CSE512 Fall 2019 - Machine Learning - Homework 3

October 7, 2019

Name: Anmol Shukla

Solar ID: 112551470

NetID: anmshukla

Email address: Anmol.shukla@stonybrook.edu

Names of people whom I have discussed the homework with: None

1 Question 1

1.1.1

1.1.1. Bayes Risk is given by : $R(a_i|x) = \sum_j L(a_i|c_j)\, p(c_j|x)$

where L is the cost of taking action a_i given a class c_j

$$\therefore R(y=1|x) = \underbrace{L(y=1|\hat{y}=1)\, P(\hat{y}=1|x)}_{\text{this term is 0 as there is no risk/loss when prediction is correct}} + L(y=1|\hat{y}=0)\, P(\hat{y}=0|x)$$

$$= \underbrace{L(y=1|\hat{y}=0)}_{\text{cost of false positive}}\, P(\hat{y}=0|x)$$

↳ prior of ~~positive~~ negative

$$= \alpha \cdot (1-\eta(x))$$

$$R(y=0|x) = L(y=0|\hat{y}=1)\, P(\hat{y}=1|x) + \underbrace{L(y=0|\hat{y}=0)P(\hat{y}=0|x)}_{\text{0 as this is a right prediction}}$$

$$= L(y=0|\hat{y}=1) \cdot P(\hat{y}=1|x)$$

$$= 1 \cdot \eta(x) \Rightarrow \eta(x)$$

Optimal bayes classifier minimizes the risk

$$\therefore \gamma^*(x) = \min(\eta(x), \alpha(1-\eta(x)))$$

1.1.2

1.1.2. Following from 1.1.1

total risk for a data point x whose nearest neighbor is "z" $= r(x)$

$r(x) =$ error we make when we make a wrong decision

a wrong decision is made when label $(x) \neq$ label (z)

$$r(x) = \underbrace{L(y=0 \mid \hat{y}=1)}_{1} P(\hat{y}=1 \mid x) P(\hat{y}=0 \mid z) + \underbrace{L(y=1 \mid \hat{y}=1)}_{\alpha} P(\hat{y}=0 \mid x) P(\hat{y}=1 \mid z)$$

$$r(x) = \eta(x)(1-\eta(z)) + \alpha (1-\eta(x))\eta(z)$$

but as $n \to \infty$ (no. of training data)

$x \to z$ or $z \to x$

$$\therefore r(x) = \eta(x)(1-\eta(x)) + \alpha\, \eta(x)(1-\eta(x))$$

$$r(x) = (1+\alpha)\,\eta(x)(1-\eta(x))$$

1.1.3

1.1.3. To prove $\gamma(x) \leq (1+\alpha)\gamma^*(x)(1-\gamma^*(x))$

from previous question, we know that

$$\gamma(x) = (1+\alpha)\eta(x)(1-\eta(x))$$

&,

$$\gamma^*(x) = \min\big(\eta(x), \alpha(1-\eta(x))\big)$$

$$\therefore \gamma^*(x) = \eta(x) \quad \text{if } \eta(x) < \alpha(1-\eta(x)) \quad -(1)$$

$$= \alpha(1-\eta(x)) \quad \eta(x) > \alpha(1-\eta(x)) \quad -(2)$$

We can prove the lower bound on $\gamma^*(x)$ when $\gamma^*(x) = \eta(x)$

from eq (1),

$$\gamma^*(x) = \eta(x)$$

$$\therefore \gamma(x) = (1+\alpha)\gamma^*(x)(1-\gamma^*(x))$$

Now, we need to prove : $\gamma(x) < (1+\alpha)\gamma^*(x)(1-\gamma^*(x))$

$$\gamma(x) < (1+\alpha)\alpha(1-\eta(x))(1-\alpha(1-\eta(x)))$$

subs. the value of $\gamma(x)$

$$\cancel{(1+\alpha)}\eta(x)\cancel{(1-\eta(x))} < \cancel{(1+\alpha)}\alpha\cancel{(1-\eta(x))}(1-\alpha(1-\eta(x)))$$

