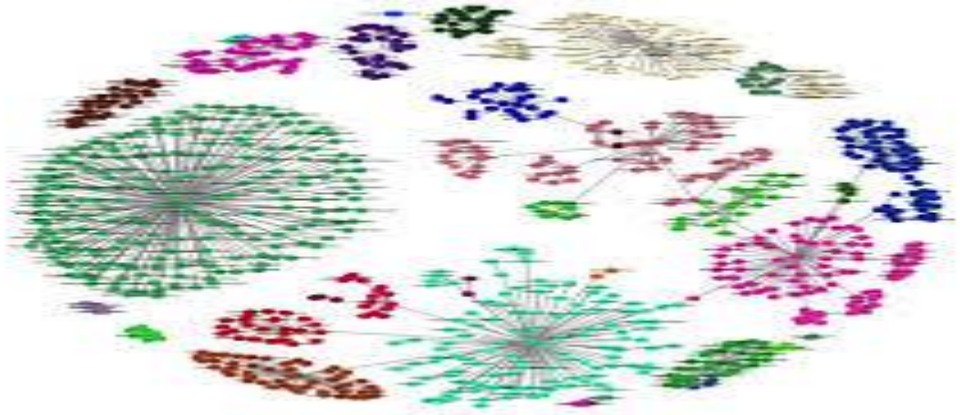
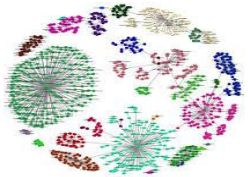


Key ML Algorithms- KNN

INTERNSHIPSTUDIO

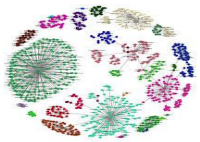
Agenda



INTERNSHIPSTUDIO

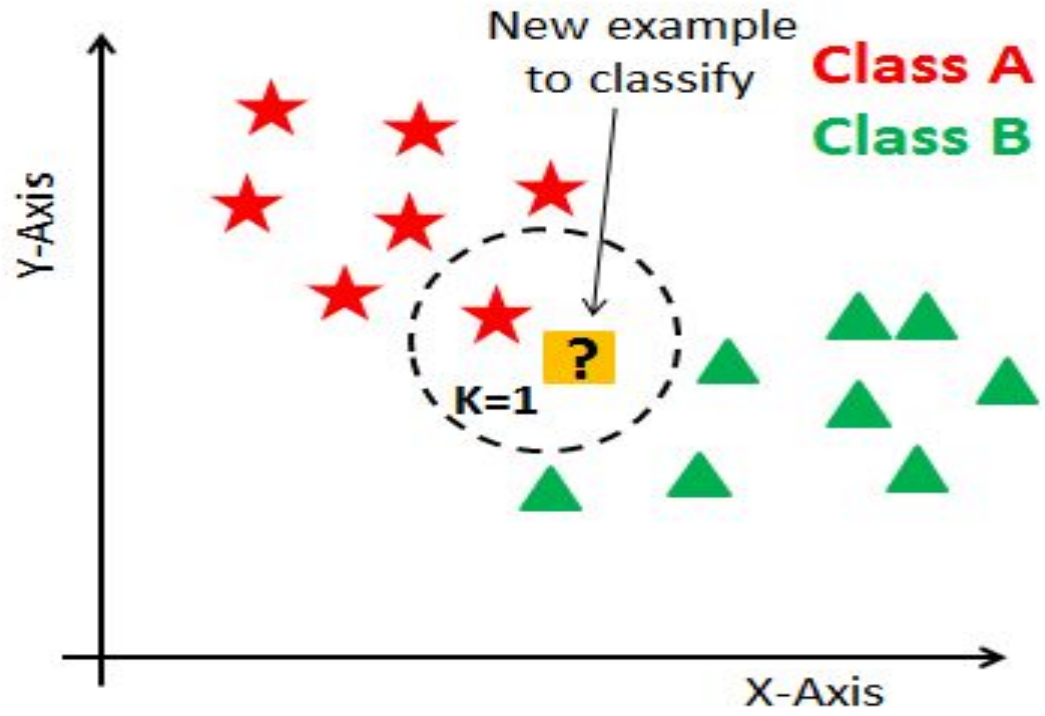
- K-nearest neighbors
- KNN Classifier
- KNN Regressor
- Dimensionality Reduction
- Principal Component Analysis
- Singular Value Decomposition

What is KNN?



