עיבוד שפה טבעית, שיעורי בית 1!

(1

(a

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

$$\frac{e^{x_i} \cdot e^{c}}{\sum_j e^{x_j} \cdot e^{c}} = \frac{e^{c} \cdot e^{x_i}}{e^{c} \sum_j e^{x_j}} = \frac{e^{x_i}}{\sum_j e^{x_j}} = \text{Softmax}(x);$$

As a result Softmax(x) = Softmax(x+c)
(it's the same math just on multiple coordinates).

(c

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

$$\sigma'(x) = \frac{0 - -e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}+1-1}{(1+e^{-x})^2} =$$

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

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

So: $\sigma'(x) = \sigma(x)(1-\sigma(x))$.

(2

$$CE(y,\hat{y}) = -\sum y_i \log(\hat{y}_i) \quad (a$$

$$\frac{\partial CE(y,\hat{y})}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log\left(\frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)}\right)\right) =$$

$$\frac{\partial}{\partial v_c}\left(-\log(\exp(u_o^T v_c)) - -\log\left(\sum_w \exp(u_w^T v_c)\right)\right) =$$

$$\frac{\partial}{\partial v_c}\left(-u_o^T v_c + \log\left(\sum_w \exp(u_w^T v_c)\right)\right) =$$

$$-u_o + \frac{1}{\sum_w \exp(u_w^T v_c)} \cdot \sum_x \exp(u_x^T v_c) \cdot u_x =$$

$$\boxed{-u_o + \sum_x p(x|c) \cdot u_x}$$

(b)

Case 1: Vector y has a 1 at location u_o

$$\frac{\partial}{\partial u_o} CE(y, \hat{y}) = \frac{\partial}{\partial u_o} \left(-\log \left(\frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)} \right) \right) =$$

$$\frac{\partial}{\partial u_o} \left(-\log(\exp(u_o^T v_c)) - \log\left(\sum_w \exp(u_w^T v_c) \right) \right) =$$

$$\frac{\partial}{\partial u_o} \left(-u_o^T v_c + \log\left(\sum_w \exp(u_w^T v_c) \right) \right) =$$

$$-v_c + \frac{1}{\sum_w \exp(u_w^T v_c)} \cdot v_c \exp(u_o^T v_c) =$$

$$-v_c + v_c p(o|c) = v_c (p(o|c) - 1).$$

Case 2: Vector y has a 1 at location k such that $u_k \neq u_o$.

$$\frac{\partial}{\partial u_k} CE(y, \hat{y}) = \frac{\partial}{\partial u_k} \left(-\log \left(\frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)} \right) \right) =$$

$$\frac{\partial}{\partial u_k} \left(-\log \left(\exp(u_o^T v_c) \right) - -\log \left(\sum_w \exp(u_w^T v_c) \right) \right) =$$

$$\frac{\partial}{\partial u_k} \left(-u_o^T v_c + \log \left(\sum_w \exp(u_w^T v_c) \right) \right) =$$

$$0 + \frac{1}{\sum_w \exp(u_w^T v_c)} \cdot v_c \exp(u_k^T v_c) =$$

$$v_c \cdot \frac{\exp(u_k^T v_c)}{\sum_w \exp(u_w^T v_c)} = \boxed{v_c \, p(k|c)}$$

C

$$\frac{\partial J_{neg\text{-}sample}}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K}\log(\sigma(-u_k^T v_c))\right) =$$

$$-\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c)(1-\sigma(u_o^T v_c)) \cdot u_o$$

$$-\sum_{k=1}^{K}\frac{\sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c))}{\sigma(-u_k^T v_c)} \cdot -u_k =$$

$$u_o(\sigma(u_o^T v_c)-1) - \sum_{k=1}^{K}(\sigma(-u_k^T v_c)-1)u_k$$

$$\frac{\partial}{\partial u_o}\left(J_{neg\text{-}sample}\right) =$$

$$\frac{\partial}{\partial u_o}\left(-\log\left(\sigma(u_o^T v_c)\right) - \sum_{k=1}^{K} \log\left(\sigma(-u_k^T v_c)\right)\right) =$$

↑ has u_o u_o

$$\frac{\partial}{\partial u_o}\left(-\log\left(\sigma(u_o^T v_c)\right)\right) =$$

$$-\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c)\left(1-\sigma(u_o^T v_c)\right) \cdot v_c =$$

$$v_c\left(\sigma(u_o^T v_c) - 1\right)$$

$$\frac{\partial}{\partial u_w}\left(J_{neg\text{-}sample}\right) =$$

$$\frac{\partial}{\partial u_w}\left(-\log\left(\sigma(u_o^T v_c)\right) - \sum_{w=1}^{k} \log\left(\sigma(-u_w^T v_c)\right)\right) =$$

u_o has no u_w

$$\frac{\partial}{\partial u_w}\left(-\sum_{w=1}^{k} \log\left(\sigma(-u_w^T v_c)\right)\right) =$$

$$-\frac{1}{\sigma(-u_w^T v_c)}\,\sigma(-u_w^T v_c)\left(1-\sigma(-u_w^T v_c)\right)\cdot -v_c =$$

$$-v_c\left(\sigma(-u_w^T v_c)-1\right).$$

(4

$$J = \sum_{-M \le j \le M,\ j \neq 0} F(w_{c+j}, v_c)$$

$$\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}\left(\sum_{-M \le j \le M,\ j \neq 0} F(w_{c+j}, v_c)\right) =$$

$$\sum_{-M \le j \le M,\ j \neq 0} \frac{\partial F(w_{c+j}, v_c)}{\partial v_c}$$

$$\frac{\partial J}{\partial w_{c+i}} = \frac{\partial}{\partial w_{c+i}}\left(\sum_{-M \le j \le M,\ j \neq 0} F(w_{c+j}, v_c)\right) =$$

$$\frac{\partial F(w_{c+i}, v_c)}{\partial w_{c+i}} \quad -M \le i \le M,\ i \neq 0$$

(h

In our picture we can

see that good and bad

wonderful and great

boring and waste

cool and sweet

has resulted similar values in the 2d

PCA based on the knn cos distance

it makes sense as wonderful and great

are similar many words, good and bad

are opposites, and boring and waste

and cool and sweet could be

used in similar situations.