#一、题目

<https://www.kaggle.com/c/digit-recognizer>

#二、kNN算法

请参考<https://www.jianshu.com/p/dddd1c348553>

#三、代码

编程语言使用python 3.6

```

from numpy import \*

import operator

import csv

import pdb

def toInt(array):

array=mat(array)

m,n=shape(array)

newArray=zeros((m,n))

for i in range(m):

for j in range(n):

newArray[i,j]=int(array[i,j])

return newArray

def nomalizing(array):

m,n=shape(array)

for i in range(m):

for j in range(n):

if array[i,j]!=0:

array[i,j]=1

return array

def loadTrainData():

l=[]

with open('train.csv') as file:

lines=csv.reader(file)

for line in lines:

l.append(line) #42001\*785

l.remove(l[0]) # remove row 0, get 42000\*785

l=array(l) # transfer list to array

label=l[:,0] # 从0行0列取到最后一行0列, 显示为1行42000列，而不是42000行1列

data=l[:,1:] # from row 0 to the last row, from col 1 to the last col, 42000\*784

return nomalizing(toInt(data)),toInt(label) #data 42000\*784 label 1\*42000

def loadTestData():

l=[]

with open('test.csv') as file:

lines=csv.reader(file)

for line in lines:

l.append(line)

#28001\*784

l.remove(l[0]) #28000\*784

data=array(l)

return nomalizing(toInt(data)) # data 28000\*784

def loadTestResult():

l=[]

with open('knn\_benchmark.csv') as file:

lines=csv.reader(file)

for line in lines:

l.append(line)

#28001\*2

l.remove(l[0]) #28000\*2

label=array(l)

return toInt(label[:,1]) # label 1\*28000

#inX:1\*n dataSet:m\*n labels:m\*1

def classify(inX, dataSet, labels, k):

inX=mat(inX)

dataSet=mat(dataSet)

labels=mat(labels)

dataSetSize = dataSet.shape[0]

diffMat = tile(inX, (dataSetSize,1)) - dataSet

sqDiffMat = array(diffMat)\*\*2

sqDistances = sqDiffMat.sum(axis=1)

distances = sqDistances\*\*0.5

sortedDistIndicies = distances.argsort()

classCount={}

for i in range(k):

voteIlabel = labels[sortedDistIndicies[i],0]

classCount[voteIlabel] = classCount.get(voteIlabel,0) + 1

sortedClassCount = sorted(classCount.items(), key=operator.itemgetter(1), reverse=True)

return sortedClassCount[0][0]

def saveResult(result):

with open('result.csv', 'w', newline = '') as myFile:

myWriter=csv.writer(myFile)

for i in result:

tmp=[]

tmp.append(i)

myWriter.writerow(tmp)

def Test():

trainData,trainLabel=loadTrainData()

testData=loadTestData()

testLabel=loadTestResult()

m,n=shape(testData)

errorCount=0

resultList=[]

for i in range(m):

print ("classify: ",i)

classifierResult = classify(testData[i], trainData[0:20000], trainLabel.transpose()[0:20000], 5)

resultList.append(classifierResult)

print ("the classifier came back with: %d, the real answer is: %d" % (classifierResult, testLabel[0,i]))

if (classifierResult != testLabel[0,i]):

errorCount += 1.0

print ("\nthe total number of errors is: %d" % errorCount)

print ("\nthe total error rate is: %f" % (errorCount/float(m)))

saveResult(resultList)

```

运行程序：

打开cmd窗口，进入上述代码knn.py所在的目录，进入python环境

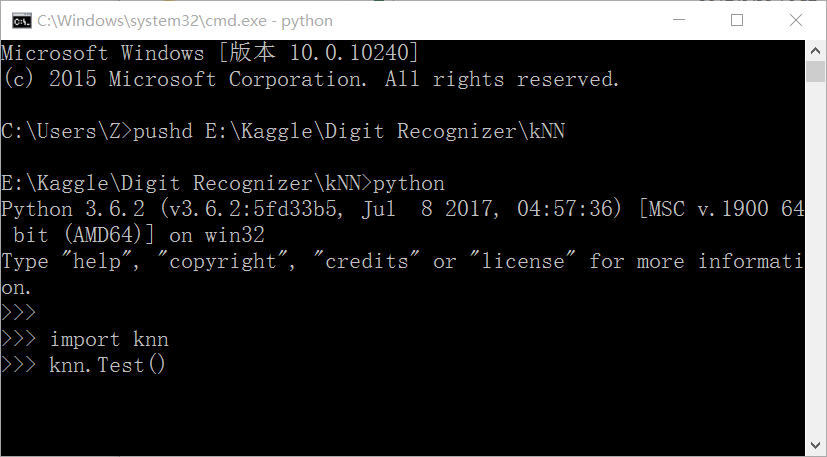
执行命令

```

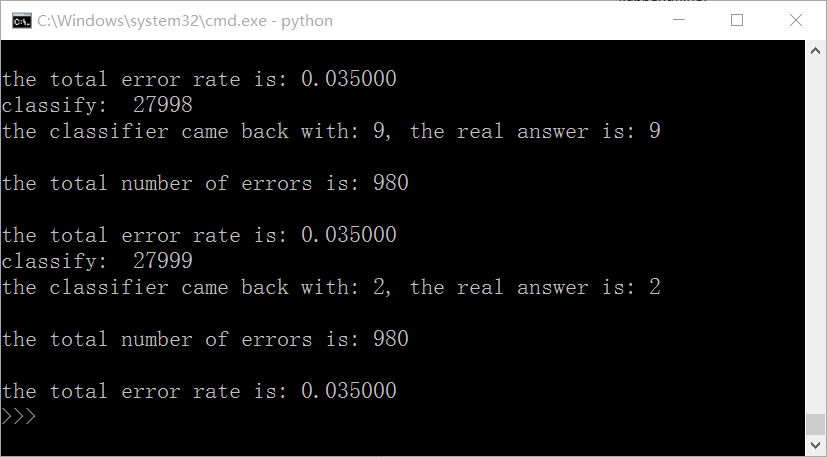
import knn

knn.Test()

```



运行结果：



#四、Github代码下载

[下载地址]( https://github.com/zhenghaishu/Kaggle/tree/master/Digit%20Recognizer)

#五、参考

<https://blog.csdn.net/u012162613/article/details/41929171>