Applied Machine Learning
Exercise 7

Prof. Dr. Amr Alanwar

November 2024

Classification Datasets

You can use any of the datasets (or both).

1. **Iris dataset** D_1: Target attribute classes: Iris Setosa, Iris Versicolour, Iris Virginica.
   Iris Dataset
2. **Wine Quality** called D_2: (use `winequality-red.csv`).
   Wine Quality Dataset

Note: Dataset D_2 can also be used for a regression problem.
You are required to pre-process the given datasets.

Exercise 1: Implement K-Nearest Neighbor (KNN) (4 Points)

Your task is to implement the KNN algorithm. To implement KNN, you have to:

- Split data into a train and test split (70% and 30% respectively).
- Implement a similarity (or a distance) measure. To begin with, you can implement the Euclidean Distance.
- Implement a function that returns the top K Nearest Neighbors for a given query (data point).
- Provide the prediction for a given query (for a classification task, you can use majority voting; for regression, you can use the mean).
- Measure the quality of your prediction. [Hint: Choose a quality criterion according to the task you are solving, i.e., a regression or classification task. Defend your choice].

Exercise 2: Optimize and Compare KNN Algorithm (6 Points)

Part A: Determine the Optimal Value of K (3 Points)

In this exercise, you need to determine the optimal value of K for the given datasets.

1. How can you choose the value of K for KNN? Provide a criterion to choose an optimal value of K.
2. Implement the criterion for choosing the optimal value of K.
3. Experimentally demonstrate that your chosen value is better than other values of K. [Hint: Run your experiment with different values of K and plot the error measure for each value].

Part B: Compare KNN Algorithm with Tree-Based Methods (3 Points)

In this task, you are allowed to use Scikit-Learn. In particular, you will use the Nearest Neighbor and Decision Tree implementations provided by Scikit-Learn.

- Use Nearest Neighbor and Decision Tree provided by Scikit-Learn to solve the classification task for the two datasets.
- Provide the optimal hyperparameters for both methods. [Hint: Use Grid Search and cross-validation and present results to support your solution].
- Present a comparison of the two methods using evaluation results on test datasets. [Hint: It is better to use cross-validation to confirm your results].

Annex

You can use libraries:

- Scikit-Learn User Guide
- sklearn.metrics Documentation
- sklearn.model_selection
- sklearn.linear_model
- sklearn.preprocessing
- `matplotlib` for plotting.