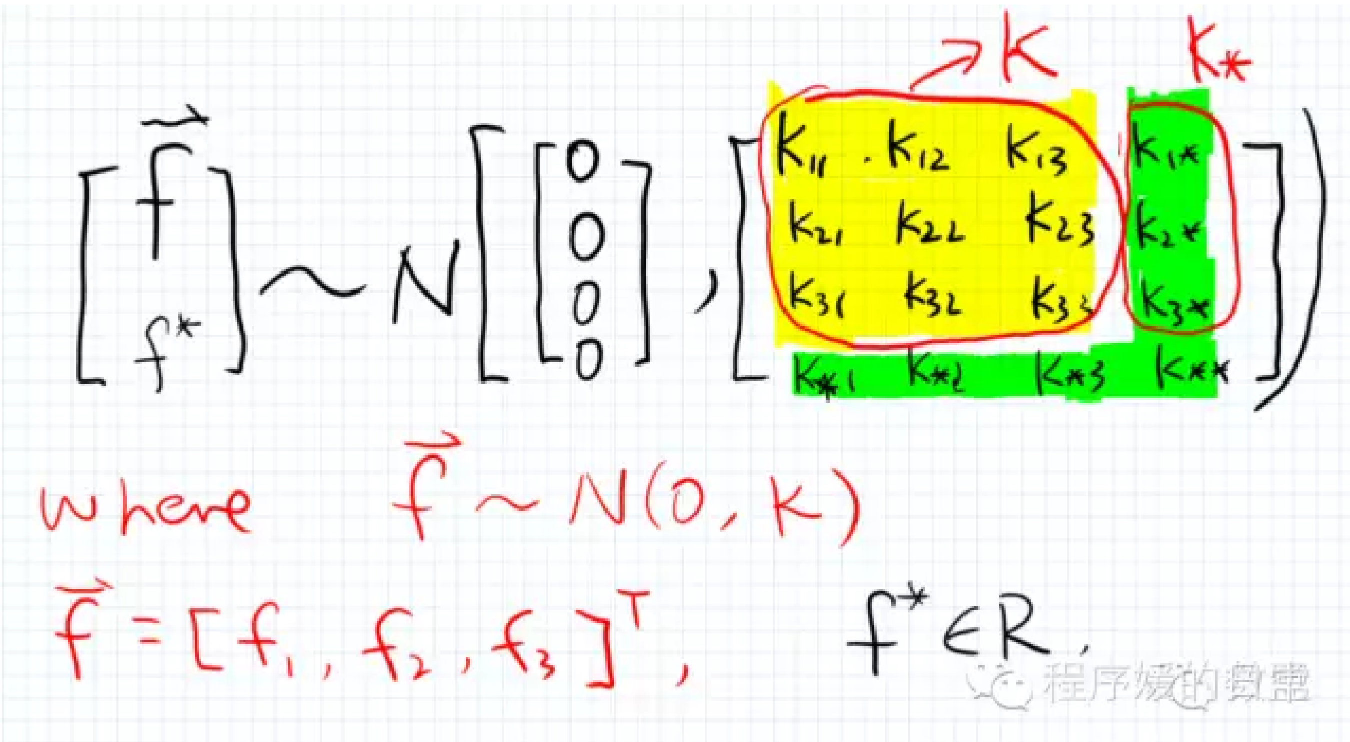
干货 | word2vec 公式推导

小S 程序媛的日常 2015-06-17

今天，答应大家把 **word2vec** 大坑结了（回复代码：GH001 和 GH006）。而且最干最干的公式推导部分。本来，小S 信心满满说要做这件事，是因为看到了好基友**牡丹**同学之前的一篇分享，



手绘风格笔记，萌，炫酷，美！我当时立即问牡丹同学这东西是怎么搞的，他说是 Surface 直接搞的。我当时心想，那我的 DPT-S1 肯定也可以的！于是乎早早爬起的小

S就开始写啊写，结果……_(:з 」∠)_

$$J(\theta)=\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\le j\le c,\\ j\neq 0}}\log p(w_{t+j}|w_t)$$

$$\Rightarrow \quad p(w_O|w_I)=\frac{\exp({v'_{w_O}}^{T} v_{w_I})}{\sum_{w=1}^{W}\exp({v'_w}^{T} v_{w_I})}$$

rewrite

$$\Rightarrow \quad \frac{\partial}{\partial v_c}\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w=1}^{W}\exp(u_w^T v_c)}\right)$$

看来手写还是太丑了，我这样自我要求低的人也不能容忍了。**于是给大家还是乖乖编辑公式！我真的好敬业！都是我写的！**

进入正题，

我们是的主要 idea 是，想通过周围词来表示中心词。周围词就是 contexts, 中心词是 central word。转化为目标函数就是，最大化给定中心词时其他周围词的概率：

$$J(\theta)=\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}|w_t)$$

我们把 上面公式中 log 里面的概率提取出来，并且把下标稍微改变一下，让它们更通用一点。因为是给定中心词，我们假设中心词就是 Input，周围词就是 Output。所以是 wI 和 wO。

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^\top v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left({v'_w}^\top v_{w_I}\right)}$$

这里也要注意，这个v_w是 word 对应的 vector representation。每个 word 实际有两个 vector！（分别作为 input 和 output 的时候）

再继续，大家都知道，我们该求导啦！

在公式推导的时候，如果我们拿不准 vectors 怎么继续求导，可以先变成 vectors 中的单个元素。在这里，我们把这个概率继续改写，因为 wI 和 wO，是 vectors，我们先求单个 word 的求导，再合成 vectors 的求导。

于是就有了，

$$\frac{\partial}{\partial v_c} p(o|c) = \frac{\partial}{\partial v_c} \log\left(\frac{\exp(u_0^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)}\right)$$

