K-Nearest Neighbour: The Distance-Based Machine Learning Algorithm.

Article Published: 29.08.2024

- **LinkedIn:** www.linkedin.com/in/uddit-7258792ab
- **GitHub:** github.com/UDDITwork
- **Website:** udditAimLblogs.com
- **Youtube:** https://rb.gy/gfagcs

TABLE OF CONTENTS

Introduction:

KNN also called K- nearest neighbour is a supervised machine learning algorithm that can be used for classification and regression problems. K nearest neighbour is one of the simplest algorithms to learn. K nearest neighbour is non-parametric i,e. It does not make any assumptions for underlying data assumptions. K nearest neighbour is also termed as a lazy algorithm as it does not learn during the training phase rather it stores the data points but learns during the testing phase. It is a distance-based algorithm

1. KNN Algorithm real-world scenario

Let's take a good look at a related real-world scenario before we get started with this awesome algorithm.

We are often notified that you share many characteristics with your peers, whether it be your thinking process, working etiquette, philosophies, or other factors. As a result, we build friendships with people we deem similar to us.

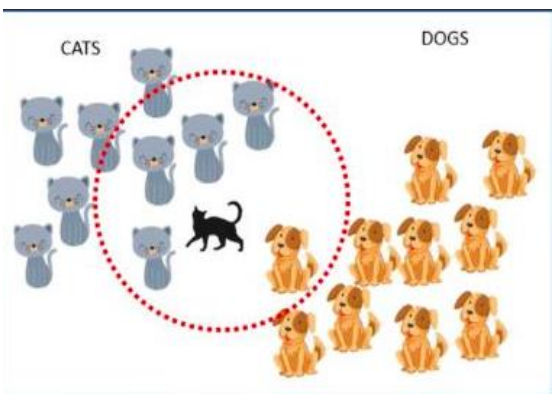
The KNN algorithm employs the same principle. Its aim is to locate all of the closest neighbors around a new unknown data point in order to figure out what class it belongs to. It's a distance-based approach.

Consider the diagram below; it is straightforward and easy for humans to identify it as a "Cat" based on its closest allies. This operation, however, cannot be performed directly by the algorithm.

KNN calculates the distance from all points in the proximity of the unknown data and filters out the ones with the shortest distances to it. As a result, it's often referred to as a distance-based algorithm.

In order to correctly classify the results, we must first determine the value of K (Number of Nearest Neighbours).

In the following diagram, the value of K is 5. Since there are four cats and just one dog in the proximity of the five closest neighbours, the algorithm would predict that it is a cat based on the proximity of the five closest neighbors in the red circle's boundaries.



Here, 'K' is the hyperparameter for KNN. For proper classification/prediction, the value of K must be fine-tuned.

But, How do we select the right value of K?

We don't have a particular method for determining the correct value of K. Here, we'll try to test the model's accuracy for different K values. The value of K that delivers the best accuracy for both training and testing data is selected.

2.DISTANCE METRICS

It is essential to choose the most appropriate distance metrics for a particular dataset. The following are the various distance metrics:

Minkowski Distance-Minkowski distance is calculated where distances are in the form of vectors that have a length and the length cannot be negative.

It would help with the value of K in algorithmic use.

Manhattan Distance-The distance between two points is the sum of the absolute differences of their Cartesian coordinates.

Euclidean Distance- It is a measure of the true straight line distance between two points in Euclidean space.

Cosine Distance-It is used to calculate the similarity between two vectors. It measures the direction and uses the cosine function to calculate the angle between two vectors.

Jaccard Distance-It is similar to cosine distance as both the methods compare one type of attribute distributed among all data. The Jaccard approach looks at the two data sets and finds the incident where both values are equal to 1.

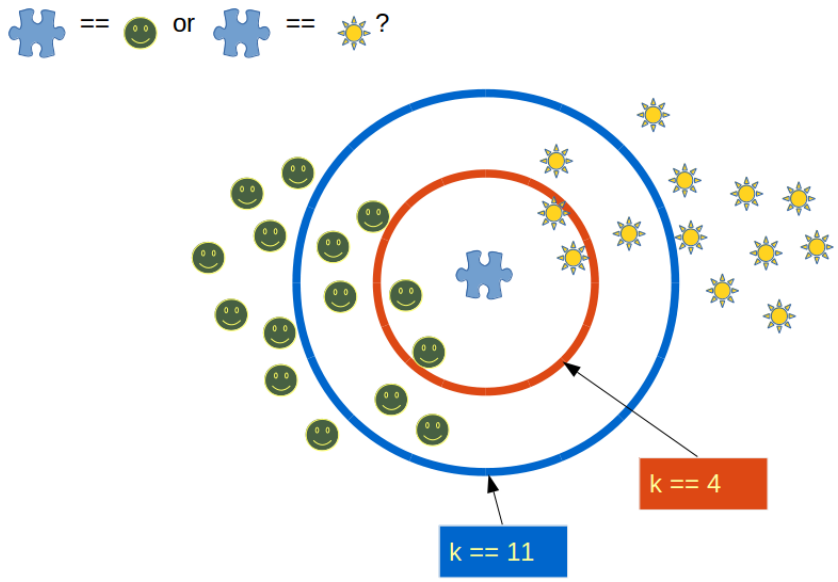
Note!!

