10/8/22

Homework - (1) Part (1)

Date:

Q1. (1.) Sigmoid function.

$$f(x) = \frac{1}{1+e^{-x}}$$

ans. ★ Gradient of $f(x)$ $\equiv$ derivative of $f(x)$ w.r.t x $\equiv \frac{df(x)}{dx}$

$$\therefore \frac{df(x)}{dx} = \frac{d\left[\frac{1}{1+e^{-x}}\right]}{dx} \longrightarrow (1)$$

★ By applying quotient rule,

$$\frac{df}{dx} = \frac{\frac{d(1)}{dx}^{0} \cdot (1+e^{-x}) - \frac{d(1+e^{-x})}{dx} \cdot (1)}{(1+e^{-x})^2}$$

$$= \frac{d(1+e^{-x})}{dx} \cdot \frac{(-1)}{(1+e^{-x})^2} \quad \cancel{\star (e^{-x}) \cdot (-1)}$$

$$= \frac{-1}{(1+e^{-x})^2}\left[\frac{d(1)}{dx}^{0} + \frac{d(e^{-x})}{dx}\right]$$

$$= \frac{-1}{(1+e^{-x})^2} \cdot e^{-x} \cdot (-1) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{(1+e^{-x}-1)}{(1+e^{-x})^2}$$

$$= f(x)\left[1-f(x)\right]$$

(2.) Softmax

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, \quad 1 \le i \le n$$

ans. ★ For simplicity, $\sum_{j=1}^{n} \equiv \sum$.

$$\Rightarrow f(x_i) = \frac{e^{x_i}}{\sum e^{x_j}}, \quad 1 \le i \le n$$

P.T.O.

$$f'(x_i) = \frac{d[f(x_i)]}{dx_j} = \frac{d\left[\frac{(e^{x_i})}{\Sigma e^{x_j}}\right]}{dx_j} \longrightarrow (1)$$

★ By applying quotient rule,

$$f'(x_i) = \frac{\frac{d(e^{x_i})}{dx_j} \cdot [\Sigma e^{x_j}] - \frac{d[\Sigma e^{x_j}]}{dx_j} \cdot (e^{x_i})}{[\Sigma e^{x_j}]^2}$$

$$= \frac{\Sigma e^{x_j} \cdot \frac{d(e^{x_i})}{dx_j} - e^{x_i}\left[0+0+\cdots+e^{x_j}+\cdots+0\right]}{[\Sigma e^{x_j}]^2} \longrightarrow (2)$$

★ if $i=j$,
eq. (2) becomes,

$$= \frac{e^{x_i} \cdot \Sigma e^{x_i} - e^{x_i} \cdot e^{x_i}}{(\Sigma e^{x_i})^2} = \frac{e^{x_i}}{\Sigma e^{x_i}} \cdot \left[1 - \frac{e^{x_i}}{\Sigma e^{x_i}}\right]$$

$$= f(x_i) \cdot [1 - f(x_i)] \longrightarrow (3)$$

NOTE:-
★★★ If we did a partial derivative w.r.t. to (x_i), $(i \neq j)$ will not exist. [this is a generalized case].

★ if $i \neq j$, eq (2) becomes,

$$= 0 - \frac{e^{x_i} \cdot e^{x_j}}{(\Sigma e^{x_j})^2} = -f(x_i) \cdot f(x_j) \longrightarrow (4)$$

Combining (3) & (4), [eq (4) is only for generalized case].

$$f'(x_i) = \begin{cases} f(x_i) \cdot (1 - f(x_i)) & , \text{if } i = j \\ -f(x_i) \cdot f(x_j) & , \text{if } i \neq j \end{cases}$$

(3) Softplus activation

$$f(x) = \frac{1}{\beta} \cdot \ln(1+e^{\beta x})$$

ans. $f'(x) = \frac{d[f(x)]}{dx} = \frac{d\left[\frac{1}{\beta} \cdot \ln(1+e^{\beta x})\right]}{dx}$

$= \frac{1}{\beta}\left[\frac{d(\ln(1+e^{\beta x}))}{dx}\right] = \frac{1}{\beta}\left[\frac{1}{1+e^{\beta x}}\right] \cdot [0+e^{\beta x} \cdot \beta]$

$= \frac{e^{\beta x}}{1+e^{\beta x}} = \frac{1}{e^{-\beta x}+1}$

Page No.:
Date:

