Assignment Number 03

Name: Mihir Unmesh Patil

Roll No: TYCOC213
Batch: C/C-3

Aim

The primary objective of this assignment is to analyze the relationship between input features and class labels using k-Nearest Neighbors (k-NN) based classification models. The study involves:

1. Implementation of k-NN classification models (Uniform k-NN, Distance-Weighted k-NN) and Nearest Centroid method to classify test results and breast cancer data.
2. Determination of the optimal k value using the elbow method for k-NN models.
3. Comparison of performance across Uniform k-NN, Distance-Weighted k-NN, and Nearest Centroid models.
4. Evaluation of model performance using accuracy, confusion matrix, and classification report metrics.
5. Visualization of k-value effects, model accuracy comparisons, and predictive behavior for better interpretability.

Objectives

- Implement Uniform and Distance-Weighted k-NN classifiers to predict class labels for given datasets.
- Apply the Nearest Centroid method as a baseline comparison for classification.
- Use the elbow method to identify the optimal k value for k-NN models.
- Compute performance metrics (accuracy, confusion matrix, precision, recall, F1-score) for model evaluation.
- Visualize the impact of k on model accuracy and compare final model performances using bar plots.
- Analyze the strengths and weaknesses of each classification approach.

Theory

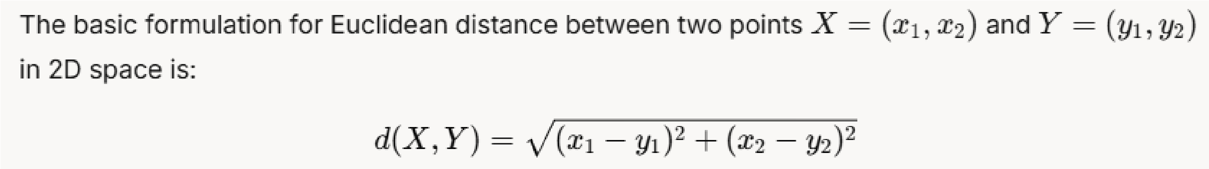
Classification Analysis

Classification is a supervised learning technique used to predict the categorical class labels of new data points based on labeled training data. It involves modeling the relationship between input features (independent variables) and discrete output classes (dependent variable). Classification is widely applied in fields such as medical diagnosis, fraud detection, and pattern recognition.

k-Nearest Neighbors (k-NN) Classification

k-NN is a simple, instance-based learning algorithm that classifies a data point based on the majority class of its k nearest neighbors in the feature space. It operates on the principle of similarity, where "nearness" is typically measured using Euclidean distance. The algorithm can be described as follows:

- For a given test point X, compute the distance to all training points.
- Select the k nearest neighbors based on the smallest distances.
- Assign the class label based on a majority vote among the k neighbors.



The basic formulation for Euclidean distance between two points $X = (x_1, x_2)$ and $Y = (y_1, y_2)$ in 2D space is:

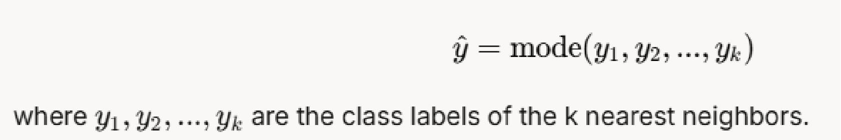
$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

The choice of k is critical:

- A small k (e.g., k=1) makes the model sensitive to noise.
- A large k smooths the decision boundary but may include irrelevant neighbors.

In this assignment, we implemented two variants of k-NN:

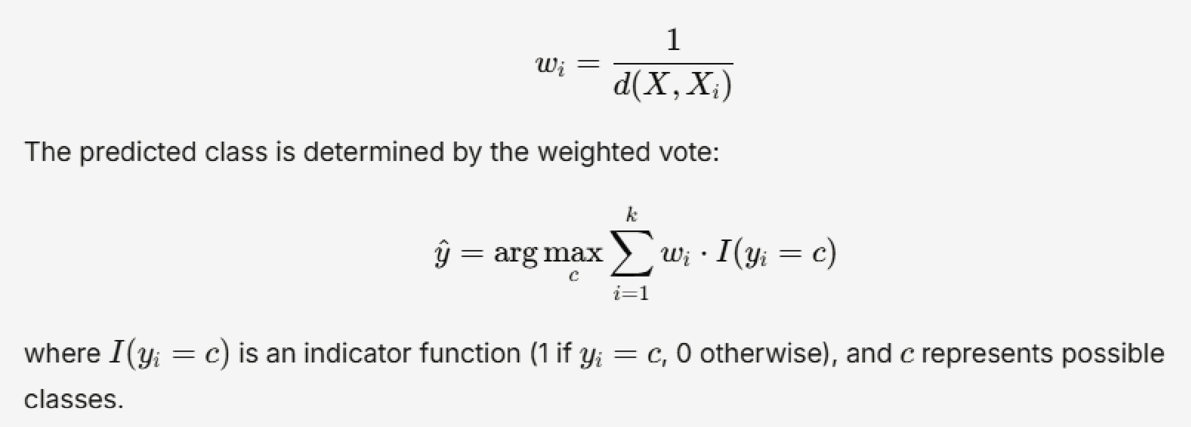
1. **Uniform k-NN**: All k neighbors contribute equally to the classification decision via a simple majority vote. The predicted class y^ for a test point X is:



$$\hat{y} = \text{mode}(y_1, y_2, ..., y_k)$$

where $y_1, y_2, ..., y_k$ are the class labels of the k nearest neighbors.

