Project 1 - Density Estimation and Classification

Prem Patel – 1217192561

1. Introduction:

In this project we were told to implement the Naïve Bayes and Logistic regression model using the MNIST dataset. There are in total 70,000 images of handwritten dataset, divided in 60,000 training images and 10,000 testing images. In this particular task we only use the dataset of digit 7 & 8.

2. Feature Extraction:

We were tasked to take the training data and testing data of the digit 7 & 8 and execute the model on that particular dataset. We were given

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

For Feature Extraction we have to find out the mean and standard deviation of the all pixel values of the image. Here the image data which we are using is having pixels 28x28. So the data would be in matrix form giving each value of pixel so the data would be contain 784 columns representing the 28x28 pixel values of one image in single row and there would be 12116 (7 & 8 – [6265+1028]+[5851+974]) rows of this pixel values given combined image data 7 & 8. Also, there would be training and testing labels in data in form of 2002x1 matrix to train & test the data respectively. Here, we assume that the two features are independent and each image (represented by a 2-D features vector) is drawn from a 2-D normal distribution.

3. Naïve Bayes Classification:

To apply Naïve Bayes model firstly the data is divided into four parts training_X, training_Y, testing_X, testing_Y in which,

- Training_X represents the image data used to train data
- Training_Y represents the image labels for training the model of the data given with respect to Training_X
- Testing_X represents the image data used to test the model
- Testing_Y represents the image labels used to verify that whether the model is able to predict the data given in Testing_X correctly or not.

So now we take the mean and standard deviation of the entire training_x dataset. Then we divide the data into two sets, one set containing only digit 7 and another set containing digit 8. Now we calculate the gaussian naïve bayes, because it is a typical assumption if the values are continuous associated with each class are distributed according to normal distribution. We use this formula to predict class probability:

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

We already knew the mean and standard deviation so implementing those values in the formula. We are applying the formula for each data set separately on digit 7 set and digit 8 set. Now we use testing data to test the data sets to predict whether that value will be classified in which dataset if probability of digit 7 greater than probability of digit 8 than it will append number 7 in the list or else number 8. This will continue for complete testing_x dataset and each data will be classified as 7 or 8 and will be stored in the list. After analysing and predicting the data we will now check the accuracy of the model by comparing the output predicted by the model that is stored in the list and the labels given in testing_y. Here, we will find two different types of accuracy:

1) Overall Accuracy: In this we will calculate the entire accuracy of the dataset i.e. total right answers predicted by the model/ total no of samples.
2) Class Accuracy: In this we will calculate the accuracy which is provided by a particular class (in our case it is set). It is calculated using the [total no of output (belonging to that class/set) / total no of samples.]

This way the Naïve Bayes Model is implemented.

4. Logistic Regression:

In logistic regression we use the same data separation techniques and instead of using separated mean and standard deviation we now combine them to make a one common matrix containing both mean and standard deviation. In this we use sigmoid function and regression formula to calculate probability of occurrence of that particular digit in the testing data inputted. In this formula there are two more input parameters which we take into consideration i.e. learning rate and epoch. The learning rate and epoch is to be defined by the user. For now, the learning rate has been set to 0.001 and epoch has been set to 100,000 (as these are the only value which are found optimal i.e. gives very good accuracy using try and error). After that we do the same procedure of finding the output values for testing_x and comparing it with the labels given in the testing_y. We are also calculating the overall accuracy and class accuracy of both the sets of digit 7 and digit 8. This way logistic regression is implemented.

5. Results:

	Overall Accuracy	Class_Accuracy_7	Class_Accuracy_8
Naïve Bayes	70.32%	74.02%	66.42%
Logistic Regression	81.66%	78.60%	84.90%

6. Summary:

In this project we got the following accuracies represented in the table above. To increase the accuracies, we can also try to extract more features for more better prediction of the data.