Project Part 1 Report
Naive Bayes Classifier

Nourhan ElNaggar
1221732434

I. Introduction

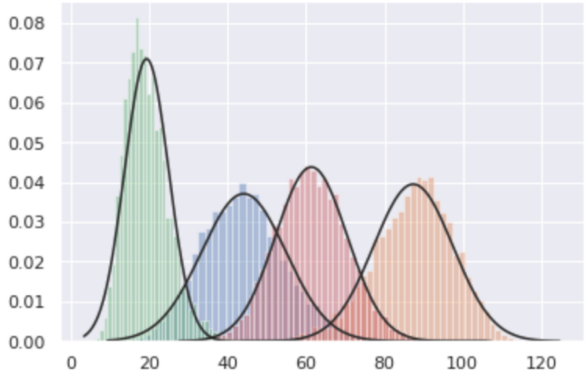
The aim of this project is to use the trainset and testset that was provided from the MNIST database (Modified National Institute of Standards and Technology database) to classify all the unknown labels for two digits (24 in my case) and predict all the labels of them.

II. Discussion and results

I extracted the features from the original trainset and testset. Two features were required to be extracted:

- **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)

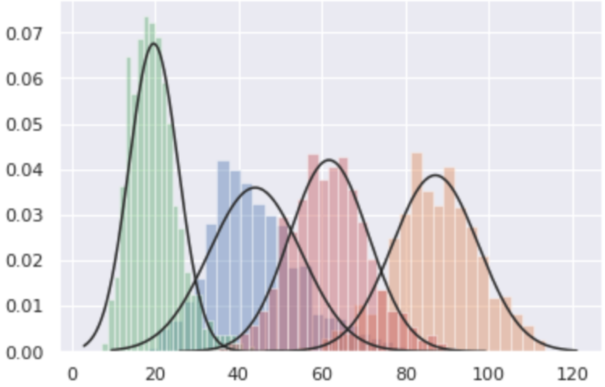


This chart represents the normal distribution chart for the extracted data:

- The train data 1 feature 1
- The train data 1 feature 2
- The train data 2 feature 1
- The train data 2 feature 2

Did the same step for testset as well, and below is the result chart:

- The test data 1 feature 1
- The test data 1 feature 2
- The test data 2 feature 1
- The test data 2 feature 2



The length of the data used is as following:

Training	*Length*	*Testing*	*Length*
0	5000	0	980
1	5000	1	1135

Based on the first task's data, I generated the below parameters for both classes:

Parameter	*Digit 0*	*Digit 1*
Mean of feature1	44.194659949	19.407065051
Variance of feature1	114.689917342	31.4895261975
Mean of feature2	87.4189676075	61.4115611832
Variance of feature2	100.735082148	82.2415746334

After collecting the above data, I used them for prediction classification of both digit 0 and digit 1 using Naive Bayes Classification. Therefore, I applied the stats.norm on both mean and variance of trained data and used the results with the Probability Density Function (PDF) -equation below- that was calculated on the data to provide the probability matrix for each parameter and used argmax to extract the maximum value of the NumPy array.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

where

μ = mean
σ = standard deviation
π = 3.14159
e = 2.71828

Finally, I used the accuracy function to compute the accuracy of the predicted vs the actual data. Below table is the results I got.

Title	*Digit 0*	*Digit 1*
Accuracy	0.917346938776	0.923348017621

Overall, this was my first time implementing Naïve Bayes Classifier on a data, it was challenging and I learned a lot from it.