CSE512 Fall 2018 - Machine Learning - Homework 3

Your Name: Astitv Nagpal

Solar ID: 112008011

NetID email address: anagpal@cs.stonybrook.edu

Names of people whom you discussed the homework with: Ayush Garg

(1.1) We are given that

$X_1 \rightarrow$ follows boolean variable

$X_2 \rightarrow$ follows continuous variables.

We know that Bayes theorem states that:

$$P(Y|X) = \frac{P(Y) \cdot P(X|Y)}{P(X)}$$

where $P(Y) \rightarrow$ Prior
$P(X|Y) \rightarrow$ Likelihood
$P(X) \rightarrow$ total Prob.

$\therefore$ We can write this as:

$$P(Y|X) = \frac{P(Y) \cdot \left(P(X_1|Y) \cdot P(X_2|Y)\right)}{P(X)} \quad \text{——— (I)}$$

Here we have $P(X)$ as the total probability which is a normalizing factor. Hence it can be ignored as $P(Y|X)$ depends only on the numerator.

$\therefore$ We have

$$\left.\begin{aligned} P(X_1|Y) &= \theta_i^{x_i}(1-\theta_i)^{1-x_i} \\ P(X_2|Y) &= \frac{1}{\sqrt{2\pi}\,\sigma_i}\exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right) \end{aligned}\right\} \text{——(i)}$$

Substituting eq^n (i) in (I) we get.

$$\boxed{P(Y_k | X_i) = \underset{Y}{\text{argmax}}\ P((Y)=k) \cdot \theta_k^{x_i} (1-\theta_k)^{1-x_i} \cdot \frac{\exp \frac{-(x-\mu_k)^2}{2\sigma_k^2}}{\sqrt{2\pi}\,\sigma_k}}$$

$\therefore$ The number of feature values for X_1 & X_2 are:

(a) $Y=0 \rightarrow \theta_0$

(b) $Y=1 \rightarrow \theta_1$

(c) $Y=0 \rightarrow \mu_0$

(d) $Y=1 \rightarrow \mu_1$

(e) $Y=0 \rightarrow \sigma_0$

(f) $Y=1 \rightarrow \sigma_1$

(g) $P(Y) \rightarrow$ Prior

Hence, we need to calculate 7 different parameters for the equation.

(1.2)

We are given that xi follows Boolean variables form & hence we can say it follows Bernoulli form dist.

$\therefore$ We have

$$\left.\begin{aligned} P(x_i=1 \mid Y=1) &= \theta_{i1} \\ P(x_i=0 \mid Y=1) &= (1-\theta_{i1}) \\ P(x_i=1 \mid Y=0) &= \theta_{i0} \\ P(x_i=0 \mid Y=0) &= (1-\theta_{i0}) \end{aligned}\right\} \quad (I)$$

$\therefore$ $P(x_i \mid Y=1)$ can be written as

$$P(x_i \mid Y=1) = \theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}$$

Using Bayes theorem we can say that :

$$P(Y=1 \mid x) = \frac{P(Y=1) \cdot P(x \mid Y=1)}{P(Y=1) \cdot P(x \mid Y=1) + P(Y=0) \cdot P(x \mid Y=0)}$$

$$= \frac{\text{Prior} \cdot \text{likelihood}}{\text{Total Probability}}$$

We know,

$\therefore$ we get by dividing the numerator.

$$P(Y=1 \mid x) = \frac{1}{1+\exp\left(\ln\left(\frac{P(Y=0)\,P(x \mid Y=0)}{P(Y=1)\,P(x \mid Y=1)}\right)\right)} \quad \text{—— (i)}$$

$$= \frac{1}{1+\exp\left(\ln \frac{P(Y=0)}{P(Y=1)} + \sum_{\text{all } i} \ln \frac{P(x_i \mid Y=0)}{P(x_i \mid Y=1)}\right)}$$

Scanned by CamScanner

We have that :

$$\left.\begin{aligned} P(Y=0) &= 1-\pi \\ P(Y=1) &= \pi \end{aligned}\right\} \quad II$$

