

Artificial Intelligence and Machine Learning :
ACCA6011
KNN and Naïve Bayes

Dr. Seena Joseph

K Nearest Neighbours (KNN)

KNN

A simplest supervised Machine learning Algorithm for Classification and Regression.

It is also named as Lazy learning algorithm

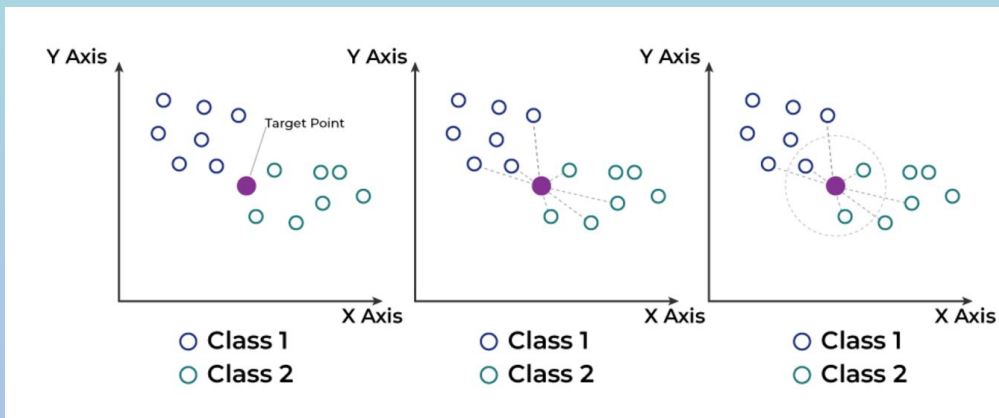
Predictions for the new data points are based on observing the neighbours.

It has no training phase.

K Nearest Neighbours (KNN)

KNN Algorithm

It is a supervised ML algorithm generally used for classification but can also be used for regression tasks. It works by finding the "k" closest data points (neighbours) to a given input and makes a predictions based on the majority class (for classification) or the average value (for regression). Since KNN makes no assumptions about the underlying data distribution it makes it a non-parametric and instance-based learning method.



Classification

1 1 0 0 0 1 0 ...

New Instance = Mode

Regression

1 99 22 53 97 ...

New Instance = Mean

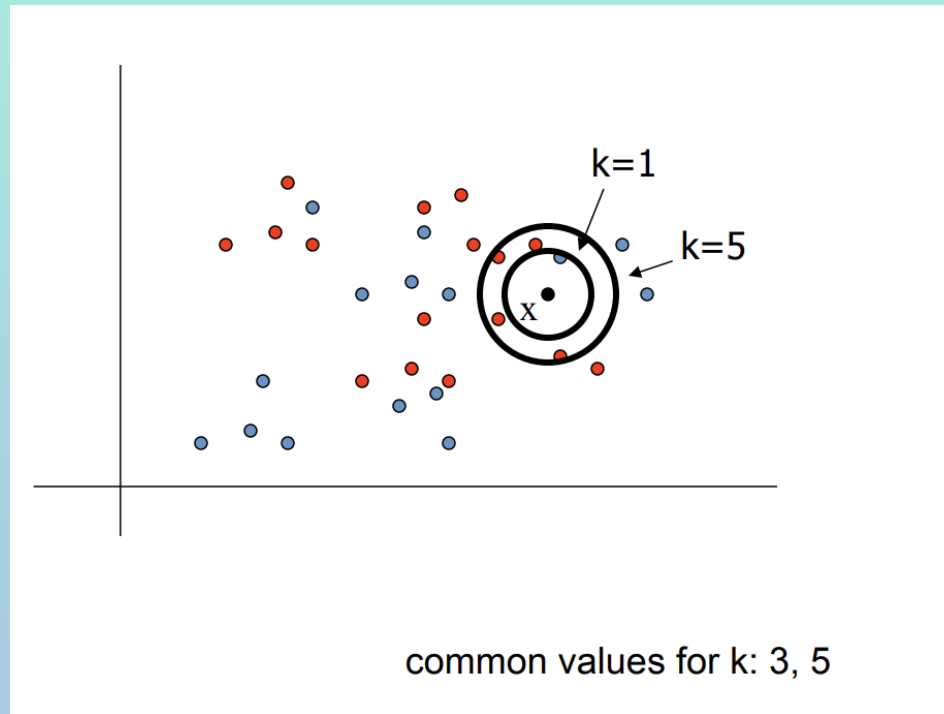
K Nearest Neighbours (KNN)

KNN Algorithm

- Plot the data points
- Locate the new test instance
- Find the distance between the new test instance and rest of the data points
- Sort the distance in ascending order
- Select the first **k** from the sorted list.
 - k is just a number that tells the algorithm how many nearby points or neighbours to look at when it makes a decision.
- For classification majority voting (mode) will be performed to determine the predicted class for the target data point.
- In the regression problem, the class label is calculated by taking average of the target values of K nearest neighbors. The calculated average value becomes the predicted output for the target data point.