INTERNSHIPSTUDIO

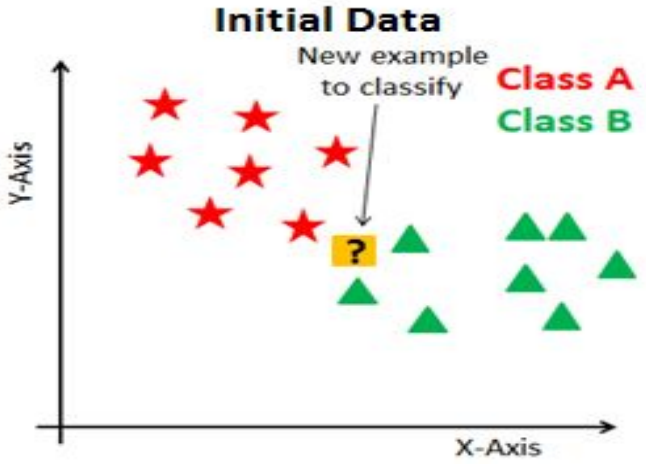
- K Nearest Neighbor is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure.
- It is mostly used to classifies a data point based on how its neighbors are classified.

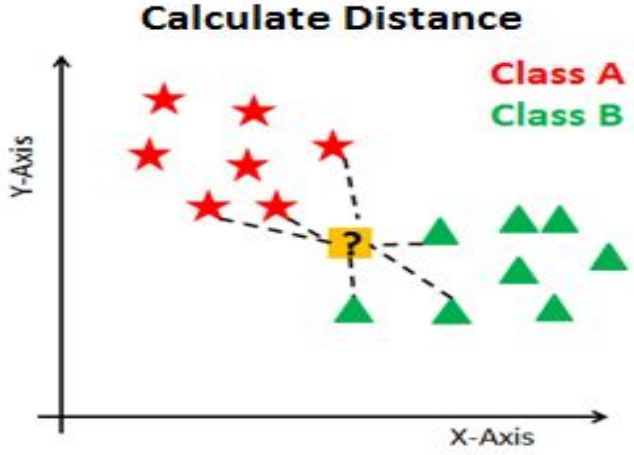


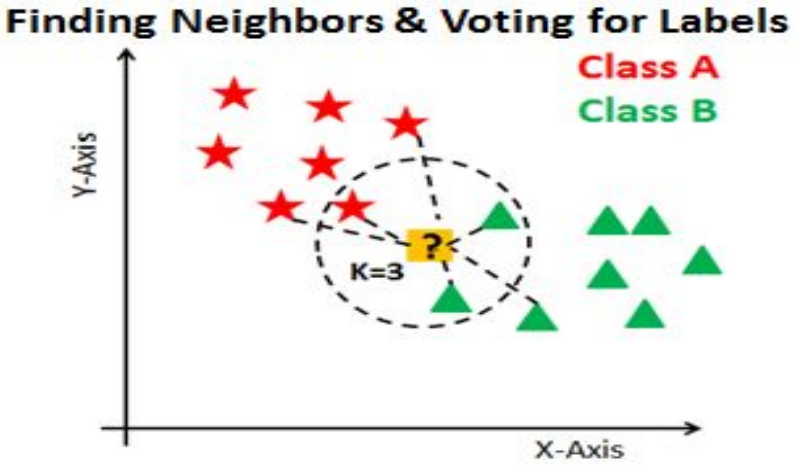
KNN Steps

KNN has the following basic steps:

