word2vec

单词w；
词典$\mathcal{D}=\{w_1, w_2, \ldots, w_N\}$，由单词组成的集合；
语料库$\mathcal{C}$，由单词组成的文本序列；
单词w的上下文$Context\,(w)$，由预料库中单词w的前c个单词和后c个单词组成的文本序列。

CBOW模型网络结构

输入层：$\mathbf{v}\left(Context(w)_1\right), \mathbf{v}\left(Context(w)_2\right), \cdots, \mathbf{v}\left(Context(w)_{2c}\right) \in \mathbb{R}^m$；

投影层：$\mathbf{x}_w=\sum_{i=1}^{2c} \mathbf{v}\left(Context(w)_i\right) \in \mathbb{R}^m$；

输出层：$T_{Huff}\left(\mathbf{x}_w\right)=s_{q\left(\mathbf{x}_w\right)}, s \in \mathbb{R}^N, q: \mathbb{R}^m \rightarrow\{1,2, \cdots, N\}$。

基于Hierarchical softmax的CBOW模型

记
$$p^w=\left(p_1^w, p_2^w \cdots, p_{l^w}^w\right)$$
为从根节点出发到达w对应的叶子结点的路径。其中，l^w为路径长度，即路径中结点数目；p_i^w为路径中的结点，p_1^w为根结点，$p_{l^w}^w$为w对应的叶子结点。

记
$$d^w=\left(d_2^w, d_3^w \cdots, d_{l^w}^w\right)$$
为w的Huffman编码。其中，$d_i^w \in\{0,1\}$为路径p^w中第i个结点对应的编码（根结点不对应编码）。

记
$$\theta^w=\left(\theta_1^w, \theta_2^w, \cdots, \theta_{l^w-1}^w\right)$$
为路径p^w中非叶子结点对应的参数向量。其中，$\theta_i^w \in \mathbb{R}^m$为路径$p^w$中第$i$个非叶子结点对应的参数向量。

条件概率
$$p\,(w \mid Context\,(w))=\prod_{j=2}^{l^w} p\left(d_j^w \mid \mathbf{x}_w, \theta_{j-1}^w\right)$$
其中
$$p\left(d_j^w \mid \mathbf{x}_w, \theta_{j-1}^w\right)=\left\{\begin{array}{l}\sigma\left(\mathbf{x}_w^{\top} \theta_{j-1}^w\right), d_j^w=0; \\ 1-\sigma\left(\mathbf{x}_w^{\top} \theta_{j-1}^w\right), d_j^w=1,\end{array}\right.$$
或者
$$p\left(d_j^w \mid \mathbf{x}_w, \theta_{j-1}^w\right)=\left[\sigma\left(\mathbf{x}_w^{\top} \theta_{j-1}^w\right)\right]^{1-d_j^w} \cdot\left[1-\sigma\left(\mathbf{x}_w^{\top} \theta_{j-1}^w\right)\right]^{d_j^w}$$

对数似然函数

$$
\begin{aligned}
\mathcal{L} &= \sum_{w \in C} \log \prod_{j=2}^{l^w} p\left(d_j^w | \mathbf{x}_w, \theta_{j-1}^w\right) \\
&= \sum_{w \in C} \sum_{j=2}^{l^w} \left\{ \left(1-d_j^w\right) \cdot \log\left[\sigma\left(\mathbf{x}_w^\top \theta_{j-1}^w\right)\right] + d_j^w \cdot \log\left[1-\sigma\left(\mathbf{x}_w^\top \theta_{j-1}^w\right)\right] \right\}
\end{aligned}
$$

对数似然函数$\mathcal{L}$关于θ_{j-1}^w的梯度

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta_{j-1}^w} &= \frac{\partial}{\partial \theta_{j-1}^w} \left\{ \sum_{w \in C} \sum_{j=2}^{l^w} \left\{ \left(1-d_j^w\right) \cdot \log\left[\sigma\left(\mathbf{x}_w^\top \theta_{j-1}^w\right)\right] + d_j^w \cdot \log\left[1-\sigma\left(\mathbf{x}_w^\top \theta_{j-1}^w\right)\right] \right\} \right\} \\
&= \left(1-d_j^w\right) \left[1-\sigma\left(\mathbf{x}_w^\top \theta_{j-1}^w\right)\right] \mathbf{x}_w - d_j^w \sigma\left(\mathbf{x}_w^\top \theta_{j-1}^w\right) \mathbf{x}_w \\
&= \left[1-d_j^w-\sigma\left(\mathbf{x}_w^\top \theta_{j-1}^w\right)\right] \mathbf{x}_w
\end{aligned}
$$

