# 10 Review Word Embedding from the Perspective of Matrix Factorization

In section 6 we have derived the objective function of skip-gram aim at noise contrast estimation(NCE). Another method named skip-gram with negative sampling(SGNS) proposed by Mikolov in 2014 [2]has only litter difference with our derivation. Something intresting is that SGNS has been proved result in the same embedding matrices with the factorisation of word-word co-occurrence matrix in later works [5] [4], latter of which is exactly the matrix global vectors training on.

We will discuss the differences of two methods in 10.1, investigate the relation between SGNS and singular value decomposition in 10.2 and prove the equivalence of SGNS and word-word co-occurrence matrix factorization in 10.3.

## 10.1 Skip-Gram with Negative Sampling and Skip-Gram with NCE

SGNS has almost the same optimal object with SG-NCE. For simplicity, we give the objective function of two methods below and discuss their differences directly.

Firstly we give some setting or notation:

We denote the collection of observed target word and context word pairs as $D$ and use $\#(a,b)$ to denote the number of times the pairs $(a,b)$ occurs in $D$. The optimal results are denoted as target words embedding matrix W and context words embedding matrix C.

Then the function, skip-gram with NCE:

$$l = \sum_{a\in V}\{ \sum_{b\in context(a)} \log\sigma(W_a^T \cdot C_b) + k\cdot \mathbb{E}_{b_N\sim P_D} \log\sigma(-W_a^T\cdot C_{b_N})\}$$

$$= \sum_{a\in V}\sum_{b\in V}\#(a,b)\cdot\log\sigma(W_a^T\cdot C_b) + \sum_{a\in V} k\cdot\mathbb{E}_{b_N\sim P_D}\log\sigma(-W_a^T\cdot C_{b_N})$$

Skip-gram with negative sampling:

$$l = \sum_{a\in V}\sum_{b\in V}\#(a,b)\cdot\{\log\sigma(W_a^T\cdot C_b) + k\cdot\mathbb{E}_{b_N\sim P_D}\log\sigma(-W_a^T\cdot C_{b_N})\}$$

$$= \sum_{a\in V}\sum_{b\in V}\#(a,b)\cdot\log\sigma(W_a^T\cdot C_b) + \sum_{a\in V}\sum_{b\in V}\#(a,b)\cdot k\cdot\mathbb{E}_{b_N\sim P_D}\log\sigma(-W_a^T\cdot C_{b_N})$$

Ignore the similarities, NCE focus on a single target word and computes the likelihood of conditional probability, that's probability of a word 'b' appear in the context in condition of a certain target word 'a' being decided. While negative sampling focus on pair of co-occurrence word, it computes the mutual information of all co-occurrence word pairs occur in the collection of corpus, each of which decides by $\frac{\#(a,b)\cdot|D|}{\#(a)\cdot\#(b)}$.

There are differences in objective function due to the distinguishing meaning given to the embedding matrix. In detail we have:

$$\sigma(W_a^T\cdot C_b) \overset{NCE}{\sim} P(b\in context(a)|a) = \frac{\#(a,b)}{\#(a)}$$

$$\sigma(W_a^T \cdot C_b) \overset{NS}{\sim} I(a,b) = \frac{\#(a,b) \cdot |D|}{\#(a) \cdot \#(b)}$$

We have carefully investigated the physical meaning of skip-gram with NCE model in section 6. Conclusion about the situation of negative sampling given here directly will be proved later. The point of this subsection is the symmetry of matrix W and matrix C is different in two methods. That is:

$$\left.\begin{array}{l} \sigma(W_a^T \cdot C_b) \overset{NCE}{=} \frac{\#(a,b)}{\#(a)} \\ \sigma(W_b^T \cdot C_a) \overset{NCE}{=} \frac{\#(a,b)}{\#(b)} \end{array}\right\} \Rightarrow W_a^T \cdot C_b \neq W_b^T \cdot C_a$$

$$\sigma(W_a^T \cdot C_b) \overset{NS}{=} \frac{\#(a,b) \cdot |D|}{\#(a) \cdot \#(b)} \overset{NS}{=} \sigma(W_b^T \cdot C_a) \Rightarrow W_a^T \cdot C_b = W_b^T \cdot C_a$$

Different symmetry means when optimizing reach the ideal optimal point, NS has $W^T = C$ while NCE hasn't.