1. **Calculate distance**
2. **Find closest neighbors**
3. **Vote for labels**

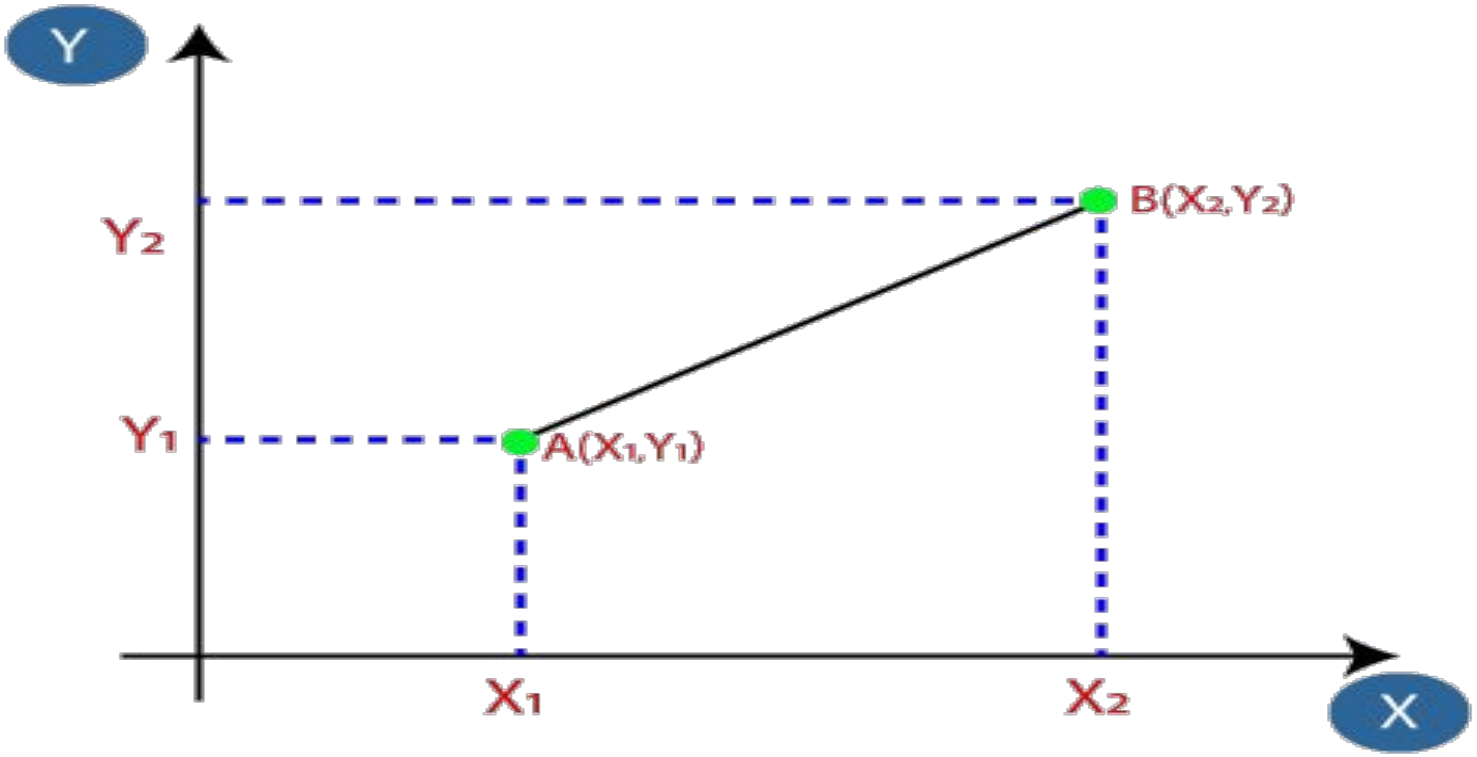






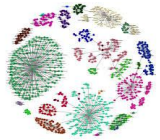
Euclidean Distance

Euclidean Distance represents the shortest distance between two points.



Euclidean Distance between A_1 and $B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

