CS224N$_A$1

Yinwei Dai

June 2019

1 Question 1

(a)

the cross entropy loss between the true (dicrete) probability p and another distribution q is $-\sum_i p_i log(q_i)$
we can know that

$$y_w = 0 w! = o$$

$$y_w = 1 w = 0$$

therefore

$$-\sum_{w=1}^{V} y_w log(\hat{y_w}) = y_o log(\hat{y_0}) = log(\hat{y_0})$$

(b)

$$\frac{\partial J_{naive_softmax}(v_c, o, U)}{\partial v_c} = -\frac{\partial log(P(O = 0|C = c))}{\partial v_c} = -\frac{u_0^T v_c}{\partial v_c} + \frac{log(\sum_{w \in Vocab} exp(u_w^T v_c))}{\partial v_c} =$$

$$-u_o + \sum_{w=1}^{V} \frac{exp(u_w^T v_c)}{\sum_{w=1}^{V} exp(u_w^T v_c)} u_w = -u_0 + \sum_{w=1}^{V} log(P(O = w|C = c) u_w = U^T(\hat{y} - y)$$

(c)

$$\frac{\partial J_{naive_softmax}(v_c, o, U)}{\partial u_w} = -\frac{u_0^T v_c}{\partial u_w} + \frac{log(\sum_{w \in Vocab} exp(u_w^T v_c))}{\partial u_w}$$

when w = 0, we can get

$$\frac{\partial J_{naive_softmax}(v_c, o, U)}{\partial u_w} = -v_c + \frac{exp(u_o^T v_c)}{\sum_{w=1}^{V} exp(u_w^T v_c)} v_c = (P(O = w|C = c) - 1) v_c$$

when w ! = 0, we have

$$\frac{\partial J_{naive_softmax}(v_c, o, U)}{\partial u_w} = \frac{exp(u_w^T v_c)}{\sum_{m=1}^{V} exp(u_m^T v_c)} v_c = (P(O = w|C = c) v_c$$

Above all, we can draw a conclusion

$$\frac{\partial J_{naive_softmax}(v_c, o, U)}{\partial u_w} = (\hat{y} - y)^T v_c$$

(d)

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

(e)

$$\frac{\partial J_{nag-sample}(v_c, o, U)}{\partial v_c} = -(1 - \sigma(u_o^T v_c))u_o + \sum_{k=1}^{K}(1 - \sigma(u_k^T v_c))u_o)$$

$$\frac{\partial J_{nag-sample}(v_c, o, U)}{\partial u_0} = \frac{\partial(-log(\sigma(u_o^T v_c)))}{\partial u_o} = -(1 - \sigma(u_o^T v_c))v_c$$

$$\frac{\partial J_{nag-sample}(v_c, o, U)}{\partial u_k} = \frac{\partial(-log(\sigma(u_k^T v_c)))}{\partial u_k} = (1 - \sigma(u_k^T v_c))v_c$$

(f)
(i)

$$\partial J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)/\partial U = \sum_{-m \le j \le m, j!=0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

(ii)

$$\partial J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)/\partial v_c = \sum_{-m \le j \le m, j!=0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

(iii)

$$\partial J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)/\partial v_w = 0$$

2 Question 2

(a)
(b)
(c)