word2vec 优化方式推导

一. Negative Sampling (负采样)

选择一个正样本和K个负样本，训练$(K+1)$个二分类器.

1. CBOW 的公式推导

假设中心词为w，周围词为context(w)，负样本集NEG(w)

定义 Indicator Function. $I^w(\tilde{w}) = \begin{cases} 1, & \tilde{w}=w \\ 0, & \tilde{w}\neq w \end{cases}$

用来表示词对 $[context(w), \tilde{w}]$ 的标签，$context(w) = \frac{\sum_{j=0}^{j\neq 0} u_{j+w}}{2j}$（周围词平均向量）$= \vec{x}_w$

似然函数 likelihood

$= \prod_{\tilde{w}\in\{w\}\cup NEG(w)} p(\tilde{w}\mid context(w))$ ①，令目标词 $\tilde{w}$ 的词向量为 $\theta^{\tilde{w}}$

其中 $p(\tilde{w}\mid context(w)) = \begin{cases} \sigma(\vec{x}_w^T\cdot\theta^{\tilde{w}}), & L^w(\tilde{w})=1 \\ 1-\sigma(\vec{x}_w^T\cdot\theta^{\tilde{w}}), & L^w(\tilde{w})=0 \end{cases}$

$\therefore$ ①式 $= \sigma(\vec{x}_w^T\cdot\theta^w)\cdot\prod_{\tilde{w}\in NEG(w)}(1-\sigma(\vec{x}_w^T\cdot\theta^{\tilde{w}})) = g(w)$

给定语料库，我们的目标 $\max\prod_{w\in C} g(w) \Leftrightarrow \max\log\prod_{w\in C} g(w)$

$\Rightarrow \max\sum_{w\in C}\left[\log\sigma(\vec{x}_w^T\cdot\theta^{\tilde{w}}) + \sum_{\tilde{w}\in NEG(w)}\log\sigma(-x_w^T\cdot\theta^{\tilde{w}})\right]$ $(1-\sigma(x)=\sigma(-x))$

令 $L(w,\tilde{w}) = I^w(\tilde{w})\log\sigma(\vec{x}_w^T\cdot\vec{\theta}^{\tilde{w}}) + (1-I^w(\tilde{w}))\log(1-\sigma(\vec{x}_w^T\cdot\vec{\theta}^{\tilde{w}}))$ ②

代表每个训练样本的 log-likelihood，使用梯度上升法

$\frac{\partial L}{\partial\theta^{\tilde{w}}} = I^w(\tilde{w})(1-\sigma(\vec{x}_w^T\cdot\theta^{\tilde{w}}))\cdot\vec{x}_w + (1-I^w(\tilde{w}))(-\sigma(\vec{x}_w^T\cdot\theta^{\tilde{w}}))\cdot\vec{x}_w$

$= (I^w(\tilde{w}) - I^w(\tilde{w})\sigma(x_w^T\cdot\theta^{\tilde{w}}) - \sigma(x_w^T\cdot\theta^{\tilde{w}}) + I^w(\tilde{w})\sigma(x_w^T\cdot\theta^{\tilde{w}}))\cdot x_w$

$= [I^w(\tilde{w}) - \sigma(x_w^T\cdot\theta^{\tilde{w}})]\cdot x_w$

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

$(\log\sigma(x))' = 1-\sigma(x)$

$[\log(1-\sigma(x))]' = -\sigma(x)$

由于θ^w和x_w^T在②式中对称，

$\frac{\partial L}{\partial \theta^{\tilde{w}}}$换成$\frac{\partial L}{\partial x_w^T} = \left[I^w(\tilde{w}) - \sigma(\vec{x}_w^T \cdot \vec{\theta}^{\tilde{w}})\right] \cdot \vec{\theta}^{\tilde{w}}$

对于K+1个样本，可使用下述公式更新梯度

$\theta^{\tilde{w}} : \theta^{\tilde{w}} + \eta\left[I^w(\tilde{w}) - \sigma(\vec{x}_w^T \cdot \vec{\theta}^{\tilde{w}})\right] \cdot \vec{x}_w$

