Eda Gür

50488

**Comp 421 HW1**

For the code part, I started with the safelog function, because we want to assume that log(0)=0. After that, I defined two variables for training and test datasets as well as a matrix to hold the estimated parameters. After that, I read the image and label data from the csv files by setting the header option to false, because I don’t want to emit any data point. Next, I have a for loop. With this for loop that runs for each 5 letters, I can get the index of the letters from the labels data and use the indices on the image data to separate the letter data points from each other. I also divide them into training and test. I also get the column means for each letter in the training set. This is the Pcd value. This is the probability that a datapoint d belonging to class c is colored.

I have two function called scoring\_function and scoring\_function\_matrix. The first one is for a single class. The second function uses a for loop to put together the scoring function vectors for the five classes into a single class. I call rbind to unite these vectors. Scoring\_function\_matrix takes pcd as parameter and passes pcd[i,] to scoring\_function method. We use the parametric classification here. rowSums(data\_matrix\*safelog(extended\_pcd) + (one\_matrix-data\_matrix)\*safelog(one\_matrix-extended\_pcd))+prior\_prob gives s the scoring function formula we have seen in the multivariate classification class on 01.10.2018. For a given data point, whichever class has the highest g(x) function is the one that we will predict the data point belongs to.

Finally, I have a function for the confusion matrix. Training and test have separate confusion matrices. With a quick glimpse, it is visible that the training data is predicted more accurately. For the implementation of the confusion, I use a double for loop to iterate over each class and each data point. I find the max scoring function and if, for example, B is classified as C, I increment the confusion[3,2] entry.

What I difficulty with: I wanted to define multiple variables to the same value for initialization, but R doesn’t allow it. Also, I couldn’t find an equivalent version of a ++ or += in R. If I could, it would come handy in the confusion function, because I have a lengthy line a=a+1 for the increment in the confusion matrix entries where a is long piece of code.