Choosing the right value for K

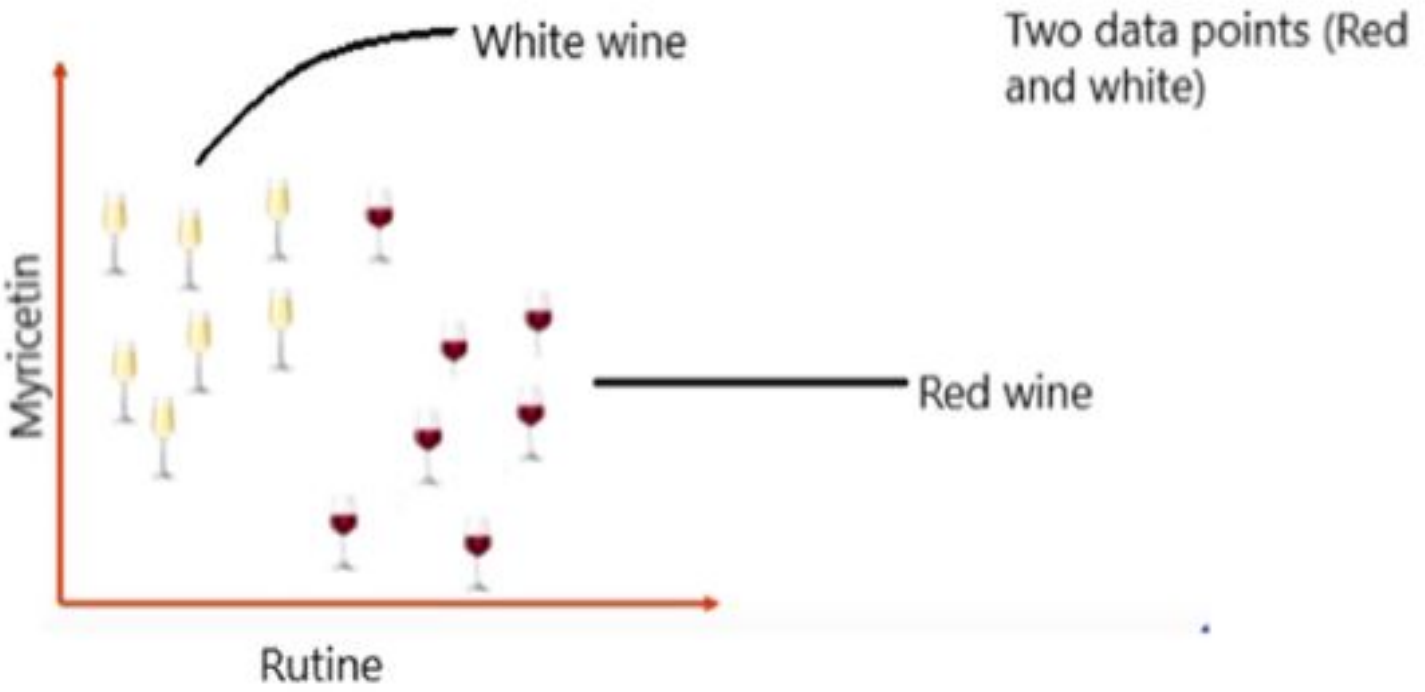


INTERNSHIPSTUDIO

- To get the right K, you should run the KNN algorithm several times with different values of K and select the one that has the least number of errors.
- As your value of K increases, your prediction becomes more stable due to the majority of voters.
- When you start receiving an increasing number of errors, you should know you are pushing your K too far.