$x_w^* : x_w^* + \sum_{\tilde{w} \in \{w\} \cup NEG(w)} \eta\left[I^w(\tilde{w}) - \sigma(\vec{x}_w^T \cdot \vec{\theta}^{\tilde{w}})\right] \cdot \vec{\theta}^{\tilde{w}}$

2. skipgram公式推导

似然函数 $= \prod_{w \in C} \prod_{u \in context(w)} \prod_{\tilde{w} \in \{u\} \cup NEG(u)} P(\tilde{w}|w)$

其中 $P(\tilde{w}|w) = \begin{cases} \sigma(\vec{v}_w^T \cdot \vec{u}_{\tilde{w}}), & I^w(\tilde{w}) = 1 \\ 1 - \sigma(\vec{v}_w^T \cdot \vec{u}_{\tilde{w}}), & I^w(\tilde{w}) = 0 \end{cases}$

Log-likelihood $= \sum_{w \in C} \sum_{u \in context(w)} \sum_{\tilde{w} \cdots} \left[I(\tilde{w}) \log \sigma(\vec{v}_w^T \cdot \vec{u}_{\tilde{w}}) + (1 - I(\tilde{w})) \log \left(1 - \sigma(\vec{v}_w^T \cdot \vec{u}_{\tilde{w}})\right)\right]$

由于公式与一(1)一样，我们可以得到

$\frac{\partial L}{\partial \vec{u}_{\tilde{w}}} = I^w(\tilde{w}) - \sigma(\vec{v}_w^T \cdot \vec{u}_{\tilde{w}}) \cdot \vec{v}_w$

$\vec{u}_{\tilde{w}} : \vec{u}_{\tilde{w}} + \eta \frac{\partial L}{\partial \vec{u}_{\tilde{w}}}$

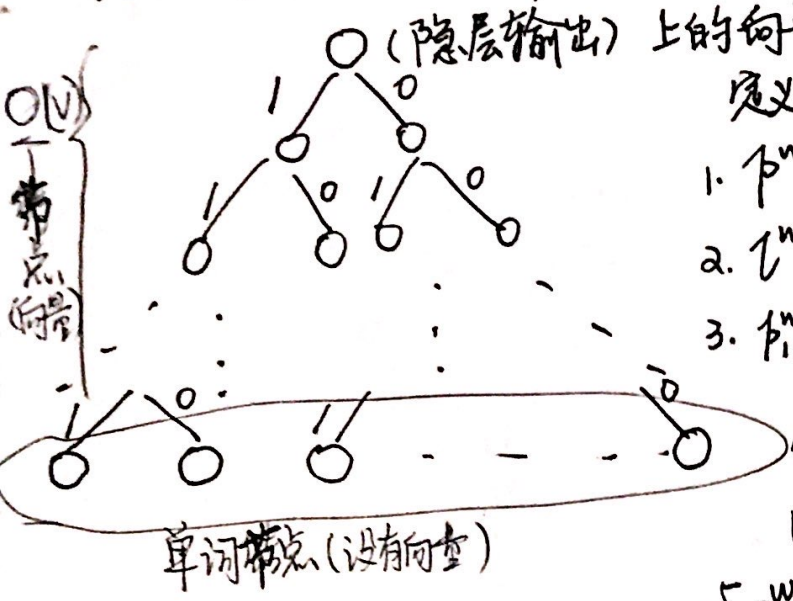
同理：$\frac{\partial L}{\partial \vec{v}_w} = \left[I^w(\tilde{w}) - \sigma(\vec{v}_w^T \cdot \vec{u}_{\tilde{w}})\right] \cdot \vec{u}_{\tilde{w}}$

对于K+1个样本：

$\vec{v}_w : \vec{v}_w + \sum_{\tilde{w} \in \{w\} \cup NEG\{u\}} \eta \frac{\partial L}{\partial \vec{v}_w^T}$

