CSC 59866 Senior Design I: Assignment 1 Summery

Introduction

This assignment focuses on the development of fundamental machine learning skills, specifically in the areas of binary classification using the k-Nearest Neighbors (K-NN) algorithm, data pre-processing, feature selection, and model evaluation. Unlike typical machine learning tasks, this assignment prohibits the use of the Scikit-Learn library for the implementation of the K-NN algorithm, challenging students to understand the underlying mechanics of this model.

Part A: Model Code

Euclidean and Manhattan Distances

The assignment begins with the implementation of functions to calculate Euclidean and Manhattan distances between two vectors. These distances serve as the foundation for measuring similarities between data points in the K-NN algorithm.

Accuracy, Generalization Error

Subsequent tasks involve creating functions to calculate accuracy, generalization error, precision, recall, F1 score, and the confusion matrix. These metrics are crucial for evaluating the performance of the K-NN classifier and understanding the trade-offs between different types of errors.

ROC Curve and Precision-Recall Curve

Functions to generate the Receiver Operating Characteristic (ROC) curve and the precision-recall curve, as well as computing the area under the curve (AUC) for the ROC curve, are also required. These are advanced metrics for evaluating classifier performance, especially in datasets with imbalanced classes.

KNN_Classifier Class

The centerpiece of Part A is the implementation of the `KNN_Classifier` class. This class includes methods for fitting the model to the training data and predicting labels for new data points, with the number of neighbors and the weighting scheme as adjustable parameters.

Part B: Data Processing

Data Preparation and Exploration

The assignment transitions to practical application with the wine quality dataset. We were tasked with reading the dataset into a Pandas DataFrame, converting the "quality" column into abinary target variable, and summarizing the dataset's features.

Data Cleaning and Partitioning

Identifying and dropping redundant features through pair plots is required to refine the dataset before splitting it into training and test sets using a custom `partition` function.

K-NN Implementation on Wine Dataset

Applying the previously developed `KNN_Classifier` on the wine quality dataset involves experimenting with raw and standardized data, as well as different weighting schemes, to compare model performances based on accuracy and F1 score.

Part C: Model Evaluation

Comprehensive Testing

The final part of the assignment requires a thorough evaluation of the K-NN model across various settings of k, distance metrics (Euclidean and Manhattan), and weighting schemes (uniform and distance). This comprehensive testing aims to find the optimal configuration for the model when applied to the wine quality dataset.

Conclusion

The assignment culminates in a critical analysis of the model's performance, with insights into how different preprocessing steps, distance metrics, and weighting schemes affect the accuracy and overall effectiveness of the K-NN classifier.