基于 Negative Sampling 的 word2vec 原理

word2vec 是将词转化为词向量，从而可以定量描述它们之间的关系。可以按照以下顺序深入的学习：(1).神经网络语言模型(Neural network language model, NNLM)，(2).传统的 CBOW(Continuous Bag-of-Words)与 Skip-Gram 两种模型，(3).基于 Hierarchical Softmax 的 CBOW 与 Skip-Gram 两种模型，(4).基于 Negative Sampling 的 CBOW 与 Skip-Gram 两种模型。

在 NNLM 中每个词语出现的概率只与这个词语前面 k（包括该词语本身）个词语有关，而 CBOW 与 Skip-Gram 中每个词语出现的概率与其上下文的 $2c$（不包括该词语本身）个词语有关，之所以有不同的版本主要是为了减小计算量。本文只针对基于 Negative Sampling 的 word2vec 原理进行介绍，它包含 CBOW 与 Skip-Gram 两种模型。

1. 基于 Negative Sampling 的 CBOW 模型

在 CBOW 模型中，通过上下文的 $2c$ 个词语 $context(\omega_0)$预测词语 ω_0。由于词语 ω_0 与 $context(\omega_0)$有关，可以把它们当做一个正样本。通过 Negative Sampling，我们可以采样到 neg 个与词语 ω_0 不同的负样本，因此我们可以得到 $neg+1$ 个样本，即：$(context(\omega_0), \omega_i)$，其中，$i=0,1,\ldots,neg$，当 $i=0$ 表示正样本，其他情况表示负样本。

假设在 CBOW 中，输入的 $context(\omega_0)$中每个词语的词向量为 $\boldsymbol{x}_k$（共 $2c$ 个），在投影层取这 $2c$ 个词语的均值 $\boldsymbol{x}_{\omega_a}$，即：

$$\boldsymbol{x}_{\omega_a} = \frac{1}{2c}\sum_{k=1}^{2c}\boldsymbol{x}_k$$

同时，每个词语 ω_i 的参数向量为$\boldsymbol{\theta}^{\omega_i}$。这样我们希望在 $context(\omega_0)$条件下词语 ω_0 发生的概率 $P\left(\omega_0/context(\omega_0)\right)$满足下式：

$$P\left(\omega_0/context(\omega_0)\right) = \sigma\left(\boldsymbol{x}_{\omega_a}^T\boldsymbol{\theta}^{\omega_0}\right)$$

上式中样本 label 为 $y_0=1$，σ 为 sigmoid 函数。同时，我们也希望在 $context(\omega_0)$条件下词语 ω_i 发生的概率 $P\left(\omega_i/context(\omega_i)\right)$满足下式：

$$P\left(\omega_i/context(\omega_0)\right) = 1-\sigma\left(\boldsymbol{x}_{\omega_a}^T\boldsymbol{\theta}^{\omega_i}\right)$$

上式中样本 label 为 $y_i=0$，其中，$i=1,\ldots,neg$，因此我们希望 CBOW 模型的似然函数达到最大，即：

$$\prod_{i=0}^{neg}\sigma\left(\boldsymbol{x}_{\omega_a}^T\boldsymbol{\theta}^{\omega_i}\right)^{y_i}\left(1-\sigma\left(\boldsymbol{x}_{\omega_a}^T\boldsymbol{\theta}^{\omega_i}\right)\right)^{1-y_i}$$

上式中，当 $i=0$ 时，$y_i=1$，否则，$y_i=0$。因此对数似然函数为：

$$L = \sum_{i=0}^{neg} y_i \ln\sigma\left(\boldsymbol{x}_{\omega_a}^T\boldsymbol{\theta}^{\omega_i}\right) + \left(1-y_i\right)\ln\left(1-\sigma\left(\boldsymbol{x}_{\omega_a}^T\boldsymbol{\theta}^{\omega_i}\right)\right)$$

由于采用的是随机梯度上升法，因此 L 对 $\boldsymbol{\theta}^{\omega_i}$ 偏导为：

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\theta}^{\omega_i}} &= y_i \frac{1}{\sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)} \sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\left[1-\sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\right]\boldsymbol{x}_{\omega_a} \\
&\quad +(1-y_i)\frac{-1}{1-\sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)} \sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\left[1-\sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\right]\boldsymbol{x}_{\omega_a} \\
&= y_i\left[1-\sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\right]\boldsymbol{x}_{\omega_a} - (1-y_i)\sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\boldsymbol{x}_{\omega_a} \\
&= \left(y_i - \sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\right)\boldsymbol{x}_{\omega_a}
\end{aligned}
$$

