CSE 575 STATISTICAL MACHINE LEARNING

PROJECT 1 – DENSITY ESTIMATION AND CLASSIFICATION USING FASHION-MNIST DATASET

NAME: VAMSI KRISHNA KANAGALA **ID: 1218608781**

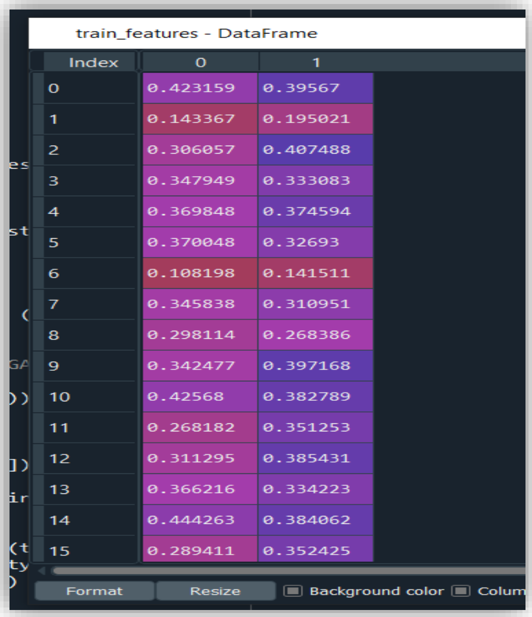
INTRODUCTION:

In this project, we were given a dataset, which is a subset from Fashion MNIST dataset. We have 14000 images of article images, which are divided into 12000 training images and 2000 testing images. Among the 12000 training images, half of them are T-shirts and the other half are Trousers. Similarly, among the 2000 testing images, half of them are T-shirts and the other half are Trousers.

FEATURE EXTRACTION:

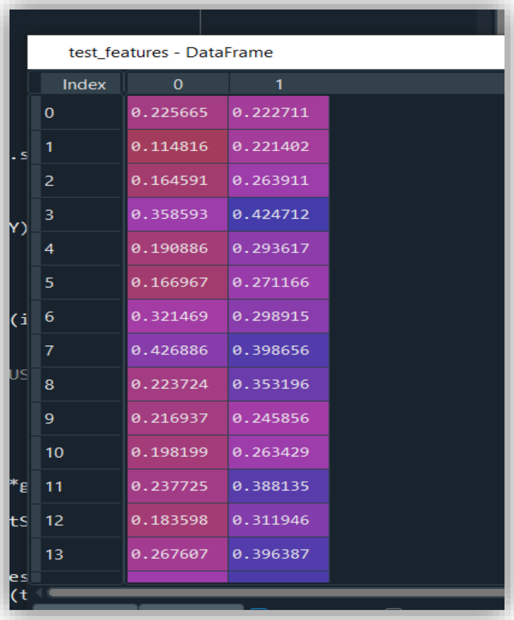
Each row in both the training and testing dataset, contains an array of 784 integer values, which are the values of each pixel. And for each image, we are required to extract two features, the average and the standard deviation of all the pixel values.

We assume that both the features are independent, and that each image is drawn from a 2-D Normal distribution. Therefore, we can implement the algorithms, Naïve Bayes Classifier and Logistic Regression on the datasets we derived. Below are the snippets of the feature matrix of both training and testing datasets .



train_features - DataFrame

Index	0	1
0	0.423159	0.39567
1	0.143367	0.195021
2	0.306057	0.407488
3	0.347949	0.333083
4	0.369848	0.374594
5	0.370048	0.32693
6	0.108198	0.141511
7	0.345838	0.310951
8	0.298114	0.268386
9	0.342477	0.397168
10	0.42568	0.382789
11	0.268182	0.351253
12	0.311295	0.385431
13	0.366216	0.334223
14	0.444263	0.384062
15	0.289411	0.352425



test_features - DataFrame

Index	0	1
0	0.225665	0.222711
1	0.114816	0.221402
2	0.164591	0.263911
3	0.358593	0.424712
4	0.190886	0.293617
5	0.166967	0.271166
6	0.321469	0.298915
7	0.426886	0.398656
8	0.223724	0.353196
9	0.216937	0.245856
10	0.198199	0.263429
11	0.237725	0.388135
12	0.183598	0.311946
13	0.267607	0.396387

NAÏVE BAYES CLASSIFICATION:

Once after finding the feature matrix for the training datasets, We have summarized the data by class. We have given the class-label as '**0**' for T-Shirts and '**1**' for Trousers. Based on the class labels, we have separated our training datasets into two separate datasets to calculate the statistics(mean and standard deviation) for each class. As the total number is same for both T-shirts and Trousers, the prior probability would be same for both the classes. Now, for each image in testing data, probability of each are calculated by calculating the posterior probability and by using the Gaussian Probability Distribution function. Below are the formulae used to find the class probabilities.