To prove: $\eta(x) < \alpha(1-\alpha(1-\eta(x)))$ $-(1)$

Since $\gamma^*(x) = \alpha(1-\eta(x))$ in the above case

it follows that $\eta(x) > \alpha(1-\eta(x))$

~~Also~~ $\cancel{\alpha > 1}$ $\therefore -\eta(x) < -\alpha(1-\eta(x))$

$\therefore \cancel{\alpha(1-\eta(x))}$

$$1-\eta(x) < 1-\alpha(1-\eta(x))$$

$$\alpha > 1$$

$$\therefore 1-\eta(x) < \alpha(1-\alpha(1-\eta(x)))$$

$$\therefore \eta(x) < \alpha(1-\alpha(1-\eta(x)))$$

$\therefore$ We have proved eq (1)

1.1.4

1.1.4. from 1.1.3 we know that $r(x) \le (1+\alpha) r^*(x)(1-r^*(x))$

taking expectation on both the sides

$$E[r(x)] \le (1+\alpha) E[r^*(x)(1-r^*(x))]$$

$$\le (1+\alpha)\left(E[r^*(x)] - E[r^*(x)^2]\right) \quad — \quad (1)$$

Now, we know that $E[x^2] = E[x]^2 + Var[x]$

$$\therefore \; E[r^*(x)^2] = E[r^*(x)]^2 + Var[r^*(x)]$$

$$E[r^*(x)^2] \ge E[r^*(x)]^2$$

$$\therefore \; R \le (1+\alpha) E[r^*(x)] - (1+\alpha) E[r^*(x)]^2$$ - from eq (1)

$$R \le (1+\alpha) R^* - (1+\alpha) R^{*2} \quad [\text{as } E[r^*(x)] = R^*]$$

$$R \le (1+\alpha) R^*(1-R^*)$$

Hence Proved.

1.2.1

1.2.1. probability of a point being positive = η

probability that at least $(K+1)/2$ out of K points are +ve.

$r(x)$ = asymptotic risk $r(x)$ for a point x in terms of $\eta(x)$ & function $g(\cdot,\cdot)$

$r(x)$ = Prob (x is positive but all $\frac{K+1}{2}$ points are not positive)

+ Prob (x is negative but all $\frac{K+1}{2}$ points are +ve)

$$= \eta(x)(1-g(\eta,K)) + (1-\eta(x))(g(\eta,K))$$

$$= \eta(x) - \eta(x)g(\eta,K) + g(\eta,K) - \eta(x)g(\eta,K)$$

$$= \eta(x) + g(\eta,K) - 2\eta(x)g(\eta,K)$$

$$= \eta(x) + (1-2\eta(x))g(\eta,K)$$

$$= \eta(x) + (1-2\eta(x))\,g(\eta,K)$$

2.1

2.1. To Prove: $\frac{\partial (P(y^i|\bar{x}^i;\theta))}{\partial\theta_c} = (\delta(c=y^i) - P(c|\bar{x}^i;\theta))\bar{x}^i$

we can write

$\log(P(y^j|x^j;\theta)) = \delta(c=y^j)P(y^j = \cancel{k}|x^j;\theta) + (1-\delta(c=y^j))P(y^j=k|x^j;\theta)$

for simplicity, we'll write $\delta(c=y^i)$ as δ & $\sum_j$ as a matrix multiplication

also,

$$P(y=i|x;\theta) = \frac{\exp(\theta_i^T\bar{x})}{1+\sum_{j=1}\exp(\theta_j^T x)}$$

$$P(y=K|x;\theta) = \frac{1}{1+\sum_{i=1}^{K-1}\exp(\theta_j^T\bar{x})}$$

$$\log(P(y|x;\theta)) = \delta \log P(y\neq K|x;\theta) + (1-\delta)P(y=K|x;\theta)$$

$$= \delta\log\left(\frac{\exp\theta^T\bar{x}}{1+\exp\theta^T\bar{x}}\right) + (1-\delta)\log\left(\frac{1}{1+\exp\theta^T\bar{x}}\right)$$

