Classification of Smokers and Never-Smokers on the basis of Genetic information using KNN Model followed by 5-fold Crossvalidation coded from Scratch

The data file consists of data from 15 current smokers and 15 never smokers. It has rows labeled by gene name, and columns labeled by patient sample; the last row gives the class label “CurrentSmoker” or “NeverSmoker”. There is also test data in GSE994-test.txt, which consists of 25 patient samples where the class labels are all “Unknown”

1.

The nearest neighbor classification algorithm is implemented to classify the unknown patients, based on the smoking status of the labeled. The code is to implement k nearest neighbor classification using a variable for the number of neighbors, majority vote among the neighbors to choose the predicted label, and Euclidean distance.

2.

Code is run for one nearest neighbor (“1NN”) and for three (“3NN”). For each of these runs, a plain text file is generated having the following information for the 25 unknown patient samples in numerical order in a tab delimited file, one patient per line (below is just an example, not correct):

PATIENT31 CurrentSmoker

PATIENT32 NeverSmoker

PATIENT33 NeverSmoker

……..

3.

Cross validation on the labeled data, 5-fold cross-validation. The same 5-folds are used as mentioned below:

fold 1: PATIENT1 - PATIENT6

fold 2: PATIENT7 - PATIENT12

fold 3: PATIENT13 - PATIENT18

fold 4: PATIENT19 - PATIENT24

fold 5: PATIENT25 - PATIENT30

4.

The class predictions (CurrentSmoker or NeverSmoker) are calculated from the cross-validation code for each patient, in the same format detailed above, with separate files for 1 and 3 nearest neighbors.

In other words, files contain the following information for the 30 known-label patient samples in numerical order in a tab-delimited text files, each of what lists one patient per line.

PATIENT1 CurrentSmoker

PATIENT2 NeverSmoker

PATIENT3 NeverSmoker