Introduction to Machine Learning

Lab - 2 Assignment
KNN with Cross Validation

Guidelines for submission

1. Perform all tasks in a single colab/.py file.
2. Create a report regarding the steps followed while performing the given tasks. The report should not include excessive unscaled preprocessing plots.
3. Try to modularize the code for readability wherever possible
4. Plagiarism will not be tolerated.

Question: [100 marks]

Dataset: Heart Disease UCI (available on Kaggle or UCI Machine Learning Repository)

The dataset is available here.

In this lab assignment, you will explore the K-Nearest Neighbors (KNN) algorithm, which is a simple yet effective classification algorithm. The dataset contains various features related to heart health, and the goal is to predict which type of heart disease does the person have.

The target column is the last column named 'num'. It has 5 classes.

Q.1 Data Preprocessing[10 marks]

- Load the dataset and handle missing values by either dropping or imputing them.
- Normalize the dataset so that all features are on the same scale.
- Encode categorical variables using suitable techniques (e.g., one-hot encoding for nominal features, label encoding for ordinal features).
  Also mentionthe category of each categorical column(Nominal, Ordinal, Binary)
- Perform exploratory data analysis to understand the distribution of features and target variable.

Q2. Train-Test Split[5 marks]

Split the preprocessed dataset into training, validation, and test sets in the ratio 70:20:10.

Q3. KNN Implementation[20 + 10 marks]

Implement the KNN algorithm from scratch without using the Sklearn library. You can refer to the following steps:

- Calculate distances between data points using a suitable distance metric (e.g., Euclidean distance).
- Sort the data points based on distances.
- Select the top K nearest neighbors.
- Assign the class label by majority vote among the neighbors.

Use the Euclidean distance metric and experiment with different values of K (e.g., 3, 5, 7, 9, 11).

Q4) K-Fold Cross Validation[20 marks]

- Implement K-Fold Cross Validation with 5 folds.
- Apply **stratified sampling** to ensure that each fold maintains the same class distribution as the original dataset.
- Additionally, perform **bootstrapping** within each fold to create a diverse training set for each fold.

Q5) Report and Analysis[45 marks]

- Present the classification accuracy for different values of K using the validation set.
- Plot a graph showing how accuracy changes with K.
- Compare the accuracy obtained from the KNN implementation with that of the Sklearn library's KNN classifier using the test set.
- Provide observations and potential reasons for differences in performance.
- Include screenshots of the code, plots, and accuracy scores in the report.
- Reflect on the impact of K, distance metrics, and preprocessing choices on the model's performance.
- Plot the decision boundary of the KNN model for the sklearn's trained KNN model, and comment your observations about the decision boundary.

Resources

1. KNN - Kaggle Notebook
2. KNN - Sklearn