$$\log P(y|\bar{x};\theta) = \delta\theta^T\bar{x} - \delta\log(1+\exp\theta^T\bar{x}) + (\delta-1)\log(1+\exp\theta^T\bar{x})$$

$$\frac{\partial \log P(y|\bar{x};\theta)}{\partial\theta} = \delta\bar{x} - \cancel{\frac{\delta.\bar{x}\exp\theta^T\bar{x}}{1+\exp\theta^T\bar{x}}} + \cancel{\frac{\delta\bar{x}.\exp\theta^T\bar{x}}{1+\exp\theta^T\bar{x}}} - \frac{\bar{x}.\exp\theta^T\bar{x}}{1+\exp\theta^T\bar{x}}$$

$$= \delta\bar{x} - \bar{x}\frac{\exp\theta^T\bar{x}}{1+\exp\theta^T\bar{x}}$$

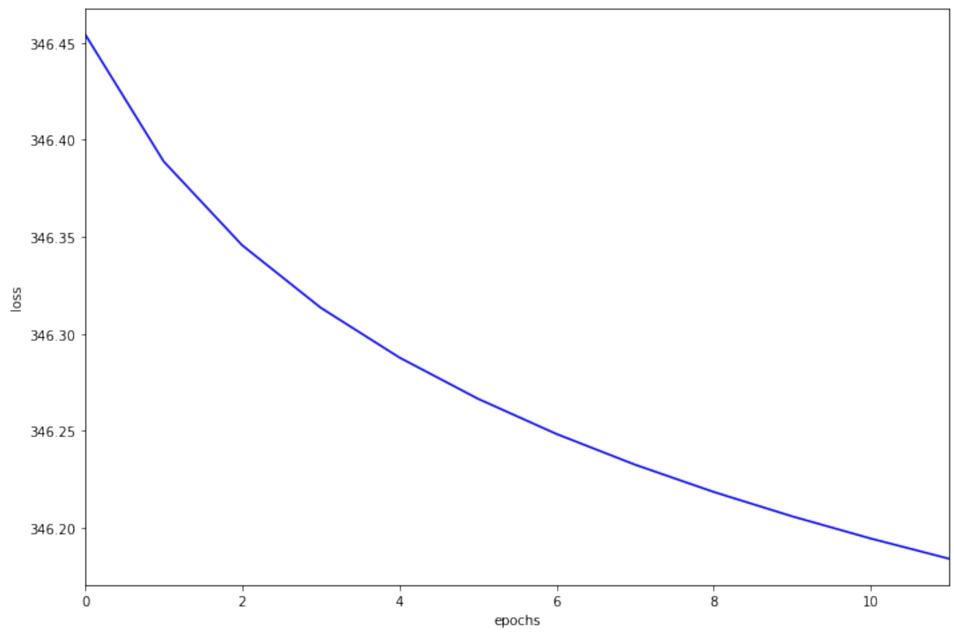
but $Pr(y=c|x;\theta) = \frac{\exp\theta^T\bar{x}}{1+\exp\theta^T\bar{x}}$

$$\therefore \frac{\partial\log P(y|\bar{x};\theta)}{\partial\theta} = \delta\bar{x} - \bar{x}P(c|\bar{x}^i;\theta)$$

