**Project Overview:**

In this part, you need to first perform parameter estimation for a given dataset (which is a subset from the MNIST dataset). The MNIST dataset contains 70,000 images of handwritten digits, divided into 60,000 training images and 10,000 testing images. We use only images for digit “7” and digit “8” in this question.

Therefore, we have the following statistics for the given dataset:

* Number of samples in the training set:  "7": 6265 ;"8": 5851.
* Number of samples in the testing set: "7": 1028;   "8": 974

You are required to extract the following two features for each image:

1. The average of all pixel values in the image
2. The standard deviation of all pixel values in the image

We assume that these two features are independent, and that each image (represented by a 2-D features vector) is drawn from a 2-D normal distribution.

You may go to the original MNIST dataset (available here[http://yann.lecun.com/exdb/mnist/ (Links to an external site.) (Links to an external site.)](http://yann.lecun.com/exdb/mnist/)) to extract the images for digit 7 and digit 8, to form the dataset for this project. To ease your effort, we have also extracted the necessary images, and store them in “mnist\_data.mat” files. The file can be downloaded [here.](https://canvas.asu.edu/courses/45793/files/11859723/download?wrap=1) A description of the file can be downloaded [herePreview the document](https://canvas.asu.edu/courses/45793/files/11859724/download?wrap=1).

You may use the following piece of code to read the dataset:

import scipy.io

Numpyfile= scipy.io.loadmat(‘mnist\_data.mat’)

The specific algorithmic tasks you need to perform for this part of the project include:

1. Extracting the features and then estimating the parameters for the 2-D normal distribution for each digit, using the training data. Note: You will have two distributions, one for each digit.
2. Use the estimated distributions for doing Naïve Bayes classification on the testing data. Report the classification accuracy for both “7” and “8” in the testing set.
3. Use the training data to train a Logistic Regression model using gradient ascent. Report the classification accuracy for both “7” and “8” in the testing set.**Note that you are not allowed to use package like sklearn to compute the boundary. You need to implement your own version for using gradient ascent to find the solution.**

**Algorithms:**

MLE Density Estimation, Naïve Bayes classification, Logistic regression

**Resources:**

A subset of MNIST dataset, download either from[http://yann.lecun.com/exdb/mnist/ (Links to an external site.) (Links to an external site.)](http://yann.lecun.com/exdb/mnist/) (requiring you to extract data corresponding to digit 7 and digit 8 only),  or from the .mat files provided.

**Workspace:**

Any Python programming environment.

**Software:**

Python environment.

**Language(s):**

Python. (MATLAB is equally fine, if you have access to it.)