Assignment 2 (written)

Chiu - 2021.1.27

$$J_{\text{naive-softmax}}(\mathbf{v}_c, \mathbf{o}, \mathbf{U}) = -\log \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)}$$

$$J_{\text{neg-sample}}(\mathbf{v}_c, \mathbf{o}, \mathbf{U}) = -\log(\sigma(\mathbf{u}_o^T \mathbf{v}_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T \mathbf{v}_c))$$

$$J_{\text{skip-gram}}(\mathbf{v}_c, \mathbf{w}_{t-m}, \ldots, \mathbf{w}_{t+m}, \mathbf{U}) = \sum_{\substack{-m \le j \le m \\ j \ne 0}} J(\mathbf{v}_c, \mathbf{w}_{t+j}, \mathbf{U})$$

Note:

- 负采样不是单纯把$w \in \text{Vocab}$改成了k从1到K，而是把分母整个改造成用sigmoid函数取log再求和
- naive-softmax是在log里求exp的和，负采样是在log外求和，log内为sigmoid
- 【注意】负采样时负样本[1, K]中不会出现正例o！

Notation:

- u, v 均为d×1的列向量
- 矩阵U维数为|V|×d，每行是u^T
- y：真实分布，是one-hot（y_o表示outside word的真实分布，即仅下标o处为1），这里约定是列向量
- $\hat{y}$：softmax或负采样后的预测分布，这里约定是列向量（$\hat{y}_o$表示下标o处的prob，即一个元素）

Naive-Softmax

（a）由实际分布为one-hot，即$y_w = \begin{cases} 1, & w = o \\ 0, & else \end{cases}$ 可得交叉熵loss=naive-softmax，即

$$CE(y, \hat{y}) = -\sum_{w \in \text{Vocab}} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$$

【注意】

- 含义：输入softmax参数θ得到预测分布$\hat{y}$，它的负log损失 <=> 对某个真实存在的分布y和这个预测分布$\hat{y} = \text{softmax}(\theta)$计算交叉熵
- 这个等式可以推导出naive-softmax关于θ的微分：（word2vec中θ=点积=Uv_c）

$$\frac{\partial J}{\partial \theta} = \frac{\partial CE(y, \hat{y})}{\partial \theta} = (\hat{y} - y)^T$$

  - 令点积=$\theta = Uv_c$，概率列向量$\hat{y}$ = Prob = softmax(θ)，则$J = CE(y, \text{softmax}(\theta))$

- 交叉熵的微分$\frac{\partial CE(y,\hat{y})}{\partial \theta} = \frac{\partial CE(y,\hat{y})}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \theta} = \sum_k \frac{\partial CE}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial \theta}$，这里softmax函数的输入θ=y
  - 由$\frac{\partial CE(y,\hat{y})}{\partial \hat{y}_w} = \begin{cases} 1/\hat{y}_o, & w = o \\ 0, & else \end{cases}$，

$$\frac{\partial \hat{y}_i}{\partial \theta_j} = \begin{cases} \frac{\exp(\theta_i)(\sum \ldots) - \exp(2\theta_i)}{(\sum \exp(\theta_k))^2} = \hat{y}_i(1 - \hat{y}_i), & i = j \\ -\frac{\exp(\theta_o) \cdot \exp(\theta_j)}{(\sum \exp(\theta_k))^2} = -\hat{y}_i \hat{y}_j, & i \ne j \end{cases}$$

  - 得$\frac{\partial CE(y,\hat{y})}{\partial \theta_j} = \begin{cases} 1 - \hat{y}_o, & j = o \\ -\hat{y}_j, & j \ne o \end{cases}$，向量表示为$y^T - \hat{y}^T$ (1×d)

- [上述推导左侧微分含有对word2vec中真实分布y的依赖，实际上可以给出一个无依赖的推导]