It is recommended to always select an odd value of K ~

When the value of K is set to even, a situation may arise in which the elements from both groups are equal. In the diagram below, elements from both groups are equal in the internal "Red" circle (k == 4).

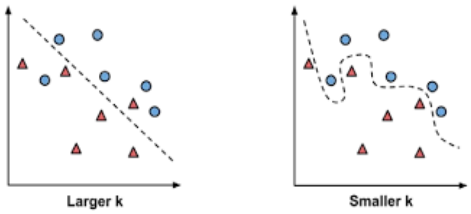
In this condition, the model would be unable to do the correct classification for you. Here the model will randomly assign any of the two classes to this new unknown data.

Choosing an odd value for K is preferred because such a state of equality between the two classes would never occur here. Due to the fact that one of the two groups would still be in the majority, the value of K is selected as odd.



The impact of selecting a smaller or larger K value on the model

- **Larger K value:** The case of underfitting occurs when the value of k is increased. In this case, the model would be unable to correctly learn on the training data.
- **Smaller k value:** The condition of overfitting occurs when the value of k is smaller. The model will capture all of the training data, including noise. The model will perform poorly for the test data in this scenario.



One of the trickiest questions to be asked is how we should choose the K value.

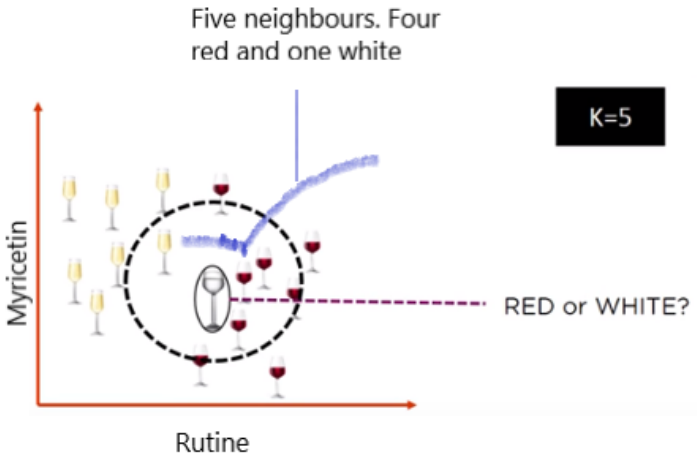
- One should not use a low value of K= 1 because it may lead to overfitting i,e during the training phase performs good but during the testing phase, the model performs badly. Choosing a high value of K can also lead to underfitting i.e it performs poorly during the training and testing phase.
- We should not use even values of K when classifying binary classification problems. Suppose we choose K=4 and the neighbouring 4 points are evenly distributed among classes i.e 2 data points belong to category 1 and 2 data points belong to category 2. In that case, the data point cannot classify as there is a tie between the classes.
- Choose K value based on domain knowledge.
- Plot the elbow curve between different K values and error. Choose the K value when there is a sudden drop in the error rate.

3.How does KNN work for 'Classification' and 'Regression' problem statements?

Classification

When the problem statement is of 'classification' type, KNN tends to use the concept of "Majority Voting". Within the given range of K values, the class with the most votes is chosen.

Consider the following diagram, in which a circle is drawn within the radius of the five closest neighbours. Four of the five neighbours in this neighbourhood voted for 'RED,' while one voted for 'WHITE.' It will be classified as a 'RED' wine based on the majority votes.



Several parties compete in an election in a democratic country like India. Parties compete for voter support during election campaigns. The public votes for the candidate with whom they feel more connected.

