21 Aug '22

HW 1

adarshs@comp.nus.edu.sg

Q.1 Solve gradients for:

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

$$\frac{\partial f}{\partial x} = -\frac{1}{(1+e^{-x})^2} \times (-e^{-x})$$

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

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

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

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

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

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

$$f(x_i) = \frac{e^{x_i}}{(e^{x_1} + e^{x_2} + \cdots e^{x_i} + \cdots e^{x_n})}$$

$$\frac{\partial f}{\partial x_i} = \frac{e^{x_i}\cdot\left(\sum_{j=1}^{n} e^{x_j}\right) + e^{x_i}\cdot\left(e^{x_i}\right)}{\left(\sum_{j=1}^{n} e^{x_j}\right)^2}$$

$$= \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} + \left(\frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \right)^2$$

$$= f(x_i)\left(1 + f(x_i)\right)$$

3) Softplus activat^n : $f(x) = \frac{1}{\beta} \cdot \ln\left(1 + e^{\beta x}\right)$

$$\frac{\partial f}{\partial x} = \frac{1}{\cancel{\beta}} \times \frac{1}{(1+e^{\beta x})} \times e^{\beta x} \times \cancel{\beta}$$

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

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

Q2 Solve gradients & do a shape check:

1) $f(\vec{x}) = \vec{x}^T(\vec{A}\vec{x} + \vec{z})$

Assuming $\dim(\vec{x}) = n \times 1$
$\dim(\vec{A}) = n \times n$
$\dim(\vec{z}) = n \times 1$

Shape check: $\dim(\bar{A}\bar{x}) = n \times 1$
$\dim(\bar{x}^T) = 1 \times n$
$\dim(\bar{A}\bar{x} + \bar{z}) = n \times 1$
$\Rightarrow \dim\left[\bar{x}^T(\bar{A}\bar{x} + \bar{z})\right] = 1 \times 1 = \text{scalar}$

Simplifying the expression:

$$f(x) = \bar{x}^T(\bar{A}\bar{x} + \bar{z})$$

$$= \bar{x}^T.\bar{A}\bar{x} + \bar{x}^T.\bar{z}$$

Given that $\bar{x}$ and $\bar{z}$ are both column vectors of dim $(n \times 1)$, we can write:

$$f(x) = \underbrace{\bar{x}^T\bar{A}\bar{x}}_{U} + \underbrace{\bar{z}^T\bar{x}}_{V}$$ (both scalars)

$$\frac{\partial V}{\partial \bar{x}} = \underbrace{\bar{z}}_{n\times 1} \quad \left(\because \partial(\bar{a}^T\bar{x})/\partial\bar{x} = \bar{a}\right)$$ (denominator layout)

$$U = \bar{x}^T\bar{A}\bar{x}$$

Now, if

$$\bar{x}^T = [x_1 \; x_2 \; \cdots \; x_n]$$

$$\bar{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & & \cdot \\ \vdots & & & & \\ a_{n1} & a_{n2} & \cdots & & a_{nn} \end{bmatrix}$$

Then,

$$\bar{A}\bar{x} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \cdots a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \cdots a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 \cdots \end{bmatrix}$$

(dim $= n \times 1$)

$$\Rightarrow \bar{x}^T\bar{A}\bar{x} = x_1(a_{11}x_1 + a_{12}x_2 + \cdots) + x_2(a_{21}x_1 + a_{22}x_2 + \cdots) + \cdots$$

$$\Rightarrow \quad \bar{x}^T \bar{A}\, \bar{x} = x_1\left(\sum_{i=1}^{n} a_{1i} x_i\right) + x_2\left(\sum_{i=1}^{n} a_{2i} x_i\right) + \cdots + x_n\left(\sum_{i=1}^{n} a_{ni} x_i\right)$$

(dim = 1×1 = scalar)

$$= \sum_{j=1}^{n} x_j\left(\sum_{i=1}^{n} a_{ji} x_i\right)$$

$$\Rightarrow \quad \underbrace{\frac{\partial u}{\partial \bar{x}}}_{n\times 1} = \frac{\partial\left(\bar{x}^T \bar{A}\, \bar{x}\right)}{\partial \bar{x}} = \begin{bmatrix} \partial u / \partial x_1 \\ \partial u / \partial x_2 \\ \vdots \\ \partial u / \partial x_n \end{bmatrix}_{n\times 1}$$

