Assignment No 04:K-Nearest Neighbors

CSE-0408 Summer 2021, FINAL EXAM

TASNOVA TASNIM
Department of Computer Science and Engineering
State University of Bangladesh (SUB)
Dhaka, Bangladesh
email:tasnimprity12@gmail.com

***Abstract*—The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.KNN is a lazy learning, non-parametric algorithm. It uses data with several classes to predict the classification of the new sample point. KNN is non-parametric since it doesn't make any assumptions on the data being studied, i.e., the model is distributed from the data.**

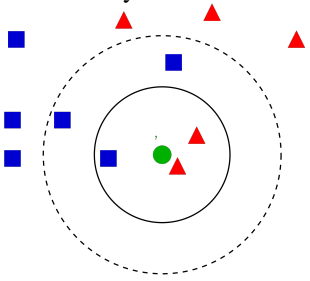
n

***Index Terms*—k-nn,python**

I. Introduction

k-nearest neighbors algorithm (k-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951,[1] and later expanded by Thomas Cover.[2] It is used for classification and regression. In both cases, the input consists of the k closest training examples in data set. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors. k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.



II. Literature Review

Thomas Cover and Peter Hart proved an upper bound error rate with multiclass kNN classifications in 1967 (twice the Bayes error rate). Research paper: Nearest neighbor pattern classification (1967) Cover Hart's discovery paved the way for a number of new rejection studies such as:

In 1970: Edward Hellman examined "the (k,k?) nearest neighbor rule with a reject option". In 1975: Fukunaga and Hostetler made refinements with respect Bayes error rate In 1976: Dudani; in 1978 Bailey and Jain published distance weighted approaches

In 1983: Adam Jozwik introduced: A learning scheme for a fuzzy k-NN rule

In 1985: James Keller et al developed FKNN (Fuzzy kNN): A fuzzy k-nearest neighbor algorithm

In 2000: Bermejo and Cabestany published: Adaptive soft k-nearest-neighbour classifiers.

There has been many improvements to kNNs since those days and new approaches continue to emerge to this day KNN algorithm applies the birds of a feather. It assumes that similar things are near to each other; that is, they are nearby.

III. Proposed Methodology

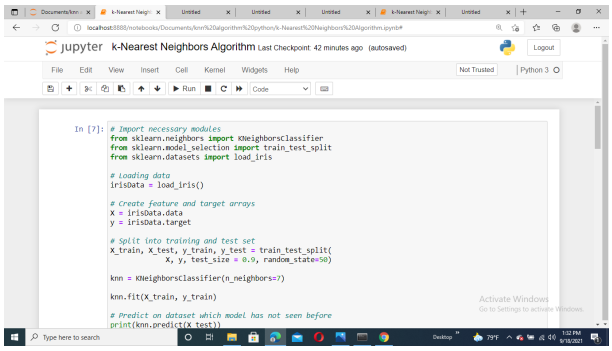
we classified a data set including large number of unlabeled data, if only utilize the few training examples available, then we can't obtain a high accuracy classifier with inadequate training examples; if we want to obtain a classifier of high performance, then labeling the unlabeled data is necessary, but labeling vast unlabeled data wastes time and consumes strength. In this paper, we propose a novel method which uses SVM cooperated with KNN for classification based on semi supervised learning theory. The general model is depicted as above (See Figure 2). To begin with, we construct a weaker classifier SVM according to the few training examples available, then using the weaker SVM classifies the remaining large number of unlabeled data in the data set, picking out n examples belonging to each class around the decision boundary by calculating Euclidean distance in the feature space, because the examples located around the boundary are easy to be misclassified, but they are likely to be the support vectors, we call them boundary vectors, so picking out these boundary vectors whose labels are fuzzy labeled by the weaker classifier SVM. Secondly we recognize these boundary vectors

