K-Nearest Neighbors

18 October 2022 11:31

What is K-Nearest Neighbors?

- K-nearest neighbors (kNN) is a supervised machine learning algorithm that can be used to solve **both classification and regression tasks**.
- kNN as an algorithm seems to be **inspired from real life**.
    - People tend to be effected by the people around them.
    - Our behaviour is guided by the friends we grew up with.
    - Our parents also shape our personality in some ways.
    - If you grow up with people who love sports, it is highly likely that you will end up loving sports.
    - There are of course exceptions.
    - kNN works in a similar fashion.

The value of a data point is determined by the data points around it.

- If you have one very close friend and spend most of your time with him/her, you will end up sharing similar interests and enjoying same things.
- That is kNN with k=1.

Similarly,

- If you always hang out with a group of 5, each one in the group has an effect on your behaviour and you will end up being the average of 5.
- That is kNN with k=5.

kNN classifier determines the class of a data point by majority voting principle.

- If k is set to 5, the classes of 5 closest points are checked. Prediction is done according to the majority class.
- Similarly, kNN regression takes the mean value of 5 closest points.

We observe people who are close but how data points are determined to be close**?**

Answer: The distance between data points is measured.

The Methods to measure the distance:

1. Euclidean distance,
2. cosine similarity measure,
3. Minkowsky,
4. correlation, and
5. Chi square

From the above list, Euclidean and Minkowsky distance(with p = 2) are most

commonly used.

Characteristics of KNN

18 October 2022 11:32

- K-Nearest Neighbors is one of the simplest supervised learning algorithms.
- kNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- kNN algorithm stores all the available data and classifies a new data point based on the similarity.
- This means when new data appears then it can be easily classified into a well suite category by using kNN algorithm.
- kNN algorithm can be used for regression as well as for classification problems.
- kNN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- kNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

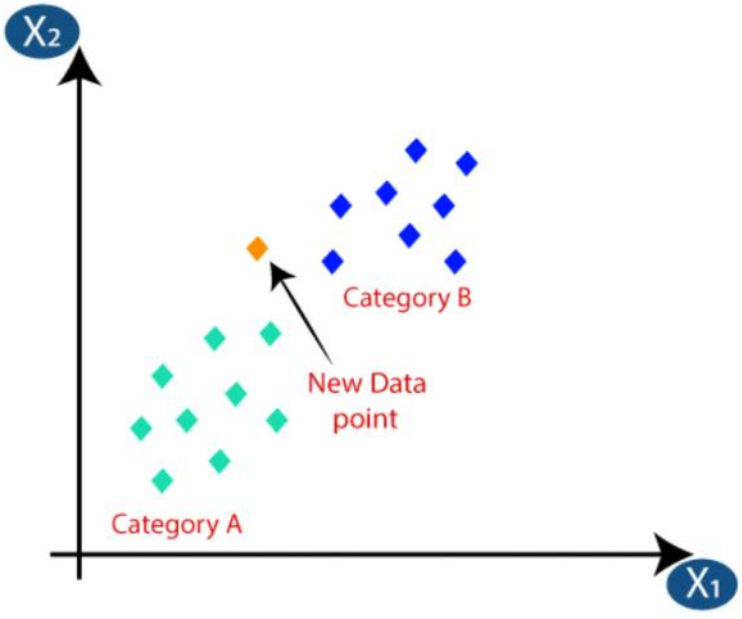
Working of KNN

18 October 2022 11:32

The kNN working can be explained on the basis of the below algorithm:

Step-1:
Select the **number K** of the neighbors.
Step-2:
Calculate the **Euclidean distance** of neighbors.
Step-3:
Take the **K nearest neighbor**s as per the calculated Euclidean distance.
Step-4:
Among these k nearest neighbors, **count the number of the data points in each category**.
Step-5:
Assign the **new data points** to that **category** for which the number of the neighbor is **maximum**.
Step-6:
Our model is ready.

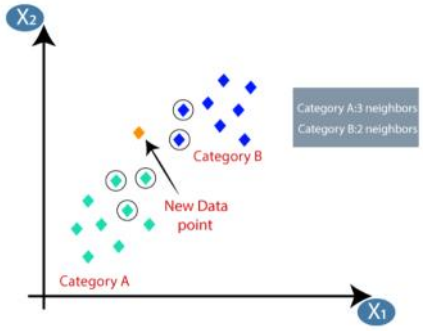
Example:



1. Step to choose the number of neighbours, k = 5.
2. Calculating the Euclidean distance between the data points.
3. Sorting out the K (i.e. 5) Nearest neighbors from the previous step

calculation.

- Notably, here we got 3 datapoints from the category A
- 2 data points from the category B.



K - Value

18 October 2022 11:32

Below are some points to remember while selecting the value of K in the K-NN algorithm:

- There is ~~no particular way~~ to determine the best value for "K", so we need to try some values to find the best out of them. **The most preferred value for K is 5.**
- A **very low value** for K such as K=1 or K=2, can be **noisy** and lead to the effects of **outliers** in the model.
- **Large values** for K are **good**, but it may find **some difficulties**.

Adv & Dis-adv of KNN

18 October 2022 11:33

Advantage	Disadvantage
• It is simple to implement. • It is robust to the noisy training data • It can be more effective if the training data is large.	• Always needs to determine the value of K which may be complex some time. • The computation cost is high because of calculating the distance between the data points for all the training samples.

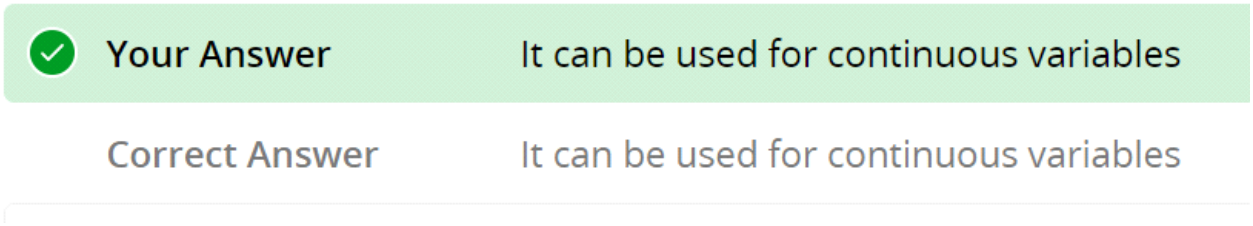
Implementation of KNN algorithm

18 October 2022 11:33

https://colab.research.google.com/drive/18pmoO09TasyPHzRusWL141pyN3vSBrCV#scrollTo=MO7VNjc02Mlo

Quiz

6 Which of the following is true about Manhattan distance?



- As the Manhattan distance are calculated between real valued features.