(denominator layout)

(dim = n × 1)

$$= \begin{bmatrix} (a_{11}x_1 + a_{12}x_2 + \cdots a_{1n}x_n) + (a_{11}x_1 + a_{21}x_2 + a_{31}x_3 + \cdots a_{n1}x_n) \\ (a_{21}x_1 + a_{22}x_2 + \cdots a_{2n}x_n) + (a_{12}x_1 + a_{22}x_2 + a_{32}x_3 + \cdots) \\ \vdots \\ \vdots \end{bmatrix}$$

$$= \bar{A}\,\bar{x} + A^T \bar{x}$$

$$= \underbrace{\underbrace{\left(\bar{A} + A^T\right)}_{n\times n} \underbrace{\bar{x}}_{n\times 1}}_{n\times 1}$$

$$\therefore \quad \underbrace{\frac{\partial f}{\partial \bar{x}}}_{n\times 1} = \frac{\partial u}{\partial \bar{x}} + \frac{\partial v}{\partial \bar{x}} = \underbrace{\left(\bar{A} + \bar{A}^T\right)\bar{x}}_{n\times 1} + \underbrace{\bar{z}}_{n\times 1}$$

2) L2 loss : $L(\omega) = \frac{1}{2}(\underbrace{\bar{\omega}^T\bar{x} - y}_{})^2$

$u = $ scalar

Using denominator layout.

$\dim(y) = 1\times 1 = $ scalar
$\dim(\bar{x}) = n\times 1$
$\dim(\bar{\omega}) = n\times 1$

$$\underbrace{\frac{\partial L}{\partial \bar{\omega}}}_{n\times 1} = \frac{1}{\cancel{2}} \times \cancel{2} \times u \times \frac{\partial u}{\partial \bar{\omega}}$$

$$\frac{\partial u}{\partial \bar{\omega}} = \frac{\partial(\bar{\omega}^T\bar{x} - y)}{\partial \bar{\omega}}$$

$$= \frac{\partial(\bar{\omega}^T\bar{x})}{\partial \bar{\omega}}$$

$$= \frac{\partial(\bar{x}^T\bar{\omega})}{\partial \bar{\omega}}$$

(given $\bar{\omega}$ & $\bar{x}$ are $n\times 1$ column vectors)

$$= \bar{x}$$

$\left(\because \frac{\partial(\bar{a}^T\bar{x})}{\partial \bar{x}} = \bar{a}\right)$

$$\Rightarrow \underbrace{\frac{\partial L}{\partial \bar{\omega}}}_{n\times 1} = u\,\bar{x} = \underbrace{\underbrace{(\bar{\omega}^T\bar{x} - y)}_{\text{scalar}}\underbrace{\bar{x}}_{n\times 1}}_{n\times 1}$$

3) L2 Loss (multiple examples)

$$L(\omega) = \frac{1}{2}\|\bar{X}\bar{\omega} - \bar{y}\|^2$$

$\dim(y) = n\times 1$
$\dim(\bar{\omega}) = m\times 1$
$\dim(\bar{X}) = n\times m$

$$= \frac{1}{2}(\bar{X}\bar{\omega} - \bar{y})^T\underbrace{(\bar{X}\bar{\omega} - \bar{y})}_{u}$$

$\dim(\bar{u}) = n\times 1$

$$= \frac{1}{2}\bar{u}^T\bar{u}$$

$(\dim(L) = 1\times 1 = $ scalar$)$

$$\Rightarrow \underbrace{\frac{\partial L}{\partial \bar{w}}}_{m\times 1} = \underbrace{\frac{\partial \bar{U}}{\partial \bar{w}}}_{m\times n} \times \underbrace{\frac{\partial L}{\partial \bar{U}}}_{n\times 1}$$

$$= \frac{\partial(\bar{X}\bar{w} - \bar{y})}{\partial \bar{w}} \times \frac{1}{\cancel{2}} \times \cancel{2} \times \bar{U}$$

$$= \frac{\partial(\bar{X}\bar{w})}{\partial \bar{w}} \times (\bar{X}\bar{w} - \bar{y})$$

$$= \underbrace{\underbrace{\bar{X}^T}_{m\times n}\underbrace{(\bar{X}\bar{w} - y)}_{n\times 1}}_{m\times 1} \quad \left(\because \frac{\partial(\bar{A}\bar{x})}{\partial \bar{x}} = \bar{A}^T\right)$$