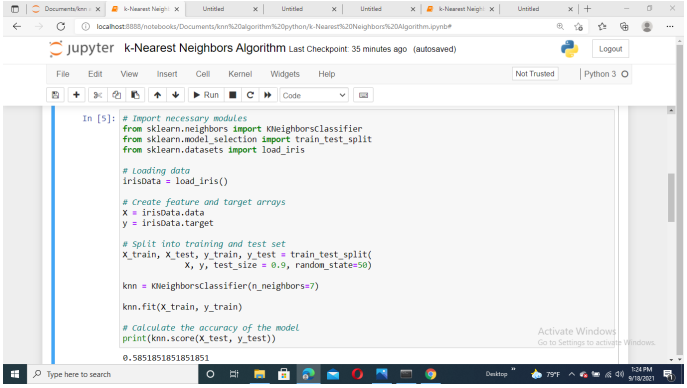
as testing set while recognize initial training examples as training set, use KNN method to classify them and recognize the results as the labels for boundary vectors. In the end, we put these boundary vectors and their labels into initial training set to enlarge the number of the training examples then retrain a SVM, iteratively until the number of the training examples is m times of the whole data set. The experimental results on three UCI data sets indicate that the final classifier SVM has significant improvement on accuracy.

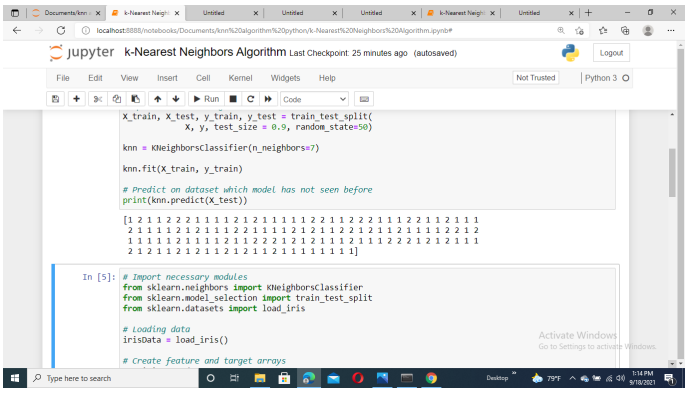
IV. Conclusion

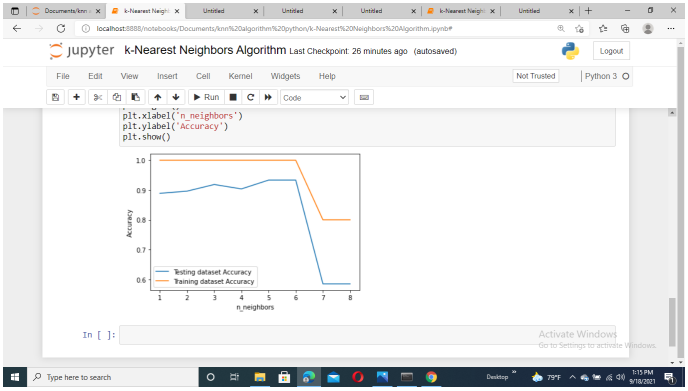
Every learning algorithm will tend to suit some problem types better than others, and will typically have many different parameters and configurations to be adjusted before achieving optimal performance on a dataset, AdaBoost (with decision trees as the weak learners) is often referred to as the best outofthe-box classifier. When used with decision tree learning, information gathered at each stage of the AdaBoost algorithm about the relative 'hardness' of each training sample is fed into the tree growing algorithm such that later trees tend to focus on harder-to-classify examples. The supervised machine learning algorithms such as decision trees and support vector machine are capable enough to deal with big data mining tasks. Even though the algorithms efficiency considerably improving there is a need for adaptive boosting process required in order to increase the predictive accuracy much more In future, what you bring in your project and the idea of your work.

V. code









Acknowledgment

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

References

[1] X.J. Zhu. Semi-supervised learning literature survey[R]. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, December, 2007.

[2] Everitt, Brian S.; Landau, Sabine; Leese, Morven; and Stahl, Daniel (2011) "Miscellaneous Clustering Methods", in Cluster Analysis, 5th Edition, John Wiley Sons, Ltd., Chichester, UK

[3] Fung, G., and Mangasarian, O. Semi-supervised support vector machines for unlabeled data classification (Technical Report 99-05). Data Mining Institute, University of Wisconsin Madison, 1999

[4] Hastie, Trevor. (2001). The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. Tibshirani, Robert., Friedman, J. H. (Jerome H.). New York: Springer.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.