Machine Learning: Foundations and Applications

Lab Test-1 Date: 14.09.2023 Time: 1 hr 30min + 15min Total Marks: 15

Problem Statement:

A floriculture research team X is studying the use of multiple measurements to distinguish three different iris flower species. The dataset contains a set of 150 records under five attributes: sepal length, sepal width, petal length, petal width and species (see Fig. 1). Develop a K Nearest Neighbour (KNN) classifier that classifies the species according to the above measurements.

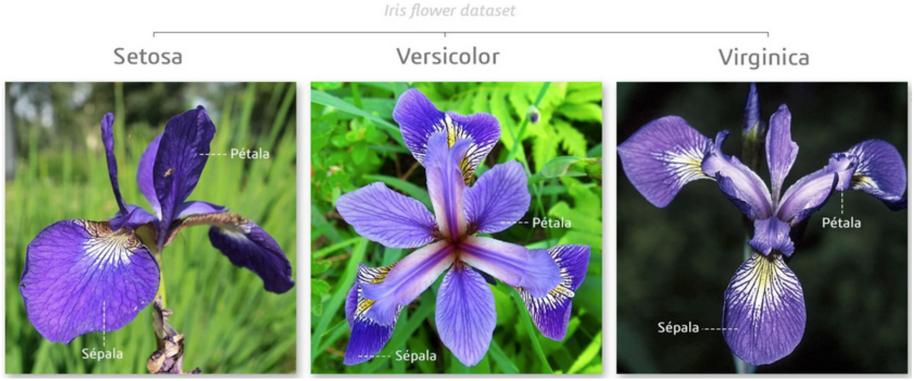

Figure 1: Different iris flower species and their attributes

Implementation: [5 Marks]

- Implementation of distance weighted K-NN algorithm (KNN_Distance) from scratch (without using builtin functions). The algorithm should be able to choose a particular K value.
- Evaluate the model using Percentage Accuracy.

****Implement [KNN_Distance] from scratch**. You may make use of the numpy library to perform basic operations (e.g., sorting).

** Use Euclidean measure for distance (d).

** For weighted distance use weight as $1/d^2$

**In general, you may use libraries to process and handle data.

****DO NOT** perform feature scaling before feeding the data in your model.

** To save computation, you may pre-compute pairwise distances of data-points and store that in a matrix.

Experiments: [5+3=8 Marks]

The dataset will be split into Train:Test with 80:20 ratio. Pl shuffle the data before splitting.

1. **Experiment 1:** Report the effect of varying K in [KNN_Distance] on Test data. Choose K values from [1, 3, 5, 10, 20]. Plot Percentage Accuracy vs K. Find the best value of the hyperparameter K.

2. **Experiment 2:** Add noise to only a fraction of the training data: consider separately 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% of the training data for noise addition. Choose a normal distribution with zero mean and standard deviation 2.0. Next, design a KNN using the optimal K found in the earlier experiment. How does the performance vary as compared to that of the noiseless case (Experiment 1)?

(For noise, one may use: numpy.random.normal(*loc=mean*, *scale=std_dev*, size=train_data.shape) with seed)

Report your observations with appropriate explanations.

Datasets:

This dataset comprises three iris species with 50 samples each as well as some properties about each flower. You can find the dataset here.

- ID: Identification number of the flower
- Sepal length: Length of sepal in cm (in real numbers)
- Sepal Width: Width of flower sepal in cm (in real numbers)
- Petal length: Length of flower petal in cm (in real numbers)
- Petal Width: Width of flower petal in cm (in real numbers)
- Species: Three iris flower species (iris-setosa, iris-versicolor, and iris-virginica)

Problem: Predict the species of an iris flower

Submission:

A .zip file containing the python source code and a PDF report file. The final name should follow the template: <Assign-No>_<Your Roll No>.zip. For example, if your roll no is 15CE30021, the filename for LabTest-1 will be: `LabTest-1_15ce30021.zip`

1. A **single python code (.py)** containing the implementations of the models and experiments with comments at function level. The first two lines should **contain your name and roll no**.

2. A report [PDF] containing **[2 Marks]**
   a. Experiment 1: Plot of Percentage Accuracy vs K. Also mention the best choice for the K and the corresponding percentage accuracy.
   b. Experiment 2: Report the performance at different noise levels. Comment on the robustness of K-NN to noise in the training dataset.