IIIS 2014 Spring: ATCS - Selected Topics in Optimization
Lecture date: ADD DATE, 2009
Instructor: Jian Li **Scribe: Feng Wenzheng**

1 Introduction

In this lecture, we introduce an effective word embedding method, skip-gram with negative-sampling (SGNS), and prove that it is implicitly factorizing a word-context matrix[1], whose elements are the pointwise mutual information(PMI) of the respective word and context pairs.

2 Skip-Gram with Negative Sampling

The skip-gram model assumes a corpus of words $w \in V_w$ and their contexts $c \in V_c$, where V_w and V_c are the word and context vocabularies. The collection of word-context pairs are denoted as D, and $\#(w,c)$ is the number of times the word-context pair (w,c) appears in D. $\#(w) = \sum_{c' \in V_c} \#(w,c')$ and $\#(c) = \sum_{w' \in V_w} \#(w',c)$ are the number of times w and c occurred in D, respectively. $w \in V_w$ is associated with a vector $\vec{w} \in \mathbb{R}^d$ and similarly $c \in V_c$ is represented as vector $\vec{c} \in \mathbb{R}^d$. We refer to the vectors $\vec{w}$ as rows in a $|V_w| \times d$ matrix W, and to the vectors $\vec{c}$ as roes in a $|V_c| \times d$ matrix C. As for a word-context pair (w,c), the probability distribution that (w,c) came from the data is modeled as:

$$P(D=1|w,c) = \delta(\vec{w} \cdot \vec{c}) = \frac{1}{1+e^{-\vec{w}\cdot\vec{c}}}$$

The objective of negative sampling is to maximize $P(D=1|w,c)$ for observed (w,c) pairs while maximize $P(D=0|w,c) = 1 - P(D=1|w,c)$ for randomly selecting a context for a given word. Then the objective function of SGNS is:

$$J = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c) \log(\delta(\vec{w},\vec{c})) + k\mathbb{E}_{C_N \sim P_D}[\log(\delta(-\vec{w},\vec{c_N}))]) \tag{1}$$

Where k is the number of "negative" samples and c_N is the sampled context, and we assume P_D is the uniform distribution $P_D(c) = \frac{\#(c)}{|D|}$.

3 Word Embedding as Matrix Factorization

Let $M = W \cdot C^T$, then SGNS can be described as factorizing the implicit matrix M of $|V_w| \times |V_c|$ dimensions into two low-rank matrices. A matrix entry M_{ij} is associated to the dot product $W_i \cdot C_j = \vec{w}_i \cdot \vec{c}_j$. Thus SGNS is factorizing a matrix in which each row corresponds to a word $w \in V_w$, each column corresponds to a context $c \in V_c$, and each cell contains a quality $f(w,c)$ reflecting the strength of association between the corresponding (w,c) pair. We can prove that $f(w,c)$ is the PMI of (w,c) with adding a global constant.

Proof:
Rewriting the equation 1:

$$\begin{aligned} J &= \sum_{w\in V_w}\sum_{c\in V_c} \#(w,c)\log(\delta(\vec{w},\vec{c})) + k\mathbb{E}_{C_N\sim P_D}[\log(\delta(-\vec{w},\vec{c_N}))]) \\ &= \sum_{w\in V_w}\sum_{c\in V_c} \#(w,c)\log(\delta(\vec{w},\vec{c})) + \sum_{w} \#(w)[k\mathbb{E}_{C_N\sim P_D}[\log(\delta(-\vec{w},\vec{c_N}))]] \\ &= \sum_{w\in V_w}\sum_{c\in V_c} \#(w,c)\log(\delta(\vec{w},\vec{c})) + \sum_{w} \#(w)k\sum_{c_N\in V_c}\frac{\#(c_N)}{|D|}\log(\delta(-\vec{w},\vec{c_N})) \end{aligned} \tag{2}$$

Denote $J(w,c)$ as the single objective for (w,c), i.e.$J=\sum_{w,c}J(w,c)$, then:

$$J(w,c) = \#(w,c)\log(\delta(\vec{w},\vec{c})) + k\#(w)\frac{\#(c_N)}{|D|}\log(\delta(-\vec{w},\vec{c_N})) \tag{3}$$

We define $x=\vec{w}\cdot\vec{c}$. For optimizing the objective, we compute the partial derivative with respect to x:

$$\frac{\partial J(w,c)}{\partial x} = \#(w,c)\delta(-x) - k\frac{\#(w)\#(c)}{|D|}\delta(x) \tag{4}$$

Let $\frac{\partial J(w,c)}{\partial x}=0$:

$$\#(w,c)\delta(-x) - k\frac{\#(w)\#(c)}{|D|}\delta(x) = 0 \tag{5}$$

$$\Rightarrow |D|\#(w,c)(1+e^{-x}) - k\#(w)\#(c)(1+e^{x}) = 0 \tag{6}$$

$$\Rightarrow e^{2x} - (\frac{|D|\#(w,c)}{k\#(w)\#(c)} - 1)e^{x} - \frac{|D|\#(w,c)}{k\#(w)\#(c)} = 0 \tag{7}$$

Let $y=e^x$, then we can solve y from the quadratic equation of it, which has two equations, $y=-1$(invalid) and :

$$y = \frac{D\#(w,c)}{k\#(w)\#(c)} \tag{8}$$

Then

$$\vec{w}\cdot\vec{c} = \log(y) = \log(\frac{|D|\#(w,c)}{\#(w)\#(c)}) - \log(k)$$

The expression $\log(\frac{|D|\#(w,c)}{\#(w)\#(c)})$ is the pointwise mutual information of (w,c). Thus we can prove the matrix M is factorizing:

$$M_{ij}^{SGNS} = W_i\cdot C_j = \vec{w}_i\cdot\vec{c}_j = PMI(w_i,c_j) - \log k \tag{9}$$

References

[1] Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization[C]//Advances in neural information processing systems. 2014: 2177-2185.