K-Nearest Neighbors (KNN) Algorithm to Predict the Species of The Plant Iris

Charvi Dave

Abstract

Classification is the process of predicting the class of a set of given data points. It classifies datasets into *respective and relevant categories. KNN is one of the simplest machine learning algorithms used for pattern recognition. It is a classifier based on supervised learning, it analyses the training data and produces an inferred function, which can be used for mapping new examples.* *This paper summarizes the KNN algorithm, how it works and also includes an experiment in which KNN is carried out on a dataset. The dataset consists of sepal and petal details of various Iris plants using which the species of the plant Iris is predicted. The results show that the KNN algorithm enhances its accuracy.*

Keywords: KNN algorithm; machine learning; classification

1. Introduction

Classification is one of the most fundamental concepts in data science. Using pre-categorized training datasets, machine learning programs use a variety of algorithms to classify future datasets into categories. Classification algorithms in machine learning use input training data to predict the likelihood/probability that the data will fall into one of the predetermined categories. Classification techniques include KNN classification, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine and so on. This paper summarizes the KNN algorithm.

This paper is structured as follows:

The second part includes the answers to the questions – ‘What is KNN?’ and ‘How does it work?’ - and also a discussion of the benefits and limitations of KNN.

The third part consists of the implementation of KNN on the dataset of the plant Iris along with visualisations, explanations and the results.

The fourth part includes a note on the applications of the KNN algorithm.

1. Introduction to KNN Algorithm
   1. What is KNN?

 K-Nearest Neighbor (KNN) is a very flexible classification algorithm which is used to divide data into classes based on the distance between the data points. KNN assumes that data points which are close to one another must be similar and hence, the data point to be classified will be grouped with the closest cluster/group.

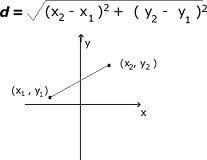
* 1. **How does KNN work?**

*Steps of KNN:*

Step 1: Choose the number of neighbors (K) to include in the 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. Most common default value of 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. Increasing k reduces variance but may increase bias.

Step 2: Take the K nearest neighbors of the new data point, according to Euclidean distance. (Manhattan distance, Minkowski distance or other distances can be used but Euclidean is the most used distance)**\***

Euclidean distance formula:



Step 3: Among these K neighbors, count the number of data points in each category.

Step 4: Assign the new data point to the category where the count of neighbors was more.

Let us take a scenario where we have two categories (or classes) in our dataset – category A (red) and category B (blue). Let there be two variables/columns in the dataset, x and y.

Now let us add a new data point to our dataset. We have to predict if it falls in category A or category B.

y Category B

BEFORE KNN

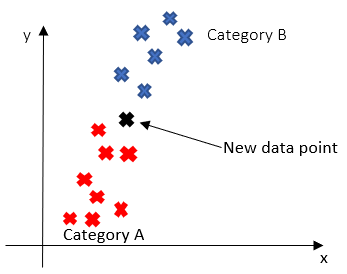
New data point

Category A

x

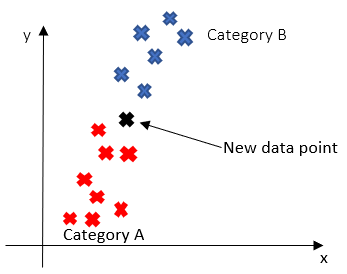
Step 1: Let us take k=5

Step 2: Taking the 5 nearest neighbors of the new data point (in the circle).



Step 3: 3 points belong to Category A and 2 point belong to Category B.

Step 4: Since Category A consisted of the greater number of points, the new data point will be assigned to Category A.