θ_{j-1}^w的更新

$$
\theta_{j-1}^w = \theta_{j-1}^w + \eta \left[1-d_j^w-\sigma\left(\mathbf{x}_w^\top \theta_{j-1}^w\right)\right] \mathbf{x}_w
$$

其中，η为学习率。

对数似然函数$\mathcal{L}$关于$\mathbf{x}_w$的梯度

$$
\frac{\partial \mathcal{L}}{\partial \mathbf{x}_w} = \sum_{j=2}^{l^w} \left[1-d_j^w-\sigma\left(\mathbf{x}_w^\top \theta_{j-1}^w\right)\right] \theta_{j-1}^w
$$

$\mathbf{v}(\tilde{w})$的更新

$$
\mathbf{v}(\tilde{w}) = \mathbf{v}(\tilde{w}) + \eta \frac{\partial \mathcal{L}}{\partial \mathbf{x}_w}
$$

其中，$\tilde{w} \in Context(w)$。

Skip-gram模型网络结构

输入层：$\mathbf{v}(w) \in \mathbb{R}^m$

输出层：$T_{Huff}\left(\mathbf{v}_w\right) = s_{q(\mathbf{v}_w)}, s \in \mathbb{R}^N, q : \mathbb{R}^m \to \{1, 2, \cdots, N\}$

基于Hierarchical softmax的Skip-gram模型

条件概率

$$
p\left(Context(w) | w\right) = \prod_{u \in Context(w)} p(u|w)
$$

其中

$$
p(u|w) = \prod_{j=2}^{l^u} p\left(d_j^u | \mathbf{v}(w), \theta_{j-1}^u\right)
$$

且

$$
p\left(d_j^u | \mathbf{v}(w), \theta_{j-1}^u\right) = \left[\sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right]^{1-d_j^u} \cdot \left[1-\sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right]^{d_j^u}
$$

对数似然函数

$$\begin{aligned}\mathcal{L} &= \sum_{w\in C} log \prod_{u\in Context(w)} \prod_{j=2}^{l^u} \left\{ \left[\sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right]^{1-d_j^u} \cdot \left[1-\sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right]^{d_j^u} \right\} \\ &= \sum_{w\in C} \sum_{u\in Context(w)} \sum_{j=2}^{l^u} \left\{ \left(1-d_j^u\right) \cdot \log\left[\sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right] + d_j^u \cdot \log\left[1-\sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right] \right\}\end{aligned}$$

对数似然函数$\mathcal{L}$关于θ_{j-1}^u的梯度

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta_{j-1}^u} &= \frac{\partial}{\theta_{j-1}^u} \left\{ \sum_{w\in C} \sum_{u\in Context(w)} \sum_{j=2}^{l^u} \left\{ \left(1-d_j^u\right) \cdot \log\left[\sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right] + d_j^u \cdot \log\left[1-\sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right] \right\} \right\} \\ &= \sum_{w\in C} \left\{ \left(1-d_j^u\right) \left[1-\sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right] \mathbf{v}(w) - d_j^u \sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right) \mathbf{v}(w) \right\} \\ &= \sum_{w\in C} \left[1-d_j^u - \sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right] \mathbf{v}(w)\end{aligned}$$

θ_{j-1}^u的更新

$$\theta_{j-1}^u = \theta_{j-1}^u + \eta \sum_{w\in C} \left[1-d_j^u - \sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right] \mathbf{v}(w)$$

其中，η为学习率。

对数似然函数$\mathcal{L}$关于$\mathbf{v}(w)$的梯度

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}(w)} = \sum_{u\in Context(w)} \sum_{j=2}^{l^u} \left[1-d_j^u - \sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right] \theta_{j-1}^u$$

$\mathbf{v}(w)$的跟新

$$\mathbf{v}(w) = \mathbf{v}(w) + \eta \sum_{u\in Context(w)} \sum_{j=2}^{l^u} \left[1-d_j^u - \sigma\left(\mathbf{v}(w)^\top \theta_{j-1}^u\right)\right] \theta_{j-1}^u$$

