Cover page for answers.pdf
CSE512 Fall 2018 - Machine Learning - Homework 3

Your Name: SAIF SULEMAN VAZIR

Solar ID: 112072061

NetID email address: saifsuleman.vazir@stonybrook.edu

Names of people whom you discussed the homework with: -

HOMEWORK 3

1) 1.1 X_1 = boolean variable
Y = boolean variable
X_2 = continuous variable

As X_1 is a boolean variable, we can formulate it in terms of a bernoulli distribution (for 1 sample) or binomial distribution (>1 samples)
We require only one parameter for X_1 for 'each' class.

For $Y=1$, we have $\theta_{i0}^{x_i}(1-\theta_{i0})^{1-x_i}$ [x_i is sample data point in X_1]

For $Y=0$, we have $\theta_{i1}^{x_i}(1-\theta_{i1})^{1-x_i}$ [x_i is 1 or true when $x_i=1$ else 0 or false]

We need 1 parameter for Y, as we have only 2 classes which can be represented as γ and $1-\gamma$.

For continuous variable X_2, we require a Gaussian distribution with $N \alpha (\mu_{ik}, \sigma_{ik}^2)$.
As we have 2 classes, $k=2$, therefore we require 4 parameters in total.

Total no. of parameters required = $2+1+4=\underline{\underline{7}}$

Now, to find $P(Y|X)$ we use Bayes theorem:

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

Let's say we want to find $P(Y=1|X)$ [class = 1] we have:

$$P(Y=1|X) = \frac{P(X|Y=1)\, P(Y=1)}{\sum_i P(Y=i) \cdot \prod_j P(X_j|Y=i)}$$

(using conditional independence for Naive Bayes)

$$P(Y=1|X) = \frac{P(Y=1) \cdot P(X_1|Y=1) \cdot P(X_2|Y=2)}{P(Y=1) \cdot P(X_1|Y=1) P(X_2|Y=1) + P(Y=0) P(X_1|Y=0) P(X_2|Y=0)}$$

$$= \frac{\gamma \cdot [\delta_1^{x_i}(1-\delta_1)^{x_i}] [\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_i-\mu_{ik})^2}{2\sigma_1^2}}]}{\gamma [\delta_1^{x_i}(1-\delta_1)^{1-x_i}) (\frac{1}{\sqrt{2\pi}\sigma_1^2} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}}) + (1-\gamma)(\delta_0^{x_i}(1-\delta)^{x_i}) \cdot (\frac{1}{\sqrt{2\pi}\sigma_0^2} e^{-\frac{(x_i-\mu_0)^2}{2\sigma_0^2}})}$$

Suppose we have total 'N' data points

~~for each~~ $\therefore |X_1| = |X_2| = N$

we can write the above f^n as:

(using only numerator as denominator becomes independent of $Y=y$ as will not affect ~~our maximization~~ our formula).

$$P(Y=1|X) \propto \gamma \cdot [\delta_1^m (1-\delta_1)^{N-m}] \cdot [\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}}]$$

If we want to maximize our function for any $Y=y_k$:

$$P(Y=y_k|X) \propto \underset{y_k}{\operatorname{argmax}}\ P(Y=y_k)\ P(X_1|Y=y_k) \cdot P(X_2|Y=y_k)$$

Taking log on both sides we get:

$$\log P(Y=y_k|X) = \underset{y_k}{\operatorname{argmax}} \log P(Y=y_k) + \log P(X_1|Y=y_k) + \log P(X_2|Y=y_k)$$

log of prior

$\log P(X_1|Y=y_k)$: MLE for binomial distribution

$\log P(X_2|Y=y_k)$: MLE for Gaussian distribution

1.2) Consider X to be a set of boolean variables, we can say that each X_i has a binomial distribution.

We use the same notations from previous answer:

$P(Y=1) = \gamma \quad P(Y=0) = 1-\gamma$

Now, as $X = \langle X_1, X_2 \ldots X_n \rangle$ all X_i are boolean R.V's themselves, we have:

$P(X_i|Y=1) = \delta_{i1}^{X_i}(1-\delta_{i1})^{1-X_i}$

$P(X_i|Y=0) = \delta_{i0}^{X_i}(1-\delta_{i0})^{1-X_i}$

Now we use the eq^n:

$$P(Y=1|X) = \frac{P(Y=1)\cdot P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$

