CSE512 Fall 2019 - Machine Learning - Homework 3

Your Name: Aveena Kottwani

Solar ID: 112689816

NetID email address: akottwani@cs.kottwani@stonybrook.edu

Names of people whom you discussed the homework with: Chaitra Hegde

1.1

1.1.1

1.

1.1

1.1.1 Show optimal Bayes risk for data point x $r^*(x)$

We know that

1) cost of false positive = α = cost of true negative

2) cost of false negative = 1 = cost of true positive

Probability of (x = positive) = $\eta(x)$

Probability of (x = negative) = $1 - \eta(x)$

Risk (x = positive) = (Cost of positive)(Probability of x = positive)

$= C_0(P(Y \neq h(x) \mid Y=1))$ $\quad P(Y \neq h(x) \mid Y=1)$

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

Risk (x = negative) = (Cost of negative)(Probability of x = negative)

$= C_1(P(Y \neq h(x) \mid Y=0))$ $\quad P(Y \neq h(x) \mid Y=0)$

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

Optimal Bayes risk is minimum of the risk of both $Y=0$ (negative) & $Y=1$ (positive)

Hence

Optimal Bayes Risk $r^*(x) = \min(1 \cdot \eta(x), \alpha(1-\eta(x)))$

1.1.2

1.1.2 Asympotic Risk of 1-NN classifier

$P(Y = +ve \mid X = x) = \eta(x)$

$P(Y = -ve \mid X = x) = 1 - \eta(x)$

The risk for x with probability $\eta(x)$ is

$$r^*(x) = [(1)\eta(x)](1 - \eta(x)) + [\alpha(1 - \eta(x))]\eta(x)$$

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

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

1.1.3

1.1.3

Prove:

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

Sol: We know that

$r(x) = \eta(x)(1-\eta(x))[1+\alpha] \rightarrow$ Eq ①

Condition 1: When $\eta(x) < \alpha(1-\eta(x))$

We know $r^*(x) = \min(\eta(x), \alpha(1-\eta(x)))$

Since $\min(\eta(x), \alpha(1-\eta(x))) = \eta(x)$

$r^*(x) = \eta(x) \longrightarrow$ eq ②

Substituting by Using eq (2) ↑ in equation ① $r(x)$

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

↳ Equation Ⓐ

Condition 2: When $\eta(x) > \alpha(1-\eta(x))$

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

$r^*(x) = \alpha(1-\eta(x)) \Rightarrow \boxed{1-\eta(x) = \frac{r^*(x)}{\alpha}}$

Substituting this value in equation ① of $r(x)$

$r(x) = \eta(x) \frac{r^*(x)}{\alpha}(1+\alpha)$

Also $\eta(x) = 1 - \frac{r^*(x)}{\alpha}$

$r(x) = \left(1-\frac{r^*(x)}{\alpha}\right) \frac{r^*(x)}{\alpha}(1+\alpha)$

Since $\alpha > 1$, ~~sub~~ removing α, decreases the value of $r(x)$

$r(x) = \left(1-\frac{r^*(x)}{\alpha}\right)\frac{r^*(x)}{\alpha}(1+\alpha)$

$\alpha r(x) = \left(1-\frac{r^*(x)}{\alpha}\right) r^*(x)(1+\alpha)$

Since $\alpha > 1$, Removing α, and substituting $\alpha r(x) \rightarrow r(x)$ and adjusting the equation sign