2. **Distance-Weighted k-NN**: Neighbors contribute to the decision based on their proximity to the test point. Closer neighbors have more influence, with weights typically defined as the inverse of distance:



$$w_i = \frac{1}{d(X, X_i)}$$

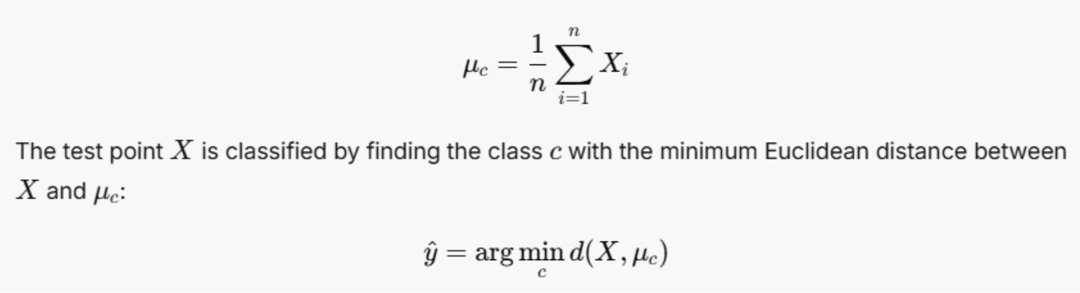
The predicted class is determined by the weighted vote:

$$\hat{y} = \arg\max_c \sum_{i=1}^{k} w_i \cdot I(y_i = c)$$

where $I(y_i = c)$ is an indicator function (1 if $y_i = c$, 0 otherwise), and c represents possible classes.

Nearest Centroid Classification

The Nearest Centroid (NC) classifier is a simple method that assigns a test point to the class whose centroid (mean of feature vectors) is closest in the feature space. For a class c with training points {X1,X2,...,Xn}, the centroid μc is:



$$\mu_c = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The test point X is classified by finding the class c with the minimum Euclidean distance between X and μ_c:

$$\hat{y} = \arg\min_c d(X, \mu_c)$$

Unlike k-NN, NC does not consider individual neighbors but relies on class prototypes, making it computationally efficient but potentially less flexible for complex data distributions.

Elbow Method for Optimal k Selection

The elbow method is a heuristic used to determine the optimal number of neighbors (k) in k-NN. It involves:

- Computing the cross-validation accuracy for a range of k values.
- Plotting accuracy against k.
- Identifying the "elbow" point where additional increases in k yield diminishing returns in accuracy.

This method balances model complexity and performance, avoiding overfitting (small k) or oversmoothing (large k).

Performance Metrics

The models were evaluated using the following metrics:

1. **Accuracy**:

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

- Measures the overall correctness of predictions. Higher accuracy indicates better performance.

2. **Confusion Matrix**:

- A table showing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for binary classification:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

- Provides insight into model performance across classes.

3. **Classification Report**:

- Includes precision, recall, and F1-score:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Precision measures prediction accuracy, recall measures coverage of true positives, and F1-score balances both.

Visualization Techniques

To enhance interpretability:

- **Accuracy vs k Plots**: Show how accuracy varies with k for each model, aiding in optimal k selection.
- **Bar Plots**: Compare test accuracy across models for both datasets, highlighting relative performance.

Application to Datasets

1. **Small Dataset**:

- Features: S1, S2 (symptom values)
- Target: Test_Class (Positive/Negative)
- Size: 6 samples
- Challenges: Limited data size affects model stability, particularly for distance-weighted k-NN.

2. **Wine Dataset**:

- Features: First two attributes from sklearn dataset

- Target: Binary class (0/1)
- Size: 178 samples
- Advantages: Larger size provides more robust training and evaluation.

Model Comparison Considerations

- **Uniform k-NN**: Simple and effective for small datasets but may struggle with noisy or complex data.
- **Distance-Weighted k-NN**: Emphasizes closer neighbors, potentially improving accuracy in larger datasets but sensitive to scale and outliers in small datasets.
- **Nearest Centroid**: Fast and interpretable but assumes linear separability of class means, which may not hold for all data.

Practical Implications

- k-NN models are non-parametric and adaptable to various data distributions, making them suitable for medical diagnosis tasks.
- Distance weighting can enhance performance when data points have varying relevance, though it may fail with very small datasets.
- Nearest Centroid offers a lightweight alternative for quick classification but may oversimplify complex relationships.
- The elbow method ensures optimal parameter selection, critical for balancing bias and variance in k-NN.

CODE:

Assignment_03_ML

Conclusion:

This theoretical framework provides the foundation for understanding the implementation, evaluation, and comparison of k-NN-based classification models in this assignment.

References:

https://pmc.ncbi.nlm.nih.gov/articles/PMC7924495/pdf/peerj-cs-05-194.pdf
https://link.springer.com/article/10.1007/BF00153759
https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12_toc.pdf.download.html