A0248330L

No.

Date

CS5242 : LECTURE 1 HOMEWORK

Question 1

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

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

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

$$= \frac{e^{-x}}{(1+e^{-x})^2}$$ [Add, subtract 1 in numerator]

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

$$= f(x) - [f(x)]^2$$

$$= f(x)[1 - f(x)]$$

Sigmoid derivative: $\frac{e^{-x}}{(1+e^{-x})^2} = f(x)[1-f(x)]$

where $f(x)$ is the Sigmoid function.

A0248330L

No.

Date

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

Computing the derivative for an arbitrary x_k

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

we know $\partial/\partial x_k \left(\sum_{j=1}^{n} e^{x_j}\right) = e^{x_k}$

If $i = k$, then we will have $\partial/\partial x_k (e^{x_i}) = e^{x_k}$ otherwise 0.

Therefore, we consider the 2 cases:

When $i = k$: (For differentiating with respect to x_i)

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

$$= \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \cdot \frac{\sum_{j=1}^{n} e^{x_j} - e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \quad [k = i]$$

$$= f(x_i) \cdot (1 - f(x_i))$$

When $i \neq k$ (If we were to differentiate for other x_k)

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

Softmax derivative $=$ $f(x_i) \cdot (1 - f(x_i))$ if $i = k$
$-f(x_k) \cdot f(x_i)$ if $i \neq k$

for an arbitrary k where $f(x)$ is the softmax function.

A0248330L

No.

Date

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

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

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

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

$$= \frac{e^{\beta x}}{1+e^{\beta x}} \quad \text{[Dividing by } e^{\beta x}\text{]}$$

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

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

A0248330L

No.

Date

<u>Question 2</u>

1) $f(x) = x^T(Ax+z)$ where

$x \in R^{n\times 1}$, $A \in R^{n\times n}$, $z \in R^{n\times 1}$

$$\frac{\partial}{\partial x}(f(x)) = \frac{\partial}{\partial x}\left(x^T(Ax+z)\right)$$

$$= \frac{\partial}{\partial x}\left(x^TAx + x^Tz\right)$$

$$= \frac{\partial}{\partial x}\left(x^TAx\right) + \frac{\partial}{\partial x}\left(x^Tz\right) \quad [\text{Shapes } R^{1\times 1}]$$

Taking individual derivatives with respect to a generic x_k

$$\frac{\partial}{\partial x_k}\left(x^TAx\right) = \frac{\partial}{\partial x_k}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j\right]$$

We can separate this into terms when $i=j$ and $i \neq j$

$$= \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{n}\left(a_{ii}x_i^2 + \sum_{j\neq i}^{n} x_i a_{ij} x_j\right)\right)$$

$$= 2a_{kk}x_k + \sum_{j\neq k} x_j a_{jk} + \sum_{j\neq k} a_{kj}x_j$$

$$= \sum_{j=1}^{n} x_j a_{jk} + \sum_{j=1}^{n} a_{kj}x_j$$

Converting to a matrix form with k rows, we get

$$= A^Tx + Ax = (A^T+A)x \quad \text{Shape } R^{n\times 1}$$

For the partial derivative of $x^T z$, we can look at a generic x_k

$$x^T z = \sum_{i=1}^{n} x_i z_i$$

$$\frac{\partial}{\partial x_k}(x^T z) = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{n} x_i z_i\right) = z_k$$

Thus in a matrix form with k rows, we get z
shape $R^{n \times 1}$

$$\frac{\partial}{\partial x}(f(x)) = (A^T + A)x + z \quad \text{with shape } R^{n \times 1}$$

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

where $w \in R^{n \times 1}$, $x \in R^{n \times 1}$, $y \in R^{n \times 1}$, $L(w) \in R^{1 \times 1}$

$$\frac{\partial}{\partial w}(L(w)) = \frac{\partial}{\partial w}\left[\frac{1}{2}(w^T x - y)^2\right]$$

Using the chain rule:

$$= \frac{1}{2} \cdot 2 \cdot (w^T x - y) \cdot \frac{\partial}{\partial w}(w^T x)$$

$$= (w^T x - y) \cdot \frac{\partial}{\partial w}(w^T x)$$

$w^T x = \sum_{i=1}^{n} w_i x_i$, the derivative of which would be x as ~~seen~~ shown from the previous question

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

Since $(w^T x - y)$ is of shape $R^{1 \times 1}$ which implies it is a scalar, it can be multiplied with x of shape $R^{n \times 1}$

$$\frac{\partial}{\partial w}(L(w)) = (w^T x - y) \cdot x$$ with shape $R^{n \times 1}$

No.