$r(x) \le (1-r^*(x))\, r^*(x)(1+\alpha)$

↳ Equation Ⓑ

from Equation Ⓐ & Ⓑ we prove

$r(x) \le (1-r^*(x))\, r^*(x)(1+\alpha)$

1.1.4

1.1.4

R is the asymptotic risk of 1-NN classifier &

R^* is bayes risk

Prove: $R \leq (1+\alpha) R^* (1-R^*)$

we know that

$$R(X) = E(r(x))$$

we know $r(x) \leq (1+\alpha) r^*(x)[1-r^*(x)]$

$$\therefore R(X) = E[(1+\alpha) r^*(x)[1-r^*(x)]]$$

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

$$= (1+\alpha) E[r^*(x) - [r^*(x)]^2]$$

$$= (1+\alpha) (E[r^*(x)] - E[[r^*(x)]^2])$$

$$= (1+\alpha) E[r^*(x)] - Var(r^*(x)) - E(r^*)^2$$

Since $E[r^*(x)^2] = Var(r^*(x)) + (E[r^*(x)])^2$

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

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

$$R(X) \geq (1+\alpha) E[r^*(x)] - E[r^*(x)]^2$$

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

$$E(r^*(x)) = R^*(X)$$

$$R(X) \geq (1+\alpha) R^*(X) [1 - R^*(X)]$$

1.2

1.2.1

1.2.

1.2.1 $P(Y = +ve \mid X = x) = \eta(x)$

$P(Y = -ve \mid X = x) = (1 - \eta(x))$

Asymptotic risk is the sum of cost & probability of x.

$$r(x) = \sum \text{Prob}(X = x) \cdot C(x)$$

$r(x)$ for point x :-

$P(Y = +ve \mid X = x) = \eta(x)$

$P(Y = -ve \mid X = x) = 1 -$ Probability that at least $(k+1)/2$ out of k points are positive

$= 1 - g(\eta, k)$

Asymptotic risk

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

Hence proved.

1.2.2. Prove $r(x) = r^*(x)\,(1 - 2r^*(x))\, g(r^*, k)$

We know

$$r(x) = \eta(x)\,(1 - g(\eta,k)) + (1 - \eta(x))\, g(\eta,k) \quad \text{(I)}$$

(1) When $\eta(x) < \frac{k}{2}$

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

$\therefore$ From equation (I)

$r(x) = r^*(x)\,(1 - g(r^*,k)) + (1 - r^*)\, g(r^*, k)$

$= r^* - g(r^*,k)\, r^* + g(r^*,k) - r^* g(r^*,k)$

$= r^* - 2g(r^*,k)\, r^* + g(r^*,k)$

(2) When $\cancel{g(\eta,k)}$ $\eta(x) > \frac{k}{2}$

$r^*(x) = g(\eta, k)$

$r(x) = \eta(x)\,(1 - r^*(x)) + [1 - \eta(x)]\, r^*(x)$

$= \eta(x) - \eta(x)\, r^* + r^* - \eta(x)\, r^*$

$r(x) = r^* + \eta(x) - 2\eta(x)\, r^*(x)$

Since $g(n,K)$ is probability at ~~tes~~ least $K+1/2$ out of K is positive

$$\eta(x) = g(n,k) \quad \text{for } K+1/2 \text{ out of } k.$$

$$r(x) = r^*(x) - 2g(r^*(x),k)r^* + g(r^*(x),k)$$

$$r(x) = r^*(x)\,[1 - 2r^*(x)]\,g(r^*(x),K)$$

Hence proved

1.2.3

1.2.3 $p(H(n) \geqslant (p+\varepsilon)n) \leq \exp(-2\varepsilon^2 n)$

To prove:

$$g(r^*(x), K) \leq \exp(-2\,(0.5 - r^*x)^2 K)$$

correspondingly we see that

$$g(r^*(x), k) = P(H(n) \geqslant (p+\varepsilon)n)$$

$$\varepsilon = \frac{1}{2} - r^*(x)$$

$$n = k$$

Hence proved.

2.1

2.1

Prove : $\frac{\partial \log(P(y^i \mid \bar{x}^i; \theta))}{\partial \theta_c} = (\delta(c = y^i) - P(c \mid \bar{x}^i; \theta))\bar{x}^i$

Conditions → 1) when $y^i \neq$ class c 2) $y^i =$ class c.

