**HW2**

2016312107 문경진



HW2\_2016312107.py for script file

result.txt for classification result (writed by script)

**Overview of Code (HW2\_2016312107.py)**

The code is consisted of five functions. Here are those usage.

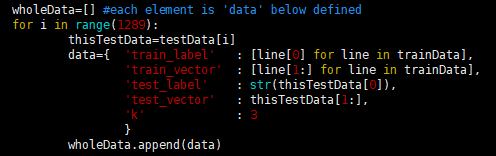
getCSVFile() : given input filename, read data and return data array. csv file address is set by using relative address.

threefold() : given data, fold it to three-fold

knn() : given each ‘metadata’(later discussed) calculate all the distance between target and return result of knn with k=3. This procedure is done by 1. Make distances’ vector, 2. Sort the vector, 3. Count the label of nearest 3 point and 4. Return major label.

Train() : given folded data, make ‘metadata’. After making metadata for whole data, parallelize to make RDD, and collect data using flatMap(knn) and collect(), which dataRDD provides. Then returns true label, predicted label, confusion matrix, scores, and elapsed time for executing train().

‘metadata” is defined like this :



For all testing data, make data dictionary. ‘data dictionary’ is composed of five objects.

‘train\_label’ : labels corresponding to train data.

Ex) [‘A’, ‘A’, ‘B’, …., ‘F’]

‘train\_vector’ : 16\*1 vector of train data. Shape is [1289\*2, 2]

Ex) [[2,4,3,5,1,6,6,5,4,3,6,5,1,2,3,4], [1,1,1,1,3,6,2,8,5,3,5,6,6,6,6,6], …]

‘test\_label’ : real label of testData. Only one-letter-string. Ex) ‘A’

‘test\_vector’ : vector of testData. Ex) [2,2,5,4,7,1,5,8,9,1,2,2,6,5,4,3]

‘k’ : k value for knn. Fixed value 3.

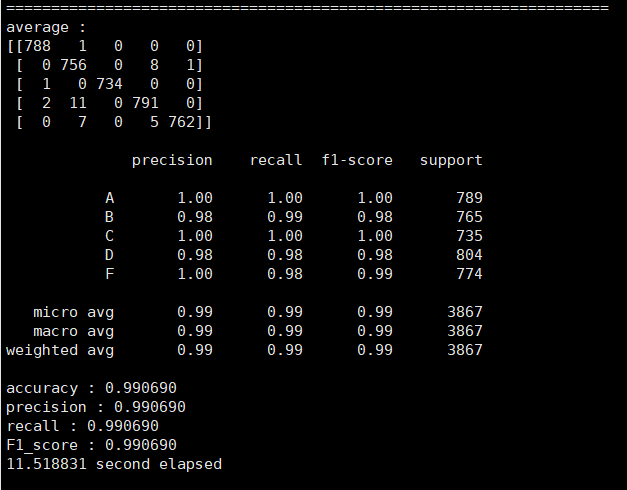
wholeData is consisted of each data dictionaries. This list will be parallelized.

**Result**

As exception occurred at thread main when running code in cluster mode, cannot calculate the executing time on cluster mode. But it could run by client mode by this command :

$ spark-submit --master local[\*] HW2\_2016312107.py

Here is the result of average scores, including confusion matrix, accuracy, precision, recall, f1-score. Result of each fold is written in result.txt.



accuracy, precision, recall and f1\_score is calculated by functions in sklearn.metrics module.

first 2-d array indicates confusion matrix. It also calculated by sklearn.metrics.confusion\_matrix()