Substituting eqn (I) & (II) in (i) we get ;

$$P(Y=1|X) = \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \sum_{\text{all } i} \ln \frac{\theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i}}{\theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}}\right)}$$

$$= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \sum_{\text{all } i} x_i \ln\frac{\theta_{i0}}{\theta_{i1}} + (1-x_i)\ln\left(\frac{1-\theta_{i0}}{1-\theta_{i1}}\right)\right)}$$

$$= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \frac{(1-\theta_{i0})}{(1-\theta_{i1})} + \sum_{\text{all } i} x_i\left(\ln\frac{\theta_{i0}}{\theta_{i1}} - \ln\left(\frac{1-\theta_{i0}}{1-\theta_{i1}}\right)\right)\right)}$$

⇒ Here we can observe that if we set w_0 & w_1 & substitute

$$\left.\begin{aligned} w_0 &= \ln\frac{1-\pi}{\pi} + \sum_{\text{all } i} \ln\left(\frac{1-\theta_{i0}}{1-\theta_{i1}}\right) \\ w_1 &= \ln\frac{\theta_{i0}}{\theta_{i1}} - \ln\frac{(1-\theta_{i0})}{(1-\theta_{i1})} \end{aligned}\right\}$$

Substituting these values in the above equation we get :

$$P(Y=1|X) = \frac{1}{1+\exp\left(w_0 + \sum_{\text{all } i} w_i x_i\right)} \quad \text{—(ii)}$$

We can observe that eq^n (ii) is the same as that we have for Linear Regression.

Hence, we can say that. Linear Regression is also the discriminative counterpart to a Naive Bayes generative classifier over Boolean features.

(2.1) We know that $L(\theta) = -\frac{1}{n}\sum_{i=1}^{n} Y_i \log P(Y_i|x_i,\theta) + (1-Y_i)\log(1-P(Y_i|x_i,\theta))$ $\longrightarrow$ (I)

$$P(Y_i|x_i,\theta) = P(Y_i=1|x_i,\theta) = \frac{1}{1+\exp\left(-\sum_{\text{all } k} \theta_k x_i^k\right)}$$

This follows sigmoid function.

$$\log(P(Y_i|x_i,\theta)) = \log(P(Y_i=1|x_i,\theta)) = \log\left(\frac{1}{1+\exp\left(-\sum_{\text{all } k}\theta_k x_i^k\right)}\right)$$

$$\log(P(Y_i|x_i,\theta)) = -\log\left(1+\exp\left(-\sum_{\text{all } k}\theta_k x_i^k\right)\right)$$

$$\frac{\partial(\log(P(Y_i=1|x_i,\theta)))}{\partial\theta} = \frac{x_i^k \exp\left(-\sum_{\text{all } k}\theta_k x_i\right)}{1+\exp\left(-\sum_{\text{all } k}\theta_k x_i^k\right)} = x_i\cdot(1-P(Y_i|x_i,\theta))$$ $\longrightarrow$ (i)

And we also have,

$$\log(1-P(Y_i|x_i,\theta)) = \log(1-P(Y_i=1|x_i,\theta))$$

$$= -\sum_{\text{all } k} \theta_k x_i^k - \log\left(1+\exp\left(-\sum_{\text{all } k}\theta_k x_i^k\right)\right)$$

We now take derivative

$$= \frac{\partial(\log(1-P(Y_i|x_i,\theta)))}{\partial\theta} = -x_i + x_i\left(1-P(Y_i|x_i,\theta)\right)$$

$$= \cancel{-x_i} + \cancel{x_i} - x_i P(Y_i|x_i,\theta)$$

$$= -x_i\left(P(Y_i|x_i,\theta)\right) \longleftarrow \text{(ii)}$$

Taking derivative of eq^n (I) we get:

$$\cancel{\frac{\partial L(\theta)}{\partial\theta} = -\frac{1}{n}\sum_{i=1}^{n} y_i}$$

$$\frac{\partial L(\theta)}{\partial(\theta)} = -\frac{1}{n}\sum_{i=1}^{n} Y_i \frac{\partial}{\partial\theta}(\log P(Y_i|x_i,\theta)) + (1-y_i)\frac{\partial}{\partial\theta}(\log 1-P(Y_i|x_i,\theta))$$

Now substituting the values of eq^n (i) & (ii) we get,

$$= -\frac{1}{n}\sum_{i=1}^{n} Y_i x_i\left(1-P(Y_i|x_i,\theta)\right) + (1-Y_i)\left(-x_i\left(P\, Y_i|x_i,\theta\right)\right)$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\Big(Y_i x_i - Y_i x_i\left(P(\cancel{Y_i}|x_i,\theta)\right) - x_i\left(P(Y_i|x_i,\theta)\right)$$

$$+ x_i Y_i\left(P(\cancel{Y_i}|x_i),\theta\right)\Big)$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \left(Y_i x_i - x_i \left(P Y_i | x_i, \theta \right) \right) \longrightarrow (\text{II})$$