$$\frac{\partial \log[P(y^i \mid \bar{x}^i; \theta)]}{\partial \theta_c} = \frac{\partial}{\partial \theta_c} \log\left[\frac{\exp(\theta_i^T \bar{x}^i)}{1 + \sum_j^{k-1} \exp(\theta_j^T \bar{x}^i)}\right]$$

$$= \frac{\partial}{\partial \theta_c}\left[\log(\exp \theta_i^T \bar{x}^i) + \log\left(1 + \sum_j^{k-1} \exp(\theta_j^T \bar{x}^i)\right)^{-1}\right]$$

$$= \frac{\partial}{\partial \theta_c}\left[\theta_i^T \bar{x}^i - \log\left[1 + \sum_j^{k-1} \exp \theta_j^T \bar{x}^i\right]\right]$$

$$= \frac{\partial \theta_i^T \bar{x}^i}{\partial \theta_c} - \frac{\partial}{\partial \theta_c} \log\left[1 + \sum_j^{k-1} \exp \theta_j^T \bar{x}^i\right]$$

$$= \bar{x}^i \frac{\partial \theta_i^T}{\partial \theta_c} - \frac{\frac{\partial}{\partial \theta_c}\left(\sum_j^{k-1} \exp \theta_j^T \bar{x}^i\right)}{\left(1 + \sum_j^{k-1} \exp \theta_j^T \bar{x}_i\right)} (\bar{x}^i) \frac{\partial(\theta_j^T)}{\partial \theta_c}$$

$$= \bar{x}^i \left[\frac{\partial \theta_i^T}{\partial \theta_c} - \frac{\sum_j^{k-1} \frac{\partial}{\partial \theta_c} \exp(\theta_j^T \bar{x}^i) \left(\frac{\partial \theta_j^T}{\partial \theta_c}\right)}{1 + \sum_j^{k-1} \exp \theta_j^T \bar{x}_i}\right]$$

When $\theta_i = \theta_c$

$$\frac{\partial}{\partial \theta_c} \theta_i^T = 1$$

else $\frac{\partial \theta_i^T}{\partial \theta_c} = 0$

Also when $\theta_i = \theta_c$.

$$\frac{\partial \sum}{\partial \theta_c} \exp(\theta_j^T \bar{x}^i) = \exp(\theta_c^T \bar{x}_i)$$

else $\frac{\partial}{\partial \theta_c} \sum \exp(\theta_j^T \bar{x}^i) = 0$

$$\frac{\partial}{\partial \theta_c} \sum_j^{k-1} \exp(\theta_j^T \bar{x}^i) = \exp(\theta_c^T \bar{x}^i)$$

$\therefore \frac{\partial}{\partial \theta_c} \log [P(Y_i | \bar{X}^i; \theta)] = L$

When $\theta_c = \theta_j$

$$L = \bar{X}^i \left[1 - \frac{\exp(\theta_c^T \bar{X}^i)}{1 + \sum_j^{k-1} \exp(\theta_j^T \bar{X}^i)} \right]$$

else

$$L = \bar{X}^i \left[0 - \frac{\exp(\theta_c^T \bar{X}^i)}{1 + \sum_j^{k-1} \exp(\theta_j^T \bar{X}^i)} \right]$$

$\therefore$ considering an indicator function $\delta(c = Y_i)$

$\delta(c = Y_i) = 1$ when $\theta_c = \theta_j$

else $= 0$ when $\theta_c \neq \theta_j$

$$\therefore \quad L = \left[\delta(c = Y_i) - \frac{\exp(\theta_c^T \bar{X}^i)}{1 + \sum_j^{k-1} \exp(\theta_j^T \bar{X}^i)} \right] \bar{X}^i$$

$$= [\delta(c = Y_i) - P(c | \bar{X}^i; \theta)] \bar{X}^i$$

Hence proved:

$$\frac{\partial}{\partial \theta_c} \log P(Y_i | \bar{X}^i; \theta) = [\delta(c = Y_i) - P(c | \bar{X}^i; \theta)] \bar{X}^i$$

2.2

1. (15 points) Run your implementation on the provided training data with max epoch = 1000, m = 16, $\eta 0 = 0.1$, $\eta 1 = 1$, $\delta = 0.00001$.

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

The number of epochs: 25

(b) Plot the curve showing $L(\theta)$ as a function of epoch.

(c) What is the final value of $L(\theta)$ after the optimization?

