**Density Estimation and Classification**

Name: Thanuja Kallamadi

Student ID: 1217132202

Email ID: [tkallama@asu.edu](mailto:tkallama@asu.edu)

**Introduction:**

The main aim of the project is to perform the parameter estimation and extraction for a given dataset and classify the digits into 7 and 8. We are supposed to find the classification accuracies of both the digits 7 and 8 of the testing set using Naïve Bayes and Logistic Regression classification algorithms.

**Dataset:**

For the project we are using the original MNIST dataset to extract images for digit 7 and digit 8. The MNIST dataset contains 70,000 images of handwritten digits which are divided into 60,000 training and 10,000 testing images.

We are considering only images of digit 7 and 8 for this project. The number of samples in the training set consists of 6265 images of digit 7 and 5851 images of digit 8. The number of samples in the testing set consists of 1028 images of digit 7 and 974 images of digit 8.

**Feature Extraction:**

Two main features mentioned below were extracted for all the images in the dataset.

1. The average of all pixel values in the image:

Mean value of all pixel values in the image is found using the below formula:

It is assumed that each images pixel values are drawn from a Gaussian distribution.

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

Standard deviation value of all pixel values in the image is found using the below formula:

It is assumed that each images pixel values are drawn from a Gaussian distribution.

It is also assumed that the two features mentioned above are linearly independent.

**Naïve Bayes Classification:**

Naïve Bayes classifier is one of the machine learning classifiers which uses Bayes theorem to calculate the probability of unknown datasets. The Naïve Bayes classifier assumes that the presence of one feature doesn’t affect the presence of other feature which implies that the features are conditionally independent of each other.

The Gaussian probability values for each digit were calculated by passing the values of covariance matrices to the below formula:

The above equation is approximately equal to the one below which is in terms of

The value which we got above was multiplied with its prior probabilities which provided us the value of probabilities of digits 7 and 8. Finally the probabilities of both digit 7 and digit 8 were compared and were predicted as digit 7 if the probability of 7 was greater than the probability of 8 and vice versa.

The accuracies of each digit were calculated by comparing the values predicted with the ones in the testing data.

**Logistic Regression Classification:**

Logistic Regression is one of the supervised learning classification technique which uses the logistic function (sigmoid function) to model the data.

The logistic function i.e. sigmoid function is given by the below formula

Here the value of z is given as z=

where is the weights vector which has a constant weight and weights associated with each vector.

By using the Gradient ascent formula, we iteratively calculate the and keep updating weights until the optimal weights which provide the maximum accuracy are found. The formula for gradient ascent is as given below:

where is the learning rate and is the gradient.

After finding the weights we are calculating the z values and the z values are passed to calculate the value of logistic function. If the values returned are greater than 0.5(boundary value as the values are varying from 0 to 1) we are considering those data as digit 8 if not, we are considering it as digit 7.

The accuracies of each digit were calculated by comparing the values predicted with the ones in the testing data.

**Results:**

1. **Naïve Bayes:**

The Naïve Bayes classifier predicted the digits 7 and 8 with an accuracy of 69.53%.

The accuracy of digit 7 is 75.97%

The accuracy of digit 8 is 62.73%

1. **Logistic Regression:**

The logistic regression classifier predicted the digits 7 and 8 with an accuracy of 73.57% in seconds when it was iterated for 10,000 iterations by varying the learning rate.

The above learning rate was taken because taking high value of may make the model converge faster and there is a chance of not reaching the exact maxima.

However, changing the number of iterations will provide us different accuracies. Some of the results are as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number of iterations | Overall Accuracy | Accuracy of predicting digit 7 | Accuracy of predicting digit 8 | Learning rate | Time taken |
| 1000 | 68.73% | 71.98% | 65.29% | 0.0005 | 60 sec |
| 5000 | 71.93% | 76.94% | 66.63% | 0.0005 | 300sec |
| 12000 | 74.42% | 80.35% | 68.17% | 0.0005 | 840sec |

**Conclusion:**

In this project two features i.e. mean and standard deviation were extracted for all the images. Then two classification algorithms Naïve Bayes and Logistic Regression were used to model and predict the data.

The probability of the class was calculated with the help of Bayes theorem in Naïve Bayes classification and logistic function in Logistic regression classification method.

Finally, accuracies of each digit 7 and 8 were calculated and printed.