Machine Learning: Project 2: K-NN Classifier

Problem Statement:

We would like to diagnose whether a patient has heart disease based on certain clinical parameters. Physicians often diagnose a patient based on the inference for similar patients. We will use the KNN classifier. Write a program to use a K-nearest neighbor it to predict class labels of test data. *Euclidean distance should be used as the distance metric. The value of parameter K should be user defined.* You may choose the best value of K. The learned classifier should be tested on test instances with unknown class labels, and the predicted class labels for the test instances should be printed as output.

Data Set Description:

Training Data Filename: *project2.csv*

Training Data File Format: There are 13 features. The last column denotes the class label. Heart disease correspond to class label 1, and no disease corresponds to 0. Features are: age, sex (0/1), cp (chest pain), trestbps (resting bp), chol (cholestorol), fbs (fasting blood sugar), restecg (rest ecg abnormality), thalach (thalium stress test maximum heart rate), exang (exercise induced angina), oldpeak (ecg ST peak), slope (ecg ST slope), ca (colored fluoroscopy).

Test Data Filename: *project2_test.csv*

Note that, there is no class label column. Each row is a test instance. There are 4 test instances. The row number corresponds to the instance number of the test instances.

Output Format: Predicted class labels (0/1) for the test data exactly in the order in which the test instances are present in the test file. Put a blank space between printed the class labels. (e.g., output 0 0 1 1, if the predicted class labels are - Test Instance 1: 0, Test Instance 2: 0, Test Instance 3: 1, Test Instance 4: 1). Output, in above format, should be printed to the file: *rollnumber_P2.out* (e.g., 20CS10001_P2.out). Strictly use this filename format.

Submission Guidelines:

You may use one of the following languages: c/C++/Java/Python. You should name your file as <rollnumber_P2.extension> (e.g., 14CS10001_P2.c). Your program should be standalone and should not use any *special purpose* library. numpy may be used. You should submit the program file only and not the output/input file. You also have to submit a readme file outlining your approach. The submitted single program file *should* have the following header comments:

Roll # Name # Assignment number # Specific compilation/execution flags (if required)

*Please submit the program in MS Teams by **March 13, 2022 midnight** (hard deadline). Copying from friends/web will lead to strict penalties.*