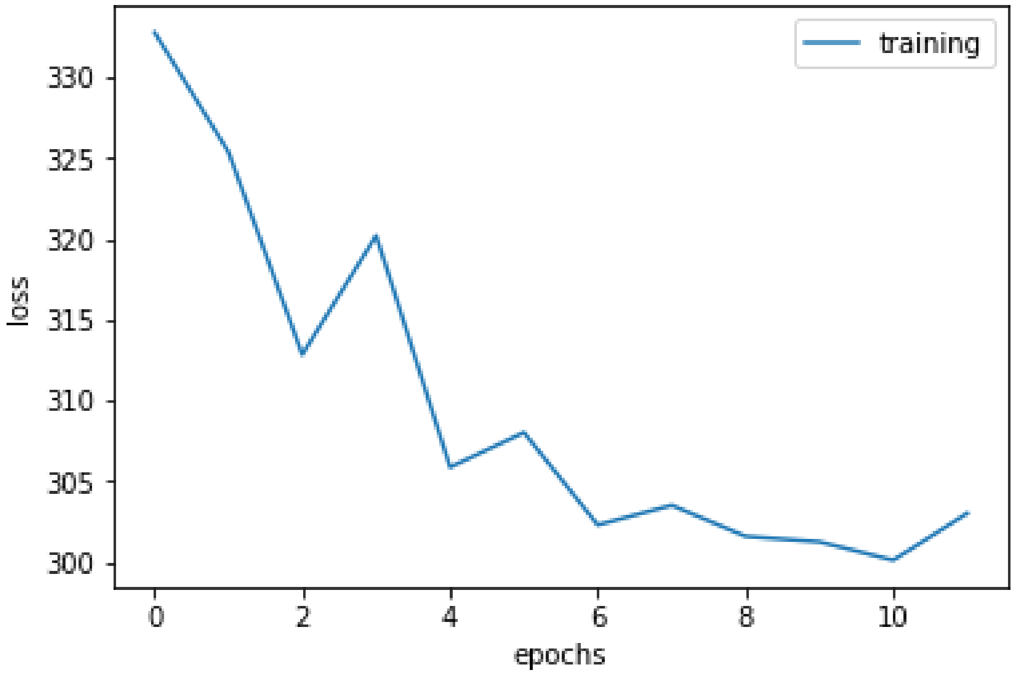
$$= (\delta(c=y) - P(c|\bar{x};\theta))\bar{x}$$

2 Question 2

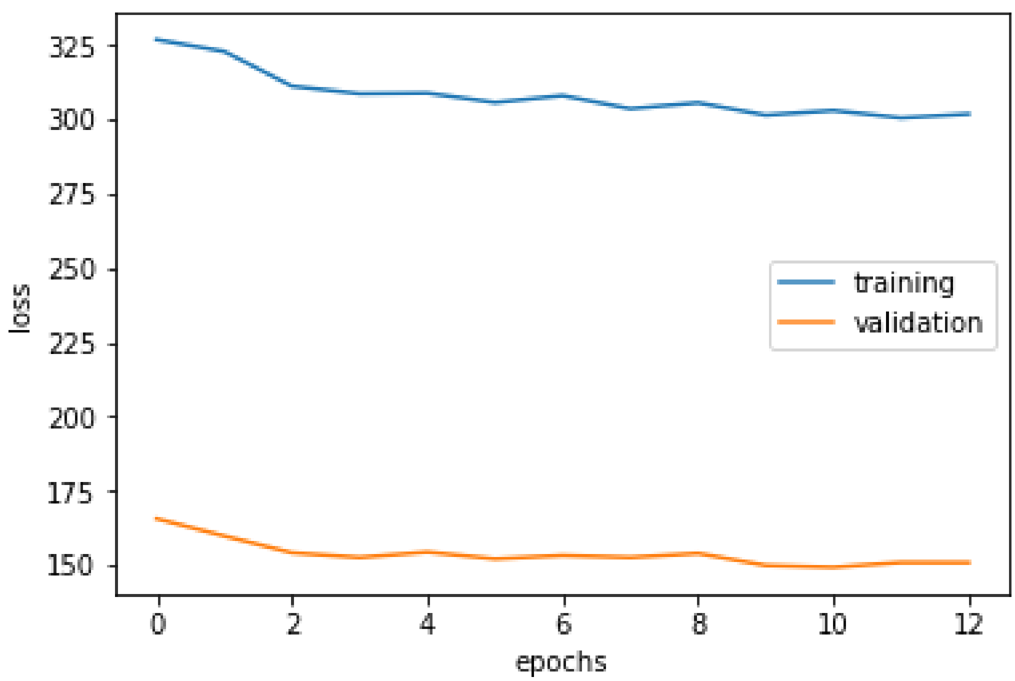
2.3 Implement Logistic Regression with SGD