We have.

$$\frac{\partial \log \left(P(Y_i | x_i, \theta) \right)}{\partial \theta} = n \cdot \frac{\partial L(\theta)}{\partial \theta}$$

Substituting this in eq^n (II) we get.

$$\frac{\partial \log \left(P(\cancel{Y_i} | x_i, \theta) \right)}{\partial \theta} = \cancel{n} \cdot \left(-\frac{1}{\cancel{n}} \right) \left(Y_i x_i - x_i \left(P x_i | Y_i, \theta \right) \right)$$

$$\frac{\partial \log \left(P(Y_i | x_i, \theta) \right)}{\partial \theta} = \left(Y_i - P(Y_i | x_i, \theta) \right) x_i$$

Hence Proved.

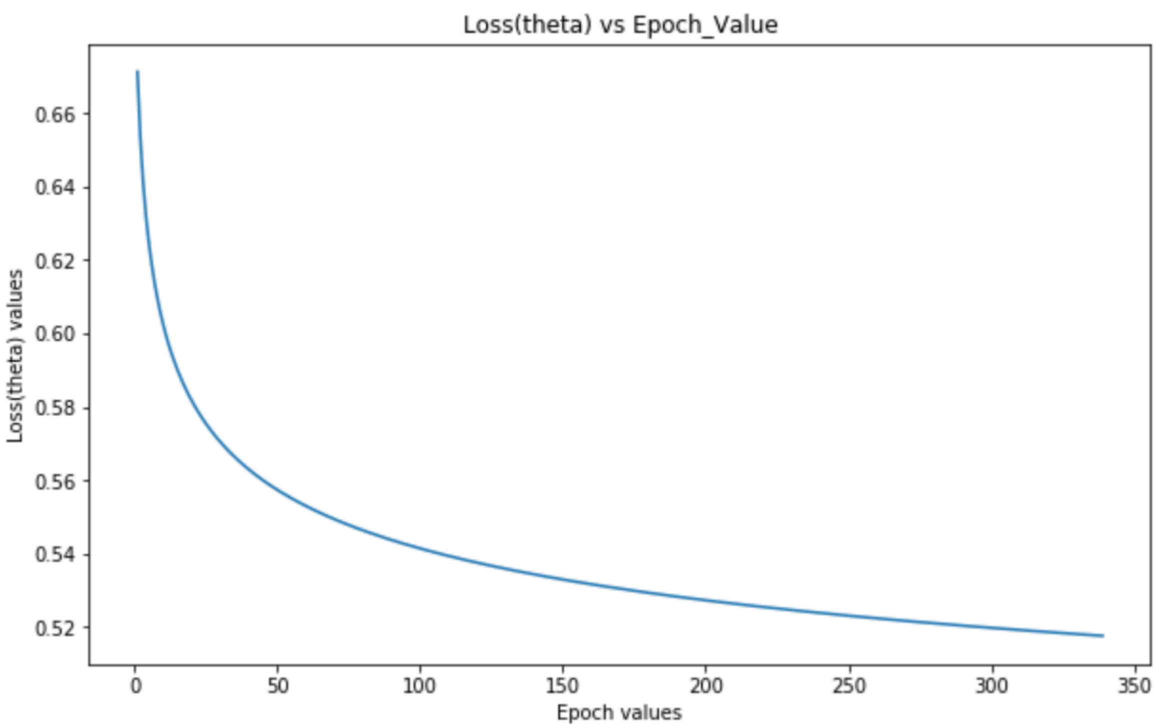
Question 2.3

1)

(a) Report the number of epochs that your algorithm takes before exiting.

339

(b) Plot the curve showing L(θ) as a function of epoch.



(c) What is the final value of L(θ) after the optimization?

0.5176151929899254

2)

(a) Report the values of (η_0, η_1). How many epochs does it take? What is the final value of L(θ)?

