Final Report On K-Nearest Neighbors

CSE-0408 Summer 2021

Sheikh Afrin
Department of Computer Science and Engineering
State University of Bangladesh (SUB)
Dhaka, Bangladesh
sheikhafrin2016@gmail.com

***Abstract*—This algorithm is used to solve the classification model problems. K-nearest neighbor or K-NN algorithm basically creates an imaginary boundary to classify the data. When new data points come in, the algorithm will try to predict that to the nearest of the boundary line.**

***Index Terms*—Python**

I. Introduction

The K-Nearest-Neighbors (KNN) is a nonparametric classification algorithm, i.e. it does not make any presumptions on the elementary dataset.It is known for it's simplicity and effectiveness. It is a supervised learning algorithm. A labeled training dataset is provided where the data points are categorized into various classes, so that the class of the unlabeled data can be predicted.In Classification, different characteristics determine the class to which the unlabeled data belongs. KNN is mostly used as a classifier.It is used to classify data based on closest or neighbouring training examples in a given region.

This method is used for its simplicity of execution and low computation time.For continuous data, it uses the euclidean distance to calculate its nearest neighbours .

II. Literature Review

Along the years, a great effort was done in the scientific community in order to solve or mitigate the imbalanced dataset problem. Specifically for KNN, there are several balancing methods based on this algorithm. This section will provide a bibliographic review about the KNN and its derivate algorithms for dataset balancing. Also, the random oversampling and undersampling methods, the class overlapping problem, and evaluation measures will be reviewed.

III. Proposed Methodology

K-Nearest Neighbors (kNN) is such a method and, despite its simplicity, continues to perform fairly well for large training sets. It essentially relies only on the most basic assumption underlying all prediction: that observations with similar characteristics will tend to have similar outcomes. Nearest Neighbor methods assign a predicted value to a new observation based on the plurality or mean of its k "Nearest Neighbors" in the training set.K-Nearest Neighbors (KNN) is a standard machine-learning method that has been extended to large-scale data mining efforts.

IV. ADVANTAGES

1. No Training Period: KNN is called Lazy Learner.It does not learn anything in the training period. It does not derive any discriminative function from the training data.
2. Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm.
3. KNN is very easy to implement. There are only two parameters required to implement KNN i.e. the value of K and the distance function.

V. DISADVANTAGES

1. Does not work well with large dataset.
2. Does not work well with high dimensions.
3. Need feature scaling: We need to do feature scaling before applying KNN algorithm to any dataset.
4. Sensitive to noisy data, missing values and outliers: KNN is sensitive to noise in the dataset.

VI. CODE

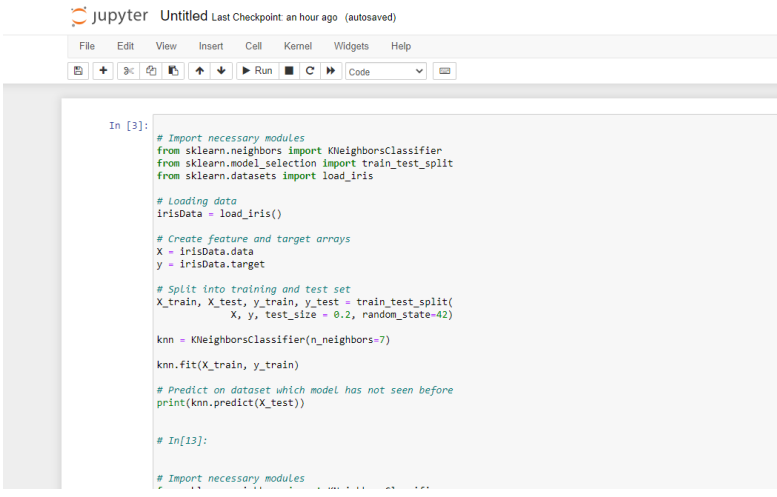

Fig. 1.