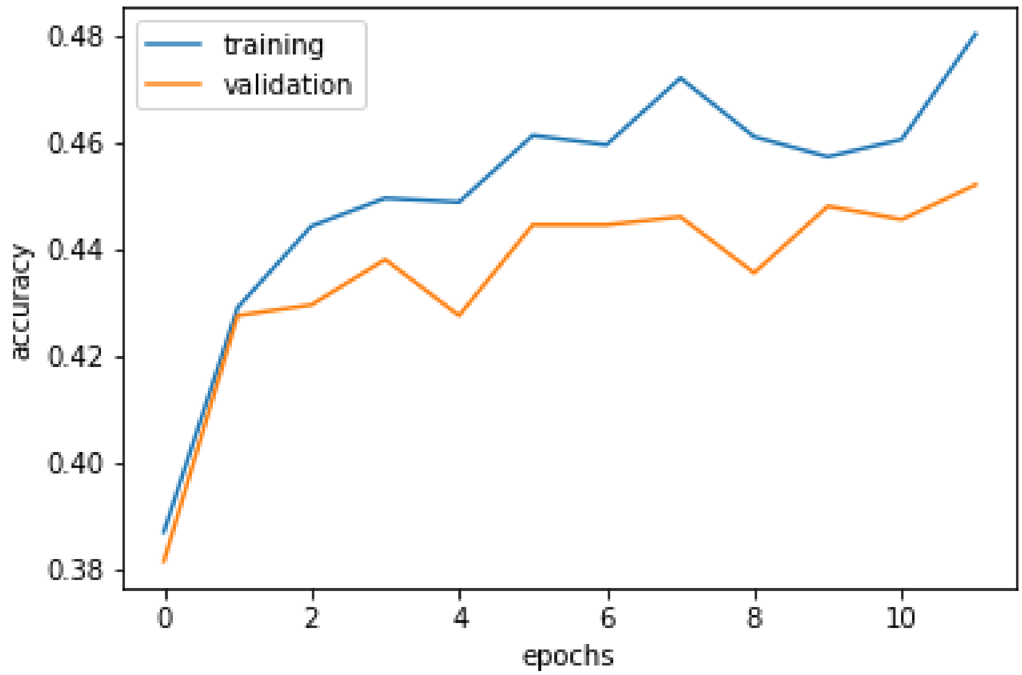
1. max_epoch = 1000, m = 16, $\eta_0 = 0.1$, $\eta_1 = 1$, $\delta = 0.00001$
   (a) Number of epochs taken before existing = 12
   (b) 

   (c) $L(\theta) = 346.14301852200146$
2. Experimenting with values of η_0, η_1
   (a) $\eta_0 = 10$, $\eta_1 = 5$, $epochs = 12$, $L(\theta) = 303.0168711323694$
   (b) The below graph shows the variation of loss vs epochs. Please note that due to high variance of loss in the few initial epochs, I have kept the minimum epochs required to check for the stopping condition is 10 epochs. Therefore, the algorithm doesn't exit at epoch = 2 as shown in the below graph.