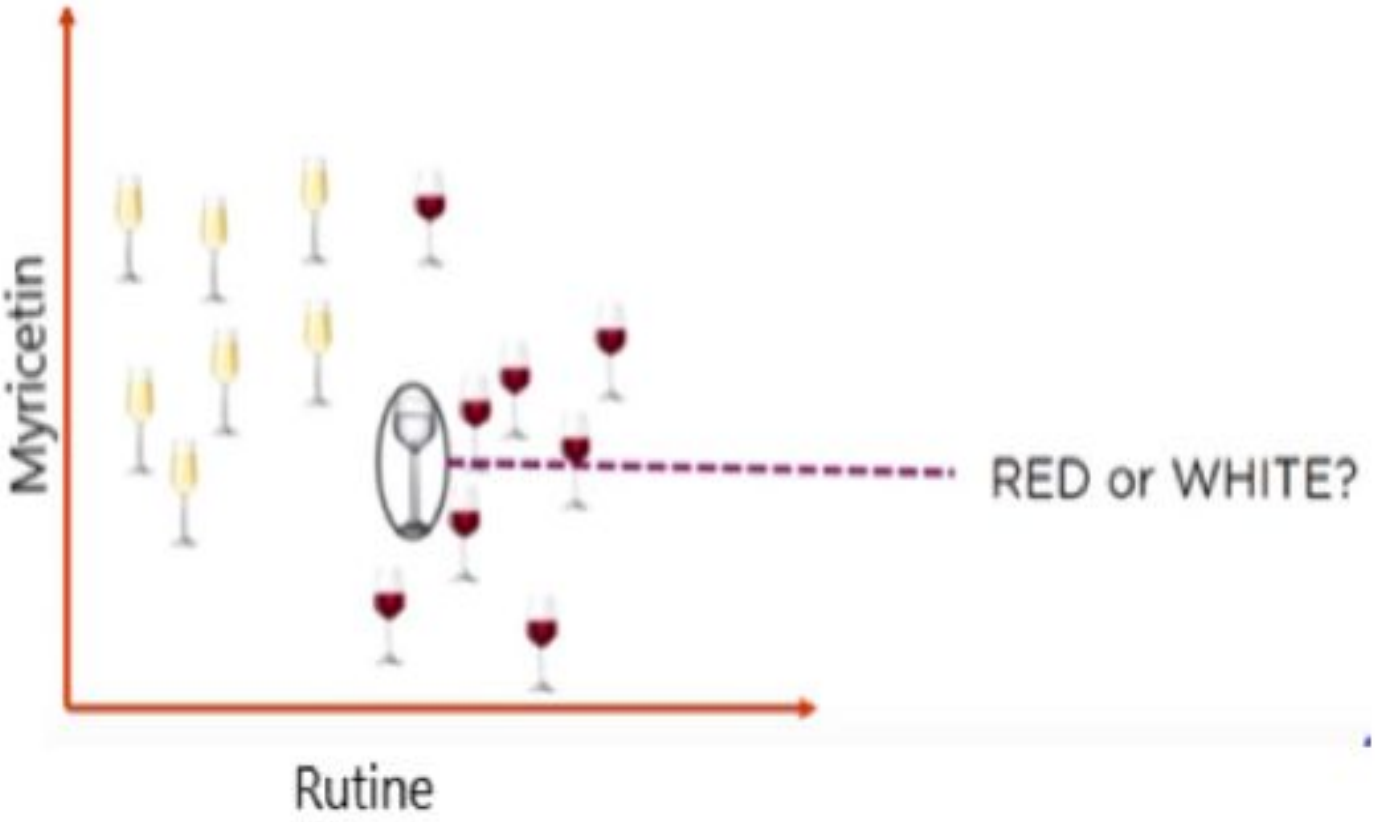
Example- Red and White wines

- Let’s take below wine example. Two chemical components called Rutine and Myricetin. Consider a measurement with two data points, **Red and White wines**.