[dividing by $P(Y=1)P(X|Y=1)$]

$$\therefore P(Y=1|X) = \frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(y=1)P(X|Y=1)}}$$

$$= \frac{1}{1 + e^{\left[\ln\left(\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}\right)\right]}}$$

$$= \frac{1}{1 + e^{\left[\ln\left(\frac{1-\gamma}{\gamma}\cdot\frac{\prod_{i=1} P(X_i|Y=0)}{\prod_{i=1} P(X_i|Y=1)}\right)\right]}}$$

$$= \frac{1}{1 + e^{\left[\ln\frac{1-\gamma}{\gamma} + \ln\prod_{i=1}\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right]}}$$

$$= \frac{1}{1 + e^{\left[\ln\frac{1-\gamma}{\gamma} + \sum_{i=1}\ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right]}}$$

$$= \frac{1}{1 + e^{\left[\ln\frac{1-\gamma}{\gamma} + \sum_{i=1}\ln\left[\frac{\delta_{i0}^{X_i}(1-\delta_{i0})^{1-X_i}}{\delta_{i1}^{X_i}(1-\delta_{i1})^{1-X_i}}\right]\right]}}$$

$$= \frac{1}{1+e^{\left[\ln\frac{1-\gamma}{\gamma} + \sum_{i=1}^{n} \ln\left(\frac{\delta_{i0}}{\delta_{i1}}\right)^{x_i} + \ln\left(\frac{1-\delta_{i0}}{1-\delta_{i1}}\right)^{(1-x_i)}\right]}}$$

$$= \frac{1}{1+e^{\left[\ln\frac{1-\gamma}{\gamma} + \sum_{i=1}^{n} x_i \ln\frac{\delta_{i0}}{\delta_{i1}} + (1-x_i)\ln\frac{1-\delta_{i0}}{1-\delta_{i1}}\right]}}$$

[split last term]

$$= \frac{1}{1+e^{\left[\ln\frac{1-\gamma}{\gamma} + \sum_{i=1}^{n} x_i \ln\frac{\delta_{i0}}{\delta_{i1}} + \ln\frac{1-\delta_{i0}}{1-\delta_{i1}} - x_i \ln\frac{1-\delta_{i0}}{1-\delta_{i1}}\right]}}$$

$$= \frac{1}{1+e^{\left[\ln\frac{1-\gamma}{\gamma} + \sum_{i=1}^{n} \ln\frac{1-\delta_{i0}}{1-\delta_{i1}} + \sum_{i=1}^{n} x_i\left[\ln\frac{\delta_{i0}}{\delta_{i1}} - \ln\frac{1-\delta_{i0}}{1-\delta_{i1}}\right]\right]}}$$

$$= \frac{1}{1+e^{\left[\ln\frac{1-\gamma}{\gamma} + \sum_{i=1}^{n} \ln\frac{1-\delta_{i0}}{1-\delta_{i1}} + \sum_{i=1}^{n} x_i \ln\frac{\delta_{i0}\cdot(1-\delta_{i1})}{\delta_{i1}(1-\delta_{i0})}\right]}}$$

Comparing with the original eq^n:

$$P(Y=1\mid X) = \frac{1}{1+e^{-\left(\sum_{i=1}^{n} \theta_i x_i + \theta_{d+1}\right)}}$$

$\therefore$ we have:

$$P(Y=1\mid X) = \frac{1}{1+e^{-\left[\ln\frac{\gamma}{1-\gamma} + \sum_{i=1}^{n} \ln\frac{1-\delta_{i1}}{1-\delta_{i0}} + \sum_{i=1}^{n} x_i \ln\frac{\delta_{i1}(1-\delta_{i0})}{\delta_{i0}(1-\delta_{i1})}\right]}}$$

[note that (-1) has been multiplied ~~2 times~~ as we have reversed the numerator and denominator in ln]

$$\therefore \theta_{d+1} = \ln\frac{\gamma}{1-\gamma} + \sum_{i=1}^{n} \ln\frac{1-\delta_{i1}}{1-\delta_{i0}}$$

$$\theta_i = \ln\frac{\delta_{i1}(1-\delta_{i0})}{\delta_{i0}(1-\delta_{i1})}$$

$\therefore$ we get:

$$P(Y=1\mid X) = \frac{1}{1+e^{-\left[\sum_{i=1}^{n} \theta_i x_i + \theta_{d+1}\right]}}$$

Scanned by CamScanner

2.1 To prove:

$$\frac{\partial}{\partial \theta} \log P(Y^i \mid \bar{X}^i; \theta) = [Y^i - P(Y=1 \mid \bar{X}^i; \theta)]\bar{X}^i$$

We can write:

$$\log P(Y^i \mid \bar{X}^i, \theta) = \underbrace{Y^i \log P(Y=1 \mid \bar{X}^i, \theta)}_{\text{(when } y=1)} + \underbrace{(1-Y^i)\log P(Y=0 \mid X^i, \theta)}_{\text{(when } y=0)}$$