同理，L 对 $\boldsymbol{x}_{\omega_a}$ 偏导为：

$$
\frac{\partial L}{\partial \boldsymbol{x}_{\omega_a}} = \sum_{i=0}^{neg}\left(y_i - \sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\right)\boldsymbol{\theta}^{\omega_i}
$$

因此，

$$
\begin{cases}
\boldsymbol{x}_{\omega_a} \leftarrow \boldsymbol{x}_{\omega_a} + \eta \sum_{i=0}^{neg}\left(y_i - \sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\right)\boldsymbol{\theta}^{\omega_i} \\
\boldsymbol{\theta}^{\omega_i} \leftarrow \boldsymbol{\theta}^{\omega_i} + \eta\left(y_i - \sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right)\right)\boldsymbol{x}_{\omega_a}
\end{cases}
$$

上式中，η 为学习率。故，在随机梯度上升法中，基于 Negative Sampling 的 CBOW 模型算法流程为：

输入：基于 CBOW 的语料训练样本，词向量的维度大小 *Mcount*，CBOW 的上下文大小 $2c$，学习率 η，负采样的个数 *neg*

输出：词汇表每个词对应的模型参数向量 $\boldsymbol{\theta}$，所有的词向量 $\boldsymbol{x}$

1.随机初始化词汇表每个词对应的模型参数向量 $\boldsymbol{\theta}$ 和所有的词向量 $\boldsymbol{x}$

2.对于每个训练样本($context(\omega_0)$, ω_0)，负采样出 *neg* 个负例中心词 ω_i，其中 i=1,2,...*neg*

3.进行梯度上升迭代，将每个训练样本($context(\omega_0)$, ω_0)扩充为每个样本的训练集 ($context(\omega_0)$, ω_0, ω_1,…ω_i)，并做如下处理：

a.计算投影层 $2c$ 个词语的均值 $\boldsymbol{x}_{\omega_a} = \frac{1}{2c}\sum_{k=1}^{2c}\boldsymbol{x}_k$ ，并令 $\boldsymbol{x}_{\omega_a}$ 的偏差 e=0；

b. for i= 0 to *neg*, 计算：

$$
\begin{aligned}
& f = \sigma\left(\boldsymbol{x}_{\omega_a}^T \boldsymbol{\theta}^{\omega_i}\right) \\
& g = \eta\left(y_i - f\right) \\
& e = e + g\boldsymbol{\theta}^{\omega_i} \\
& \boldsymbol{\theta}^{\omega_i} = \boldsymbol{\theta}^{\omega_i} + g\boldsymbol{x}_{\omega_a}
\end{aligned}
$$

c.对 $context(\omega_0)$中的每个词向量 $\boldsymbol{x}_k$（共 $2c$ 个）进行更新：

$$
\boldsymbol{x}_k = \boldsymbol{x}_k + e
$$

4.如果梯度收敛或者达到迭代设定值，则结束梯度迭代，否则返回步骤 3。

2. 基于 Negative Sampling 的 Skip-Gram 模型

在 Skip-Gram 模型中，通过词语 ω_0 预测上下文的 $2c$ 个词语 $context(\omega_0)$。对于 $context(\omega_0)$ 中的每个词语 ω_{0k}($\omega_{0k} \in context(\omega_0)$，其中，$k$=1,…$2c$)，都要进行一次 Negative Sampling，因此对于上下文的每个词语 ω_{0k} 都有 *neg* 个与词语 ω_{0k} 不同的负样本，故在 ω_0 条件下上下文

$context(\omega_0)$发生的概率为：

$$P\left(context\left(\omega_0\right)/\omega_0\right)=\prod_{k=1}^{2c}\prod_{i=0}^{neg}P\left(\omega_{0k}^{i}/\omega_0\right)$$

上式中，$\omega_{0k}^{0}\in context(\omega_0)$，其中，$\omega_{0k}^{i}$ (i=1,…neg)是 ω_{0k}^{0} 的 neg 个负样本。又因为：

$$P\left(\omega_{0k}^{i}/\omega_0\right)=\begin{cases}\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{0}}\right) & i=0, y_{0k}^{0}=1\\ 1-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right) & i\neq 0, y_{0k}^{i}=0\end{cases}$$

上式中，$\boldsymbol{x}_{\omega_0}$ 表示词语 ω_0 的词向量，$\boldsymbol{\theta}^{\omega_{0k}^{0}}$ 表示上下文 $context(\omega_0)$中第 k 个词语对应的模型参数向量，$\boldsymbol{\theta}^{\omega_{0k}^{i}}$ 表示上下文 $context(\omega_0)$中第 k 个词语第 i 个负样本(i=1,…neg)对应的模型参数向量，当 i=0 时，上下文 $context(\omega_0)$中第 k 个词语样本 label 为 $y_{0k}^{0}=1$，当 i=1,…neg 时，上下文 $context(\omega_0)$中第 k 个词语第 i 个负样本 label 为 $y_{0k}^{i}=0$。因此，在 ω_0 条件下上下文 $context(\omega_0)$发生的似然函数为：

$$P\left(context\left(\omega_0\right)/\omega_0\right)=\prod_{k=1}^{2c}\prod_{i=0}^{neg}\left\{\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)^{y_{0k}^{i}}\left[1-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\right]^{1-y_{0k}^{i}}\right\}$$

