**Name : Sushmitha Nagarajan**

**ID : 1001556348**

**Assignment2 - Data Mining**

**Questions 1:**

**Batch Training Descent for Logistic Regression :**

1. Initialize all weights in 1 to some random small numbers.

2. Calculate the bias values or label them as 1, 0

3. calculate and stacks arrays horizontally.

3. Set the max number of iterations and iterate them

4. calculate score value based on dot function of attributes an weights.

5. calculate the sigmoid value of scores.

6. Calculate the entropy cost and it reduces over each iteration.

7. Based on the cost function of cross entropy, calculate the gradient descent. with learning rate.

8. Update the weights again over the iterations.

**Learning Rate : 1**

Iterations : Iteration numbers= 1 | cross\_entropy\_errors= 1.5388491863352955

Edge Weights Earned : [-0.79905387 -1.37562404 -2.7160066 ]

**Learning Rate : 0.1**

Iterations : Iteration numbers= 1 |cross\_entropy\_errors= 1.5388491863352955

Edge Weights Earned : [0.82009461 0.7624376 0.62839934]

**Learning Rate : 0.01**

Iterations : Iteration numbers= 1 |cross\_entropy\_errors= 1.5388491863352955

**Edge Weights Earned :** [0.98200946 0.97624376 0.96283993]

Accuracy of training data in comparison with Sklearn model

Accuracy from scratch: 0.595

Accuracy from sk-learn: 0.9615

Accuracy of testing data in comparison with Sklearn model

Accuracy from scratch: 0.592

Accuracy from sk-learn: 0.965

2. **Online Training**

**Learning rate: 0.01**

Edge Weights earned: [ 0.58317099 0.29033243 -0.31410469]

**Learning rate : 0.1**

Edge Weights earned: [ 0.58317099 0.29033243 -0.31410469]

**Learning rate : 1**

Edge Weights earned:[nan,nan,nan]

**No of iterations training set went through :** Evaluvation for iterations

Iteration numbers= 3680 cross\_entropy\_errors= 0.5570595758098248

**Accuracy :**

Accuracy comparison between Sklearn model and Online training data

Accuracy in Online training data: 0.6425

Accuracy in sk-learn: 0.963

Accuracy comparison between Sklearn model and Online Testing data

Accuracy in Online Testing data: 0.656

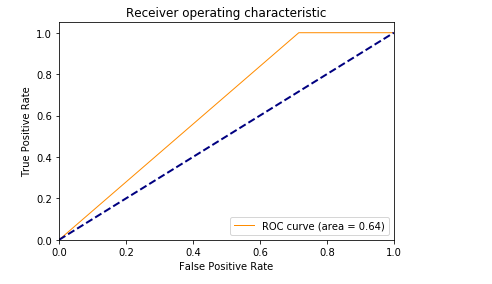
Accuracy in sk-learn: 0.962

Accuracy in Online Training is more , they are not same.

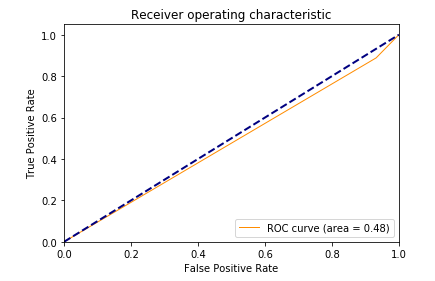
These are more accurate than the batch training model as it calculates and makes the weight efficiently over all the data.

3.

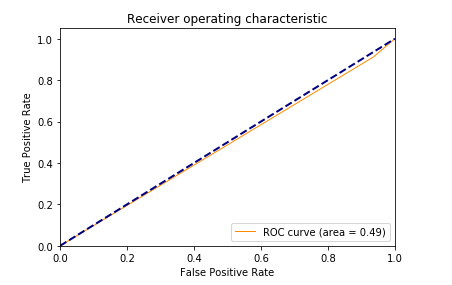
**ROC curve for Online training :**



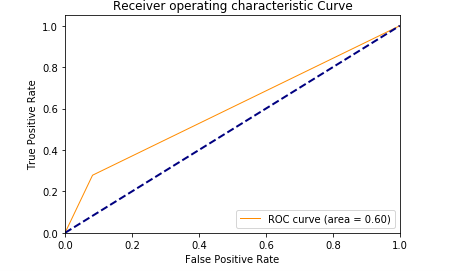
ROC curve for Batch training learning rate :0.01



ROC curve for Batch training learning rate :0.1



ROC curve for Batch training learning rate :1



**How to Run:**

* Type – “jupyter notebook” in the anaconda terminal.
* Open Jupyter Notebook in the browser tab and run the program cell by cell, until all the cells are executed and the final output has been achieved.
* Python version – 3.7.