The final value of $L(\theta)$ after the optimization : 1.36701623

2. (10 points) Keep m = 16, $\delta = 0.00001$, experiment with different values of $\eta 0$ and $\eta 1$. Can you find a pair of parameters ($\eta 0$, $\eta 1$) that leads to faster convergence?

(a) Report the values of ($\eta 0$, $\eta 1$). How many epochs does it take? What is the final value of $L(\theta)$?

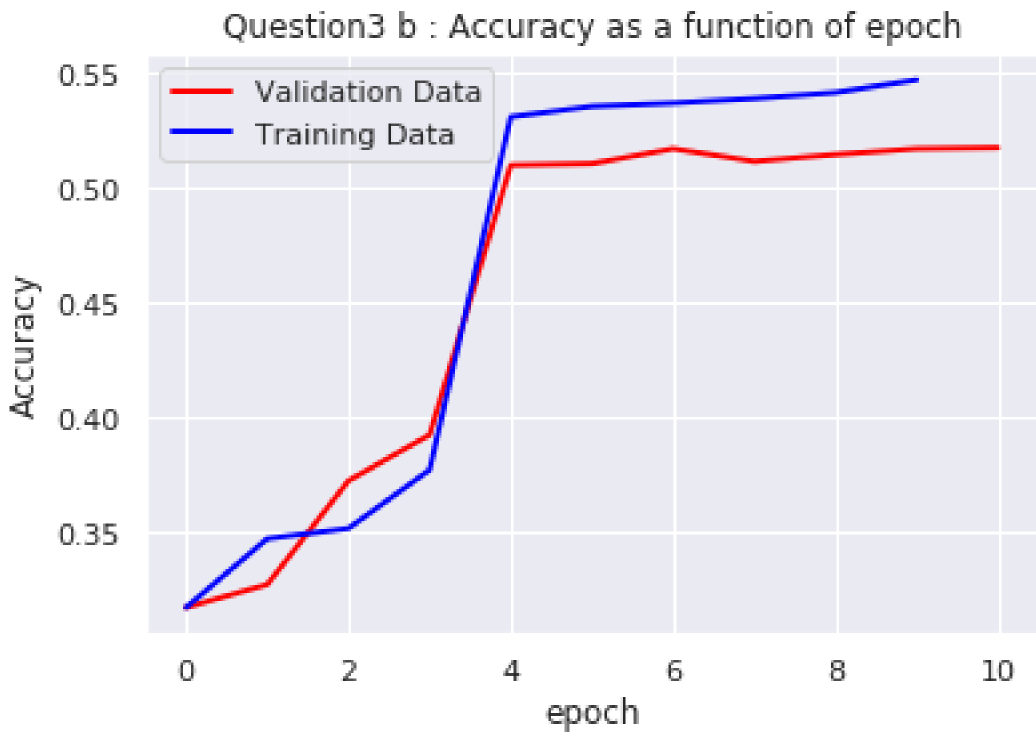
1. ($\eta 0$, $\eta 1$) =(0.65,1) epochs=29 $L(\theta)$= 1.33895
2. ($\eta 0$, $\eta 1$) =(0.5,1) epochs=19 $L(\theta)$= 1.300086
3. ($\eta 0$, $\eta 1$) =(0.9,1) **epochs=14** $L(\theta)$= 1.36866335
4. ($\eta 0$, $\eta 1$) =(0.7,1) epochs=16 $L(\theta)$= 1.318225

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

3. (10 points) Evaluate the performance on validation data

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



4. (5 points) With the learned classifier:

(a) Report the confusion matrices on the validation and the training data.
Train:

	Predicted Class 1	Predicted Class 2	Predicted Class 3	Predicted Class 4
Actual Class 1	402	41	39	301
Actual Class 2	271	147	340	511
Actual Class 3	330	176	157	143
Actual Class 4	132	334	324	342

Validation:

	Predicted Class 1	Predicted Class 2	Predicted Class 3	Predicted Class 4
Actual Class 1	57	59	67	33
Actual Class 2	324	136	73	127
Actual Class 3	127	51	127	133
Actual Class 4	300	61	361	87

2.4.

2 Accuracy from Kaggle: 0.28166