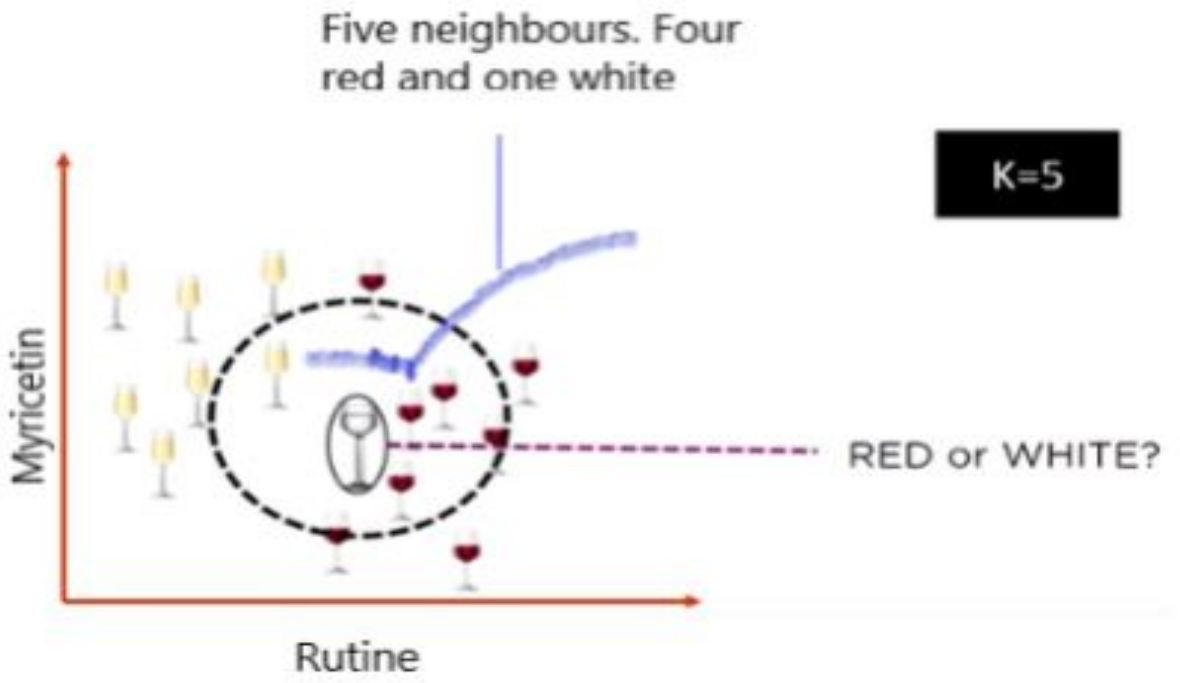
Example- Red and White wines

- Suppose, if we add a new glass of wine in the dataset. We would like to know whether the new wine is red or white?



Example- Red and White wines

- So, we need to find out what the neighbours are in this case. Let's say k = 5 and the new data point is classified by the majority of votes from its five neighbours.



- The new point would be classified as red since four out of five neighbours are red.

More Example

K-NN IN ACTION

- Consider the following data:
  A={weight,color}
  G={Apple(A), Banana(B)}

- We need to predict the type of a fruit with:
  weight = 378
  color = red

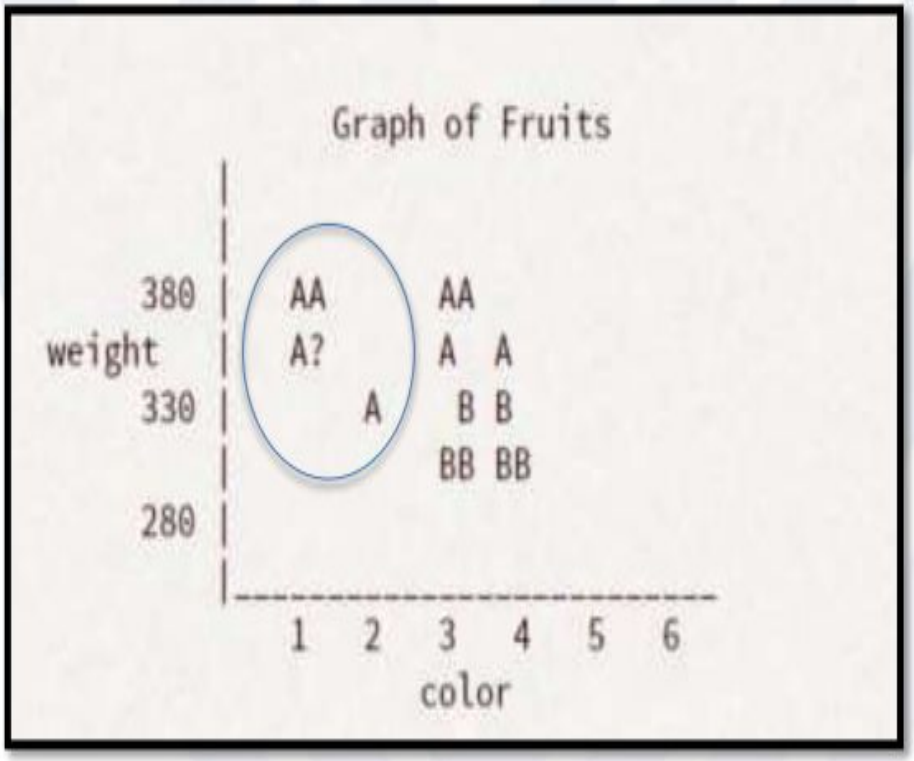
weight (g)	color	Type of fruit
303	3	Banana
370	1	Apple
298	3	Banana
277	3	Banana
377	4	Apple
299	3	Banana
382	1	Apple
374	4	Apple
303	4	Banana
309	3	Banana
359	1	Apple
366	1	Apple
311	3	Banana
302	3	Banana
373	4	Apple
305	3	Banana
371	3	Apple

More Example

INTERNSHIPSTUDIO

PLOTTING

- Using K=3,
  Our result will be,



KNN Advantages & Disadvantages

Advantages :-

- Quick calculation time
- Simple algorithm – to interpret
- Versatile – useful for regression and classification
- High accuracy – you do not need to compare with better-supervised learning models
- No assumptions about data – no need to make additional assumptions, tune several parameters, or build a model. This makes it crucial in nonlinear data case.