Date

3) $L(w) = \frac{1}{2m} \| Xw - y \|^2$ where

$X \in R^{n \times n}$, $w \in R^{n \times 1}$, $y \in R^{n \times 1}$ [Using vectorization]

$$\frac{\partial}{\partial w}(L(w)) = \frac{\partial}{\partial w}\left(1/2m \, \|Xw - y\|^2 \right)$$

Let $\tilde{y} = Xw$ with shape $R^{n \times 1}$

$$= \frac{\partial}{\partial w}\left(\frac{1}{2m} \|\tilde{y} - y\|^2 \right)$$

$$= \frac{1}{2m} \frac{\partial}{\partial w}\left((\tilde{y} - y)^T (\tilde{y} - y) \right)$$

Let $u = \tilde{y} - y$ with shape $R^{n \times 1}$

$$= \cancel{\frac{1}{2m}\left(\frac{\partial}{\partial w}\left(Xw \right.\right.}$$

$$\frac{\partial u}{\partial w} = \frac{\partial}{\partial w}(\tilde{y} - y) = \frac{\partial (Xw - y)}{\partial w} = X^T$$

$$\frac{\partial}{\partial w}(L(w)) = \frac{1}{2m}\left(\frac{\partial}{\partial w}(u^T u) \right)$$

$$= \frac{1}{2m}\left(\frac{\partial u}{\partial w} \cdot u + \frac{\partial u}{\partial w} \cdot u \right) = \frac{1}{m} \cdot X^T u$$

$$\frac{\partial}{\partial w}(L(w)) = \frac{1}{m} \cdot X^T (\tilde{y} - y)$$ with shape $R^{n \times 1}$

A0248330L

No.

Date

Question 3

$$z = Wx + b \quad \text{where} \quad W \in R^{n \times n} \quad b \in R^{n \times 1}$$
$$x \in R^{n \times 1} \quad z \in R^{n \times 1}$$

$$L = \|z - y\|^2 \quad \text{where} \quad y \in R^{n \times 1} \quad L \in R^{1 \times 1}$$

$$\frac{\partial L}{\partial W} = \frac{\partial}{\partial W} \|z - y\|^2$$

$$= \frac{\partial}{\partial W} (z-y)^T (z-y) \qquad \text{Let } u = z - y \text{ with shape } R^{n \times 1}$$

$$= \frac{\partial}{\partial W} (u^T u)$$

$$= \cancel{\frac{\partial}{\partial W}} \left(\frac{\partial u}{\partial W} \cdot u + \frac{\partial u}{\partial W} \cdot u \right)$$

$$= 2u \cdot \frac{\partial u}{\partial W} = 2u \cdot \frac{\partial}{\partial W}(z - y)$$

$$= 2u \cdot \frac{\partial}{\partial W}(Wx + b - y)$$

$$= 2u \left(\frac{\partial}{\partial W}(Wx) + 0 \right) \quad \left[\partial/\partial W(Wx) = x^T \text{ due to vectorization} \right]$$

$$= 2ux^T$$

$$\frac{\partial L}{\partial W} = 2ux^T \quad \text{with shape } R^{n \times n}$$

A0248330L

No.

Date

Question 4

Linear Regression $y = xw$

Loss L2 => $L(w) = \frac{1}{2}(w^T x - y)^2$

Initial $w=0$, $x=1$, $y=100$.