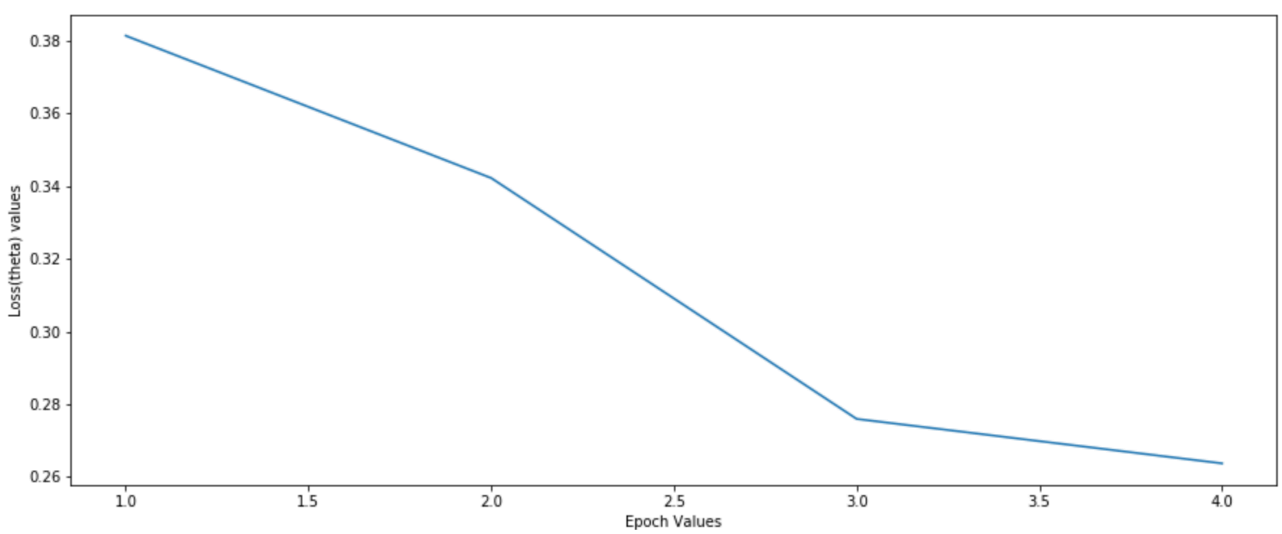
eta_0 = 40

eta_1 = 0.2

number of epochs = 4

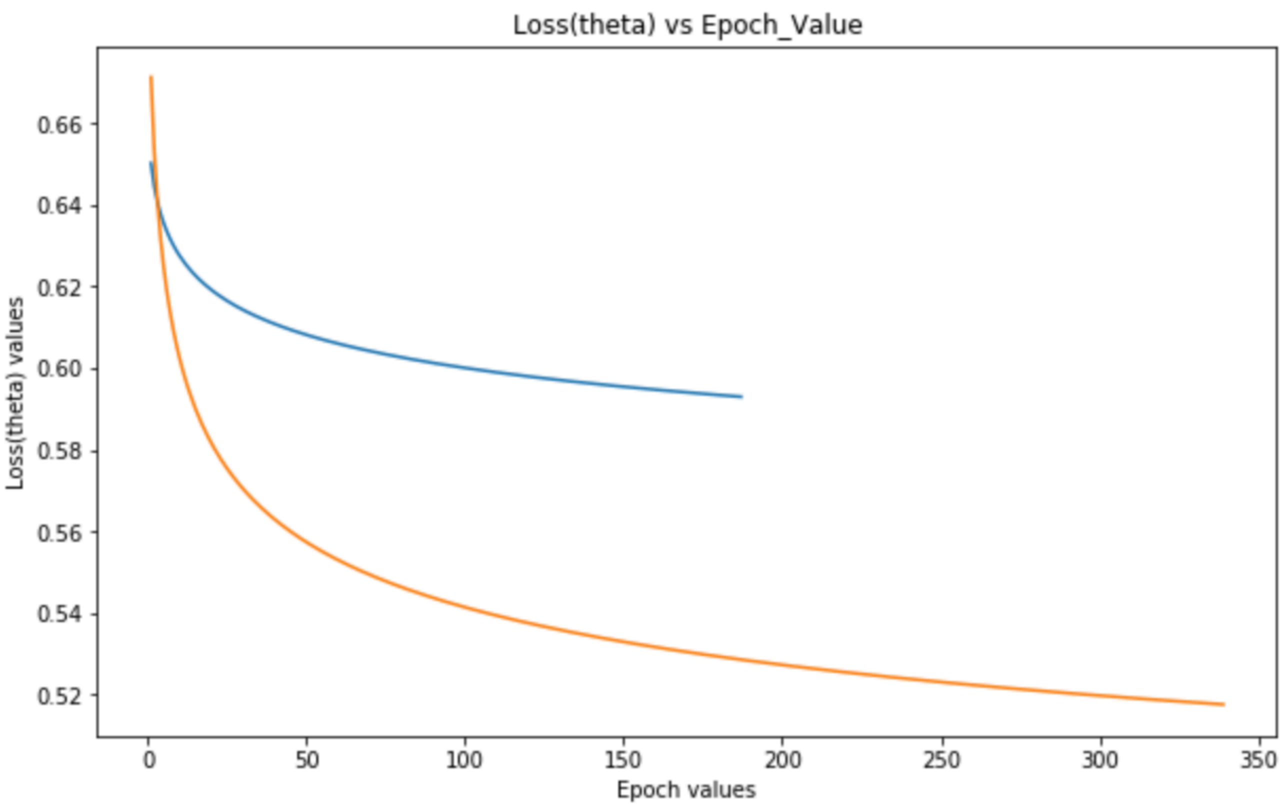
Final value of L(theta) = 0.26155505584268873

