10.1.2 跳字模型

跳字模型假设基于某个词来生成它在文本序列周围的词。举个例子，假设文本序列是“the”“man”“loves”“his”“son”。以“loves”作为中心词，设背景窗口大小为2。如图10.1所示，跳字模型所关心的是，给定中心词“loves”，生成与它距离不超过2个词的背景词“the”“man”“his”“son”的条件概率，即

$$P(\text{“the”},\text{“man”},\text{“his”},\text{“son”} \mid \text{“loves”}).$$

假设给定中心词的情况下，背景词的生成是相互独立的，那么上式可以改写成

$$P(\text{“the”} \mid \text{“loves”}) \cdot P(\text{“man”} \mid \text{“loves”}) \cdot P(\text{“his”} \mid \text{“loves”}) \cdot P(\text{“son”} \mid \text{“loves”}).$$

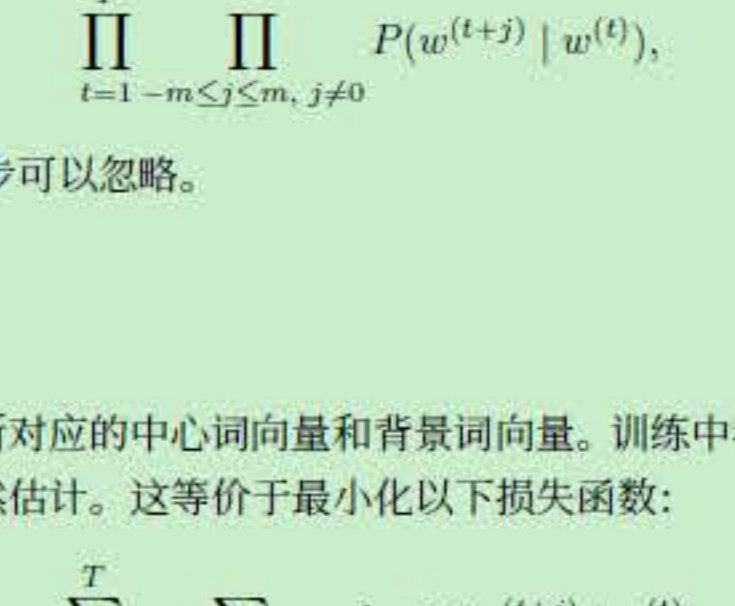

图 10.1: 跳字模型关心给定中心词生成背景词的条件概率

在跳字模型中，每个词被表示成两个d维向量，用来计算条件概率。假设这个词在词典中索引为i，当它为中心词时向量表示为$\boldsymbol{v}_i \in \mathbb{R}^d$，而为背景词时向量表示为$\boldsymbol{u}_i \in \mathbb{R}^d$。设中心词$w_c$在词典中索引为$c$，背景词$w_o$在词典中索引为$o$，给定中心词生成背景词的条件概率可以通过对向量内积做softmax运算而得到：

$$P(w_o \mid w_c) = \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{i \in \mathcal{V}} \exp(\boldsymbol{u}_i^\top \boldsymbol{v}_c)},$$

其中词典索引集$\mathcal{V} = \{0, 1, \ldots, |\mathcal{V}|-1\}$。假设给定一个长度为$T$的文本序列，设时间步$t$的词为$w^{(t)}$。假设给定中心词的情况下背景词的生成相互独立，当背景窗口大小为m时，跳字模型的似然函数即给定任一中心词生成所有背景词的概率

$$\prod_{t=1}^{T} \prod_{-m \le j \le m,\ j \ne 0} P(w^{(t+j)} \mid w^{(t)}),$$

这里小于1和大于T的时间步可以忽略。

训练跳字模型

跳字模型的参数是每个词所对应的中心词向量和背景词向量。训练中我们通过最大化似然函数来学习模型参数，即最大似然估计。这等价于最小化以下损失函数：

$$-\sum_{t=1}^{T} \sum_{-m \le j \le m,\ j \ne 0} \log P(w^{(t+j)} \mid w^{(t)}).$$

如果使用随机梯度下降，那么在每一次迭代里我们随机采样一个较短的子序列来计算有关该子序列的损失，然后计算梯度来更新模型参数。梯度计算的关键是条件概率的对数有关中心词向量和

背景词向量的梯度。根据定义，首先看到

$$\log P(w_o \mid w_c) = \boldsymbol{u}_o^\top \boldsymbol{v}_c - \log\left(\sum_{i \in \mathcal{V}} \exp(\boldsymbol{u}_i^\top \boldsymbol{v}_c)\right)$$

通过微分，我们可以得到上式中$\boldsymbol{v}_c$的梯度

$$\begin{aligned}
\frac{\partial \log P(w_o \mid w_c)}{\partial \boldsymbol{v}_c} &= \boldsymbol{u}_o - \frac{\sum_{j \in \mathcal{V}} \exp(\boldsymbol{u}_j^\top \boldsymbol{v}_c)\boldsymbol{u}_j}{\sum_{i \in \mathcal{V}} \exp(\boldsymbol{u}_i^\top \boldsymbol{v}_c)} \\
&= \boldsymbol{u}_o - \sum_{j \in \mathcal{V}} \left(\frac{\exp(\boldsymbol{u}_j^\top \boldsymbol{v}_c)}{\sum_{i \in \mathcal{V}} \exp(\boldsymbol{u}_i^\top \boldsymbol{v}_c)}\right) \boldsymbol{u}_j \\
&= \boldsymbol{u}_o - \sum_{j \in \mathcal{V}} P(w_j \mid w_c) \boldsymbol{u}_j.
\end{aligned}$$

它的计算需要词典中所有词以w_c为中心词的条件概率。有关其他词向量的梯度同理可得。

训练结束后，对于词典中的任一索引为i的词，我们均得到该词作为中心词和背景词的两组词向量$\boldsymbol{v}_i$和$\boldsymbol{u}_i$。在自然语言处理应用中，一般使用跳字模型的中心词向量作为词的表征向量。

10.1.3 连续词袋模型

连续词袋模型与跳字模型类似。与跳字模型最大的不同在于，连续词袋模型假设基于某中心词在文本序列前后的背景词来生成该中心词。在同样的文本序列“the”“man”“loves”“his”“son”里，以“loves”作为中心词，且背景窗口大小为2时，连续词袋模型关心的是，给定背景词“the”“man”“his”“son”生成中心词“loves”的条件概率（如图10.2所示），也就是

$$P(\text{“loves”} \mid \text{“the”},\text{“man”},\text{“his”},\text{“son”}).$$