1) Learning rate $\alpha = 0.5$

$\left[\frac{\partial J(w)}{\partial(w)} = \frac{\partial L(w)}{\partial w} = (w^T x - y) \cdot x\right]$

i) $\partial J/\partial w = -100$

$w_1 = 0 - 0.5 * (-100) = 50$ $[w = w - \alpha(\partial J/\partial w)]$

ii) $\partial J/\partial w = (50-100) * 1 = -50$

$w_2 = 50 - 0.5 * (-50) = 75$

iii) $\partial J/\partial w = (75-100) * 1 = -25$

$w_3 = 75 - 0.5 * (-25) = 87.5$

iv) $\partial J/\partial w = (87.5-100) * 1 = -12.5$

$w_4 = 87.5 - 0.5 * (-12.5) = 93.75$

v) $\partial J/\partial w = (93.75-100) * 1 = -6.25$

$w_5 = 93.75 - 0.5 * (-6.25) = 96.875$

Learning rate $\alpha = 1.5$

i) $\partial J/\partial w = -100$

$w_1 = 0 - 1.5 * (-100) = 150$

ii) $\partial J/\partial w = 150 - 100 = 50$

$w_2 = 150 - 1.5 * (50) = 75$

iii) $\partial J/\partial w = 75 - 100 = -25$

$w_3 = 75 - 1.5 * (-25) = 112.5$

iv) $\partial J/\partial w = 112.5 - 100 = 12.5$

$w_4 = 112.5 - 1.5 * (12.5) = 93.75$

v) $\partial J/\partial w = 93.75 - 100 = -6.25$

$w_5 = 93.75 - 1.5 * (-6.25) = 103.125$

Learning Rate $\alpha = 2.5$

i) $\partial J / \partial w = 0 - 100 = -100$
$w_1 = 0 - 2.5 * (-100) = 250$

ii) $\partial J / \partial w = 250 - 100 = 150$
$w_2 = 250 - 2.5 * (150) = -125$

iii) $\partial J / \partial w = -125 - 100 = -225$
$w_3 = -125 - 2.5 * (-225) = 437.5$

iv) $\partial J / \partial w = 437.5 - 100 = 337.5$
$w_4 = 437.5 - 2.5 * (337.5) = -406.25$

v) $\partial J / \partial w = -406.25 - 100 = -506.25$
$w_5 = -406.25 - 2.5 * (-506.25) = 859.375$

Learning Rate \ Weight Iteration	0	1	2	3	4	5
0.5	0	50	75	87.5	93.75	96.875
1.5	0	150	75	112.5	93.75	103.125
2.5	0	250	-125	437.5	-406.25	859.375

Learning Rate	Converges?	Oscillates?
0.5	Yes	No
1.5	Yes	Yes
2.5	No	Yes

A0248330L

No.

Date

2) For the given linear regression problem, we can see that the gradient descent can converge as long as the alpha i.e. learning rate is not too large.

We use the stopping criteria when distance between w and optimal point is < 1.

From our expeiments, it seems

$0 \le a \le 1 \Rightarrow$ gradient descent converges without oscillating.

~~⇒ the higher the~~

~~$1 \le a \le 2 \Rightarrow$ gradient descent~~

$1 < a < 2 \Rightarrow$ gradient descent converges, but oscillates

$a \geq 2 \Rightarrow$ gradient descent does not converge and oscillates.

No.

Date

3) In the following given linear regression problem, the weight update is as follows:

$$w_{t+1} = w_t + \Delta w \quad \text{where}$$

$$\Delta w = -\alpha \frac{\partial L}{\partial w} = -\alpha \left(w_t^T x - y\right) \cdot x$$

Substituting values, we have

$$\Delta w = \alpha \left[100 - w_t\right]$$

The quantity $|100 - w_t|$ gives us the magnitude by which the weight has to shift along the gradient descent curve towards the optimal point.

With a value of $0 \leq \alpha \leq 1$, the weight change happens in the direction towards the optimal point without overshooting in the other direction. Thus it converges without oscillating.

When $1 < \alpha < 2$, the weight change moves to the other side of the gradient descent curve, by a magnitude smaller than $|100 - w_t|$ after the update. ~~After~~ Thus it oscillates, but converges.

When $\alpha \geq 2$, then the magnitude of the new weight grows larger than the original, thus exploding the gradient descent and not converging.