由于，我们希望在 ω_0 条件下上下文 $context(\omega_0)$发生的似然函数最大，因此对上式取对数得：

$$L=\sum_{k=1}^{2c}\sum_{i=0}^{neg}\left\{y_{0k}^{i}\ln\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)+\left(1-y_{0k}^{i}\right)\ln\left[1-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\right]\right\}$$

因为采用的是随机梯度上升法，因此 L 对 $\boldsymbol{\theta}^{\omega_{0k}^{i}}$ 偏导为：

$$\begin{aligned}\frac{\partial L}{\partial\boldsymbol{\theta}^{\omega_{0k}^{i}}}&=y_{0k}^{i}\frac{1}{\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)}\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\left[1-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\right]\boldsymbol{x}_{\omega_0}\\&\quad+\left(1-y_{0k}^{i}\right)\frac{-1}{1-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)}\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\left[1-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\right]\boldsymbol{x}_{\omega_0}\\&=y_{0k}^{i}\left[1-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\right]\boldsymbol{x}_{\omega_0}-\left(1-y_{0k}^{i}\right)\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\boldsymbol{x}_{\omega_0}\\&=\left(y_{0k}^{i}-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\right)\boldsymbol{x}_{\omega_0}\end{aligned}$$

同理，L 对 $\boldsymbol{x}_{\omega_0}$ 偏导为：

$$\frac{\partial L}{\partial\boldsymbol{x}_{\omega_0}}=\sum_{k=1}^{2c}\sum_{i=0}^{neg}\left(y_{0k}^{i}-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\right)\boldsymbol{\theta}^{\omega_{0k}^{i}}$$

因此，

$$\begin{cases}\boldsymbol{x}_{\omega_0}\leftarrow\boldsymbol{x}_{\omega_0}+\eta\sum_{k=1}^{2c}\sum_{i=0}^{neg}\left(y_{0k}^{i}-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\right)\boldsymbol{\theta}^{\omega_{0k}^{i}}\\ \boldsymbol{\theta}^{\omega_{0k}^{i}}\leftarrow\boldsymbol{\theta}^{\omega_{0k}^{i}}+\eta\left(y_{0k}^{i}-\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\right)\boldsymbol{x}_{\omega_0}\end{cases}$$

上式中，η 为学习率。故，在随机梯度上升法中，基于 Negative Sampling 的 Skip-Gram 模型

算法流程为：
输入：基于 Skip-Gram 的语料训练样本，词向量的维度大小 *Mcount*，Skip-Gram 的上下文大小 2*c*，学习率 η，*context*(ω_0)中每个词语负采样的个数 *neg*
输出：词汇表每个词对应的模型参数向量 $\boldsymbol{\theta}$，所有的词向量 $\boldsymbol{x}$
1.随机初始化词汇表每个词语对应的模型参数向量 $\boldsymbol{\theta}$ 和所有的词向量 $\boldsymbol{x}$
2.对于每个训练样本(ω_0, *context*(ω_0))，对 *context*(ω_0)中每个词语负采样出 *neg* 个负例中心词 ω_{0k}^{i}，其中 *i*=1,2,...*neg*，*k*=1,…2*c*
3.进行梯度上升迭代，将每个训练样本(ω_0, *context*(ω_0))扩充为 2*c* 个样本的训练集 $\left(\omega_0,\omega_{0k}^{0},\omega_{0k}^{1},\dots,\omega_{0k}^{neg}\right)$，并做如下处理：

a. for *k*= 1 to 2*c*, 计算：

(i).令词语 ω_0 词向量 $\boldsymbol{x}_{\omega_0}$ 的偏差 *e*=0；

(ii). for *i*= 0 to *neg*, 计算：

$$
\begin{aligned}
&f=\sigma\left(\boldsymbol{x}_{\omega_0}^{T}\boldsymbol{\theta}^{\omega_{0k}^{i}}\right)\\
&g=\eta\left(y_i-f\right)\\
&e=e+g\boldsymbol{\theta}^{\omega_{0k}^{i}}\\
&\boldsymbol{\theta}^{\omega_{0k}^{i}}=\boldsymbol{\theta}^{\omega_{0k}^{i}}+g\boldsymbol{x}_{\omega_0}
\end{aligned}
$$