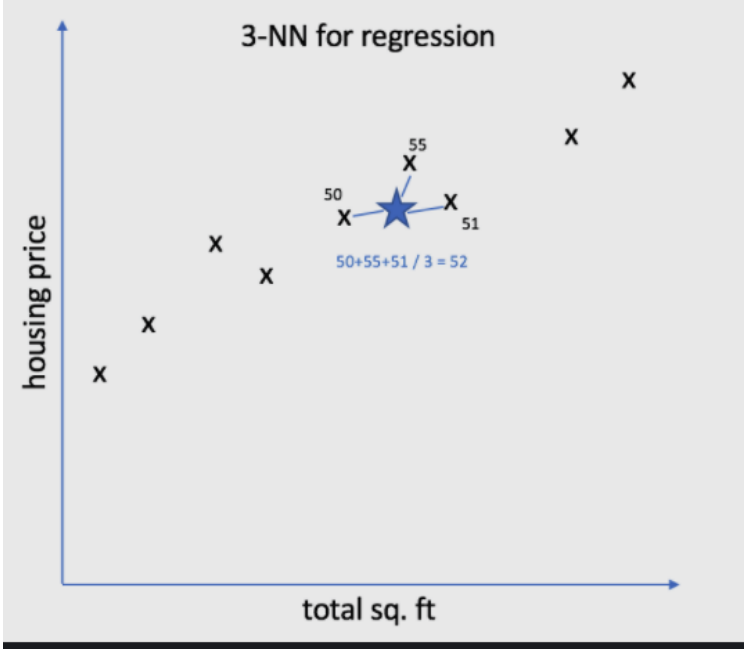
When the votes for all of the candidates have been recorded, the candidate with the most votes is declared as the election's winner.

Regression

KNN employs a mean/average method for predicting the value of new data. Based on the value of K, it would consider all of the nearest neighbours.

The algorithm attempts to calculate the mean for all the nearest neighbours' values until it has identified all the nearest neighbours within a certain range of the K value.

Consider the diagram below, where the value of k is set to 3. It will now calculate the mean (52) based on the values of these neighbours (50, 55, and 51) and allocate this value to the unknown data.



3.Impact of Imbalanced dataset and Outliers on KNN

Imbalanced dataset

When dealing with an imbalanced data set, the model will become biased. Consider the example shown in the diagram below, where the “Yes” class is more prominent.

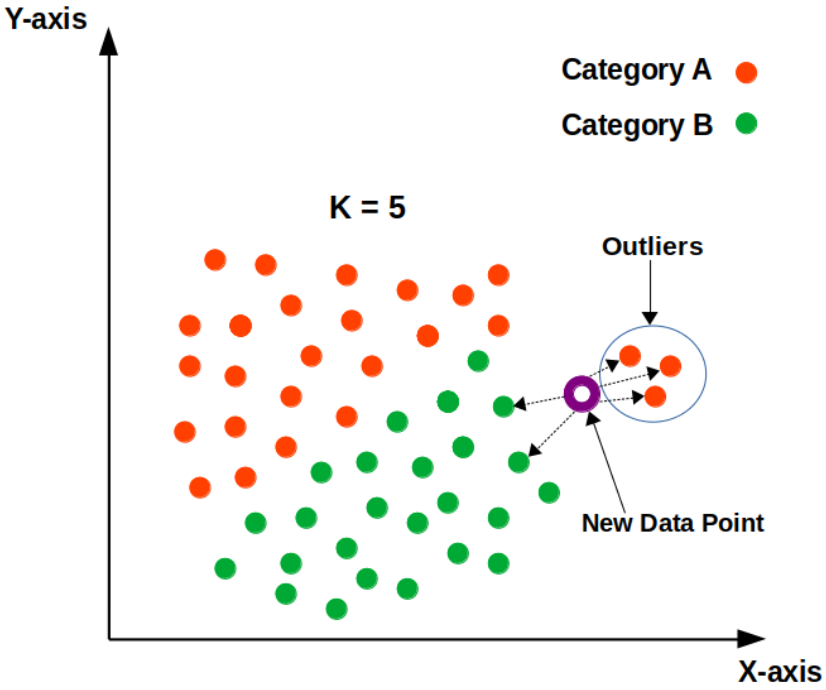
As a consequence, the bulk of the closest neighbours to this new point will be from the dominant class. Because of this, we must balance our data set using either an “Upscaling” or “Downscaling” strategy.

Outliers

Outliers are the points that differ significantly from the rest of the data points.

The outliers will impact the classification/prediction of the model. The appropriate class for the new data point, according to the following diagram, should be “Category B” in green.

The model, however, would be unable to have the appropriate classification due to the existence of outliers. As a result, removing outliers before using KNN is recommended.



4.Importance of scaling down the numeric variables to the same level

Data has 2 parts: –

1) Magnitude

2) Unit

For instance; if we say 20 years then “20” is the magnitude here and “years” is its unit.

