Title

Author

June 25, 2023

1 Theory

1.1 Question A

$$-y \times log(\hat{y}) - (1-y) \times log(1-\hat{y}) = -log\ P(O=o|C=c)$$

where y is the probability of getting o after c.
this as you can see is cross entropy loss, thats because, we already know that $y = 1$, and $\hat{y}$ is the predicted value of the probability of gettin o. so the naive softmax loss is just a simplified version of the cross entropy loss

1.2 Question B

so

$$J(v_c) = -logP(v_c)$$
$$P(v_c) = softmax(x_i)$$
$$x_i = u_i * v_c$$

so differentiating

$$\frac{\partial J}{\partial v_c} = \frac{\partial J}{\partial P} \times \frac{\partial P}{x_o} \times \frac{\partial x_o}{v_c}$$

so we get

$$\frac{\partial J}{\partial v_c} = \frac{1}{P(v_c)} \times softmax_i \cdot (1\{i=j\} - softmax_j) \times u_o^T$$

now using the shape convention u_o^T should be transposed again, thus the answer is

$$\frac{\partial J}{\partial v_c} = \left(\frac{1}{P(v_c)} \times softmax_i \cdot (1\{i=j\} - softmax_j)\right) \odot u_o$$

1.3 Question C

$$J(v_c) = -logP(v_c)$$
$$P(v_c) = softmax(x_i)$$
$$x_i = u_w * v_c$$

so for some u_w where $w = o$ differentiating as in this case the softmax function only has u_w in its denominator and numerator, this case is similar to the previous case

$$\frac{\partial J}{\partial u_w} = \frac{\partial J}{\partial P} \times \frac{\partial P}{x_o} \times \frac{\partial x_o}{u_w}$$
$$\frac{\partial J}{\partial u_w} = \frac{1}{P(v_c)} \times softmax_i \cdot (1\{i=j\} - softmax_j) \times v_c$$
$$\frac{\partial J}{\partial v_c} = \left(\frac{1}{P(v_c)} \times softmax_i \cdot (1\{i=j\} - softmax_j)\right) \odot v_c$$

.

now for the case where $w \neq o$ in this case the softmax has u_w in the denominator

$$\frac{\partial J}{\partial u_w} = \frac{\partial J}{\partial P} \times \frac{\partial P}{x_o} \times \frac{\partial x_o}{u_w}$$
$$\frac{\partial J}{\partial u_w} = \frac{1}{P(v_c)} \times softmax_i \times -\frac{1}{\sum_j exp(x_j)} \times v_c$$

1.4 Question D

1.5 Question E

$$f(x) = max(0, x)$$

so for $x > 0$

$$f'(x) = \frac{dx}{dx} = 1$$

and for $x < 0$

$$f'(x) = \frac{d0}{dx} = 0$$

1.6 Question F

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

using product rule

$$\frac{d\sigma(x)}{dx} = e^x \times \frac{d}{dx}\frac{1}{1+e^x} + \frac{1}{1+e^x} \times \frac{d}{dx}e^x$$

$$\frac{d\sigma(x)}{dx} = e^x \times \frac{d}{dx}(1+e^x)^- + \frac{1}{1+e^x} \times e^x$$

$$\frac{d\sigma(x)}{dx}\sigma(x) = e^x \times \frac{1}{-(1+e^x)^2} \times e^x + \frac{e^x}{1+e^x}$$

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

$$\frac{d\sigma(x)}{dx} = \sigma(x) - \sigma(x)^2$$

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

1.7 Question G

1.8 Question H

1.9 Question I