K Nearest Neighbours (KNN)

KNN Algorithm



K Nearest Neighbours (KNN)

KNN Algorithm

Example 1: Consider an example where you are working on a regression problem that uses KNN. The distance of the new instance from all other datapoints are listed below:

Distance	Target
3.9	1
4.5	4
7.6	2
2.3	5
3.6	3

What will be the predicted value for the new data point if $k=3$?

K Nearest Neighbours (KNN)

Sort the distance in ascending order. So it is: 2.3, 3.6 and 3.9. So the predicted value is the average of neighbours, 1, 3 and 5 which is 3.

Distance	Target
3.9	1
4.5	4
7.6	2
2.3	5
3.6	3

K Nearest Neighbours (KNN)

KNN Algorithm

Example 2: Consider an example where you are working on a classification problem that uses KNN. The distance of the new instance from all other datapoints are listed below:

Distance	Target
2.3	1
4.7	0
7.6	1
5.8	0
2.5	1

What will be the predicted value for the new data point if $k=3$?

K Nearest Neighbours (KNN)

KNN Algorithm

Sort the distance in ascending order. So it is: 2.3, 2.5 and 4.7. So the predicted value is based on the majority voting (mode value) of the three neighbors which are 1, 1, and 0. So the predicted class will be 1.

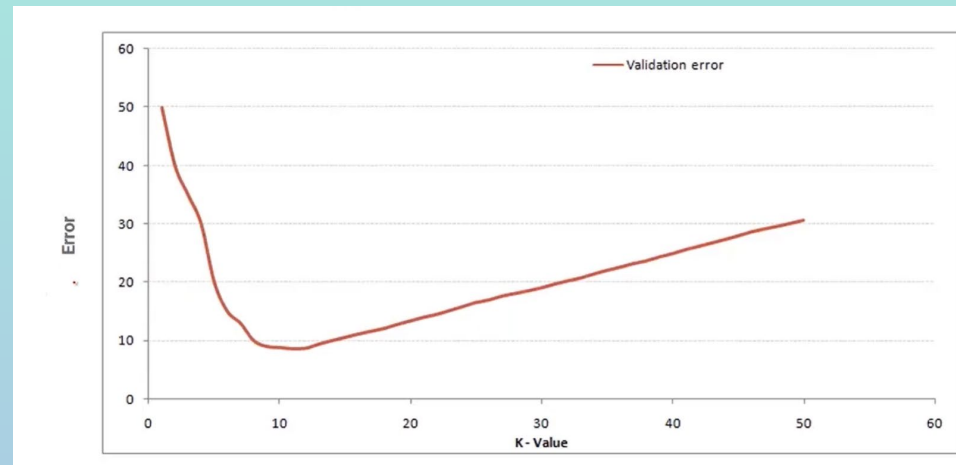
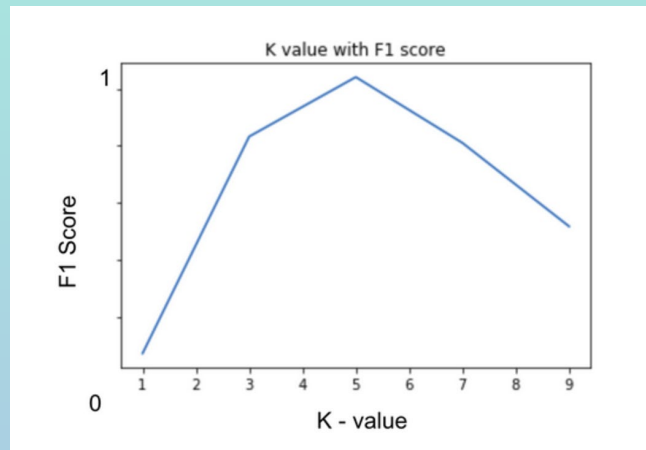
Distance	Target
2.3	1
4.7	0
7.6	1
5.8	0
2.5	1

What will be the predicted value for the new data point if $k=3$?

K Nearest Neighbours (KNN)

KNN Algorithm

Choosing the optimal k value – Elbow method

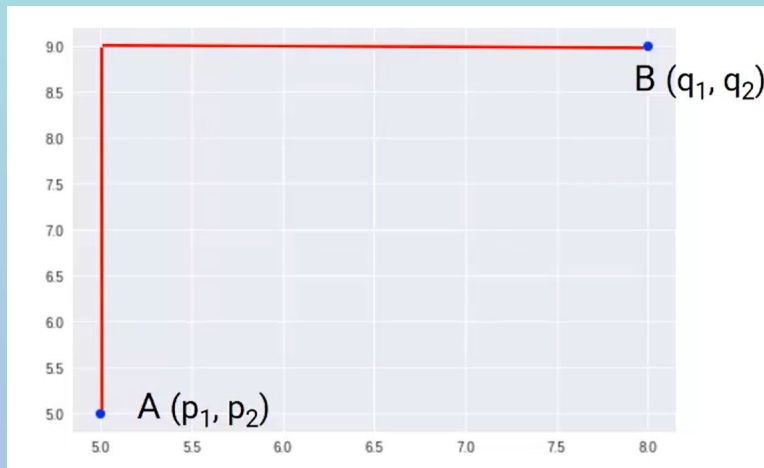


K Nearest Neighbours (KNN)

