Natural Language Processing Written Homework

Hadi Sheikhi 97521369

a)

a. The true probabilities are y = p and estimated probabilities are vector $\hat{y}$ = q.

$$J_{CE} = -\sum_{w=1}^{|V|} p(w)\log q(w) = -p(o)\log q(o) = -\log\big(P(o|c)\big) = J_{naive-softmax}$$

b) I considered that dimensionality of y is (1, |V|). By discussion of the question, dimension of u and v is (d, 1) because they are columns of U and V matrices.

$$\frac{\partial}{\partial v_c} J_{naive-softmax} = -\frac{\partial}{\partial v_c}\log\big(P(o|c)\big) = -\frac{\partial}{\partial v_c}\log\frac{\exp(u_o^T.v_c)}{\sum_{w=1}^{|V|}\exp(u_w^T.v_c)}$$

$$= -\frac{\partial}{\partial v_c}\log\exp(u_o^T.v_c) + \frac{\partial}{\partial v_c}\log\sum_{w=1}^{|V|}\exp(u_w^T.v_c)$$

$$= -\frac{\partial}{\partial v_c}u_o^T.v_c + \frac{\partial}{\partial v_c}\log\sum_{w=1}^{|V|}\exp(u_w^T.v_c) = -u_o + \frac{\partial}{\partial v_c}\log\sum_{w=1}^{|V|}\exp(u_w^T.v_c)$$

$$= -u_o + \frac{\sum_{w=1}^{|V|}u_w\exp(u_w^T.v_c)}{\sum_{x=1}^{|V|}\exp(u_x^T.v_c)} = -u_o + \sum_{w=1}^{|V|}u_w\left(\frac{\exp(u_w^T.v_c)}{\sum_{x=1}^{|V|}\exp(u_x^T.v_c)}\right)$$

$$= -u_o + \sum_{w=1}^{|V|}u_w P(w\,|c) = -\underbrace{u_o}_{d\times 1} + \underbrace{U}_{d\times|V|}.\underbrace{\hat{y}^T}_{|V|\times 1} = -u_o + U\hat{y}^T$$

c)

case I: $w \neq o$

$$\frac{\partial}{\partial u_w} J_{naive-softmax} = -\frac{\partial}{\partial u_w}\log\frac{\exp(u_o^T.v_c)}{\sum_{k=1}^{|V|}\exp(u_k^T.v_c)}$$

$$= -\frac{\partial}{\partial u_w}\log\exp(u_o^T.v_c) + \frac{\partial}{\partial u_w}\log\sum_{k=1}^{|V|}\exp\big(u_k^T.v_c\big) = 0 + \frac{v_c\exp(u_w^T.v_c)}{\sum_{x=1}^{|V|}\exp(u_x^T.v_c)}$$

$$= \underbrace{v_c}_{d\times 1}.\underbrace{\hat{y}_w}_{1\times 1} = v_c.\hat{y}_w$$

case II: $w = o$

$$\frac{\partial}{\partial u_w} J_{naive-softmax} = -\frac{\partial}{\partial u_w}\log\frac{\exp(u_o^T.v_c)}{\sum_{k=1}^{|V|}\exp(u_k^T.v_c)} = -\frac{\partial}{\partial u_w}(u_w^T.v_c) + \frac{\partial}{\partial u_w}\log\sum_{k=1}^{|V|}\exp\big(u_k^T.v_c\big)$$

$$= -v_c + \frac{v_c\exp(u_w^T.v_c)}{\sum_{x=1}^{|V|}\exp(u_x^T.v_c)} = -\underbrace{v_c}_{d\times 1} + \underbrace{v_c}_{d\times 1}.\underbrace{\hat{y}_w}_{1\times 1} = -v_c + v_c.\hat{y}_w$$

d)

$$\frac{\partial}{\partial U} J_{naive-softmax} = \left[\frac{\partial J(v_c,o,U)}{\partial u_1} \quad \frac{\partial J(v_c,o,U)}{\partial u_2} \quad \frac{\partial J(v_c,o,U)}{\partial u_3} \quad \dots \quad \frac{\partial J(v_c,o,U)}{\partial u_{|VOCAB|}}\right]$$

e)

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

Hadi Sheikhi 97521369

f)

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

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

$$= -\frac{\sigma(u_o^T.v_c) - \sigma^2(u_o^T.v_c)}{\sigma(u_o^T.v_c)}\frac{\partial}{\partial v_c}(u_o^T.v_c)$$

$$-\sum_{k=1}^{K}\left(\frac{\sigma(-u_k^T.v_c) - \sigma^2(-u_k^T.v_c)}{\sigma(-u_k^T.v_c)}\frac{\partial}{\partial v_c}(-u_k^T.v_c)\right)$$

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

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

$$= -\frac{1}{\sigma(u_o^T.v_c)}\frac{\partial}{\partial u_o}\left(\sigma(u_o^T.v_c)\right) - 0 = -\frac{\sigma(u_o^T.v_c) - \sigma^2(u_o^T.v_c)}{\sigma(u_o^T.v_c)}\frac{\partial}{\partial u_o}(u_o^T.v_c) =$$