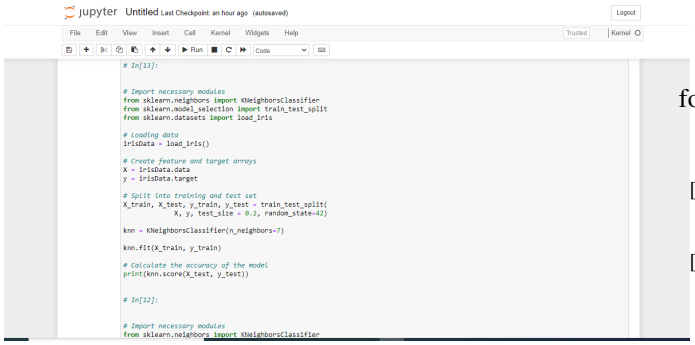

Fig. 2.

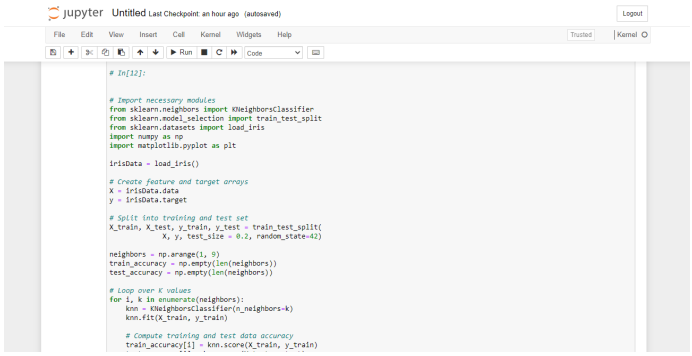

Fig. 3.

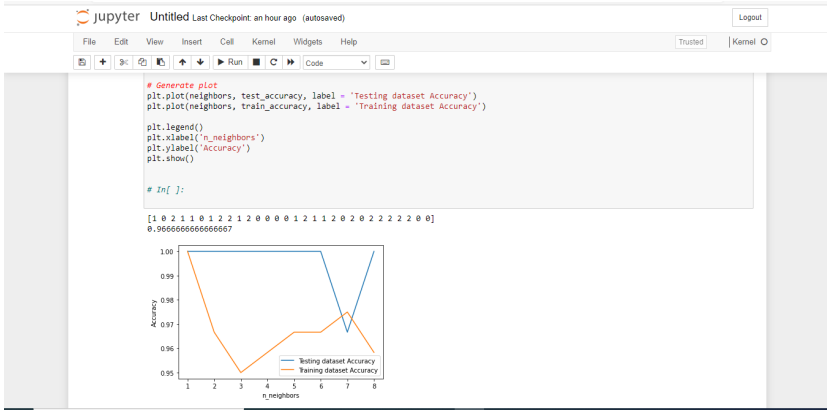

Fig. 4.

VII. Conclusion

Machine learning algorithms have improved with the increase in research and data mining tools. K- nearest neighbour algorithm is a simple but high accuracy algorithm that has proven effective in several cases.The nearest neighbour algorithm works by classifying the new unlabeled data by examining the classes of it's nearest neighbours.In KNN algorithm, a constant number of nearest neighbours determine the classification of an unlabeled data which is assigned by K, where K is a positive integer.

Acknowledgment

I would like to thank my honourable **Khan Md. Hasib Sir** for his time, generosity and critical insights into this project.

References

[1] Beckmann, M., Ebecken, N. F., & de Lima, B. S. P. (2015). A KNN undersampling approach for data balancing. Journal of Intelligent Learning Systems and Applications, 7(04), 104.

[2] Makkar, T., Kumar, Y., Dubey, A. K., Rocha, Á., & Goyal, A. (2017, December). Analogizing time complexity of KNN and CNN in recognizing handwritten digits. In 2017 Fourth International Conference on Image Information Processing (ICIIP) (pp. 1-6). IEEE.

[3] Hazra, T. K., Singh, D. P., & Daga, N. (2017, August). Optical character recognition using KNN on custom image dataset. In 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON) (pp. 110-114). IEEE.