Derivation of skip-gram with negative sampling

Matthias Gallé

August 2017

Note: Follows closely Goldberg and Levy [2014]

Given a set of pairs of words $W = (x_i, y_i)$ with y_i in the context of x_i (note that this is a symmetric property), we want to find:

$$\arg\max_{\theta} = \prod_{(x,y)\in W} \left(P(I=1|x,y,\theta) \prod_{z\in ns(x)} P(I=0|x,z,\theta) \right) \tag{1}$$

θ are the parameter giving the word and context embeddings and $ns(x)$ returns a set of negative context for word x.

$P(I=1|x,y,\theta)$ will be modeled as $\frac{1}{1+e^{-x\cdot y}}$.

Equation 1 then becomes (taking log to go to sum):

$$\arg\max_{\theta} = \sum_{(x,y)\in W} \left(\log \frac{1}{1+e^{-x\cdot y}} + \sum_{z\in ns(x)} \log\left(1 - \frac{1}{1+e^{-x\cdot z}}\right) \right) \tag{2}$$

$$\sum_{(x,y)\in W} \left(\log \frac{1}{1+e^{-x\cdot y}} + \sum_{z\in ns(x)} \log \frac{1}{1+e^{x\cdot z}} \right) \tag{3}$$

$$\sum_{(x,y)\in W} \left(\log \sigma(x\cdot y) + \sum_{z\in ns(x)} \log \sigma(-x\cdot z) \right) \tag{4}$$

$$\tag{5}$$

For deriving the gradient, note that $\frac{\partial \log \sigma(xy)}{\partial x} = \frac{y}{1+e^{xy}}$

References

Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.