Since it is a distance-dependent algorithm, KNN selects the neighbours in the closest vicinity based solely on the magnitude of the data. Have a look at the diagram below; the data is not scaled, so it can not find the closest neighbours correctly. As a consequence, the outcome will be influenced.

The data values in the previous figure have now been scaled down to the same level in the following example. Based on the scaled distance, all of the closest neighbours would be accurately identified.

I have demonstrated the use of KNN and a project based on it in one of my Youtube videos in the Machine Learning lecture series.

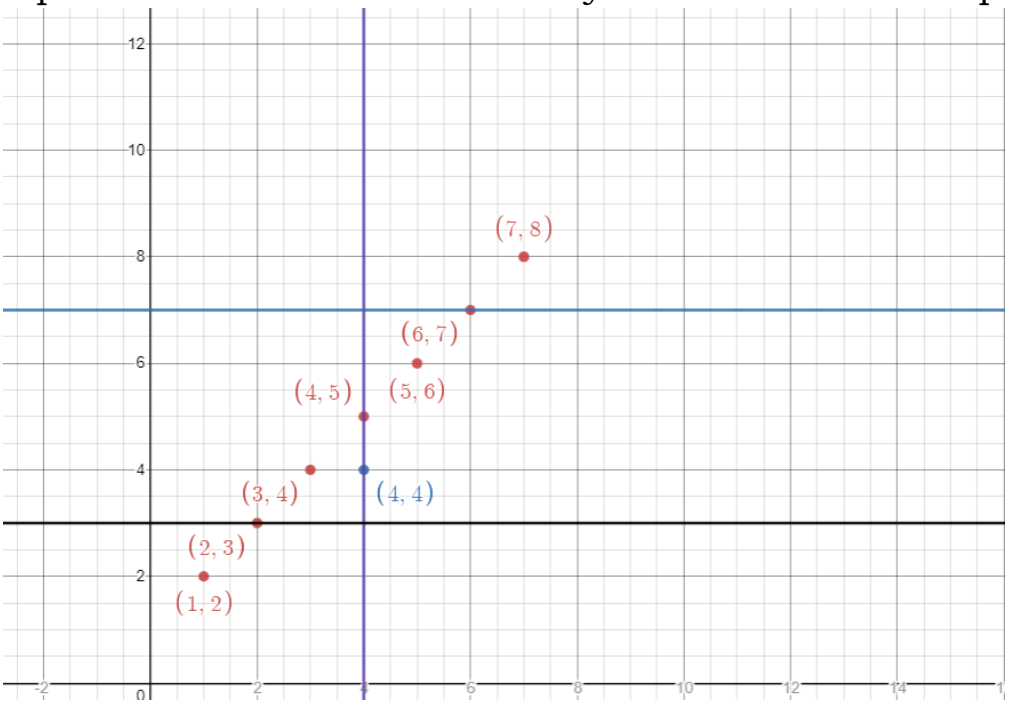
5.Different KNN Algorithms:

Before going forward learning different algorithms of KNN it is important to know what a tree is. A tree is. A tree is a non-linear data structure used to store collections of objects which are nodes that are linked together to represent the hierarchy.,

There are four different algorithms in KNN namely kd_tree,ball_tree, auto, and brute.

kd_tree=kd_tree is a binary search tree that holds more than x,y value in each node of a binary tree when plotted in XY coordinate. To classify a test point when plotted in XY coordinate we split the training data points in a form of a binary tree. We can choose to split the root node through X coordinate by taking its median.For example training set have points [(1,2),(2,3),(3,4),(4,5),(5,6),(6,7),(7,8)].By splitting the tree through X-axis we see the point (4,5) forms the root. In the next layer of the binary tree we split the data point by taking the median but rather than splitting it based on the X-axis we split it concerning the Y-axis. On the left-hand side, we get (1,2),(2,3),(3,4) and on the right hand, we get (5,6),(6,7),(7,8).On the left-hand side taking the median concerning y axis[2,3,4] we get (2,3).On the right-hand side taking the median concerning the Y-axis we get (6,7).[Note that y points on the left-hand side are[6,7,8]. We take the median of these points and not the X-axis of it].

The process carries on alternatively until all nodes are split.



https://www.linkedin.com/posts/uddit-7258792ab_python-ai-ml-activity-7223279853649756161-ljRi?utm_source=share&utm_medium=member_desktop

Above is the Link to summary explaination

Dated :29.07.2024 : IST-03:00 hrs

Malaviya National Institute of Technology Jaipur 302017

Uddit

2021UMT1791

Click the hyperlink below

Email me directly for faster doubt resolution @

or write to me **: udditalerts247@gmail.com**

APAth.ID:29072024423107702465 @LinkedIn