[using denominator layout]

Q.3 Solve for gradient $\frac{\partial L}{\partial \bar{w}}$ & do shape check

$$\bar{z} = \bar{W}\bar{x} + \bar{b}$$

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

$\dim(\bar{y}) = m \times 1$

$\dim(\bar{b}) = m \times 1$

$\dim(\bar{x}) = n \times 1$

$\dim(\bar{W}) = m \times n$

$$\left(\bar{w}_i = \underbrace{[w_{i1} \; w_{i2} \; w_{i3} \; \cdots \; w_{in}]}_{1 \times n} \right)$$

$$\bar{z} = \bar{W}\bar{x} + \bar{b}$$

let $$\bar{W} = \begin{bmatrix} \text{———} & \bar{w}_1 & \text{———} \\ \text{———} & \bar{w}_2 & \text{———} \\ & \vdots & \\ \text{———} & \bar{w}_m & \text{———} \end{bmatrix}$$

$$\bar{z} = \begin{bmatrix} \bar{w}_1 \bar{x} + b_1 \\ \bar{w}_2 \bar{x} + b_2 \\ \vdots \\ \bar{w}_m \bar{x} + b_m \end{bmatrix}_{m \times 1}$$

$$\bar{z} - \bar{y} = \begin{bmatrix} \bar{w}_1 \bar{x} + b_1 - y_1 \\ \bar{w}_2 \bar{x} + b_2 - y_2 \\ \vdots \\ \bar{w}_m \bar{x} + b_m - y_m \end{bmatrix}_{m \times 1}$$

$$L = \|\bar{z} - \bar{y}\|^2 = \sum_{i=1}^{m} \left(\bar{w}_i \bar{x} + b_i - y_i \right)^2$$

where $\bar{w}_i \bar{x} = w_{i1} x_1 + w_{i2} x_2 + w_{i3} x_3 \cdots w_{in} x_n$

$$\underbrace{\frac{\partial L}{\partial \bar{w}}}_{m \times n} = \begin{bmatrix} \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{12}} & \cdots & \frac{\partial L}{\partial w_{1n}} \\ \partial L / \partial w_{21} & \cdots & \cdots & \frac{\partial L}{\partial w_{2n}} \\ \vdots & & & \\ \frac{\partial L}{\partial w_{m1}} & \cdots & \cdots & \frac{\partial L}{\partial w_{mn}} \end{bmatrix}_{m \times n}$$

Solving just for

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial \left(\sum_{i=1}^{m} (\bar{w}_i \bar{x} + b_i - y_i)^2 \right)}{\partial w_{11}}$$

$$= \frac{\partial \left[(\bar{w}_1 \bar{x} + b_1 - y_1)^2 \right]}{\partial w_{11}}$$

$$= 2 (\bar{w}_1 \bar{x} + b_1 - y_1) \times \frac{\partial (\bar{w}_1 \bar{x} + b_i - y_i)}{\partial w_{11}}$$

$$= 2 (\bar{w}_1 \bar{x} + b_1 - y_1) x_1$$

Similarily:

$$\frac{\partial L}{\partial w_{12}} = 2 (\bar{w}_1 \bar{x} + b_1 - y_1) x_2$$

$\Rightarrow$ in general

$$\frac{\partial L}{\partial w_{ij}} = 2 (\bar{w}_i \bar{x} + b_i - y_i) x_j$$

which can be rewritten as:

$$\boxed{\underbrace{\frac{\partial L}{\partial \bar{W}} = 2 \left[\underbrace{\underbrace{\bar{W}}_{m \times n} \underbrace{\bar{x}}_{n \times 1}}_{m \times 1} + \underbrace{\bar{b} - \bar{y}}_{m \times 1} \right] \underbrace{\bar{x}^T}_{1 \times n}}_{m \times n}}$$