## 10.2 SGNS as Symmetric Factorization of Matrix

The negative sampling follows distribution which denoted as $P_D$ is decided by $P_D(a) = \frac{\#(a)}{|D|}$. Rewrite the objective for s specific pair (a,b):

$$l(a,b) = \#(a,b) \cdot \{\log \sigma(W_a^T \cdot C_b) + k \cdot \mathbb{E}_{b_N \sim P_D} \log \sigma(-W_a^T \cdot C_{b_N})\}$$

$$= \#(a,b) \cdot \log \sigma(W_a^T \cdot C_b) + k \cdot \#(a,b) \cdot \sum_{b_N \in V_C} \frac{\#(b_N)}{|D|} \cdot \log \sigma(-W_a^T \cdot C_{b_N})$$

$$= \#(a,b) \cdot \log \sigma(W_a^T \cdot C_b) + k \cdot \sum_{b \in V_C} \#(a,b) \cdot \frac{\#(b)}{|D|} \cdot \log \sigma(-W_a^T \cdot C_b)$$

$$= \#(a,b) \cdot \log \sigma(W_a^T \cdot C_b) - k \cdot \frac{\#(a) \cdot \#(b)}{|D|} \cdot \log \sigma(-W_a^T \cdot C_b)$$

For $W_a$ and $C_b$ with sufficiently large dimensionality, we can treat $W_a^T \cdot C_b$ as an independent value from any other pair. We define $x = W_a^T \cdot C_b$, derive objective's partial derivative respect to $x$:

$$\frac{\partial l(a,b)}{\partial x} = \#(a,b) \cdot \log \sigma(-x) - k \cdot \frac{\#(a) \cdot \#(b)}{|D|} \cdot \log \sigma(x)$$

Compare the derivative to zero, we have:

$$\Rightarrow e^{2x} - (\frac{\#(a,b)|D|}{k \cdot \#(a)\#(b)} - 1)e^{2x} - \frac{\#(a,b)|D|}{k \cdot \#(a)\#(b)} = 0$$

By solving this quadratic equation, we have: $e^x = -1$ or:

$$e^x = \frac{\#(a,b)|D|}{\#(a)\#(b) \cdot k}$$

Discard the constant solution we have:

$$W_a^T \cdot C_b = \log\left(\frac{\#(a,b)|D|}{\#(a)\#(b) \cdot k}\right)$$

The result agree with the conclusion we gave in 10.1. Although it's not important to our main topic, we still give the result derived from NCE objective:

$$W_a^T \cdot C_b = \log\left(\frac{\#(a,b)}{\#(a)\cdot k}\right)$$

The derivation above means when optimizing reach the perfect point, skip-gram with negative sampling result in two embedding matrix which's product is word-word mutual information matrix of corpus. We denote the latter as $M$. $M$ is a real symmetric matrix and can be always factorised as:

$$W^T \cdot C = M = U^T \Sigma U$$

or:

$$W^T \cdot C = M = U^T(\sqrt{\Sigma})^T \cdot \sqrt{\Sigma}U = (\sqrt{\Sigma}U)^T \cdot (\sqrt{\Sigma}U)$$

While there is no theoretically demonstrate of $W^T = \sqrt{\Sigma}U$, it works indeed in downstream tasks when we use the symmetric factorization of mutual information matrix as embedding matrix.

## 10.3 Equivalence of SGNS and Matrix Factorization

Following the work above, a method of explicit matrix factorization was proposed in 2015 and proved to be equal to the effect of SGNS optimizing [5]. I decide to omit the detail of this part because the derivation in original paper was perfectly complete, so that I can't give any supplement or modification. Compare to copy their demonstration without dropping any word, I would rather highly recommend that readers of this article should appreciate the paper I cited above.