Q2. (1) $f(x) = x^T(Ax+z)$, $x, z \in R^{n\times 1}$, $A \in R^{n\times n}$

ans. $f(x) = x^TAx + x^Tz$

Shape of $f(x)$ $= (1\times n)(n\times n)(n\times 1) + (1\times n)(n\times 1)$
$= (1\times 1) + (1\times 1)$
$\longrightarrow (1\times 1) \rightarrow$ scalar.

$$\frac{df(x)}{dx} = \frac{d[x^TAx + x^Tz]}{dx}$$
$$= \frac{d(x^TAx)}{dx} + \frac{d(x^Tz)}{dx}$$
$$= \frac{d(x^TA)(x)}{dx} + \frac{(x^TA)d(x)}{dx} + z$$
$$= Ax + (x^TA)^T + z$$
$$= Ax + A^Tx + z = (A+A^T)x + z$$

Shape check for $f'(x)$: $(n\times n)(n\times 1) + (n\times 1)$
$= (n\times 1) + (n\times 1)$
$= (n\times 1) \rightarrow$ column vector.
$\longrightarrow$ same shape as (x).

(2) $L(w) = \frac{1}{2}(w^Tx - y)^2$, $w, x \in R^{n\times 1}$, $y \in R^{1\times 1}$

ans. Shape of $[w^Tx - y]$
$= (1\times n)(n\times 1) - (1\times 1)$
$= (1\times 1) - (1\times 1) = (1\times 1) \rightarrow$ scalar.
$\therefore$ Shape of $L(w)$ = scalar.

Let $z = w^Tx - y$

$$\therefore L(w) = \frac{1}{2} z^2$$

$$\therefore \frac{dL(w)}{dw} = \frac{dL(w)}{dz} \cdot \frac{dz}{dw}$$

$$= \frac{1}{2} \times 2z \times \frac{d(w^T x - y)}{dw}$$

$$= z \cdot \left[\frac{d(w^T x)}{dw} - \frac{dy}{dw} \right]$$

$$= z \cdot x = (w^T x - y) \cdot x$$

Shape check for $f'(w)$ = $(1\times1)(n\times1)$

→ since z is a scalar, $zx = xz$

$\therefore (n\times1)(1\times1) \longrightarrow (n\times1)$ → column vector

(3.) $L(w) = \frac{1}{2m} \| xw - y \|^2$, $w, y \in R^{n\times1}$, $X \in R^{n\times n}$

ans. Shape of $L(w) = (1\times1)$ [∵ norm of a vector is taken].

$$L(w) = \frac{1}{2m} (xw - y)^T \cdot (xw - y)$$

Let $u = (xw - y)$

$$\Rightarrow L(w) = \frac{1}{2m} u^T \cdot u$$

$$\frac{dL(w)}{dw} = \frac{1}{2m} \left[\frac{du^T}{dw} \cdot u + u^T \cdot \frac{du}{dw} \right]$$

$$= \frac{1}{2m} \left[\frac{du^T}{dw} \cdot u + \frac{du^T}{dw} \cdot u \right] = \frac{1}{2m} \left[2 \cdot \frac{du^T}{dw} \cdot u \right]$$

$$= \frac{1}{m}\left[\frac{d(xw-y)}{dw}\right] \cdot u$$

$$= \frac{1}{m}\left[\frac{d(xw)}{dw} - \frac{dy}{dw}\right] \cdot u$$

$$= \frac{1}{m}\left[x^T\right] \cdot u$$

$$\Rightarrow \frac{dL(w)}{dw} = \frac{1}{m}\left[x^T\right] \cdot [xw-y]$$

Shape check for $L'(w) = (n \times n)[(n \times n)(n \times 1) - (n \times 1)]$

$= (n \times n)[(n \times 1)]$

$= (n \times 1)$ ⟶ column vector

Q3. $z = Wx + b$

$L = \|z-y\|^2$

$W \in R^{n \times n}$

$x, b, y \in R^{n \times 1}$, $dL/dW = ?$

ans: Shape of L ⟶ scalar since norm is taken.