  令Softmax :=S, CE := L，则

$$\begin{aligned}\frac{\partial L}{\partial \theta_j} &= -\sum_k y_k \frac{\partial \log S_k}{\partial \theta_j} \\ &= -\sum_k y_k \frac{1}{S_k}\frac{\partial S_k}{\partial \theta_j} \\ &= -\sum_{k=j} y_k \frac{1}{S_k} S_k(1-S_k) - \sum_{k\neq j} y_k \frac{1}{S_k}(-S_k S_j) \\ &= -y_j(1-S_j) + \sum_{k\neq j} y_k S_j \\ &= -y_j + S_j \sum_k y_k \quad (\text{由Softmax归一化知和为}1) \\ &= S_j - y_j\end{aligned}$$

(b)

$$\frac{\partial J_{\text{naive}-\text{softmax}}}{\partial v_c} = -u_o^T + \frac{\sum_{w\in \text{Vocab}} u_w^T \exp(u_w^T v_c)}{\sum_{w\in \text{Vocab}} \exp(u_w^T v_c)} = -u_o^T + \sum_{w\in \text{Vocab}} u_w^T P(O=w \mid C=c)$$

【发现】

- $\hat{y}_o$ = P(O=o|C=c) = y_o^T · softmax (Uv$_c$)，即取出softmax中o对应行的元素
- 右部分是outside vectors（u是列向量，转置成为行向量）的加权平均，权重就是估计它为context的概率
- 左部分可以看成估计它为context的概率是-1（？）
- 最后得到的是一个**行**向量（与shape convention相反）

(c)

$$\frac{\partial J_{\text{naive}-\text{softmax}}}{\partial u_w} = \begin{cases} -v_c^T + \frac{v_c^T \exp(u_o^T v_c)}{\sum_{w\in \text{Vocab}} \exp(u_w^T v_c)} = -v_c^T + v_c^T P(O=o|C=c), & w=o \\ \frac{v_c^T \exp(u_w^T v_c)}{\sum_{w\in \text{Vocab}} \exp(u_w^T v_c)} = v_c^T P(O=w|C=c), & w\neq o \end{cases}$$

【推导】

$$\text{观察法：} \frac{\partial J_{\text{naive}-\text{softmax}}}{\partial v_c} = -u_o^T + Prob^T \cdot U = -y^T U + \text{softmax}(Uv_c)^T U = (-y+\hat{y})^T U$$

$$\frac{\partial J_{\text{naive}-\text{softmax}}}{\partial U} = -yv_c^T + Prob \cdot v_c^T = (\hat{y}-y)v_c^T$$

$$\text{链式法则：} \frac{\partial J_{\text{naive}-\text{softmax}}}{\partial v_c} = \frac{\partial J}{\partial \theta}\cdot\frac{\partial \theta}{\partial v_c} = (\hat{y}-y)^T U$$

$$\frac{\partial J_{\text{naive}-\text{softmax}}}{\partial U} = \delta^T x^T = (\hat{y}-y)v_c^T$$

Neg-Sample

(d) $\sigma(x) = \frac{1}{1+e^{-x}}$，得到$\sigma'(x) = \sigma(x)(1-\sigma(x))$

(e)

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

$$\frac{\partial J_{\text{neg-sample}}}{\partial u_o} = -v_c^T[1-\sigma(u_o^T v_c)]$$

$$\frac{\partial J_{\text{neg-sample}}}{\partial u_k} = -[1-\sigma(-u_k^T v_c)](-v_c^T)$$

- 负采样中梯度的计算是O(K)的，比原来naive-softmax的O(|V|)明显减小，不需要遍历整个vocab；
- 可以利用sigmoid函数微分后是原输出的函数的特性，进行计算结果复用，计算量减小。

(f) 注意 J 是任意的loss term (naive-softmax / neg-sample)，且 $w \neq c$

$$\frac{\partial J_{\text{skip-gram}}}{\partial U} = \sum_{\substack{-m\le j\le m \\ j\neq 0}} \frac{\partial J}{\partial U}$$

$$\frac{\partial J_{\text{skip-gram}}}{\partial v_c} = \sum_{\substack{-m\le j\le m \\ j\neq 0}} \frac{\partial J}{\partial v_c}$$

$$\frac{\partial J_{\text{skip-gram}}}{\partial v_w} = 0 \quad (w \neq c)$$

【注意】

- J 是C维列向量，即有C个output
- U 是|V|×d维的矩阵
- $\partial J/\partial U$ 由于shape convention，形状和U相同