在这里，o和c分别就是wO和wI的单个元素（word）。c 就是 central word 的简写，vc 是 central word vector。注意，这里这个概率展开就是 softmax。其实求概率并不是只有 softmax 一种方法，但是 softmax 是相对计算最简单的一种。

现在就该把这个分数给拆了，拆成两部分**（下图公式丢了个大括号！！！！）**：

$$\frac{\partial}{\partial v_c} p(o|c) = \frac{\partial}{\partial v_c} \log\left(\frac{\exp(u_0{}^T v_c)}{\sum_{w=1}^{W} \exp(u_w{}^T v_c)}\right)$$

$$= \frac{\partial}{\partial v_c} \log \exp(u_0{}^T v_c) - \log \sum_{w=1}^{W} \exp(u_w{}^T v_c)$$

分别表示成1，2 **（下图公式丢了个大括号！！！！）**：

$$\frac{\partial}{\partial v_c} p(o|c) = \frac{\partial}{\partial v_c} \log\left(\frac{\exp(u_0{}^T v_c)}{\sum_{w=1}^{W} \exp(u_w{}^T v_c)}\right)$$

$$= \underbrace{\boxed{\frac{\partial}{\partial v_c} \log \exp(u_0{}^T v_c)}}_{1} - \underbrace{\boxed{\log \sum_{w=1}^{W} \exp(u_w{}^T v_c)}}_{2}$$

然后我们分别继续计算，先看1：

$$1: \frac{\partial}{\partial v_c} \log \exp(u_0{}^T v_c)$$

$$= \frac{\partial}{\partial v_c} u_0{}^T v_c = u_0$$

1 很简单，继续看2。2 就需要用到求导中的一个很重要的链式法则了，chain rule。
先看 chain rule 是什么:

$$chain\_rule: \frac{\partial}{\partial v_c} f(g(v_c)) = \frac{\partial f}{\partial z} \frac{\partial z}{\partial v_c}$$

让我们把 chain rule 套用到刚才的第二部分上去，什么是 f() 什么是 g() 呢。我们把 f() 看成 log()，把 log() 里面的部分看成 g()。
也就是说，

$$f() = \log()$$

$$g() = \sum_{w=1}^{W} \exp(u_w{}^T v_c)$$

$$\frac{\partial}{\partial v_c} \log \sum_{w=1}^{W} \exp(u_w{}^T v_c) = \frac{\partial}{\partial v_c} f(g(v_c)) = \frac{\partial f}{\partial z} \frac{\partial z}{\partial v_c}$$

$$= \frac{1}{\sum_{w=1}^{W} \exp(u_w{}^T v_c)} \frac{\partial}{\partial v_c} \sum_{x=1}^{W} \exp(u_x{}^T v_c)$$

我们来看最后一个等号后面，我们再拆出来一个第3部分，同时请注意 x 出现了！这是等价替换，求导的时候避免出错。

$$f() = \log()$$

$$g() = \sum_{w=1}^{W} \exp(u_w{}^T v_c)$$

$$\frac{\partial}{\partial v_c} \log \sum_{w=1}^{W} \exp(u_w{}^T v_c) = \frac{\partial}{\partial v_c} f(g(v_c)) = \frac{\partial f}{\partial z} \frac{\partial z}{\partial v_c}$$

$$= \frac{1}{\sum_{w=1}^{W} \exp(u_w{}^T v_c)} \boxed{\underbrace{\frac{\partial}{\partial v_c} \sum_{\boxed{x=1}}^{W} \exp(u_{\boxed{x}}{}^T v_c)}_{3}}$$

接下来，就变成了再把 3 展开，依然是 chain rule。

$$3: \frac{\partial}{\partial v_c} \sum_{x=1}^{W} \exp(u_x{}^T v_c)$$

$$= \sum_{x=1}^{W} \exp(u_x{}^T v_c) \bullet \frac{\partial}{\partial v_c} u_x{}^T v_c$$

$$= \sum_{x=1}^{W} \exp(u_x{}^T v_c) u_x$$

全部 3 个部分展开完毕，带入原式：

$$
\begin{aligned}
&\frac{\partial}{\partial v_c} p(o \mid c) \\
&= u_0 - \frac{1}{\sum_{w=1}^{W} \exp(u_w{}^T v_c)} \sum_{x=1}^{W} \exp(u_x{}^T v_c) u_x \\
&= u_0 - \sum_{x=1}^{W} \frac{\exp(u_x{}^T v_c)}{\sum_{w=1}^{W} \exp(u_w{}^T v_c)} \cdot u_x \\
&= u_0 - \sum_{x=1}^{W} p(x \mid c) \cdot u_x
\end{aligned}
$$

OK，大功告成。这里，我们继续分析：

$$
\begin{aligned}
&\frac{\partial}{\partial v_c} p(o|c) \\
&= u_0 - \frac{1}{\sum_{w=1}^{W} \exp(u_w{}^T v_c)} \sum_{x=1}^{W} \exp(u_x{}^T v_c) u_x \\
&= u_0 - \sum_{x=1}^{W} \frac{\exp(u_x{}^T v_c)}{\boxed{\underbrace{\sum_{w=1}^{W} \exp(u_w{}^T v_c)}_{\text{cost!!!}}}} \bullet u_x \\
&= u_0 - \sum_{x=1}^{W} p(x|c) \bullet u_x
\end{aligned}
$$

最耗费计算的就是红框的部分。所以需要 subsampling 的技术，subsampling 的话分为两种，一种是 approximate 的方法，一种是 nagetive sampling，也是 word2vec 源码中采用的方法。大家有兴趣去查就好了。

终于大坑结束！
请大家鼓励一下！

下次再见！！！！！