CSCI 5521 – Introduction to Machine Learning

Homework – 4

Mourya Karan Reddy Baddam – 5564234
Email – badda004@umn.edu

1.) We have $z^t = Wx^t$ —— ①

where $x^t \in \mathbb{R}^D$

$z^t \in \mathbb{R}^d$

W is $d \times D$ matrix.

We know W is orthonormal matrix with orthonormal rows.

Hence $W \cdot W^T = I_d$ & $W^T \cdot W \neq I_D$

a.) $v^t = W^T z^t$ —— ②

$= W^T W x^t$

and since $W^T W \neq I_D$.

$\therefore v^t = W^T W x^t \neq x^t$.

Professor HighLowHigh's claim is wrong.

b) $\sum_{t=1}^{N} \|x^t - v^t\|_2^2 = \sum_{t=1}^{N} \{(x^t - v^t)^T (x^t - v^t)\}$ (since $\|a\|_2^2 = a^T \cdot a$)

$= \sum_{t=1}^{N} \{((x^t)^T - (v^t)^T)(x^t - v^t)\}$

$= \sum_{t=1}^{N} \{(x^t)^T x^t - (x^t)^T v^t - (v^t)^T x^t + (v^t)^T v^t\}$

$= \sum_{t=1}^{N} \{(x^t)^T x^t + (v^t)^T v^t - (x^t)^T v^t - (v^t)^T x^t\}$

Using ②

$= \sum_{t=1}^{N} \{(x^t)^T x^t + (v^t)^T v^t - (x^t)^T W^T z^t - (W^T z^t)^T x^t\}$

$$= \sum_{t=1}^{N} \left\{ (x^t)^T x^t + (v^t)^T v^t - (W x^t)^T z^t - (z^t)^T (W^T)^T x^t \right\}$$

(since $(AB)^T = B^T A^T$)

$$= \sum_{t=1}^{N} \left\{ (x^t)^T x^t + (v^t)^T v^t - (z^t)^T z^t - (z^t)^T W x^t \right\}$$

(since $(A^T)^T = A$)

$$= \sum_{t=1}^{N} \left\{ (x^t)^T x^t + (v^t)^T v^t - 2 (z^t)^T z^t \right\}$$

(from ①)

$$= \sum_{t=1}^{N} \left\{ (x^t)^T x^t + (v^t)^T v^t - 2 (z^t)^T W . W^T z^t \right\}$$

(since $W . W^T = I_d$ & $z^t \in \mathbb{R}^d$)

$$= \sum_{t=1}^{N} \left\{ (x^t)^T x^t + (v^t)^T v^t - 2 (W^T z^t)^T W^T z^t \right\}$$

(since $(W^T)^T = W$ & $(z^t)^T (W^T)^T = (W^T z^t)^T$)

$$= \sum_{t=1}^{N} \left\{ (x^t)^T x^t + (v^t)^T v^t - 2 (v^t)^T v^t \right\}$$

(from ②)

$$= \sum_{t=1}^{N} \left\{ (x^t)^T x^t - (v^t)^T v^t \right\}$$

$$= \sum_{t=1}^{N} \|x^t\|_2^2 - \sum_{t=1}^{N} \|v^t\|_2^2$$

(since $a^T . a = \|a\|_2^2$

$$\therefore \sum_{t=1}^{N} \|x\|_2^2 - \sum_{t=1}^{N} \|v^t\|_2^2 = \sum_{t=1}^{N} \|x^t - v^t\|_2^2$$

Hence Professor HighlowHigh's claim is <u>correct</u>

2) Given $z_h^t = g(a_h^t)$

$$a_h^t = \sum_{j=1}^{d} w_{h,j}^t x_j^t + w_{h,0}^t$$

$$y_i^t = g(a_i^t)$$

$$a_i^t = \sum_{h=1}^{H} v_{i,h}^t z_h^t + v_{i,0}^t$$

Loss function $\rightarrow E(w,v/z) = \sum_{t=1}^{N} \sum_{i=1}^{K} L(r_i^t, y_i^t)$

a) In SGD, we take 1 point in each iteration. Hence at iteration t, we use point t

We have $v_{i,h}^{t+1} = v_{i,h}^t - \eta \frac{\partial E(w,v/z)}{\partial v_{i,h}^t}$ — ①

where $\eta \rightarrow$ learning rate.

