① Naive-Softmax

$= -\log \hat{y}_o$ 简写符号 $\hat{y}$

$= -\log P(O=o \mid C=c)$

月 日 星期

$J_{naive-softmax} = -\log \frac{\exp(u_o^T v_c)}{\sum \exp(u_w^T v_c)}$ 对 v_c, $u_{w\neq o}$, u_o 求偏导.

① $\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c} \left(\log \sum \exp(u_w^T v_c) - \log \exp(u_o^T v_c) \right)$

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

$= \frac{1}{\sum \exp(u_w^T v_c)} \cdot \frac{\partial}{\partial v_c} \sum [e^{u_w^T v_c}] - u_o$

const 可放入Σ

$= \frac{1}{\sum \exp(u_w^T v_c)} \cdot \sum [u_w \exp(u_w^T v_c)] - u_o$

$= \sum_w [u_w \exp(u_w^T v_c) / \sum \exp(u_w^T v_c)] - u_o$

$= \sum_w [u_w P(O=w \mid C=c)] - u_o$

$= \sum_w [u_w \hat{y}_w] - u_o$ 用简写符代入，公式简洁

② $\frac{\partial J}{\partial u_{w\neq o}} = \frac{\partial}{\partial u_{w\neq o}} \left(\log \sum \exp(u_w^T v_c) - \log \exp(u_o^T v_c) \right)$

$= \frac{1}{\sum \exp(u_w^T v_c)} \cdot \sum \left[\frac{\partial}{\partial u_{w\neq o}} \exp(u_w^T v_c) \right] - \frac{\partial}{\partial u_{w\neq o}} u_o^T v_c$

$= \sum_{w\neq o} \left[\frac{\exp(u_w^T v_c)}{\sum \exp(u_w^T v_c)} \cdot v_c \right] - 0$

$= \sum_{w\neq o} (\hat{y}_w \cdot v_c) = \hat{y}_w \cdot v_c$ ($\because$ w is specific)

③ $\frac{\partial J}{\partial u_{w=o}} = \frac{\partial J}{\partial u_o} = \frac{\partial}{\partial u_o} \left(\log \sum \exp(u_w^T v_c) - u_o^T v_c \right)$

$= \frac{1}{\sum \exp(u_w^T v_c)} \cdot \sum \left[\frac{\partial}{\partial u_o} \exp(u_w^T v_c) \right] - \frac{\partial}{\partial u_o} u_o^T v_c$

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

$= (\hat{y}_o - 1) \cdot v_c$

Sigmoid 的快速求导 $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$

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

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

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

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

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

② Neg-sampling.

negtive-sampling 还是 skip-gram

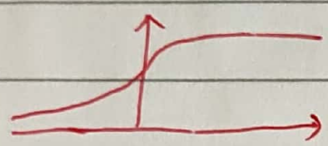
考虑中心词 c 与上下文 o ：令 $P(D=1|o,c)$ 为 (o,c) 来自语料库，$P(D=0|o,c)$ 表示 (o,c) 不来自语料库.

$$P(D=1|w,c) = \boxed{\sigma(v_c^T u_w)}$$

$$P(D=0|w,c) = 1-P(D=1|w,c)$$

$v_c^T u_w$ 表示了其相似性，越相似，数越大，

不相似就很小，可能会负.



用 sigmoid 约束到 $(0,1)$.

我们要最大化正确上下文的 $P(D=1)$，最小化错误上下文的 $P(D=1)$（即最大化错误上下文的 $P(D=0)$）

$$\max \prod_{(w,c)\in D} P(D=1|w,c) \prod_{(w,c)\notin D} P(D=0|w,c)$$

在语料库中　　　不在语料库中

取log：$\max \sum_{(w,c)\in D} \log P(D=1) + \sum_{(w,c)\notin D} \log(1-P(D=1))$

$$= \cancel{\max}\ \min -\sum_{(w,c)\in D} \log \sigma(v_c^T u_w) \cancel{+} - \sum_{(w,c)\notin D} \log(1-\sigma(v_c^T u_w))$$

定义一个来自语料库的 w，$(w,c)\in D$　$w=c-m+j$，窗口大小 $2m$

$$J = -\log \sigma(u_{c-m+j}^T v_c) - \sum_{sampled} \log \sigma(-\tilde{u}_k^T v_c)$$

上式中 $\{\tilde{u}_k \mid k=1\ldots K\}$ 是从 D 中抽样，抽样概率为实际概率 ~~[illegible]~~ $3/4$ 次方：$0.9^{3/4} \approx 0.92$，$0.09^{3/4} \approx 0.16$，

可以提高罕见词被抽中的概率.

对neg-sampling loss 求导:

$$J_{neg\text{-}sampling}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \overbrace{\sum_{k=1}^{K}}^{sampling} \log(\sigma(-u_k^T v_c))$$

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

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

{ You should know $1-\sigma(x) = \sigma(-x)$)

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

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

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

② $\frac{\partial J}{\partial u_o} = -v_c\sigma(-u_o^T v_c) + v_c \sum_{k=1}^{K} \sigma(u_k^T v_c) \Big|_{u_k=u_o}$, but $u_k \neq u_o$

$= -v_c\sigma(-u_o^T v_c) + \cancel{v_c\sigma(u_o^T v_c)} = 0$

$= -v_c\sigma(-u_o^T v_c)$

③ $\frac{\partial J}{\partial u_k} = v_c\sigma(u_k^T v_c)$

对 {naive-softmax / neg-sampling 求 $J_{skip\text{-}gram}$ 的导数

设中心 $c = w_t$, 窗口大小为 m: $[w_{t-m}, \cdots, w_{t+m}]$

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

中心词对所有窗口词轮询一遍

① $\frac{\partial J_{skip\text{-}gram}}{\partial v_c} = \sum \frac{\partial}{\partial v_c} J_{neg\text{-}sampling/naive\text{-}softmax}$

② $\frac{\partial J_{skip\text{-}gram}}{\partial v_{w \neq c}} = \frac{\partial}{\partial v_w} J = 0$, $w \neq c$

③ $\frac{\partial J_{skip\text{-}gram}}{\partial U} = \cancel{\frac{\partial}{\partial u_o} J + \sum_{k \neq o} \frac{\partial J}{\partial u_k}} = \sum \frac{\partial J}{\partial U}$