$\therefore \Rightarrow$ we know that $P(Y=1 \mid \bar{X}^i, \theta) = \dfrac{1}{1+e^{-\theta^T \bar{X}_i}} = \dfrac{e^{\theta^T \bar{X}_i}}{1+e^{\theta^T \bar{X}_i}}$

$$P(Y=0 \mid X^i, \theta) = 1 - P(Y=1 \mid X^i, \theta)$$

$$= 1 - \frac{e^{\theta^T \bar{X}^i}}{1+e^{\theta^T \bar{X}_i}} = \frac{1}{1+e^{\theta^T \bar{X}_i}}$$

$\therefore$ we get:

$$Y^i \log \frac{e^{\theta^T \bar{X}_i}}{1+e^{\theta^T \bar{X}_i}} + (1-Y_i)\log \frac{1}{1+e^{\theta^T \bar{X}_i}}$$

$$Y^i \log e^{\theta^T \bar{X}_i} + \cancel{Y^i \log \frac{1}{1+e^{\theta^T X_i}}} + \log \frac{1}{1+e^{\theta^T \bar{X}_i}} - \cancel{Y_i \log \frac{1}{1+e^{\theta^T \bar{X}_i}}}$$

$$\Rightarrow Y^i \log e^{\theta^T \bar{X}^i} - \log 1+e^{\theta^T \bar{X}^i}$$

$$\Rightarrow Y^i \theta^T \bar{X}^i - \log 1+e^{\theta^T \bar{X}^i}$$

differentiating w.r.t θ we get:

$$\frac{\partial P(Y^i \mid \bar{X}^i, \theta)}{\partial \theta} = Y^i \bar{X}^i + \frac{1}{1+e^{\theta^T X^i}} \cdot e^{\theta^T \bar{X}_i} \cdot \bar{X}^i$$

$$= \left[Y^i + \frac{e^{\theta^T \bar{X}^i}}{1+e^{\theta^T \bar{X}_i}}\right] X^i$$

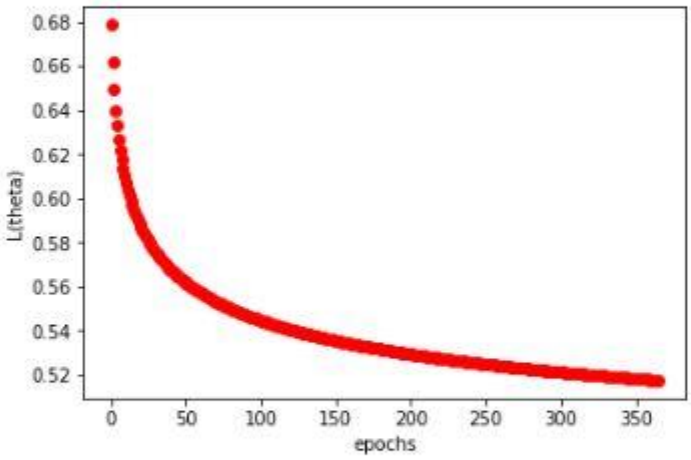
But $P(Y=1 \mid \bar{X}^i, \theta) = \dfrac{e^{\theta^T X^i}}{1+e^{\theta^T \bar{X}_i}}$

$$\therefore \frac{\partial}{\partial \theta} P(Y^i \mid \bar{X}^i, \theta) = [Y^i - P(Y=1 \mid \bar{X}^i, \theta)]\bar{X}^i$$

Scanned by CamScanner

2.3 1) a) Number of epochs till termination: 365

b) Plot:



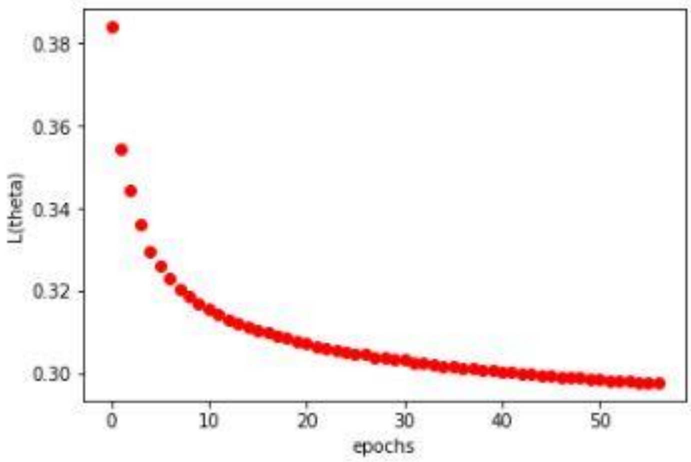
c) Final value of L(Θ) = 0.5175504365325073

2) a) Best value of η_0, η_1 = (2.5,0.1)