$$P(Y|X) = \left(\pi\ P(Xi|Y)\right) * \frac{P(Y)}{P(X)}$$

Y – Class, X_i – i^{th} test data features

$$P(Xi|Y) = \frac{1}{\left(\sqrt{2\pi}\right) * \sigma} * e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

X_i - i^{th} test data features

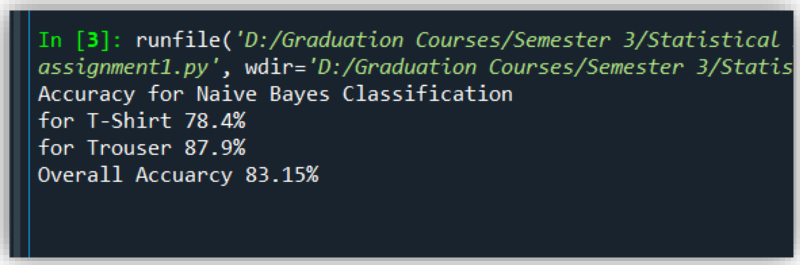
σ - Standard Deviation for x

μ – Mean for x

Once after finding the class probabilities for a given image, if the class probability of T-Shirt is greater than the class probability of Trousers, then the image is classified as T-Shirt. Similarly, if the class probability of Trouser is greater than the class probability of T-Shirt for a given image, then it is classified as Trouser.

Thus, for the given dataset, by using Naïve Bayes Algorithm we can classify a given image as a T-shirt or Trouser.

Results obtained are as follows:



```
In [3]: runfile('D:/Graduation Courses/Semester 3/Statistical
assignment1.py', wdir='D:/Graduation Courses/Semester 3/Statis
Accuracy for Naive Bayes Classification
for T-Shirt 78.4%
for Trouser 87.9%
Overall Accuarcy 83.15%
```

LOGISTIC REGRESSION: Logistic Regression is a statistical technique which models the probabilities for classification problems with two possible outcomes. It is also an extention to the linear regression model which is used for classification problems.

Here, we assume that **P(Y|X)** takes the form of a logistic sigmoid function. So, we can say that Logistic regression uses the logistic function for modelling P(Y|X), considering only the case of y belongs to **{0,1}**.

Below are the formula we used in this algorithm,

$$\text{Sigmoid Function} \quad - \quad \sigma(t) = \frac{1}{1+e^{-t}}$$

Given a sample x, we classify it as 0 (i.e., predicting y=0) if **P(y=0|x) >= P(y=1|x).**

The model parameters we used in this algorithm are weights, learning rate and number of epochs. As we have two classes, we use three weights w0, w1, w2. Initially these weights are initialized to 0. These weights will be updated in every iteration (epochs) by using gradient ascent technique. Learning rate decides the rate at which our model updates the weights. Inorder to train the model efficiently, it is better to take a slow learning rate. Number of iterations depends on how the weights are updating in each iteration, and will iterate until the weights converge.

We will update the weights as per the below formula,

$$w^{(k+1)} = w^{(k)} + \eta \nabla_{w^{(k)}} l(w)$$

$\eta > 0$ is called as the learning rate

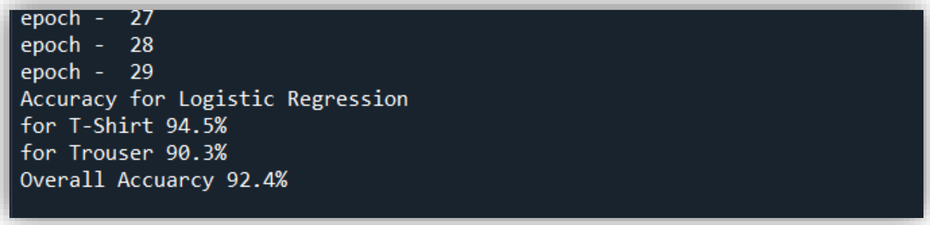
$l(w)$ is the conditional likelihood

$\nabla_{w^{(k)}}$ is the error obtained while using the weights w(k)

Once after finding the class probabilities for a given image, if the class probability of T-Shirt is greater than 0.5, then the image is classified as T-Shirt. Similarly, if the class probability of Trouser is greater than 0.5, then it is classified as Trouser.

Thus, for the given dataset, by using Naïve Bayes Algorithm we can classify a given image as a T-shirt or Trouser.

By using learning rate as 0.3 and number of epochs as 30, the results obtained are as follows



```
epoch -  27
epoch -  28
epoch -  29
Accuracy for Logistic Regression
for T-Shirt 94.5%
for Trouser 90.3%
Overall Accuarcy 92.4%
```