Solution to cs224 assignment2(written)

Mhttx

February 24, 2019

1 Notations

1. d : Vector dimension
2. n : Vocabulary size
3. $\mathbf{U} \in \mathbb{R}^{d\times n}$: The colums of $\mathbf{U}$ are all the 'outside' vectors $\mathbf{u}_w \in \mathbb{R}^{d\times 1}$
4. $\mathbf{V} \in \mathbb{R}^{d\times n}$: The columms of $\mathbf{V}$ are all the 'center' vector $\mathbf{v}_w \in \mathbb{R}^{d\times 1}$
5. $\mathbf{y}, \hat{\mathbf{y}}$: The true and predicted distribution

$$\mathbf{z} = \mathbf{U}^\intercal \mathbf{v}_c \in \mathbb{R}^{n\times 1}$$
$$\hat{\mathbf{y}} = softmax(\mathbf{z}) \in \mathbb{R}^{n\times 1}$$
$$J_{naive\_softmax} = CE(\mathbf{y}, \hat{\mathbf{y}})$$

$$\delta = \frac{\partial J_{native\_softmax}}{\partial \mathbf{z}} = (\hat{\mathbf{y}} - \mathbf{y})^\intercal \in \mathbb{R}^{1\times n} \tag{1}$$

2 Answers

1. Answer to 1-(b)
$$\frac{\partial J_{native\_softmax}}{\partial \mathbf{v}_c} = (\hat{\mathbf{y}} - \mathbf{y})^\intercal \mathbf{U}^\intercal \in \mathbb{R}^{1\times d} \tag{2}$$
2. Answer to 1-(c)
$$\frac{\partial J_{native\_softmax}}{\partial \mathbf{U}} = \delta \frac{\partial \mathbf{z}}{\partial \mathbf{U}} = (\delta^\intercal \mathbf{v}_c^\intercal)^\intercal = \mathbf{v}_c \delta = \mathbf{v}_c (\hat{\mathbf{y}} - \mathbf{y})^\intercal \in \mathbb{R}^{d\times n} \tag{3}$$
3. Answer to 1-(d)
$$\sigma'(\mathbf{x}) = \sigma(\mathbf{x}) \circ (1 - \sigma(\mathbf{x})) \tag{4}$$
4. Answer to 1-(e)
$$\begin{aligned}\frac{\partial J_{neg\_sample}}{\partial \mathbf{v}_c} &= -\frac{\sigma'(\mathbf{u}_o^\intercal \mathbf{v}_c)}{\sigma(\mathbf{u}_o^\intercal \mathbf{v}_c)} \frac{\partial (\mathbf{u}_o^\intercal \mathbf{v}_c)}{\partial \mathbf{v}_c} - \sum_{k=1}^{K} \frac{\sigma'(-\mathbf{u}_k^\intercal \mathbf{v}_c)}{\sigma(-\mathbf{u}_k^\intercal \mathbf{v}_c)} \frac{\partial (-\mathbf{u}_k^\intercal \mathbf{v}_c)}{\partial \mathbf{v}_c} \\ &= -(1 - \sigma(\mathbf{u}_o^\intercal \mathbf{v}_c))\mathbf{u}_o^\intercal + \sum_{k=1}^{K} (1 - \sigma(-\mathbf{u}_k^\intercal \mathbf{v}_c))\mathbf{u}_k^\intercal\end{aligned} \tag{5}$$

$$\frac{\partial J_{neg\_sample}}{\partial \mathbf{u}_o} = -(1-\sigma(\mathbf{u}_o^\intercal \mathbf{v}_c))\mathbf{v}_c^\intercal \tag{6}$$

$$\frac{\partial J_{neg\_sample}}{\partial \mathbf{u}_k} = (1-\sigma(-\mathbf{u}_k^\intercal \mathbf{v}_c))\mathbf{v}_c^\intercal \tag{7}$$

Computing of $J_{naive\_softmax}$ needs the inner product between $\mathbf{v}_c$ and all n vocabulary vectors, while J_{neg_sample} only $k+1$ vectors.

5. Answer to 1-(f)

$$\frac{\partial J_{skip\_gram}(\mathbf{v}_c, w_{t-m}, ..., w_{t+m}, \mathbf{U})}{\partial \mathbf{U}} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial J(\mathbf{v}_c, w_{t+j}, \mathbf{U})}{\partial \mathbf{U}} \tag{8}$$

$$\frac{\partial J_{skip\_gram}(\mathbf{v}_c, w_{t-m}, ..., w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_c} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial J(\mathbf{v}_c, w_{t+j}, \mathbf{U})}{\partial \mathbf{v}_c} \tag{9}$$

$$\frac{\partial J_{skip\_gram}(\mathbf{v}_c, w_{t-m}, ..., w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_w} = 0 \quad when \quad w \ne c \tag{10}$$