Kenny Ostyn Hanson

16-720

HW 5

Q1.1 Prove that softmax is invariant to translation

$$\text{softmax}(x_i) = \text{softmax}(x+c) \quad \forall c \in \mathbb{R}$$

$$\text{softmax}(x+c) = \left(\frac{e^{x_i+c}}{\sum_j e^{x_j+c}} \right)_i$$

$$= \left(\frac{e^{x_i} e^{c}}{\sum_j e^{x_j} e^{c}} \right)_i$$

$$= \left(\frac{e^{x_i}}{\sum_j e^{x_j}} \right)_i \equiv \text{softmax}(x_i) \quad \text{(Proven)}$$

softmax(x) is invariant to translation

For $c=0$, $e^{x_i+c} = e^{x_i} : (0 \to +\infty)$

For $c = -\max x_i$ $e^{x_i+c} : (0 \to 1)$

Using $c = -\max x_i$ will simplify the process, avoiding overflow.

Q1.2 Softmax → 3 step process

- Range for each element : 0 - 1
  Sum for all elements : 1

- Softmax takes an arbitary real valued vector x and turns it into a probability distribution of values in x.

- ① $S_i = e^{x_i}$ → Calculate the exponential value of each element
  ② $S = \sum S_i$ → Calculate the sum of exponential values
  ③ $\text{softmax}(x_i) = \frac{1}{S} S_i$ → Calculate the probability distribution of the values.

Q1.4 Derive gradient of sigmoid function

Show that gradient $\rightarrow \sigma(x)$ (without accessing x)

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

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

$$= \boxed{\frac{1}{e^{x}(1+e^{-x})^{2}}} \quad \text{(Derived)}$$

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

$$\boxed{\therefore \sigma'(x) = \sigma(x) - \sigma^{2}(x)} \quad \text{(shown)}$$

Q1.5 Show in getting $\frac{dy}{dw}$, $\frac{dJ}{dx}$ & $\frac{dJ}{db}$ ⇝ In scalars & reform matrix

$y = Wx + b$ or $y_i = \sum_{j=1}^{d} x_j W_{ij} + b_i$

$\frac{dJ}{dy} = \delta \in \mathbb{R}^{k \times 1}$ $\quad W \in \mathbb{R}^{k \times d}$ $\quad x \in \mathbb{R}^{d \times 1}$ $\quad b \in \mathbb{R}^{k \times 1}$

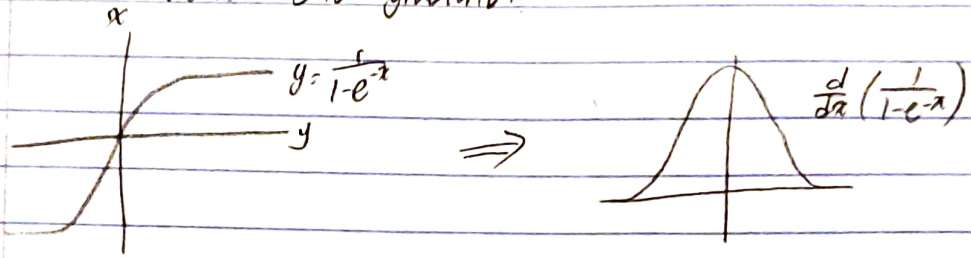
$\frac{dy_i}{dw_{ij}} = x_j \Rightarrow \frac{dJ}{dw} = \frac{dJ}{dy} \cdot \frac{dy}{dw} = \underbrace{\delta x_j}_{\substack{k \times 1 \quad 1 \times d \\ d \times k \text{ or } k \times d}} = \begin{bmatrix} \delta_1 x_1 & \cdots & \delta_k x_1 \\ \vdots & & \\ \delta_1 x_d & & \delta_k x_d \end{bmatrix} \in \mathbb{R}^{d \times k}$

$\frac{dJ}{dx} = \frac{dJ}{dy} \cdot \frac{dy}{dx} = \underbrace{\delta \cdot W_{ij}}_{\substack{k \times 1 \quad k \times d \\ d \times 1}} = \begin{bmatrix} \sum_{j=1}^{k} \delta_j W_{1j} \\ \vdots \\ \sum_{j=1}^{k} \delta_j W_{dj} \end{bmatrix} \in \mathbb{R}^{d \times 1}$

$\frac{dJ}{db} = \frac{dJ}{dy} \cdot \frac{dy}{db} = \underset{k \times 1}{\delta} \cdot \underset{k \times k}{1} = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_k \end{bmatrix} \in \mathbb{R}^{k \times 1}$

↓ const

Q1.6 1. The sigmoid activation function poses to have the "vanishing gradient" problem that can be explained through the plot. As more layers use it, the graph plateaus indicating an eventual zero gradient.



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

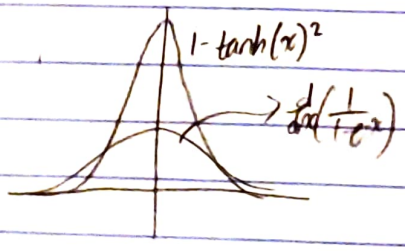
$= 1 - \frac{2e^{-2x}}{1+e^{-2x}}$

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

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

Sigmoid $= \frac{1}{1-e^{-x}} \in (0, 1)$

Tanh has a symmetric output until -1 unlike sigmoid. This can help eliminate bias in the gradient. Also, tanh is less likely to have the "vanishing gradient" problem because of its more significant gradient.

3. 

tanh has higher derivative values than sigmoid, which reduces the vanishing problem chance.

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

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

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

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

$= -(1 - 2\sigma(2x))$

$= \frac{1}{2\sigma(2x)-1}$

1. HELLO → good

HELLO → bad

HELLO ← HELLO

2. HELLO → good

HELLO → bad