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

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

$$= 0 - \frac{1}{\sigma(-u_k^T.v_c)}\frac{\partial}{\partial u_k}\left(\sigma(-u_k^T.v_c)\right) = -\frac{\sigma(-u_k^T.v_c) - \sigma^2(-u_k^T.v_c)}{\sigma(-u_k^T.v_c)}\frac{\partial}{\partial u_k}(-u_k^T.v_c)$$

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

As we see, the most iteration we should go through is [1..K] but in naïve softmax we should iterate over all vocabulary.

g) Suppose S is the set of indices such that for $s \in S$, we have $w_s = w_k$, so we can write $u_k = u_s$

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

$$= -\sum_{s=1}^{|S|}\left(\frac{1}{\sigma(-u_k^T.v_c)}\frac{\partial}{\partial u_k}\left(\sigma(-u_k^T.v_c)\right)\right)$$

$$= -\sum_{s=1}^{|S|}\left(\frac{\sigma(-u_k^T.v_c) - \sigma^2(-u_k^T.v_c)}{\sigma(-u_k^T.v_c)}\frac{\partial}{\partial u_k}(-u_k^T.v_c)\right) = \sum_{s=1}^{|S|}\left(\left(1-\sigma(-u_k^T.v_c)\right)v_c\right)$$

$$= |S|\left(1-\sigma(-u_k^T.v_c)\right)v_c$$

Hadi Sheikhi 97521369

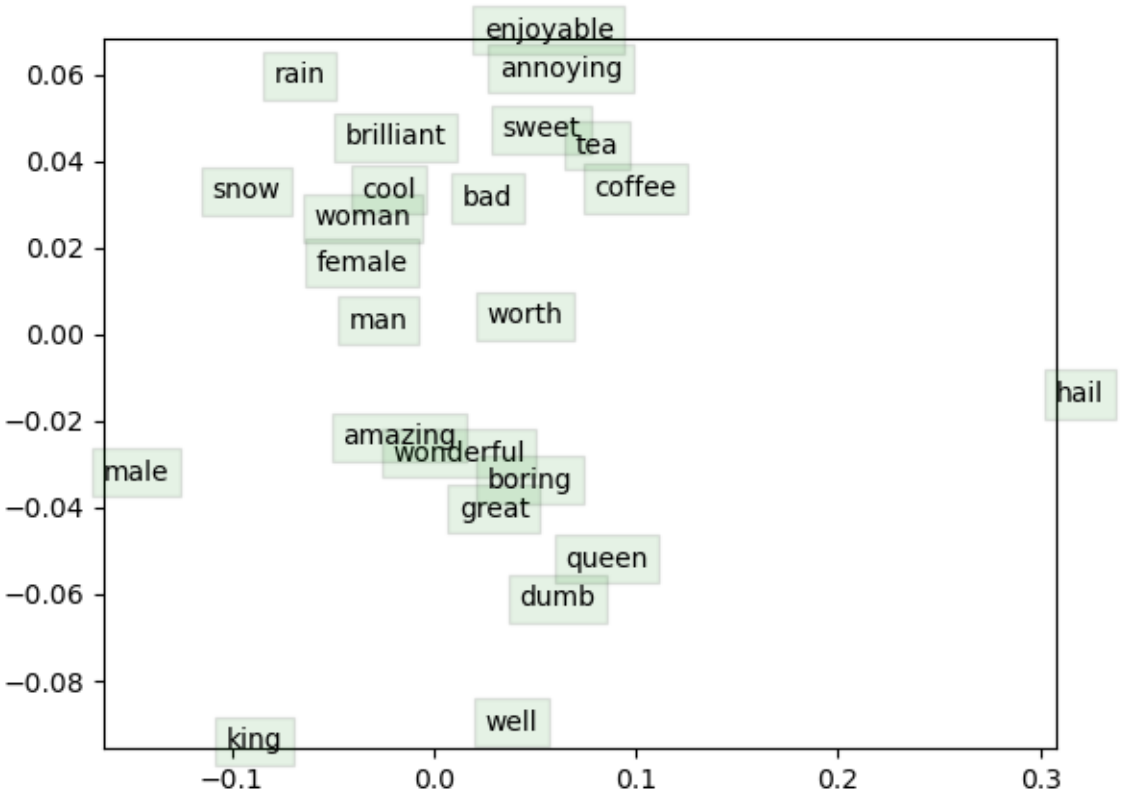
h)

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

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

$$\frac{\partial}{\partial v_w} J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}, U) = \frac{\partial}{\partial v_w} \sum_{\substack{-m \le j \le m \\ j \neq 0}} J(v_c, w_{t+j}, U) = \frac{\partial}{\partial v_w} J(v_c, w, U)$$

2



As we can see words like amazing, wonderful and great and boring have same usage and approximate meaning and they are close. If we investigate the analogy between (king, man) and (queen, woman), it's approximately true. Overall, the words which have same meaning are close to each other, but not very precise. I think it depends on the dataset we use for training.