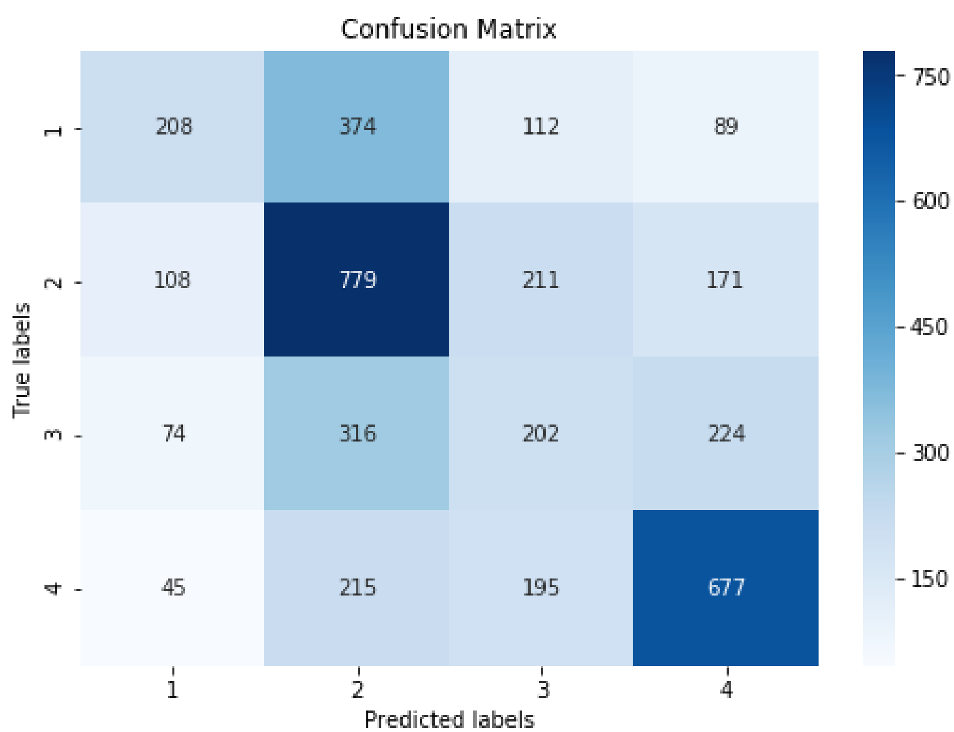
3. Evaluation on validation set

(a) 

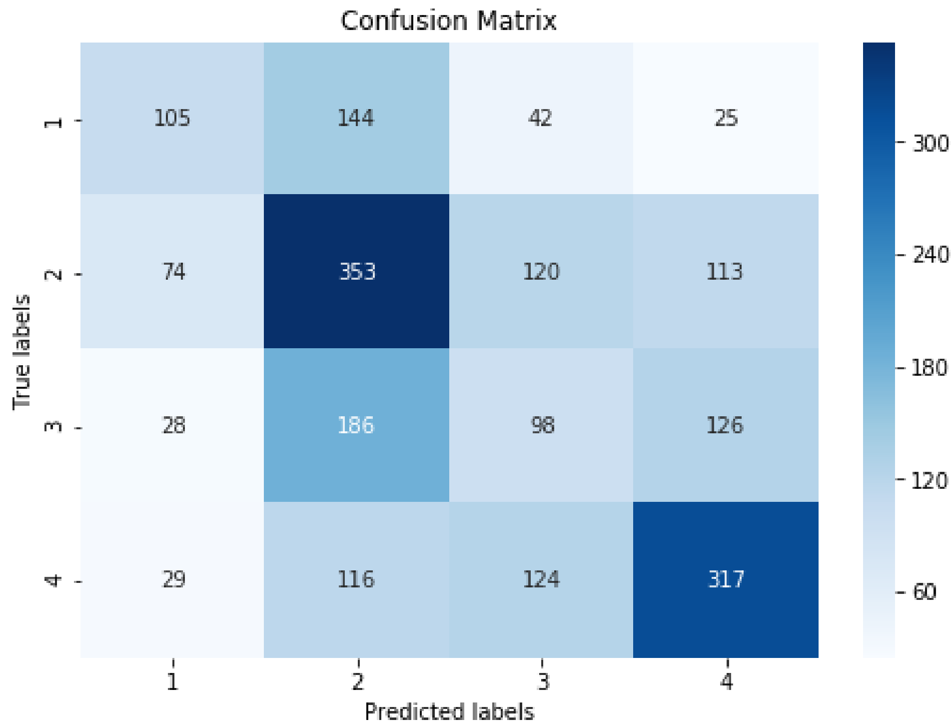
(b) 

4. Confusion Matrices

(a) Training data



(b) Validation data



2.4 Kaggle Challenge

Accuracy from kaggle - 0.47333