AFTER KNN

assigned to Category A

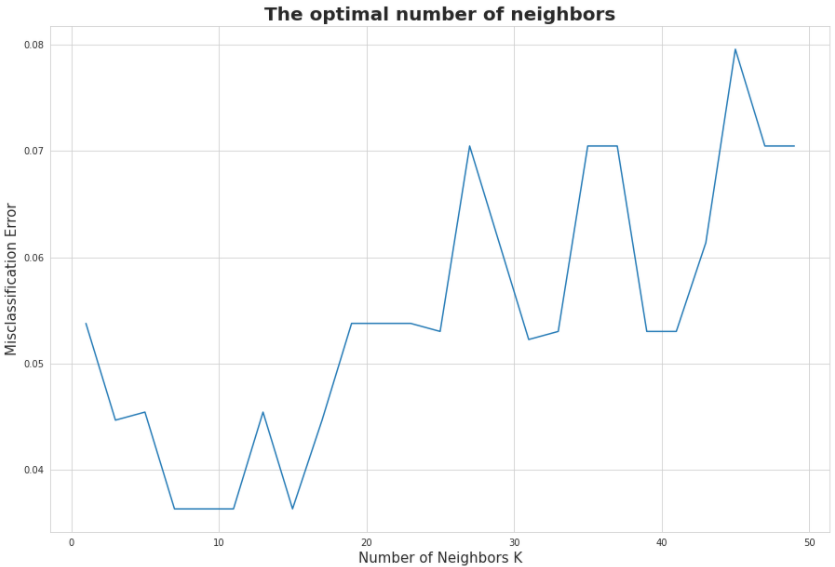
\*note: Theoretically, we are taking a small volume of the space of variables, centred at x, and with radius the distance to the Kth nearest neighbor. Then the maximum likelihood estimators of the probability that a point in this small volume belongs to each class are given by the proportion of training points in this volume that belong to each class. The KNN method assigns a new point to the class that has the largest estimated probability.

* 1. Benefits and Limitations
     1. Benefits/Advantages of KNN
* It is easy to implement.
* Training is done in a faster manner.
* It is very efficient.
  + 1. Limitations of KNN
* It requires large storage space.
* It is sensitive to noisy, missing data and outliers.
* **Does not work well with high dimensionality**as this will complicate the distance calculating process to calculate distance for each dimension.

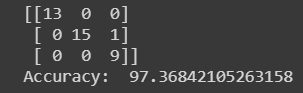
1. Implementation of KNN on the dataset

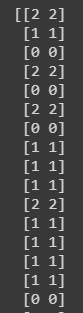
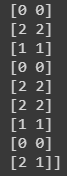
The dataset consists of sepal and petal details of different Iris plants. The three species of Iris plants that we have to predict are setosa, versicolor and virginica.

Cross-validation is performed and the misclassification error is calculated. Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. 10 fold cross-validation is performed, the input data is split into 10 subsets (folds) of data.



Plotting the performance criterion (the misclassification error, for example) against K, helps in finding the best K. At K=7, the misclassification error is the least hence K=7 is optimal. After splitting the dataset into the training set and test set, a KNN classifier is built and trained on the training set. Minkowski distance is used and the number of neighbors taken is 7 in this case. After this, the test set results are predicted and the confusion matrix is made.



Training the KNN classifier on the Iris plant dataset and using it to predict the species of each Iris plant gives us an accuracy of 97.37%.

[predicted test set result, actual test set result]

1. **Applications of KNN**

It is observed that KNN is an efficient algorithm for classification and it also provides a great accuracy. KNN is desirable in areas where there is even less knowledge of the data set than there would otherwise be. It is used in a variety of applications such as economic forecasting, data compression and genetics. For example, KNN was leveraged in a 2006 study of functional genomics for the assignment of genes based on their expression profiles.

**CONCLUSION**

In this paper I have presented an explanation on KNN and its implementation a dataset of Iris plants. To overcome the problems of low efficiency, a 10-fold cross validation process is conducted on the training dataset. Then, I have used the optimal K which is decided after performing cross validation, it is the K at which the misclassification error is the least.

**References**

1. D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)
2. M. Shouman, T. Turner, and Stocker, “Applying k-nearest neighbour in diagnosing heart disease patients,” Int. J. Inf. Educ. Technol., vol. 2, no. 3, p. 220, Jun. 2012.
3. Jingwen Sun,Weixing Du,Niancai Shi : “A Survey of KNN Algorithm” Information Engineering and Applied Computing (2018)