二. Hierarchical Softmax（层次softmax）（哈夫曼树）

叶子结点代表的是词典中的所有词$|V|$，通往目标词的路径上的点（隐层输出）上的向量是我们要优化的。



定义：

1. p^w：根结点到单词结点（叶结点）的路径
2. l^w：路径p^w中结点的个数
3. $p_1^w, p_2^w, \dots, p_{l^w}^w$：路径上的点，$p_1^w$根结点，$p_{l^w}^w$：$w$结点
4. $d_2^w, d_3^w, \dots d_{l^w}^w \in \{0,1\}$：词$w$的哈夫曼编码，由$(l^w-1)$位二进制编码构成
5. $\theta_1^w, \theta_2^w, \dots \theta_{l^w-1}^w$对应$p^w$中非叶子节点的向量

1. CBOW的公式推导

$$\max \prod_{w \in C} \prod_{j=2}^{l^w} p(d_j^w \mid x_w, \theta_{j-1}^w)$$

（x_w为隐层输出）$\vec{x_w} = \frac{\sum context(w)}{|context(w)|}$，有 $p = \begin{cases} \sigma(\vec{x_w}^T \cdot \theta_{j-1}^w), & d_j^w = 0 \text{（向右走）} \\ 1-\sigma(\vec{x_w}^T \cdot \theta_{j-1}^w), & d_j^w = 1 \text{（向左走）} \end{cases}$

$$①$$

对数似然 $\Rightarrow \log ① = \sum_{w \in C} \sum_{j=2}^{l^w} \left((1-d_j^w) \log \sigma(\vec{x_w}^T \cdot \theta_{j-1}^w) + d_j^w \log(1-\sigma(\vec{x_w}^T \cdot \theta_{j-1}^w)) \right)$

$$= L(j,w)$$

对于每个样本

$$\frac{\partial L(j,w)}{\partial \theta_{j-1}^w} = (1-d_j^w)(1-\sigma(x_w^T \cdot \theta_{j-1}^w)) \cdot x_w + d_j^w (-\sigma(x_w^T \cdot \theta_{j-1}^w)) \cdot x_w$$

$$= (1-d_j^w - \sigma(x_w^T \cdot \theta_{j-1}^w)) \cdot x_w$$

$\theta_{j-1}^w : \theta_{j-1}^w + \eta \frac{\partial L}{\partial \theta_{j-1}^w}$

对称性 $\Rightarrow \frac{\partial L}{\partial x_w} = (1-\sigma(x_w^T \theta_{j-1}^w) - d_j^w) \cdot \theta_{j-1}^w$

根据 x_w 更新 v_w $\Rightarrow$ $v_m : v_m + \eta \sum_{j=2}^{l^w} \frac{\partial L}{\partial x_w}$

2. skipgram 的公式推导

非常类似 CBOW，多了一个遍历 context(w)

$$\max \prod_{w \in C} \prod_{u \in context(w)} \prod_{j=2}^{l^w} p(d_j^w \mid x_w, \theta_{j-1}^w) \quad (x_w \text{ 隐层输出})$$

对数似然 ~~$L(w,u,j)$~~ $L = \sum_{w \in C} \sum_{u \in context(w)} \sum_{j=2}^{l^w} \log p(d_j^w \mid x_w, \theta_{j-1}^w)$

$$= \sum \sum \sum \underbrace{(1-d_j^w)\log\sigma(x_w^T \cdot \theta_{j-1}^w) + d_j^w \log(1-\sigma(x_w^T \cdot \theta_{j-1}^w))}_{L(w,j)}$$

参考上节：$\frac{\partial L}{\partial \theta_{j-1}^w} = \left(1 - d_j^w - \sigma(x_w^T \cdot \theta_{j-1}^w)\right) \cdot x_w$

$$\theta_{j-1}^w := \theta_{j-1}^w + \eta \frac{\partial L}{\partial \theta_{j-1}^w}$$

$$\frac{\partial L}{\partial x_w} = \left(1 - d_j^w - \sigma(x_w^T \cdot \theta_{j-1}^w)\right) \cdot \theta_{j-1}^w$$

$$x_w := x_w + \eta \sum_{u \in context(w)} \sum_{j=2}^{l^w} \frac{\partial L}{\partial x_w}$$