(b) Plot the curve showing L(θ) as a function of epoch.



3)

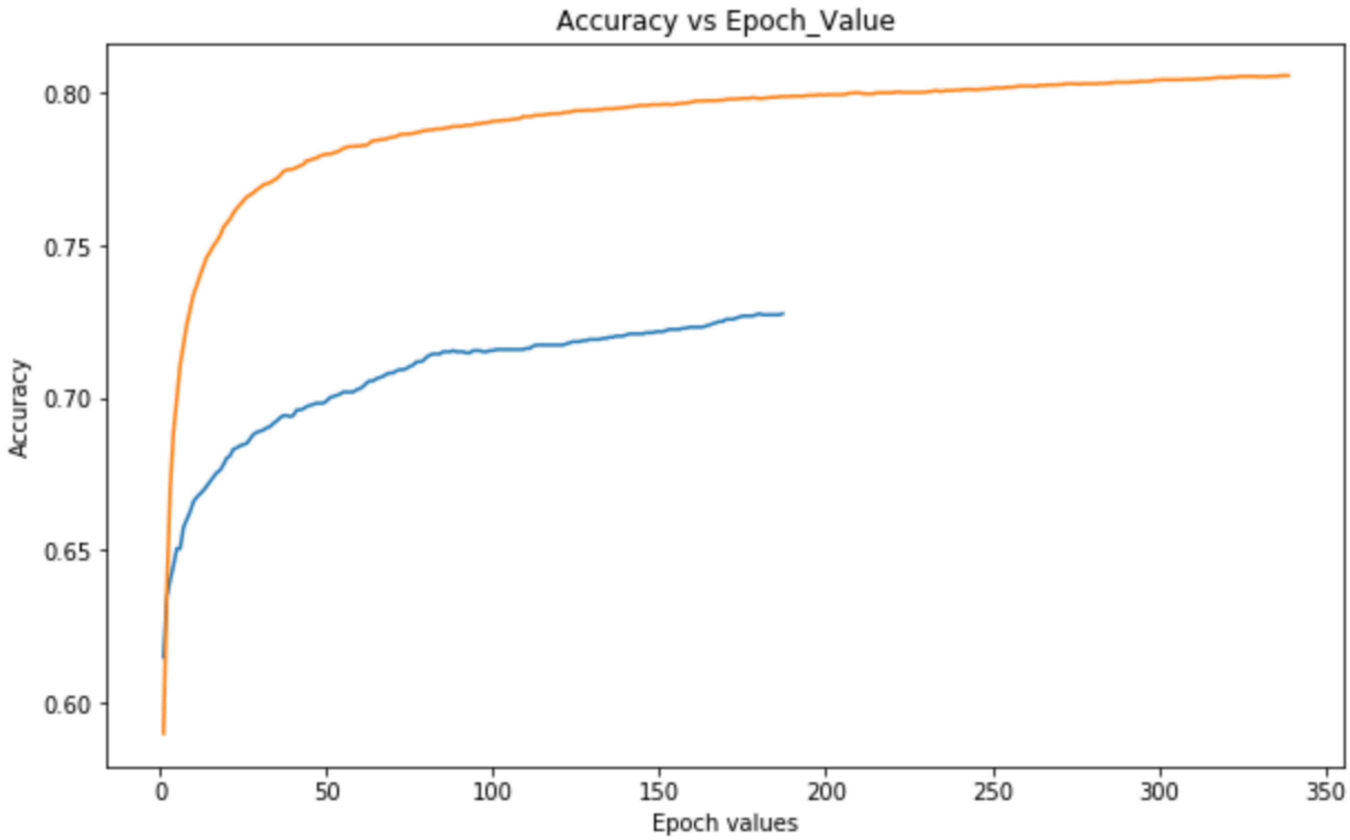
(a) Plot L(θ) as a function of epoch. On the same plot, show two curves, one for training and one for validation data.



