DS 501: STATISTICAL & MATHEMATICAL METHODS FOR DATA SCIENCE
FALL 2019
INDIVIDUAL ASSIGNMENT 4

DUE: Tuesday, November 05

DEVELOP AN OCR SYSTEM USING RIDGE REGRESSION

DATASET

This data is a subset of OCR data taken from
http://cmp.felk.cvut.cz/cmp/software/stprtool/index.html

View an image

You have 4 files: train2_5.txt, train2_5Labels.txt, test2_5.txt, test2_5Labels.txt.

Each row of train and test contains the image of a digit (either 2 or 5). Each image is a 16x16 image, stored in one row as a 256 dimensional vector. You can view any images as follows:

```
X = matrix(read.table('train2_5.txt'))
r = X[1,]       #first digit image, i.e., image in row 1
im = matrix(r,nrow=16,byrow=TRUE)   #convert vector to image
image(im[,ncol(im):1])                    #view image
```

Similarly you can view any image you like for any row of train and test matrices

MODEL BUILDING & EVALUATION

Training part

Write a function for training:

regressionCoefficients <- function(X,Y,lambda)

The above function should return the regression co-efficients when given the training data X, response values Y and the ridge constant lambda. This function should not have any loops. It should just implement one mathematical formula. Don't forget to add a column of ones to X.

Test part

Write a test function for getting predictions

predictions <- function(testX,regressionCoefficients)

This function should also be written using one mathematical formula involving no loops.

Evaluation part

Read about the confusion matrix (YOU HAVE TO READ THIS AS IT IS PART OF THE COURSE):
https://en.wikipedia.org/wiki/Confusion_matrix

Write an evaluation function that makes a confusion matrix:

Matrix <- function(prediction,actualLabels)

The confusion matrix tells us which class is classified correctly and how many mistakes we are making:

For example if you are given the following results:

Your prediction	Actual
2	2
2	5
2	2
2	2
2	2
5	5
5	5
5	2
5	2
5	2
5	5

The confusion matrix for the above result would look like this (assuming 2 is called positive class and 5 is called negative class):

	Actual 2	Actual 5	Total
Predicted 2	TP = true positives = 4	FP = false positives = 1	7
Predicted 5	FN = false negatives = 3	TN = true negatives = 3	4
Total ->	5	6	11

Main script: Bring it all together

Write a main script that

a. reads the training data and builds a regression model. Next it gets predictions from the regression model using the training set as well as the test set.

b. Find a way of mapping the OCR labels to the predictions. So for example if your prediction is 10, then how will you map it to a label?

c. For the training data as well as the test data, make the confusion matrix for different values of lambda, as given in the report.

d. Find out a value of lambda which gives you good results.

TO SUBMIT

1. Source code in R on slate
2. Hard copy of the report

NOTE

Penalty for plagiarism from the internet or for cheating amongst yourselves will NOT be tolerted