Project Part 1 [10 points] Feature Extraction, Density Estimation and Bayesian Classification (Due October 6, 11:59pm)

This part of the project uses a subset of images (with modifications) from the MNIST dataset. The original MNIST dataset (http://yann.lecun.com/exdb/mnist/) contains 70,000 images of handwritten digits, divided into 60,000 training images and 10,000 testing images. We use only images for digit "3" and digit "7" in this project. The data is stored in ".mat" files. You may use the following piece of code to read the dataset in Python (or you may use the load filename command in Matlab, since these are .mat files):

```
import scipy.io

data = scipy.io.loadmat('matlabfile.mat')
```

Following are the statistics for the data you are going to use:

Number of samples in the training set: "3": 5713; "7": 5835

Number of samples in the testing set : "3": 1428; "7": 1458

You will practice doing the following three tasks in this project:

Task 1. Feature extraction and normalization

In the .mat file, each image is stored as a 28x28 array of pixels. The pixel values range from 0 to 255. For each image i, compute two features:

1. Skewness of image k_i – This is measure of symmetry of the pixel values for i. For more information, visit https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm
2. Ratio of brighter to darker pixels r_i using a threshold *T*.

We further normalize the features in the following way, before starting any subsequent tasks. Using the feature representations of all the training images from both classes (each image i is now viewed as a 2-d vector, $X_i = [k_i, r_i]^t$, as explained), compute the mean M_j and standard deviation S_j, j=1,2, for the first and the second feature, respectively. M_j and S_j will be used to normalize all the feature vectors (both training and testing): for each feature vector X_i representing image i, a normalized feature vector Y_i will be computed as

$$Y_i = [\, y_{1i}, y_{2i}]^t = [\, (k_i - M_1)/S_1, (r_i - M_2)/S_2]^t$$

This Y_i is the final feature representation for image i and will be used for subsequent steps.

Task 2. Density estimation

We assume in the 2-d feature space of Y_i defined above, samples from each class follow a normal distribution. Using the MLE method, you will need to estimate the parameters for the 2-d normal distribution for each class/digit, using the respective training data for that class/digit. Note: You will have two distributions, one for each digit.

Task 3. Bayesian Decision Theory for optimal classification

Use the estimated distributions for doing minimum-error-rate classification, for the following two cases respectively:

Case 1: Assume that the prior probabilities are the same (i.e., P(3) = P(7) =0.5).

Case 2: Assume that the prior probabilities are: P(3) = 0.3, P(7) = 0.7.

For both cases, report the error rate of the optimal classifier, for the training set and the testing set respectively where error rate is defined as ratio of number of incorrect predictions to total number of samples.

Task 4. Try different threshold values for the feature r_i

Do all the above steps for two values of threshold *T* for r_i feature-

Case 1: *T* = 150

Case 1: *T* = 200

What to submit:

1. Your code for doing the above.

2. A report summarizing the results, e.g., the estimated parameters of the distributions, the training and testing error probabilities. Include in your report any intermediate results that you deem helpful for illustrating the partial results.

Note: There is no minimum or maximum length requirement for the report. Writing the report is the opportunity for you to reflect on your understanding of the problems/tasks through organizing your results.

The data files for the project are uploaded in the Files/Assignments folder:

train_data.mat

test_data.mat