
Where (x1, y1) = (57, 170) whose class we have to classify

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

Calculation

Now, lets calculate the nearest neighbor at k=3

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2



k = 3

57 kg	170 cm	?

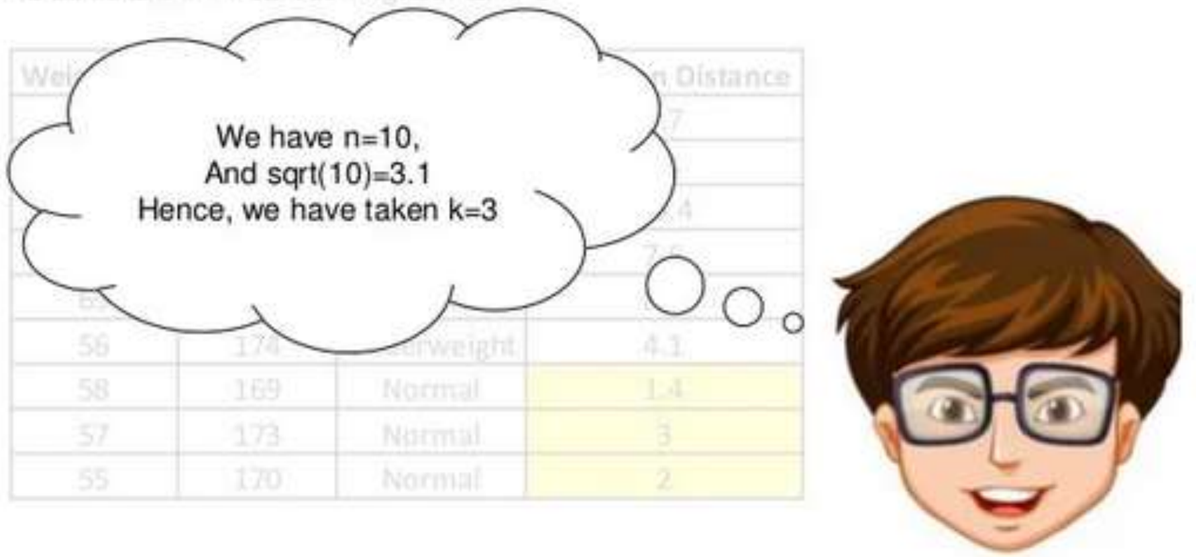
Calculation

Now, lets calculate the nearest neighbor at k=3

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

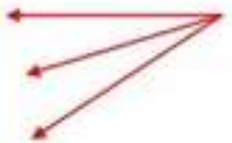
k = 3

57 kg	170 cm	?



Calculation

Class	Euclidean Distance
Underweight	6.7
Normal	13
Normal	13.4
Normal	7.6
Normal	8.2
Underweight	4.1
Normal	1.4
Normal	3
Normal	2



k = 3

So, majority neighbors are pointing towards '*Normal*'

Hence, as per KNN algorithm the class of (57, 170) should be 'Normal'

Steps in Knn

Step 1: Select the value of K neighbors(say k=5)

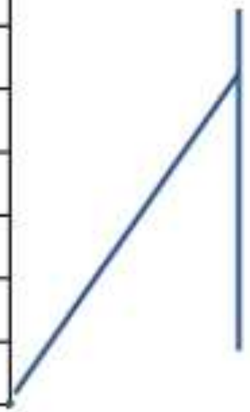
Step 2: Find the K (5) nearest data point for our new data point based on Euclidean distance(which we discuss later)

Step 3: Among these K data points count the data points in each category

Step 4: Assign the new data point to the category that has the most neighbors of the new data point

Workout-knn

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

Workout-k means

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

References

1. https://www.simplilearn.com/tutorials/machine-learning-tutorial/knn-in-python
2. https://www.youtube.com/watch?v=HZT0lxD5h6k