CSE 575: Statistical Machine Learning

Project Part 1: Density Estimation and Classification

Name: Kalp Patel
ASU ID: 1217202272

1.INTRODUCTION

In this project we are provided MNIST dataset which contains 70,000 images of handwritten digits which is divided into 1:6 proportion, given 60,000 training images and 10,000 test images. For this project we are using images for only digit ”7” and digit “8” datasets. By using only these two datasets we are given:
Number of samples in the training set: "7": 6265;"8": 5851.
Number of samples in the testing set: "7": 1028; “8": 974

We have used mainly two features for this dataset rather than using whole pixel array of the images.

1. Average value of all the pixel values of the image.
2. Standard deviation of all the pixel values of the image.

By using these two features as an input features to statistical machine learning algorithms we are required to predict the digit written in the image by using **Naïve Bayes classification** and **Logistic Regression**.

2.FEATURE EXTRACTION

As we are given dataset:
Training set: "7": 6265;"8": 5851.
Testing set: "7": 1028; “8": 974

Two features used for this project are average value and standard deviation value of all the pixel values of the image.
Since all the pixel values are given in a single row, I’ve calculated mean and standard deviation for all the rows present is a training and testing dataset. By this operation initially each row’s dimension was 784(28*28) and were reduced to just 2 dimensions.

Training dataset dimension: 12116 X 2
Test dataset dimension: 2002 X 2
Here we have assumed that these two features are independent and that each image is drawn from a 2-D normal distribution.

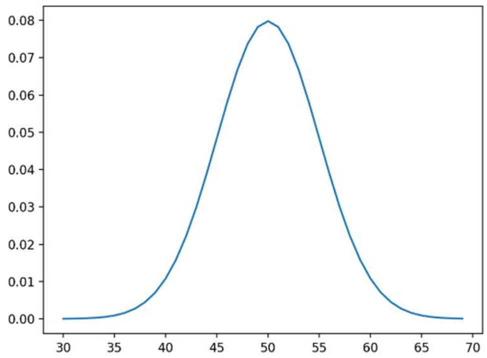
3.NAIVE BAYES CLASSIFIER

This model is a probabilistic statistical machine learning model used for classification. This method is based on Bayes theorem which is stated as below.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

By using Bayes theorem using B as given input features we can find probability for the output label A. Here A will we probability of image being “7” or “8” and B will be input features which are mean and standard deviation.

Here input variables are having continuous values and not discrete there for assumption we have considered is that these values are normally distributed.



Now, to calculate probability of any output variable given continuous variable distributed normally is given by following equation.

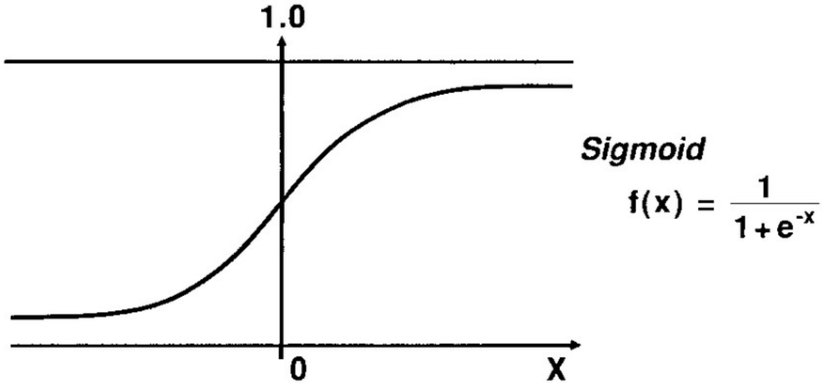
$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

Main assumption is naïve Bayes classifier is that all the input variables are conditionally independent i.e. for both the input variables y1(mean) and y2(std) we can simply multiply both of their probability by P(x|y)= P(x|y1) * P(x|y2).

We are computing this equation for the both the digit “7” and “8” and the classifier will give output as maximum of both.

4.LOGISTIC REGRESSION

For any given input features logistic regression uses a model that assigns a probability that each input belongs to a particular category.
For generating probabilities logistic regression uses a function that generates output between 0 and 1 for input variables. For this project I have used sigmoid function which is given as below.



Functions are having weights and our task is to find optimal weight values. And to compute the measurement of how well these weights are performing we use Loss function. Initially weights are assigned randomly and then they are optimized in each iteration to reduce loss function.

Algorithm used for reducing loss function is gradient descent. Derivative of loss function with respect to weights shows how loss function would change if parameters are modified.

$$\frac{\delta J(\theta)}{\delta \theta_j} = \frac{1}{m} X^T(g(X\theta) - y)$$

After calculating gradient, we will be updating weights by subtracting to them the gradient times the learning rate.

weights=weights - learning rate*gradient

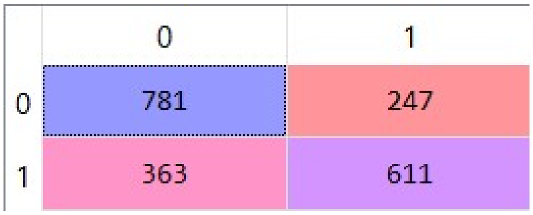
This step is repeated several times to get optimal weight values.

Now, for predicting a label for an unseen data we’ll be calling sigmoid function to get the probability for input features belonging to a class 1. If we’re taking threshold as 0.5 then if this probability gives greater than 0.5 value than output label is predicted as class 1 otherwise class 0. For the better accuracy threshold, I have used is 0.48.

5.RESULTS

1. Naïve Bayes Classifier

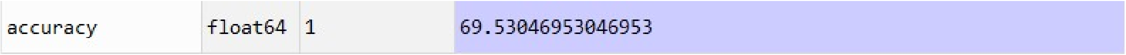
Confusion Matrix:



	0	1
0	781	247
1	363	611

Here confusion matrix is given for the NBC where row labels indicates true values and columns labels indicates predicted values. 0 label is for image of digit “7” and 1 label stands for image of digit “8”.

Accuracy:

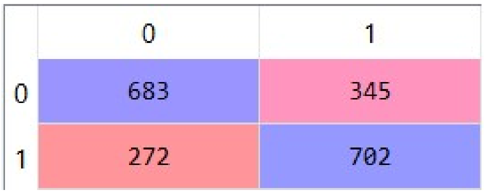


accuracy	float64	1	69.53046953046953

2. Logistic Regression

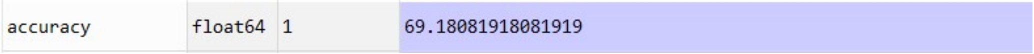
Configurations of the Logistic regression classifier.
Learning Rate=0.04
Number of iterations= 15000
Threshold value for Probability=0.48

Confusion Matrix:



	0	1
0	683	345
1	272	702

Accuracy:



accuracy	float64	1	69.18081918081919