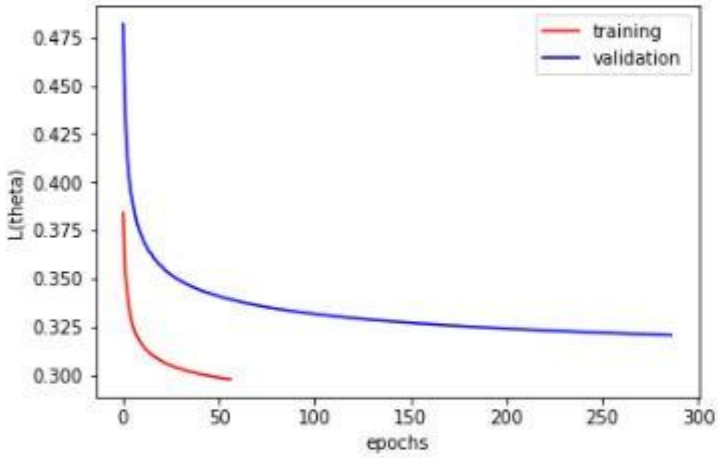
Number of epochs= 69

Final value of L(Θ)= 0.29680946550912407

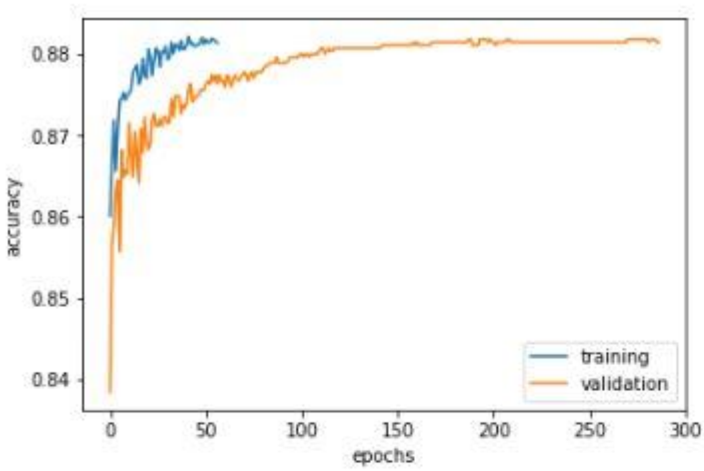
b) Plot:



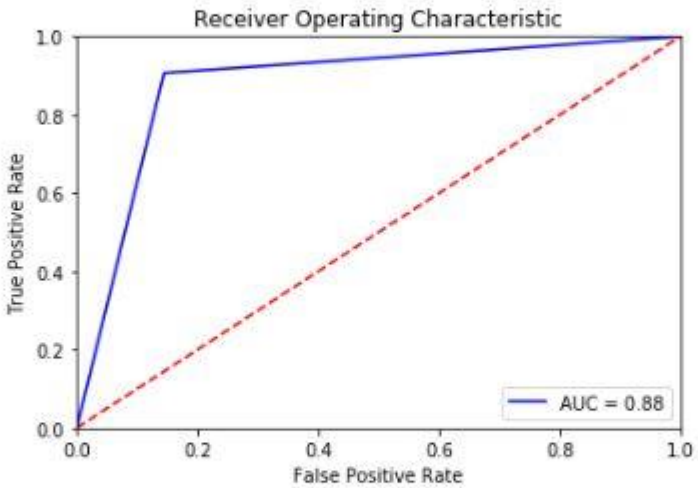
3) a) Plot for validation+training:



b) Plot for accuracy of model on training and validation:

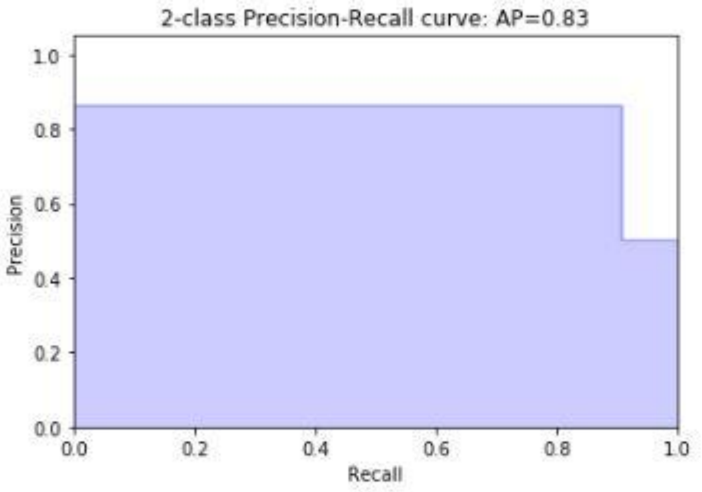


4) a) Plot for ROC curve on validation data:



Area under the curve= 0.88

b) Plot for precision-recall curve on validation data:



Average precision= 0.83

2.4 1) Best accuracy for Kaggle submission = 87.333%