Disadvantages :-

- Accuracy depends on the quality of the data
- With large data, the prediction stage might be slow
- Sensitive to the scale of the data and irrelevant features
- Require high memory – need to store all of the training data
- Given that it stores all of the training, it can be computationally expensive

K Nearest Neighbor

INTERNSHIPSTUDIO

K-NN VARIATIONS

- Weighted K-NN: Takes the weights associated with each attribute. This can give priority among attributes.

Ex: For the data,

Weight: $w(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\lambda |\mathbf{x} - \mathbf{x}_i|_2^2\right)$

Probability: $\Pr(y|\mathbf{x}) = \frac{\sum_{i=1}^{n} w(\mathbf{x}, \mathbf{x}_i)\delta(y, y_i)}{\sum_{i=1}^{n} w(\mathbf{x}, \mathbf{x}_i)}$

Where, $\delta(y, y_i) = \begin{cases} 1 & y = y_i \\ 0 & y \neq y_i \end{cases}$

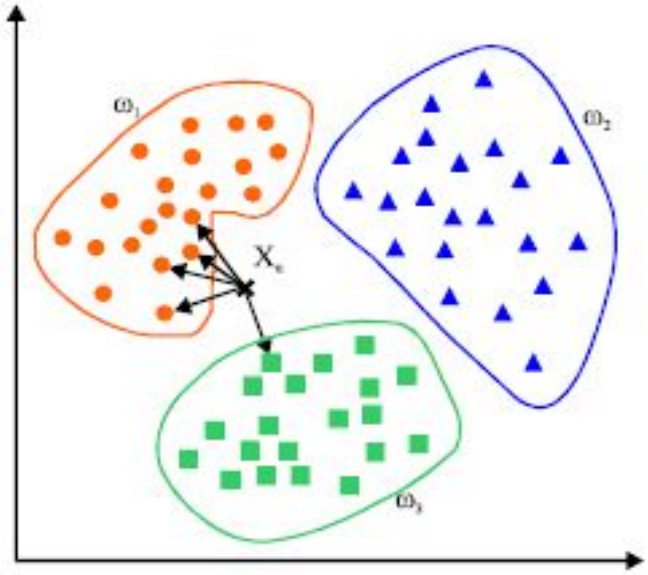
	$d(x_i, x)$	w_i
x_1	2	0.5
x_2	2	0.5
x_3	2	0.5
x_4	2	0.5
x_5	0.7	1/0.7
x_6	0.8	1/0.8

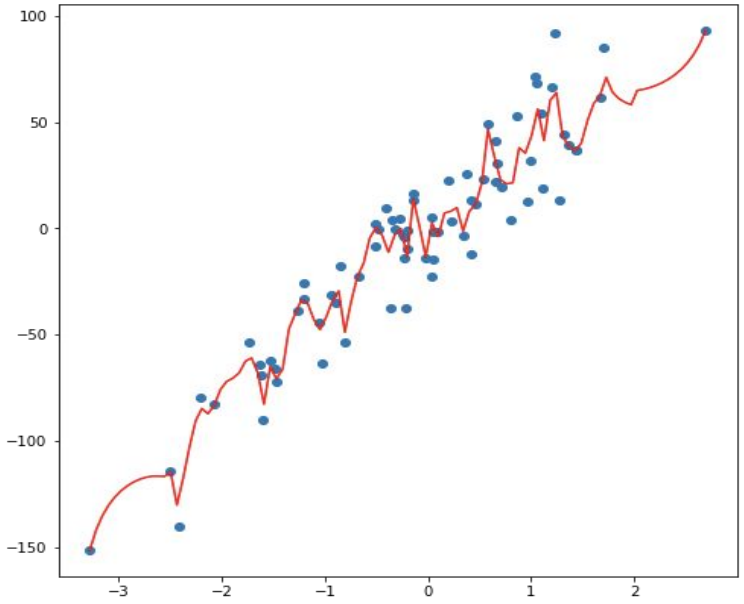
Above is the resulting dataset

KNN Classifier vs KNN Regressor

The key differences are:

1. KNN regression tries to predict the value of the output variable by using a local average.
2. KNN classification attempts to predict the class to which the output variable belong by computing the local probability.





INTERNSHIPSTUDIO

- Q.1 What is KNN?
- Q.2 What is the need of KNN?
- Q.3 What is the origin of KNN?
- Q.4 Can we use KNN in classification problems?