基于Negative Sampling的CBOW模型

设$Context(w)$的负样本子集为

$$NEG(w) \neq \emptyset$$

对于$\forall \tilde{w} \in \mathcal{D}$，定义

$$L^w(\tilde{w}) = \begin{cases} 1, \tilde{w} = w \\ 0, \tilde{w} \neq w \end{cases}$$

表示词$\tilde{w}$的标签，正样本标签为1，负样本标签为0。

关于字典$\mathcal{D}$的子集$\{w\} \bigcup NEG(w)$的似然函数

$$g(w) = \prod_{u \in \{w\} \bigcup NEG(w)} p(u|Context(w)) = \sigma\left(\mathbf{x}_w^\top \theta^w\right) \prod_{u \in NEG(w)} \left[1 - \sigma\left(\mathbf{x}_w^\top \theta^w\right)\right]$$

其中

$$p(u|Context(w)) = \begin{cases} \sigma\left(\mathbf{x}_w^\top \theta^u\right), L^w(u) = 1 \\ 1 - \sigma\left(\mathbf{x}_w^\top \theta^u\right), L^w(u) = 0 \end{cases}$$

或者

$$p(u|Context(w)) = \left[\sigma\left(\mathbf{x}_w^\top \theta^u\right)\right]^{L^w(u)} \cdot \left[1 - \sigma\left(\mathbf{x}_w^\top \theta^u\right)\right]^{1-L^w(u)}$$

$\mathbf{x}_w$为$Context(w)$词向量之和，$\theta^u \in \mathbb{R}^m$为模型参数。

关于语料库$\mathcal{C}$的对数似然函数

$$\begin{aligned} \mathcal{L} &= \log \prod_{w \in \mathcal{C}} g(w) = \sum_{w \in \mathcal{C}} \log g(w) \\ &= \sum_{w \in \mathcal{C}} \log \prod_{u \in \{w\} \bigcup NEG(w)} \left\{\left[\sigma\left(\mathbf{x}_w^\top \theta^u\right)\right]^{L^w(u)} \cdot \left[1 - \sigma\left(\mathbf{x}_w^\top \theta^u\right)\right]^{1-L^w(u)}\right\} \\ &= \sum_{w \in \mathcal{C}} \sum_{u \in \{w\} \bigcup NEG(w)} \left\{L^w(u) \cdot \log\left[\sigma\left(\mathbf{x}_w^\top \theta^u\right)\right] + \left[1 - L^w(u)\right] \cdot \log\left[1 - \sigma\left(\mathbf{x}_w^\top \theta^u\right)\right]\right\} \end{aligned}$$

对数似然函数$\mathcal{L}$关于θ^u的梯度

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta^u} &= \frac{\partial}{\partial \theta^u}\left\{\sum_{w \in \mathcal{C}} \sum_{u \in \{w\} \bigcup NEG(w)} \left\{L^w(u) \cdot \log\left[\sigma\left(\mathbf{x}_w^\top \theta^u\right)\right] + \left[1 - L^w(u)\right] \cdot \log\left[1 - \sigma\left(\mathbf{x}_w^\top \theta^u\right)\right]\right\}\right\} \\ &= L^w(u)\left[1 - \sigma\left(\mathbf{x}_w^\top \theta^u\right)\right] \mathbf{x}_w - \left[1 - L^w(u)\right] \sigma\left(\mathbf{x}_w^\top \theta^u\right) \mathbf{x}_w \\ &= \left[L^w(u) - \sigma\left(\mathbf{x}_w^\top \theta^u\right)\right] \mathbf{x}_w \end{aligned}$$

θ^u的更新

$$\theta^u = \theta^u + \eta\left[L^w(u) - \sigma\left(\mathbf{x}_w^\top \theta^u\right)\right] \mathbf{x}_w$$

对数似然函数$\mathcal{L}$关于$\mathbf{x}_w$的梯度

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}_w} = \sum_{u \in (w) \bigcup NEG(w)} \left[L^w(u) - \sigma\left(\mathbf{x}_w^\top \theta^u\right)\right] \theta^u$$

$\mathbf{v}(\tilde{w})$的更新

$$\mathbf{v}(\tilde{w}) = \mathbf{v}(\tilde{w}) + \eta \frac{\partial \mathcal{L}}{\partial \mathbf{x}_w}$$

其中，$\tilde{w} \in Context(w)$。

基于Negative Sampling的Skip-gram模型

关于字典$\mathcal{D}$的子集$\{w\}\bigcup NEG^{\tilde{w}}(w)$的似然函数

$$g(w)=\prod_{\tilde{w}\in Context(w)}\prod_{u\in\{w\}\bigcup NEG^{\tilde{w}}(w)}p(u|\tilde{w})$$

