Exercises
Deep Learning
Fall 2018

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Fernando Pérez-Cruz
Web http://www.da.inf.ethz.ch/teaching/2018/DeepLearning/

Series Monday, Nov 27, 2017
(Deep Learning, Exercise series 8 - solutions)

Solution 1 (Word Embedding):

a.

$$P(D_{v,w} = 1|\boldsymbol{x}_v, \boldsymbol{z}_w) = \sigma(\boldsymbol{x}_v^\top \boldsymbol{z}_w), \quad \sigma(\beta) := 1/(1 + \exp(-\beta)) \tag{1}$$

b.

$$\begin{aligned} P(\mathcal{S}, \bar{\mathcal{S}}|\boldsymbol{x}_{v_i}s, \boldsymbol{z}_{w_i}s) &= \prod_{(v_i,w_i)\in\mathcal{S}} P(D_{v_i,w_i} = 1|\boldsymbol{x}_{v_i}, \boldsymbol{z}_{w_i}) \prod_{(v_i,w_i)\in\bar{\mathcal{S}}} P(D_{v_i,w_i} = 0|\boldsymbol{x}_{v_i}, \boldsymbol{z}_{w_i}) && (2) \\ &= \prod_{(v_i,w_i)\in\mathcal{S}} \sigma(\boldsymbol{x}_{v_i}^\top \boldsymbol{z}_{w_i}) \prod_{(v_j,w_j)\in\bar{\mathcal{S}}} (1 - \sigma(\boldsymbol{x}_{v_j}^\top \boldsymbol{z}_{w_j})) && (3) \\ &= \prod_{(v_i,w_i)\in\mathcal{S}} \sigma(\boldsymbol{x}_{v_i}^\top \boldsymbol{z}_{w_i}) \prod_{(v_j,w_j)\in\bar{\mathcal{S}}} (\sigma(-\boldsymbol{x}_{v_j}^\top \boldsymbol{z}_{w_j})) && (4) \end{aligned}$$

Note that in last step we used the fact that

$$1 - \sigma(\beta) = 1 - 1/(1 + \exp(-\beta)) = \exp(-\beta)/(1 + \exp(-\beta)) = 1/(1 + \exp(\beta)) = \sigma(\beta). \tag{5}$$

Finally, we derive the logarithm of the likelihood function as

$$\log(P(\mathcal{S}, \bar{\mathcal{S}}|\boldsymbol{x}_{v_i}s, \boldsymbol{z}_{w_i}s)) = \sum_{(v_i,w_i)\in\mathcal{S}} \log(\sigma(\boldsymbol{x}_{v_i}^\top \boldsymbol{z}_{w_i})) + \sum_{(v_j,w_j)\in\bar{\mathcal{S}}} \log(\sigma(-\boldsymbol{x}_{v_j}^\top \boldsymbol{z}_{w_j})) \tag{6}$$

c. Objective L is easier to optimize because of the sampling over the set of negative samples, i.e. $\bar{\mathcal{S}}$. This set can be relatively larger than set $\mathcal{S}$. The sampling over the negative samples is according to the frequency of words and parameter α determines the weight of the frequency in negative sampling.