KNN Algorithm

Distance Measures

- Manhattan Distance
 - Sum of absolute difference between two points across all dimensions
 - $|8 - 5| + |9 - 5| = 3 + 4 = 7$



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

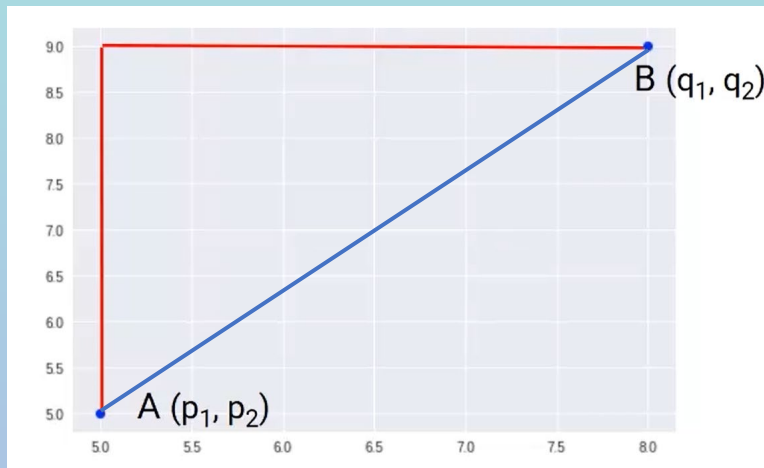
K Nearest Neighbours (KNN)

KNN Algorithm

Distance Measures

- Euclidean Distance

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



K Nearest Neighbours (KNN)

KNN Algorithm

Distance Measures

Example for distance measures for categorical variables in KNN
?

K Nearest Neighbours (KNN)

KNN Algorithm

Advantages

Simple to use: Easy to understand and implement.

No training step: No need to train as it just stores the data and uses it during prediction.

Few parameters: Only needs to set the number of neighbours (k) and a distance method.

Works well with small datasets

Naturally multiclass - KNN directly supports classification with more than two classes, without any modification.

Versatile: Works for both classification and regression problems.

Robust to noisy data (when k is chosen properly)

Disadvantages

Slow with large data: Needs to compare every point during prediction.

Struggles with many features: Accuracy drops when data has too many features.

Poor performance - when the data is imbalance

Can Overfit: It can overfit especially when the data is high-dimensional or not clean.

K Nearest Neighbours (KNN)

KNN Algorithm

Applications

- Recommendation system
- Spam detection
- Customer segmentation
- Medical diagnosis
- Document classification

Naïve Bayes (NB) classifier

NB classifier – It is a Machine Learning Algorithm used for classification task.

- Probabilistic classifier that applies Bayes theorem.
- It is based on the data and some prior knowledge.
- It assumes that all features in the data are independent.
- Each feature is given the same importance.
- It is known for simplicity, efficiency and effectiveness.

Bayes Theorem : Probability of A given that B occurred. A is the hypothesis and B is the evidence.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Naïve Bayes (NB) classifier

Bayes Theorem :

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

$P(A|B)$ is the Probability of A given that B occurred.

$P(B|A)$ is the probability of B given that A is True.

$P(A)$ and $P(B)$ are the independent probabilities of A and B.

Naïve Bayes (NB) classifier

NB classifier Example

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast	0	4
Rainy	3	2
Sunny	2	3
Grand total	5	9

Likelihood Table				
Weather	No	Yes		
Overcast	0	4	=4/14	=.29
Rainy	3	2	=5/14	=.36
Sunny	2	3	=5/14	=.36
Grand total	5	9		
	=5/14	=9/14		
	=.36	.64		

Problem: Players will play if the weather is sunny. Is this statement correct?

Naïve Bayes (NB) classifier

NB classifier

Problem: Players will play if the weather is sunny. Is this statement correct?

Example:

$$P(\text{Yes} \mid \text{Sunny}) = P(\text{Sunny} \mid \text{Yes}) * P(\text{Yes}) / P(\text{Sunny})$$

$$P(\text{Sunny} \mid \text{Yes}) = 3/9 = 0.33$$

$$P(\text{Yes}) = 9/14 = 0.64$$

$$P(\text{Sunny}) = 5/14 = 0.36$$

$$P(\text{Yes} \mid \text{Sunny}) = (0.33 * 0.64) / 0.36 = 0.59 \text{ which has higher probability.}$$

Naïve Bayes (NB) classifier

NB classifier

It is commonly used in text classification (NLP).

Class Activity

- Implement KNN and Naïve Bayes Classifiers

Extra resources

- [Guide to K-Nearest Neighbors \(KNN\) Algorithm \[2025 Edition\]](#)
- [K-Nearest Neighbor\(KNN\) Algorithm - GeeksforGeeks](#)
- [Naive Bayes Classifier in Machine Learning | Analytics Vidhya](#)
- [Naive Bayes Algorithm: Theory, Assumptions & Implementation](#)
- [Naive Bayes Classifiers – GeeksforGeeks](#)