$$\frac{\partial E(w,v/z)}{\partial v_{i,h}^t} = \frac{\partial E(w,v/z)}{\partial y_i^t} \times \frac{\partial y_i^t}{\partial a_i^t} \times \frac{\partial a_i^t}{\partial v_{i,h}^t} \quad \text{(By chain rule)}$$

$$= \frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \times \frac{\partial g(a_i^t)}{\partial a_i^t} \times \frac{\partial \sum_{h=1}^{H} v_{i,h}^t z_h^t + v_{i,0}^t}{\partial v_{i,h}^t}$$

$$\frac{\partial E(w,v/z)}{\partial v_{i,h}^t} = \frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \times g'(a_i^t) \times z_h^t \quad (\text{where } z_0^t = 1)$$

If $\Delta_i^t = g'(a_i^t)\left(-\frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t}\right)$ & $\Delta v_{i,h} = \eta \Delta_i^t z_h^t$

$$\Rightarrow -\eta \frac{\partial E(w,v/z)}{\partial v_{i,h}^t} = \Delta v_{i,h}$$

$$\Rightarrow \underline{\underline{v_{i,h}^{new} = v_{i,h}^{old} + \Delta v_{i,h}}} \quad (\text{from } ①)$$

b) We have $w_{h,j}^{t+1} = w_{h,j}^{t} - \eta \frac{\partial E(w,v|z)}{\partial w_{h,j}^{t}}$ — ②

$$\frac{\partial E(w,v|z)}{\partial w_{h,j}^{t}} = \frac{\partial E(w,v|z)}{\partial y_i^t} \times \frac{\partial y_i^t}{\partial a_i^t} \times \frac{\partial a_i^t}{\partial z_h^t} \times \frac{\partial z_h^t}{\partial a_h^t} \times \frac{\partial a_h^t}{\partial w_{h,j}^t}$$

$$= \sum_{i=1}^{k} \left\{ \frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \times \frac{\partial g(a_i^t)}{\partial a_i^t} \times \frac{\partial \sum_{h=1}^{H} v_{i,h}^t \cdot z_h^t + v_{i,0}^t}{\partial z_h^t} \right\} \times \frac{\partial g(a_h^t)}{\partial a_h^t} \times \frac{\partial \sum_{j=1}^{d} w_{h,j}^t x_j^t + w_{h,0}^t}{\partial w_{h,j}^t}$$

$$= \sum_{i=1}^{k} \left\{ \frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t} \times g'(a_i^t) \times v_{i,h}^t \right\} \times g'(a_h^t) \times x_j^t \quad (\text{where } x_0^t = 1)$$

If $\Delta_i^t = g'(a_i^t)\left(-\frac{\partial L(r_i^t, y_i^t)}{\partial y_i^t}\right)$ & $\Delta_h^t = g'(a_h^t)\left(\sum_{i=1}^{k} \Delta_i^t v_{i,h}^t\right)$

& $\Delta w_{h,j}^t = \eta \Delta_h^t x_j^t$

$$\Rightarrow -\eta \frac{\partial E(w,v|z)}{\partial w_{h,j}^t} = \eta \Delta_h^t x_j^t = \Delta w_{h,j}$$

$$\Rightarrow \underline{\underline{w_{h,j}^{new} = w_{h,j}^{old} + \Delta w_{h,j}}} \quad (\text{from } ②)$$

EC1:- $L(r_i^t, y_i^t) = (r_i^t - y_i^t)^2$

a) $g(a) = \max(0, a)$

$$\Rightarrow g(a) = \begin{cases} 0 & ; a \le 0 \\ a & ; a > 0 \end{cases}$$

$$\Rightarrow g'(a) = \begin{cases} 0 & ; a \le 0 \\ 1 & ; a > 0 \end{cases} \Rightarrow g'(a) = \frac{g(a)}{a}.$$

$$\frac{\partial L(\hat{y}_i^t, y_i^t)}{\partial y_i^t} = \frac{\partial (\hat{y}_i^t - y_i^t)^2}{\partial y_i^t} = -2(\hat{y}_i^t - y_i^t)$$

