连续词袋模型（CBOW)

与skip-gram 模型相反，CBOW是由背景词推测中心词的模型。

如果输入矩阵是I而输出矩阵是O。矩阵的形状都是$K \times H$，K是代表词典，也词典大小，H是我们设定的中间向量长度。

假设背景词t的独热向量$\hat{t}$，则:

$$
\begin{aligned}
h &= I^T \cdot \hat{t} \qquad (H \times 1) \qquad (1) \\
\hat{o} &= O \cdot h \qquad (K \times 1) \qquad (2)
\end{aligned}
$$

将(1)带入到(2)得到$O \cdot I^T \cdot \hat{t}$，是$K$维向量，记作$\hat{o}$，我们将$\hat{o}$做`softmax`处理，得到背景词对于词典的似然估计$P(t|K)$:

$$
P(t|K) = \begin{bmatrix} \frac{e^{\hat{o}_1}}{\sum_{d=1}^{K} e^{\hat{o}_d}} & \cdots & \frac{e^{\hat{o}_K}}{\sum_{d=1}^{K} e^{\hat{o}_d}} \end{bmatrix}^T
$$

其中对目标中心词r的似然估计，我们通过$P(t|r) = \hat{r}^T \cdot P(t|K)$，$\hat{r}$是中心词的独热向量

如果背景词t在词典K中的索引是j，那么$I^T \cdot \hat{t}$就是输入矩阵I的第j行，也就是说：

$$
h = I^T \cdot \hat{t} = I_j^T
$$

，而

$$
\hat{o} = O \cdot h = \begin{bmatrix} O_1 \cdot I_j^T & \cdots & O_K \cdot I_j^T \end{bmatrix}^T
$$

，其中的$O_1 \cdots O_K$是矩阵O的各行，那么:

$$
P(t|K) = \begin{bmatrix} \frac{e^{O_1 \cdot I_j^T}}{\sum_{d=1}^{K} e^{O_d \cdot I_j^T}} \cdots \frac{e^{O_i \cdot I_j^T}}{\sum_{d=1}^{K} e^{O_d \cdot I_j^T}} \cdots \cdots \frac{e^{O_K \cdot I_j^T}}{\sum_{d=1}^{K} e^{O_d \cdot I_j^T}} \end{bmatrix}
$$

假设中心词r在词典中的所以是i，则:

$$
p(t|r) = \frac{e^{O_i \cdot I_j^T}}{\sum_{d=1}^{K} e^{O_d \cdot I_j^T}} \qquad (3)
$$

需要注意这里的O_i和I_j都是行向量

因为背景词往往是多个(假设中心词是r，在文章中的位置是p，窗口尺寸是w，则背景词分别是$t_{p-w}, t_{p+1-w} + \cdots, t_{p-1}, t_{p+1}, t_{p+2} + \cdots + t_{p+w}$。$2w$个背景词对应的词典索引是$j_1$到$j_{2w}$，对应独热向量分别是：$\hat{t}_{j_1} \cdots \hat{t}_{j_{2w}}$)，所以我们这里的背景词应该采用背景词的向量的平均值，

也就是说

$$
\begin{aligned}
I_j^T &= \frac{1}{2w} I^T \cdot (\hat{t}_{j_1} + \cdots + \hat{t}_{j_{2w}}) \\
&= \frac{1}{2w} (I_{j_1}^T + \cdots + I_{j_{2w}}^T) \qquad (4)
\end{aligned}
$$

我们已知中心词对背景词的似然估计，通过交叉熵损失函数得到损失函数：

$$
\begin{aligned}
l(\Theta) &= -\log P(t|r) \\
&= -log \frac{e^{O_i \cdot I_j^T}}{\sum_{d=1}^K e^{O_d \cdot I_j^T}} \qquad (5)
\end{aligned}
$$

将公式（4）带入到公式（5）：

$$
\begin{aligned}
l(\Theta) &= -log \frac{e^{\frac{1}{2w} \cdot O_i \cdot (I_{j_1}^T + \cdots + I_{j_{2w}}^T)}}{\sum_{d=1}^K e^{\frac{1}{2w} \cdot O_d \cdot (I_{j_1}^T + \cdots + I_{j_{2w}}^T)}} \\
&= -\Big(\log e^{\frac{1}{2w} \cdot O_i \cdot (I_{j_1}^T + \cdots + I_{j_{2w}}^T)} - \log \sum_{d=1}^K e^{\frac{1}{2w} \cdot O_d \cdot (I_{j_1}^T + \cdots + I_{j_{2w}}^T)} \Big) \qquad (6)
\end{aligned}
$$

我们把公式（6）中$\frac{1}{2w} \cdot (I_{j_1}^T + \cdots + I_{j_{2w}}^T)$用$\overline{I_j^T}$替代，公式（6）简化成：

$$
l(\Theta) = -(O_i \cdot \overline{I_j^T} - \log \sum_{d=1}^K e^{O_d \cdot \overline{I_j^T}}) \qquad (7)
$$

公式（7）是损失函数，我们对其求$\overline{I_j^T}$的偏微分，得到$\overline{I_j^T}$的梯度：

$$
\begin{aligned}
\frac{\partial l(\Theta)}{\partial \overline{I_j^T}} &= -\Big(O_i - \frac{e^{O_1 \cdot \overline{I_j^T}} \cdot O_1 + \cdots + e^{O_K \cdot \overline{I_j^T}} \cdot O_K}{\sum_{d=1}^K e^{O_d \cdot \overline{I_j^T}}} \Big) \\
&= -\Big(O_i - \sum_{d=1}^K \frac{e^{O_d \cdot \overline{I_j^T}} \cdot O_d}{\sum_{d=1}^K e^{O_d \cdot \overline{I_j^T}}} \Big) \\
&= \sum_{d=1}^K P(t|r_d) \cdot O_d - O_i \qquad (8)
\end{aligned}
$$