Programming Assignment 5: k Nearest Neighbors

Instructions:

- The aim of this assignment is to give you an initial hands-on regarding real-life machine learning application.
- Use separate training and testing data as discussed in class.
- You can only use Python programming language and Jupyter Notebook.
- You can only use **numpy**, **matplotlib** and are not allowed to use **NLTK, scikit-learn or any other machine learning toolkit**.
- **Submit your code as one notebook file (.ipynb) on LMS. The name of file should be your roll number.**
- Deadline to submit this assignment is: **Sunday 3rd May, 2020 11:55 p.m.**

Problem:

The purpose of this assignment is to get you familiar with k nearest neighbor classification. You are given with Iris Data Set that contains information of three different species of iris flower. Your task is to implement k nearest classifier and use it for predicting the flower species given measurements of iris flowers.

Dataset:

The data set contains 150 instances with 5 attributes (4 input variables and 1 output class).

Attribute Information:

- Sepal length in cm
- Sepal width in cm
- Petal length in cm
- Petal width in cm
- Class
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

The data set has been divided into two sets.

- train.csv: 135 instances (45 per class)
- test.csv: 15 instances (5 per class)

Implementation:

Implement kNN keeping in view all the discussions from the class lectures. Specifically, follow the steps shown in figure below.

The KNN Algorithm

Input: Training samples $D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)\}$, Test sample $d = (\vec{x}, y)$, k. Assume $\vec{x}$ to be an m-dimensional vector.

Output: Class label of test sample d

1. Compute the distance between d and every sample in D
2. Choose the K samples in D that are nearest to d; denote the set by $S_d \in D$
3. Assign d the label y_i of the majority class in S_d

Use Euclidean as your distance metric. You can either use sorting or Quickselect to choose k nearest neighbors. Make sure you code in generic enough that it can run with any value of k. Use the procedural programming style and comment your code thoroughly (just like programming assignment 1).

Evaluation:

Report classification accuracy on test set with k = {1, 3, 5}.