Blue Line - Validation Data

Orange Line - Training Data

(b) Plot the accuracy as a function of epoch. On the same plot, show two curves, one for training and one for validation data.

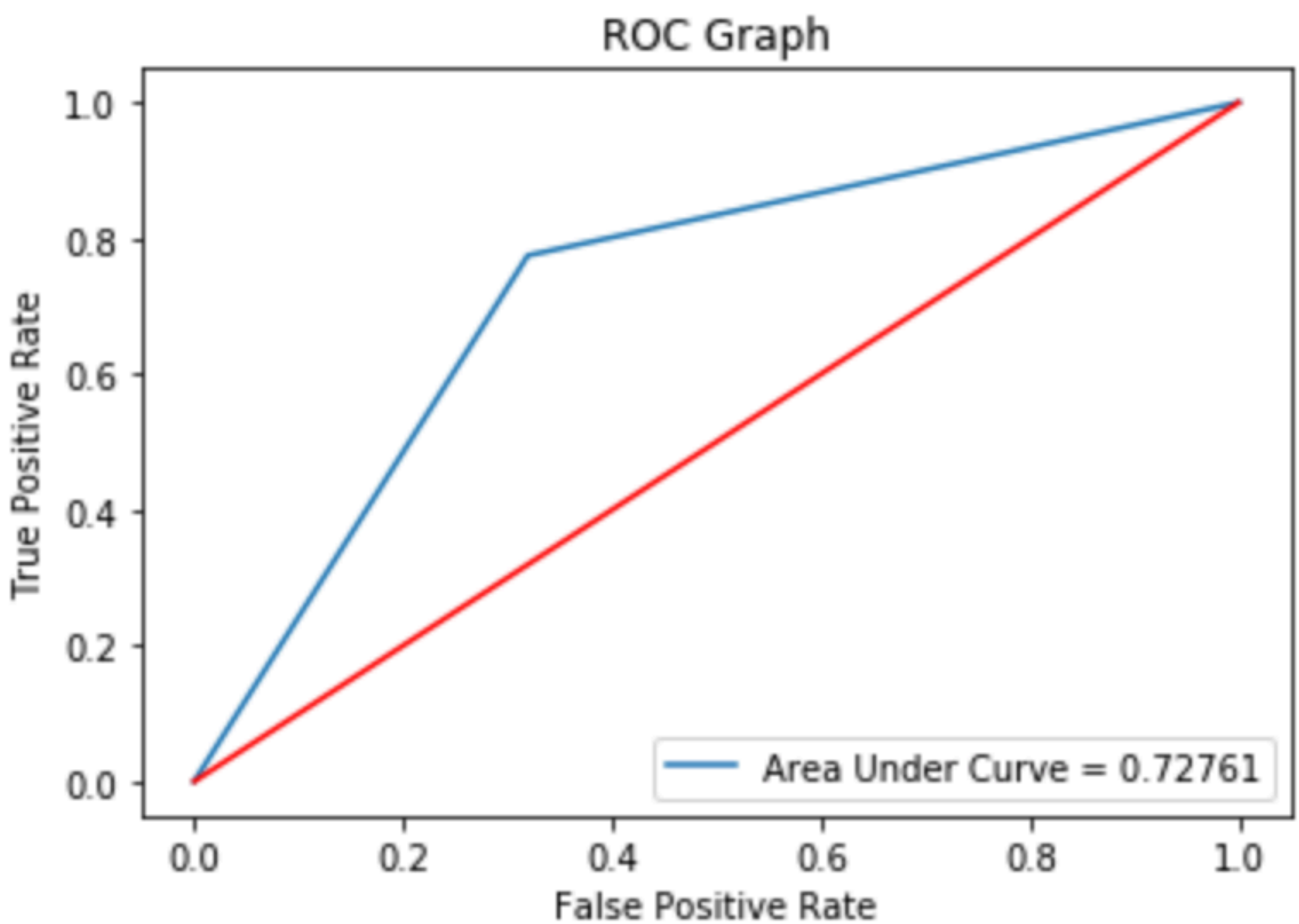


Blue Line - Validation Data

Orange Line - Training Data

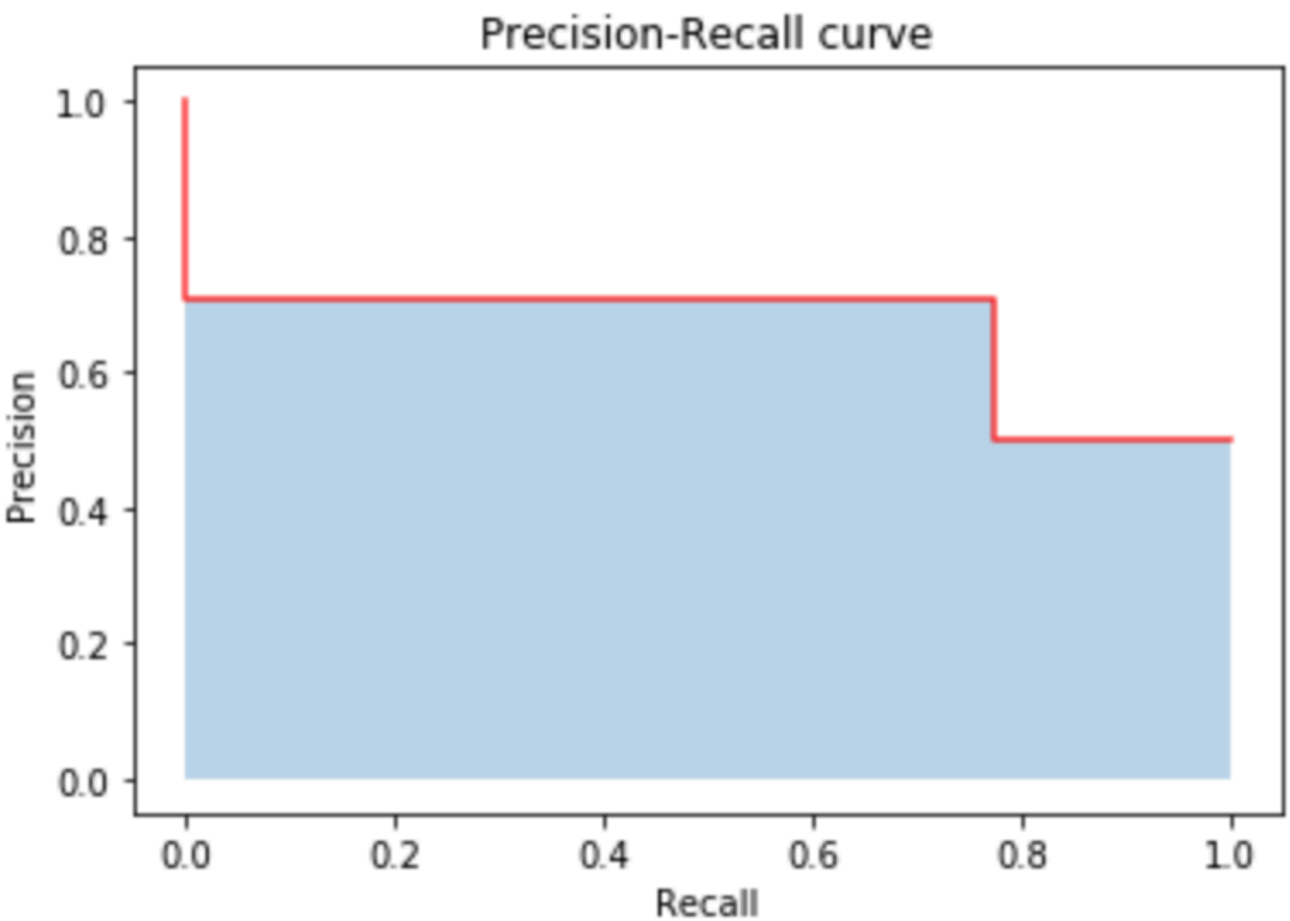
4)

(a) Plot the ROC curve on validation data. Report the area under the curve.



Area under curve = 0.7276064610866374

(b) **Plot the Precision-Recall curve on validation data. Report the average precision.**



Average Precision = 0.6611575949304713

Question 2.4

Accuracy from leader board - **0.89029**