$L = (z-y)^T \cdot (z-y)$

Let $u = (z-y)$

$\Rightarrow L = u^T \cdot u$

$$\therefore \frac{dL}{dW} = \frac{d(u^T \cdot u)}{dW} = \frac{d(u^T)}{dW} \cdot u + u^T \cdot \frac{d(u)}{dW}$$

$$= \frac{d(u^T)}{dW} \cdot u + \frac{d(u^T)}{dW} \cdot u$$

$$= \frac{2\,d(u^T) \cdot u}{dW} \longrightarrow (1)$$

Page No.:
Date:

Now, $\frac{d(u^T)}{dW} = \frac{d[Wx+b-y]^T}{dW} = \frac{d[x^TW^T]}{dW} + \frac{db^T}{dW} - \frac{dy^T}{dW}$

$\longrightarrow$ (2)

We vectorize matrix $W \in R^{n \times n}$ into a row vector of n features, of length (n) each.

$\Rightarrow W = [w_1 \quad w_2 \quad \dots \quad w_n]$

$\therefore \dim(W) = (n \times n)$

Using this in eq (2).,

$$\frac{d(u^T)}{dW} = x^T + 0 - 0 = x^T$$

Combining (1) & (3),

$$\frac{dL}{dW} = 2(Wx+b-y) \cdot x^T$$

Shape check of $\frac{dL}{dW} = [(n \times n)(n \times 1) + (n \times 1) - (n \times 1)](1 \times n)$

$= (n \times 1) \cdot (1 \times n)$

$= (n \times n)$

Page No.:
Date:

Q4. (i) Gradient descent for $\alpha = 0.5, 1.5, 2.5$

Given :-

$\hat{y} = xw$, $x \in R$, $w \in R$

L2 loss $= \frac{1}{2}(xw - y)^2 = J$

$w_0 = 0$

$[x = 1, y = 100]$

$$\frac{dJ}{dw} = \frac{1}{2} \times 2(xw - y) \cdot x = \boxed{(xw - y)x = \frac{dJ}{dw}} \rightarrow (1)$$

$= (w - 100)$

(i) $\alpha = 0.5$.

ans. Iteration 1

$$w_1 = w_0 - \alpha \left.\frac{dJ}{dw}\right|_{w_0}$$

$= 0 - 0.5(1 \times 0 - 100)(1)$

$= 50$

$L_1 = \frac{1}{2}(1 \times 50 - 100)^2 = \frac{2500}{2} = 1250$

Iteration 2

$w_2 = 50 - 0.5(50 - 100)$

$= 50 + 25 = 75$

$L_2 = \frac{1}{2}(75 - 100)^2 = \frac{625}{2} = 312.5$

Iteration 3

$w_3 = 75 - 0.5(75 - 100)$

$= 75 + 0.5 \times 25 = 87.5$

$L_3 = \frac{1}{2}(87.5-100)^2 = 78.125$

Iteration 4

$w_4 = 87.5 - 0.5(87.5-100)$

$= 93.75$

$L_4 = \frac{1}{2}(93.75-100)^2 = 19.53$

Iteration 5

$w_5 = 93.75 - 0.5(93.75-100)$

$= 96.875$

$L_5 = \frac{1}{2}(96.875-100)^2 = 4.88$

Iteration 6

$w_6 = 96.875 - 0.5(96.875-100)$

$= 98.4375$

$L_6 = \frac{1}{2}(1.5625)^2 = 1.22$

(iii) Iteration 7

$w_7 = 98.4375 - 0.5(98.4375-100)$

$= 99.21875$

$L_7 = \frac{1}{2}(99.21875-100)^2 = 0.305$

$\because$ Loss is < 1, we can stop.

optimal $w = 99.21875$

(ii) $\alpha = 1.5$

Iteration 1

$w_1 = 0 - 1.5(0 - 100) = 150.0$

$J_1 = \frac{1}{2}(150 - 100)^2 = 1250$

Iteration 2

$w_2 = 150 - 1.5(150 - 100)$

$= 75$

$J_2 = \frac{(75 - 100)^2}{2} = 312.5$

Iteration 3

$w_3 = 75 - 1.5(75 - 100)$

$= 112.5$

$J_3 = \frac{(112.5 - 100)^2}{2} = 78.125$

Iteration 4

$w_4 = 112.5 - 1.5(112.5 - 100)$

$= 93.75$

$J_4 = 19.53$

Iteration 5

$w_5 = 93.75 - 1.5(93.75 - 100)$

$= 103.125$

$J_5 = 4.88$

Iteration 6

$w_6 = 103.125 - 1.5(103.125 - 100)$

$= 98.43$

$J_6 = 1.22$

** w is oscillating around 100. Stopping for $w_6 = 98.43$.

(iii) $\alpha = 2.5$

Iteration 1

$w_1 = 0 - 2.5(0 - 100)$

$= 250$

$J_1 = \frac{(250-100)^2}{2} = 11250$

Iteration 2

$w_2 = 250 - 2.5(250 - 100)$

$= -125$

$J_2 = \frac{(-125-100)^2}{2} = 25312.5$

Iteration 3

$w_3 = -125 - 2.5(-125 - 100)$

$= 437.5$

$J_3 = 56953.125$

∴ Since learning rate is too high, Gradient descent is showing divergent behaviour.
⇒ It will take a greater no. of iterations to converge.

