K-Nearest Neighbors (KNN) Classification: An Evaluation of Different K Values and Distance Metrics

Raghda Al taei

2024

1 Introduction

In this study, we aim to evaluate the performance of the K-Nearest Neighbors (KNN) classification algorithm using different values of K and various distance metrics. KNN is a simple yet effective algorithm used for classification tasks in machine learning. The objective is to determine how the choice of K and the distance metric affects the accuracy of the classification results.

2 Methodology

2.1 Dataset

The experiment was conducted using a dataset with a set of features used for classification. The dataset was divided into training and testing sets, with 10-fold cross-validation implemented to ensure robust accuracy evaluation.

2.2 KNN Implementation

The KNN algorithm was implemented without using any pre-built libraries. We tested the algorithm with different values of K (1, 3, 5, 7, and 10) to observe how the accuracy varied.

2.3 Distance Metrics

In addition to varying K, we also evaluated the algorithm's performance using different distance metrics:

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance (with $p = 4$)

- Minkowski Distance (with $p = \frac{1}{2}$)
- Cosine Distance

For each distance metric, the KNN algorithm was run with $K = 5$.

2.4 Evaluation

The accuracy for each configuration was computed and recorded.

3 Results

The results of the KNN classification are summarized in the following tables:

3.1 Accuracy for Different K Values

K Value	Accuracy (%)
1	92.86
3	92.86
5	93.81
7	93.33
10	92.86

tableAccuracy for Different K Values

3.2 Accuracy for Different Distance Metrics (K=5)

Distance Metric	Accuracy (%)
Euclidean Distance	93.81
Manhattan Distance	90.95
Minkowski (p=4) Distance	93.81
Minkowski (p=1/2) Distance	90.95
Cosine Distance	91.43

tableAccuracy for Different Distance Metrics (K=5)

4 Discussion

The results indicate that the choice of K has a significant impact on the accuracy of the KNN algorithm. The highest accuracy was achieved with $K = 5$ at 93.81%. Other values of K also performed well, particularly $K = 3$ and $K = 7$.

When examining the distance metrics, the Euclidean and Minkowski distances (with $p = 4$) both yielded the highest accuracy of 93.81%. In contrast, Manhattan and Minkowski distances (with $p = \frac{1}{2}$) resulted in lower accuracy rates around 90.95%. The Cosine distance also demonstrated competitive performance with an accuracy of 91.43%.

Overall, the findings highlight the importance of selecting appropriate values for K and the choice of distance metrics to optimize classification accuracy in KNN.

5 Conclusion

In conclusion, our experiment demonstrates that both the selection of K and the distance metric are crucial factors in determining the performance of the KNN algorithm. Future work may involve exploring additional distance metrics or refining the dataset for improved accuracy.