Project Part-1 Report

**Density Estimation and classification using Fashion mnist**

**Abstract:**

In this part of the project, we have extracted the features for each image and using these features we have used naïve bayes classification and logistic regression to classify them in to two classes T shirts and Trousers. In order to use the classifiers, we took mean and standard deviation of the image pixels as features of our dataset.

**Dataset:**

In this project we used Fashion-MNIST dataset containing 12000 samples of training data and 2000 samples of test data. Training data contains 6000 samples for each T shirts and Trousers class. Test data contains 1000 samples for each T shirts and Trousers class. Labels 0 and 1 represent ‘t shirts’ and ‘trousers’ classes respectively.

**Feature Extraction:**

We must extract two features for each image in order to classify them. The two features for each image are the average of the pixels of each image and standard deviation of pixels of each image.

**Parameter Estimation:**

We need to estimate the parameters for using the naïve bayes classifier. For each class we need to calculate the µ and σ for each feature and for each class.

As we consider that each feature is independent, we can consider each feature as a 1-D gaussian distribution and calculate the values of µ1, µ2 and σ1, σ2 for each class.

µ1 for class 0 = mean of means of each image.

µ2 for class 0 = mean of standard deviations of each image.

σ1 for class 0 = variance of means of each image.

σ2 for class 0 = variance of standard deviations of each image.

Similarly,

µ1 for class 1 = mean of means of each image.

µ2 for class 1 = mean of standard deviations of each image.

σ1 for class 1 = variance of means of each image.

σ2 for class 1 = variance of standard deviations of each image.

Parameters for class 0:

[µ1, µ2]= [0.3256077664399094, 0.3200360871033624]

[σ1, σ2]= [0.012856013931536936, 0.007742265243564327]

Parameters for class 1:

[µ1, µ2]= [0.22290531462584984, 0.33394171202721934]

[σ1, σ2]= [0.003243958057139923, 0.0032532239122850352]

**Naïve Bayes Classifier:**

We will use the formula P(Y/X) = P(X/Y).P(Y)

P(Y) here is the prior. As we know that the no. of samples for both class 0 and class 1 are equal so the prior becomes 0.5.

P(X/Y) = P(X1/Y). P(X2/Y)

We need to use the PDF to calculate the value of P(X1) and P(X2) values for each image.

The PDF to calculate **P(Xi)= 1 e-1/2[(x-µ)/σ]^2**

**σ√2π**

We calculate the value of P(Y/X) value for both classes for each image and compare them.

If P(Y=0/X)>P(Y=1/X) then we classify it as class 0 else as class 1.

We follow this process to classify each image to either class 0 or class 1.

After applying the mentioned steps in my algorithm, the accuracy are as follows.

**Accuracy on train data: 82.38%**

**Accuracy on test data: 83.15%**

The confusion matrix on test data: **array([[784, 216],[121, 879]])**

For class 0: 784 images are accurately classified and 216 are predicted incorrectly.

For class 1: 879 images are accurately classified and 121 are predicted incorrectly.

**Logistic Regression :**

In Logistic regression we first initialize the weight vector.

We follow the following steps iteratively for 10000 times.

Step1: Calculating the sigmoid function.

Sigmoid Function: **σ(t)=1/1+e-t**

We calculate σ(wtX) to predict the label.

**Z= σ(wtX)**

Step2: Calculating Loglikelihood

**L(w)= log zi + (1-)log(1-zi)**

Step 3 : calculating gradient ascent

Updated wt=wtj+η.(Y-Z).Xj

Where η is the learning rate.

In our calculation we took η=3.5

Initially the bias value will be 0.

We use the above-mentioned steps on training data to calculate the weight vector and bias values.

Later we calculate W=w1x1+w2x2+bias on test class

If W≥0.5 then class=1 else if W<0.5 the class=0

In our algorithm we got the following values:

**Weights= [-47.89848536 , 48.96969578]**

**bias= -3.2268922026246405**

**learning rate η=3.5**

**No. of iterations=10000**

**Accuracy achieved using logistic regression: 92.25%**

**Conclusion:**

According to the predicted accuracy discriminative model logistic regression performed better than the generative model naïve bayes classifier. Naïve bayes considers the features to be independent which impacts the prediction accuracy.

­­­­­­­­­­­­­­­­­­