Q.4

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

sample : $x = 1$, $y = 100$

$w_0 = 0$

$$\text{Loss } f^n = L(w) = \frac{1}{2}(wx - y)^2$$ (∵ only one example)

$$\frac{\partial L}{\partial w} = \frac{1}{2} \times 2(wx - y) \times x$$

$$= (wx - y)x$$

1) Doing gradient descent

$$w_{i+1} = w_i - \alpha \left(\frac{\partial L}{\partial w}\right)\Big|_{w_i}$$

$\alpha = 0.5$

$$w_1 = 0 - (0.5) \times (-100)$$
$$= 50$$
$$w_2 = 50 - (0.5)(-50)$$
$$= 75$$
$$w_3 = 75 - (0.5)(-25)$$
$$= 100$$
$$w_4 = 100 - (0.5)(0)$$
$$= 100$$

$w_4 = w_3$

slope $\left|\partial L / \partial w\right|_{w=100} = 0$

$\Rightarrow$ GD converges

$\alpha = 1.5$

$$w_1 = 0 - 3/2 \times (-100)$$
$$= 150$$
$$w_2 = 150 - \frac{3}{2} \times (50)$$
$$= 75$$
$$w_3 = 75 - \frac{3}{2}(-25)$$
$$= 112.5$$
$$w_4 = 112.5 - 3/2\,(12.5)$$
$$= 93.75$$
$$w_5 = 93.75 - \frac{3}{2} \times (-6.25)$$
$$= 103.125$$

⋮

slope $|\partial L / \partial w|$ keeps getting smaller, suggesting we are moving closer to an optima.

$\alpha = 2.5$

$$w_1 = 0 - \frac{5}{2} \times (-100)$$
$$= 250$$
$$w_2 = 250 - \frac{5}{2}(150)$$
$$= -125$$
$$w_3 = -125 - 5/2\,(-225)$$
$$= 687.5$$
$$w_4 = 687.5 - \frac{5}{2}(587.5)$$
$$= -781.25$$

slope $\partial L / \partial w$ oscillates b/w +ve & -ve with increasing magnitudes

This suggests α is too large & GD won't converge here.

2) We have, from Q4.1,

$$\frac{\partial L}{\partial w} = (wx - y)x$$

$$\Rightarrow w_{i+1} = w_i - \alpha \left.\frac{\partial L}{\partial w}\right|_{w_i} \quad (i = \text{iterat}^n \text{ no.})$$

$$= w_i - \alpha (w_i x - y) x$$

for the solo exaple of $x = 1$ & $y = 100$

$$w_{i+1} = w_i - \alpha(w_i - 100)$$

$$\Rightarrow w_{i+1} = w_i(1-\alpha) + 100\alpha$$

for GD to converge, we must have $|\partial L/\partial w|$ decreasing through each iterat^n.

$$\Rightarrow \left| \left.\frac{\partial L}{\partial w}\right|_{w_{i+1}} \right| < \left| \left.\frac{\partial L}{\partial w}\right|_{w_i} \right|$$

$$\Rightarrow |(w_{i+1} x - y)x| < |(w_i x - y)x|$$

$$\Rightarrow |w_{i+1} - 100| < |w_i - 100|$$

$$\Rightarrow |w_i(1-\alpha) + 100\alpha - 100| < |w_i - 100|$$

$$\Rightarrow |(1-\alpha)(w_i - 100)| < |w_i - 100|$$

for $w_i \neq 100$,

$$|1-\alpha| < 1$$

$$\Rightarrow -1 < 1 - \alpha < 1$$

$$\Rightarrow -2 < -\alpha < 0$$

$$\Rightarrow \boxed{2 > \alpha > 0}$$

We note that $w_i = 100$ is the optima for the curve $\left(\frac{\partial L}{\partial w} = 0\right)$, at which point GD converges anyway.

Therefore,

GD converges for $\alpha \in (0, 2)$

GD diverges for $\alpha \in (-\infty, 0) \cup (2, \infty)$

And for $\alpha = \{0, 2\}$,

$\alpha = 0$	$\alpha = 2$
$w_0 = t$	$w_0 = t$
$w_1 = t(1-\alpha) + 100\alpha$	$w_1 = t(1-\alpha) + 100\alpha$
$= t \quad (= w_0)$	$= -t + 200$
$w_2 = t \quad (= w_0)$	$w_2 = (200-t)(1-2) + 200$
$\vdots$	$= t \quad (= w_0)$
	$\vdots$

Thus, for $\alpha \in \{0, 2\}$, GD oscillates & neither converges nor diverges.

3) The above part proves conditions for α for any stopping threshold ε $(|w_{i+1} - w_i|)$.