Update in ①:-

$$\Delta v_{i,h} = \eta \Delta_i^t z_h^t \quad \text{, where } z_h^t = g(a_h^t) = \begin{cases} 0 & ; a_h^t \le 0 \\ a_h^t & ; a_h^t > 0 \end{cases}$$

Where

$$\Delta_i^t = g'(a_i^t) \left(-\frac{\partial L(\hat{y}_i^t, y_i^t)}{\partial y_i^t} \right)$$

$$\Rightarrow \Delta_i^t = \frac{g(a_i^t)}{a_i^t} \cdot 2 \cdot (\hat{y}_i^t - y_i^t) \quad \text{where } g(a_i^t) = \begin{cases} 0 & ; a_i^t \le 0 \\ a_i^t & ; a_i^t > 0 \end{cases}$$

$$\Rightarrow \Delta_i^t = \begin{cases} 0 & ; a_i^t \le 0 \\ 2(\hat{y}_i^t - y_i^t) & ; a_i^t > 0 \end{cases}$$

b.) $\alpha \in [0, 1]$

$$g(a) = \max(0, a) + \alpha \min(0, a)$$

$$\Rightarrow g(a) = \begin{cases} a & ; a > 0 \\ \alpha a & ; a \le 0 \end{cases}$$

$$g'(a) = \begin{cases} 1 & ; a > 0 \\ \alpha & ; a \le 0 \end{cases} \quad \text{or } g'(a) = \frac{g(a)}{a}.$$

c) If $\alpha = 1$, the two layer perceptron will be a linear model. $\Rightarrow \underline{g(a) = a}$

Let $w_h^t = [w_{h,0}^t, w_{h,1}^t, \dots w_{h,d}^t]_{1 \times d+1}$; $x^t = [x_0^t, x_1^t, \dots x_d^t]^T_{1 \times d+1}$ where $x_0^t = 1$

$v_i^t = [v_{i,0}^t, v_{i,1}^t, \dots v_{i,H}^t]_{1 \times H+1}$; $z^t = [z_0^t, z_1^t, \dots z_H^t]^T_{1 \times H+1}$ where $z_0^t = 1$

Let $W^t = \begin{bmatrix} \leftarrow W_0^t \rightarrow \\ \leftarrow W_1^t \rightarrow \\ \vdots \\ \leftarrow W_H^t \rightarrow \end{bmatrix}_{H+1 \times d+1}$ $\quad W_0^t = [1, 0, 0, \cdots 0]_{1 \times d+1}$

$$V^t = \begin{bmatrix} \leftarrow V_1^t \rightarrow \\ \leftarrow V_2^t \rightarrow \\ \vdots \\ \leftarrow V_k^t \rightarrow \end{bmatrix}_{k \times H+1}$$

$\Rightarrow$

$$a^t = W^t x^t \quad \text{where } x^t = [x_0^t, x_1^t \cdots x_d^t]^T$$

$$z^t = g(a^t) = a^t = W^t x^t$$

Similarly

$$y^t = V^t z^t$$

$$\boxed{\therefore\ y^t = V^t W^t x^t}$$

Hence $\alpha = 1$, the two layer perception is linear.

Q3: Summary of Error Rates:

Summary: MySVM2 with m = 40 for Boston50

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
17.65%	18.81%	20.79%	17.82%	22.77%	19.57%	1.95%

Summary: MySVM2 with m = 200 for Boston50

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
17.65%	19.8%	14.85%	18.81%	24.75%	19.17%	3.25%

Summary: MySVM2 with m = n for Boston50

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
23.53%	16.83%	13.86%	16.83%	22.77%	18.77%	3.75%

Summary: LogisticRegression with Boston50

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
15.69%	11.88%	12.87%	12.87%	17.82%	14.23%	2.2%

Summary: MySVM2 with m = 40 for Boston75

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
20.59%	18.81%	17.82%	24.75%	21.78%	20.75%	2.43%

Summary: MySVM2 with m = 200 for Boston75

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
22.55%	19.8%	16.83%	20.79%	19.8%	19.96%	1.86%

Summary: MySVM2 with m = n for Boston75

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
22.55%	20.79%	17.82%	18.81%	20.79%	20.15%	1.66%

Summary: LogisticRegression with Boston75

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
5.88%	9.9%	9.9%	12.87%	6.93%	9.1%	2.47%