# Project Description

This project involves implementing supervised, unsupervised, and deep learning techniques for

density estimation and classification. The project focuses on a subset of the MNIST dataset

containing images of digits "0" and "1". The project involves four tasks: feature extraction, parameter

calculation, implementation of Naïve Bayes classifiers, and prediction of labels for the test data using

the classifiers. Finally, calculating the accuracy of the predictions.

# Preparation

In the 'Project1’' Jupyter notebook code, you will load the trainset and testset for digit0 and

digit1 respectively (Please read the code and you will understand). Both trainset and testset are

sub-dataset 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 a part of images

for digit “0” and digit “1” in this question.

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

● Number of samples in the training set: "0": 5000 ;"1": 5000.

● Number of samples in the testing set: "0": 980; "1": 1135

We assume that the prior probabilities are the same (P(Y=0) = P(Y=1) =0.5), although you may have

noticed that these two digits have different numbers of samples in testing sets.

In the existing code, myID is a 4-digit string; train0 is your trainset for digit0; train1 is your trainset for digit1; test0 is your testset for digit0; and test1 is your testset for digit1. They are all Numpy Arrays. You can also convert them into python arrays if you like.

Please do not change any existing code and just write your own logic with the existing code.

You may go to the original MNIST dataset (available here http://yann.lecun.com/exdb/mnist/) to

extract the images for digit 0 and digit 1, to form the dataset for this project. To ease your effort, we

have also extracted the necessary images, and store them in “.mat” files. You may use the following

piece of code to read the dataset:

● import scipy.io

● Numpyfile= scipy.io.loadmat(‘matlabfile.mat’)

# Programming

For your code logic, you have 4 tasks to do:

**Task 1:**

You need to first extract features from your original trainset in order to convert the original data arrays

to 2-Dimensional data points.

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

● Feature1: The average brightness of each image (average all pixel brightness values within a

whole image array)

● Feature2: The standard deviation of the brightness of each image (standard deviation of all

pixel brightness values within a whole image array)

We assume that these two features are independent and that each image is drawn from a normal

distribution.

**Task 2:**

You need to calculate all the parameters for the two-class naive bayes classifiers respectively, based

upon the 2-D data points you generated in Task 1 (In total, you should have 8 parameters).

● (No.1) Mean of feature1 for digit0

● (No.2) Variance of feature1 for digit0

● (No.3) Mean of feature2 for digit0

● (No.4) Variance of feature2 for digit0

● (No.5) Mean of feature1 for digit1

● (No.6) Variance of feature1 for digit1

● (No.7) Mean of feature2 for digit1

● (No.8) Variance of feature2 for digit1

**Task 3**:

Since you get the NB classifiers' parameters from Task 2, you need to implement their calculation

formula according to their Mathematical Expressions. Then you use your implemented classifiers to

classify/predict all the unknown labels of newly coming data points (your test data points converted

from your original testset for both digit0 and digit1). Thus, in this task, you need to work with the

testset for digit0 and digit1 (2 Numpy Arrays: test0 and test1 mentioned above) and you need to

predict all the labels of them.

Note: Remember to first convert your original 2 test data arrays (test0 and test1) into 2-D data points

as exactly the same way you did in Task 1.

**Task 4:**

In Task 3 you successfully predicted the labels for all the test data, now you need to calculate the

accuracy of your predictions for testset for both digit0 and digit1 respectively.

# Deliverables

**Results & Output**

As the result from your Notebook of Project 1, you should have your Id(string), 8 components for computed parameters and 2 components for accuracy. The order of these 11 components should be a list and look like the following:

['Id', Mean\_of\_feature1\_for\_digit0, Variance\_of\_feature1\_for\_digit0,

Mean\_of\_feature2\_for\_digit0, Variance\_of\_feature2\_for\_digit0 ,

Mean\_of\_feature1\_for\_digit1, Variance\_of\_feature1\_for\_digit1,

Mean\_of\_feature2\_for\_digit1, Variance\_of\_feature2\_for\_digit1,

Accuracy\_for\_digit0testset, Accuracy\_for\_digit1testset]

**Report**

Draft a report to go with your Results Submission. The report must contain a detailed description of your observations and analysis of the project (1-2 pages).