STUDYING THE EFFECTS OF VARYING NUMBER OF NEAREST NEIGHBORS IN THE K-NEAREST NEIGHBORS ALGORITHM

CMSC 191 - MACHINE LEARNING

Harold R. Mansilla
Department of Physical Sciences and Mathematics
College of Arts and Sciences
University of the Philippines Manila
hrmansilla@up.edu.ph

Virgilio M. Mendoza III
Department of Physical Sciences and Mathematics
College of Arts and Sciences
University of the Philippines Manila
vmmendoza1@up.edu.ph

May 15, 2019

1 Introduction

K-nearest-neighbor (kNN) classification is one of the most fundamental and simple classification methods. Developed from the need to perform discriminant analysis when reliable parametric estimates are difficult to determine, it is one of the first choices for classification when there is little to no prior knowledge about the data [1].

2 The Dataset

The dataset was retrieved from [2]. It contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Table 1: Features in the dataset

Feature	**Type**	**Description**
Age of Patient	Numerical	Age of the patient at the time of the operation
Patient's Year of Operation	Numerical	Year the patient was operated in the 20th century (1900)
Axillary Nodes	Numerical	Number of positive axlllary nodes
Survival Status	Numerical	Entropy of image
Class	Categorical	Outcome of operation

The target value is the `Survival Status` column wherein the possible values are 1 for survived 5 years or longer and 2 for died within 5 years. From [2], the number of observations for each class is not balanced. There are 306 observations in total.

3 KNN Classifier Construction

For the purpose of this study, the experimentation of the effects of the varying number of nearest neighbors, k, to the overall accuracy of a kNN classifier, a cross validation with 10-folds was used where the value of k ranges from 1 to 25.

4 Results

Figure 1 shows accuracy of the kNN classifier against the values of $k_neighbors$. The highest accuracy, 75.49%, was achieved when there were 12 neighbors. The sudden drop in accuracy can be seen when there were 14 neighbors.

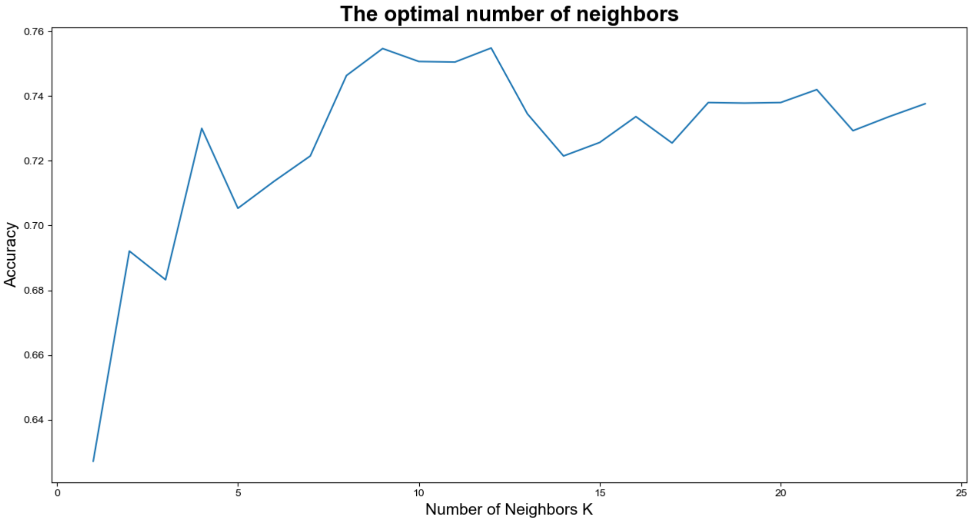

Figure 1: Plot of Accuracy against the `k_neighbors`

5 Conclusion

The effect of the number of neighbors in KNN classifier performance was experimented on. For this dataset, 0 to 14 neighbors proved optimal to the performance of the classifier, resulting in a perfect accuracy.However, increasing this number (particularly from 15 onwards), results in lower performance. However, with the range of accuracy values starting from around 0.993, it can be said that this is negligible.

For the dataset used in this study, it can be concluded that the optimal number of neighbors is 12. The lowest scoring number is 1 which scored 62.71%.

References

[1] L. E. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[2] T.-S. Lim, *Haberman’s survival data set*. [Online]. Available: `https://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival`.