(2) For optimal position, loss should be 0.

[can also be derived by finding minima]

$\Rightarrow J = \frac{1}{2}(wx - y)^2 = 0$

$\Rightarrow |wx - y| = 0$

Since there is only 1 data point,

$|w(1) - 100| = 0$

$\Rightarrow \boxed{|w - 100| = 0} \rightarrow (1)$

$\hookrightarrow$ for optimal w.

(i) Condition for α when G.D. starts oscillating around optimal position.

ans: Given:

OPTIMAL POINT = 100.

* Oscillation around optimal point means that simultaneous w_s are on either side of 100.

$\Rightarrow$ Case 1: $\begin{cases} w_n > 100 \\ w_{n-1} < 100 \end{cases}$

We know,

$w_n = w_{n-1} - \alpha(w_{n-1} - 100)$

> 100

$\Rightarrow (w_{n-1}) - 100 > \alpha(w_{n-1} - 100)$ $\hookrightarrow (1)$

Given that $(w_{n-1} - 100 < 0)$

$\therefore \dfrac{(w_{n-1} - 100)}{(w_{n-1} - 100)} < \alpha$

$\Rightarrow \boxed{\alpha > 1} \rightarrow (2)$

Case 2: $\begin{cases} w_n < 100 \\ w_{n-1} > 100 \end{cases}$

Ily,

$w_n = w_{n-1} - \alpha(w_{n-1} - 100)$

< 100

$\Rightarrow (w_{n-1} - 100) \not< \alpha(w_n - 100)$ $\hookrightarrow (3)$

Given that $(w_{n-1} - 100 > 0)$

$\dfrac{(w_{n-1} - 100)}{(w_{n-1} - 100)} < \alpha$

$\Rightarrow \boxed{\alpha > 1} \rightarrow (4)$

Combining (2) & (4),

$\boxed{\alpha > 1}$ for oscillation

(ii) Condition for α that GD can converge.

ans. For convergence, assume stopping criteria to be s.t $|w_t - 100| < 1$

$\Rightarrow -1 < (w_t - 100) < 1$

$\Rightarrow \boxed{99 < w_t < 101} \longrightarrow (1)$

We know,

$w_t = w_{t-1} - \alpha(w_{t-1} - 100) \longrightarrow (2)$

$\Rightarrow 99 < w_{t-1} - \alpha(w_{t-1} - 100) < 101$

$\Rightarrow (w_{t-1} - 99) > \alpha(w_{t-1} - 100) > (w_{t-1} - 101) \longrightarrow (3)$

Since (w_{t-1}) is the weight of a previous iteration

$(w_{t-1}) \leq 99$ or $(w_{t-1}) \geq 101$

★ Case 1: $(w_{t-1}) \leq 99$

$\Rightarrow$ From (3),

$$0 \leq \frac{(w_{t-1} - 99)}{(w_{t-1} - 100)} < \alpha < \frac{(w_{t-1} - 101)}{(w_{t-1} - 100)} \leq 2$$

$\boxed{\Rightarrow 0 < \alpha < 2} \longrightarrow (4)$ $\quad \frac{(w_{t-1} - 101)}{(w_{t-1} - 100)} \geq 1$

★ Case 2: $(w_{t-1}) \geq 101$

$$2 \geq \left(\frac{w_{t-1} - 99}{w_{t-1} - 100}\right) > \alpha > \frac{(w_{t-1} - 101)}{(w_{t-1} - 100)} \geq 0$$

$\Rightarrow \boxed{2 > \alpha > 0} \longrightarrow (5)$

Page No.:
Date:

∴ Combining case ① & ②,

$\boxed{0 < \alpha < 2}$ → for convergence.

③ (i) For $\alpha = 0.5$, it converges in iteration 7, for $w = 99.21875$

(ii) For $\alpha = 1.5$, it oscillates around the optimal point [100] starting from iteration ①

(iii) For $\alpha = 2.5$, loss is monotonically increasing & never converges.