(iii).对词语 ω_0 词向量 $\boldsymbol{x}_{\omega_0}$ 进行更新(这里面并不是等所有的 ω_{0k}^{0} 更新完后更新 $\boldsymbol{x}_{\omega_0}$，而是每处理一个 ω_{0k}^{0} 更新一次 $\boldsymbol{x}_{\omega_0}$)：

$$\boldsymbol{x}_{\omega_0}=\boldsymbol{x}_{\omega_0}+e$$

其实这一步可以更新 *context*(ω_0)中的每个词语 ω_{0k}^{0} 的词向量 $\boldsymbol{x}_{\omega_{0k}^{0}}$，因为在词语 ω_0 的条件下，*context*(ω_0)中的每个词语 ω_{0k}^{0} 发生的概率等于在 *context*(ω_0)中的每个词语 ω_{0k}^{0} 发生的条件下，词语 ω_0 发生的概率，所以这一步可以对 *context*(ω_0)中的每个词语 ω_{0k}^{0} 的词向量 $\boldsymbol{x}_{\omega_{0k}^{0}}$（共 2*c* 个）进行更新，即：

$$\boldsymbol{x}_{\omega_{0k}^{0}}=\boldsymbol{x}_{\omega_{0k}^{0}}+e$$

4.如果梯度收敛或者达到迭代设定值，则结束梯度迭代，否则返回步骤 3。

3. Negative Sampling 的负采样方法

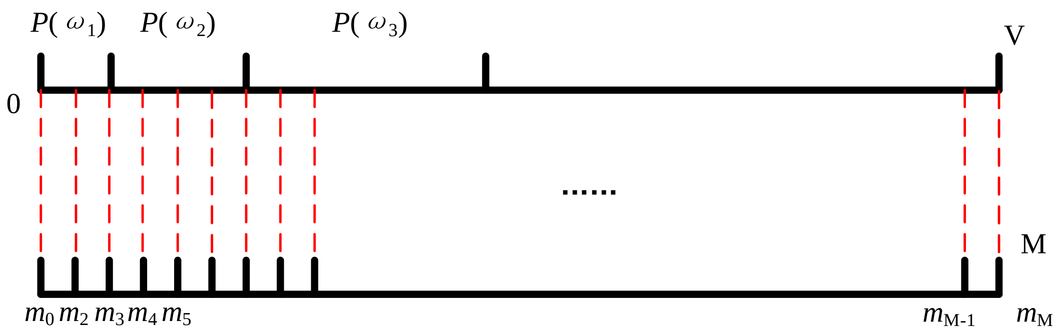
以上介绍的是基于 Negative Sampling 的 CBOW 和 Skip-Gram 模型的算法原理和流程，下面主要介绍如何进行负采样。如果词汇表是 *vocab*，每个词语 ω 的频率为：

$$p(\omega)=\frac{count(\omega)}{\sum_{u\in vocab}count(u)}$$

上式中，$count(\omega)$表示词语 ω 的数量，$count(u)$表示词汇表 *vocab* 中词语 u 的数量。但是在 word2vec 中，每个词语 ω 的频率利用如下公式计算：

$$p(\omega)=\frac{count(\omega)^{3/4}}{\sum_{u\in vocab} count(u)^{3/4}}$$

在实际采样中，假设词汇表的大小为 V，我们将一个长度为 1 的线段按照上述的频率分成 V 段，每段的长度代表上述的频率；同时我们还将这段为 1 的线段分成 M 等分，这里 M>>V。在采样的时候，我们只需要从 M 个位置中采样出 *neg* 个位置就行(如果与正样本相同，需要重新采样)，此时采样到的每一个位置对应到的线段所属的词就是我们的负例词，如下图所示：



通常情况下，在 word2vec 中，$M=10^8$。

4. 高频词采样

实际应用中，总会存在一些频率较高的词语，这些词被认为不会提供太多的信息，因此需要对这些词语进行采样过滤，下面是两种采样的方法：

方法一：

$$keepProbability=\left(\sqrt{\frac{p(\omega)}{t}}+1\right)\times\frac{t}{p(\omega)}=\sqrt{\frac{t}{p(\omega)}}+\frac{t}{p(\omega)}$$

方法二：

$$keepProbability=\sqrt{\frac{t}{p(\omega)}}$$

上述两种方法中，*keepProbability* 表示词语被保留下来的概率，t 表示采样设定的阈值，通常为 0.001。

注：

如果以上看不懂，可以参考以下两篇博客：

1.http://www.cnblogs.com/pinard/p/7249903.html

2.http://www.cnblogs.com/ooon/p/5558119.html