其中

$$p(u|\tilde{w})=\begin{cases}\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right),L^w(u)=1\\1-\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right),L^w(u)=0\end{cases}$$

或者

$$p(u|\tilde{w})=\left[\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]^{L^w(u)}\cdot\left[1-\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]^{1-L^w(u)}$$

$NEG^{\tilde{w}}(w)$为处理词$\tilde{w}$时生成的负样本子集。

关于语料库$\mathcal{C}$的对数似然函数

$$\begin{aligned}\mathcal{L}&=\log\prod_{w\in C}g(w)=\sum_{w\in C}\log g(w)\\&=\sum_{w\in C}\log\prod_{\tilde{w}\in Context(w)}\prod_{u\in\{w\}\bigcup NEG^{\tilde{w}}(w)}\left\{\left[\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]^{L^w(u)}\cdot\left[1-\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]^{1-L^w(u)}\right\}\\&=\sum_{w\in C}\sum_{\tilde{w}\in Context(w)}\sum_{u\in\{w\}\bigcup NEG^{\tilde{w}}(w)}\left\{L^w(u)\cdot\log\left[\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]+\left[1-L^w(u)\right]\cdot\log\left[1-\sigma\left(\mathbf{v}(\tilde{w})\top\theta^u\right)\right]\right\}\end{aligned}$$

对数似然函数$\mathcal{L}$关于θ^u的梯度

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\theta^u}&=\frac{\partial}{\partial\theta^u}\left\{\sum_{w\in C}\sum_{\tilde{w}\in Context(w)}\sum_{u\in\{w\}\bigcup NEG(w)}\left\{L^w(u)\cdot\log\left[\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]+\left[1-L^w(u)\right]\cdot\log\left[1-\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]\right\}\right\}\\&=L^w(u)\left[1-\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]\mathbf{v}(\tilde{w})-\left[1-L^w(u)\right]\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\mathbf{v}(\tilde{w})\\&=\left[L^w(u)-\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]\mathbf{v}(\tilde{w})\end{aligned}$$

θ^u的更新

$$\theta^u=\theta^u+\eta\left[L^w(u)-\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]\mathbf{v}(\tilde{w})$$

对数似然函数$\mathcal{L}$关于$\mathbf{v}(\tilde{w})$的梯度

$$\frac{\partial\mathcal{L}}{\partial\mathbf{v}(\tilde{w})}=\sum_{u\in(w)\bigcup NEG^{\tilde{w}}(w)}\left[L^w(u)-\sigma\left(\mathbf{v}(\tilde{w})^\top\theta^u\right)\right]\theta^u$$

$\mathbf{v}(\tilde{w})$的更新

$$\mathbf{v}(\tilde{w})=\mathbf{v}(\tilde{w})+\eta\frac{\partial\mathcal{L}}{\partial\mathbf{v}(\tilde{w})}$$

负采样算法

设词典$\mathcal{D}$中词w_i对应线段$l(w_i)$，长度为

$$len(w_i)=\frac{counter(w_i)}{\sum_{u\in D}counter(u)}$$

其中，$counter(\cdot)$为词在语料$\mathcal{C}$中的出现次数。可将线段$l(w_1)\cdots l(w_N)$拼接为长度为1的单位线段。

记

$$
\begin{aligned}
l_0 &= 0 \\
l_k &= \sum_{j=1}^{k} len\left(w_j\right), k = 1, 2, \cdots, N
\end{aligned}
$$

则以$\{l_j\}_{j=0}^{N}$为剖分点可得到区间$[0, 1]$上的一个非等距剖分

$$I_i = (l_{i-1}, l_i], i = 1, 2, \cdots, N$$

在区间$[0, 1]$上以剖分点$\left\{m_j\right\}_{j=0}^{M}$做等距剖分，其中$M \gg N$。

将等距剖分的内部点$\left\{m_j\right\}_{j=1}^{M-1}$投影到非等距剖分。则可建立$\left\{m_j\right\}_{j=1}^{M-1}$与区间$\left\{I_j\right\}_{j=1}^{N}$的映射，进一步建立与词$\left\{w_j\right\}_{j=1}^{M}$之间的映射

$$w_k = Table\left(i\right), m_i \in I_k, i = 1, 2, \cdots, M-1$$