Question 01: Binary Classification using K-Nearest Neighbors

1 Introduction

This experiment focuses on binary classification using the K-Nearest Neighbors (KNN) algorithm implemented completely from scratch. The objective is to classify tumors as malignant or benign using numerical features extracted from medical data. The study evaluates the effect of different distance metrics and values of K on classification performance.

2 Dataset Description

The dataset consists of patient diagnostic records containing multiple real-valued features. Each sample is labeled as either malignant (M) or benign (B). The target variable was encoded numerically, where malignant samples were assigned label 1 and benign samples label 0.

Irrelevant columns such as identifiers and empty attributes were removed before training.

3 Data Preprocessing

All input features were scaled using Min-Max normalization to ensure uniform contribution during distance computation. The dataset was randomly shuffled and split into training and testing sets using an 80:20 ratio.

4 K-Nearest Neighbors Algorithm

The KNN classifier predicts the class of a test sample by computing distances between the test point and all training samples. The K nearest neighbors are selected, and the most frequent class among them is assigned as the predicted label.

This implementation does not rely on any pre-built machine learning libraries.

5 Distance Metrics

The following distance metrics were implemented and compared:

- **Euclidean Distance**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Manhattan Distance**

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Minkowski Distance** ($p = 3$)

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

- **Cosine Distance**

$$d(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|}$$

- **Hamming Distance**

$$d(x, y) = \sum_{i=1}^{n} \mathbb{I}(x_i \neq y_i)$$

6 Experimental Setup

The classifier was evaluated for multiple values of $K = \{3, 4, 9, 20, 47\}$. For each distance metric and value of K, prediction accuracy was computed on the test set. Accuracy trends were visualized using an Accuracy vs. K plot.

7 Evaluation Metrics

Performance was measured using the following metrics:

7.1 Accuracy

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

7.2 Confusion Matrix

The confusion matrix consists of:

- True Positives (TP)
- True Negatives (TN)
- False Positives (FP)
- False Negatives (FN)

7.3 Precision and Recall

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

8 Results and Analysis

Experimental results indicate that Euclidean and Manhattan distance metrics consistently achieve the highest classification accuracy. The best-performing configurations were:

- Euclidean distance with $K = 4$
- Manhattan distance with $K = 3$ and $K = 4$

Other distance metrics such as Hamming and Cosine performed comparatively worse, highlighting their limited suitability for continuous numerical medical data.

9 Observations

- Feature scaling significantly improves KNN performance.
- Small values of K are sensitive to noise but capture local patterns.
- Larger values of K increase bias and reduce sensitivity.
- Distance metric choice has a strong impact on classifier accuracy.

10 Conclusion

This study demonstrates that a KNN classifier implemented from scratch can effectively perform binary classification on medical datasets. Euclidean and Manhattan distance metrics yield superior performance when combined with appropriate values of K. Despite its simplicity, KNN serves as a strong baseline classifier, though its computational cost increases with dataset size.