3: word2vec

(a) Problem: Given a predicted word vector v_c for center word c for skip-gram and softmax prediction, find the derivative of the cross entropy cost wrt v_c

Answer: The soft-max prediction gives:

$$\hat{y}_o = p(o|c) = \frac{\exp\left(\boldsymbol{u}_o^T \boldsymbol{v}_c\right)}{\sum_{w=1}^{V} \exp\left(\boldsymbol{u}_w^T \boldsymbol{v}_c\right)} \tag{1}$$

Cross entropy loss is given by

$$\mathrm{CE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = J = -\sum_{i=1}^{V} y_i \log \hat{y}_i \tag{2}$$

where $\hat{\boldsymbol{y}}$ is a $1 \times V$ matrix/vector whose components are the softmax results, and $\boldsymbol{y}$ is a $1 \times V$ matrix/vector that is 1-hot encoded with 1 at the oth entry. So, we can write $\hat{\boldsymbol{y}}$ as

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{z}) \tag{3}$$

where $\boldsymbol{z}$ comprises $z_i = \boldsymbol{u}_i^T \boldsymbol{v}_c$.

We want to calculate, by chain rule

$$\frac{\partial J}{\partial \boldsymbol{v}_c} = \frac{\partial J}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{v}_c} \tag{4}$$

where we are multiplying a $1 \times V$ matrix with a $V \times V$ matrix.

From previous results we have

$$\frac{\partial J}{\partial \boldsymbol{z}} = \hat{\boldsymbol{y}} - \boldsymbol{y} \tag{5}$$

If we consider $\boldsymbol{z}$ a column vector, and adopt numerator layout for derivatives we get

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{v}_c} = \begin{bmatrix} \boldsymbol{u}_1 & \dots & \boldsymbol{u}_V \end{bmatrix} = \boldsymbol{U} \tag{6}$$

This gives

$$\frac{\partial J}{\partial \boldsymbol{v}_c} = (\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{U} \tag{7}$$

(b) Problem: Find the partial derivative of the above cost function wrt $\boldsymbol{u}_k$

Answer: We need to calculate

$$\frac{\partial J}{\partial \boldsymbol{u}_k} = \frac{\partial J}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{u}_k} \tag{8}$$

We use results from the preceding sub-problem for $\frac{\partial J}{\partial \boldsymbol{z}}$ and note that

$$\frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{u}_k} = \boldsymbol{v}_c \delta_{ik} \tag{9}$$

Because of the Kronecker delta we can write

$$\frac{\partial J}{\partial \boldsymbol{u}_k} = \frac{\partial J}{\partial \boldsymbol{z}_i} \frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{u}_k} \tag{10}$$

This gives

$$\frac{\partial J}{\partial \boldsymbol{u}_k} = (\hat{\boldsymbol{y}} - \boldsymbol{y})_k \boldsymbol{v}_c \tag{11}$$

(c) Problem: Repeating **(a)** for a negative sampling cost function

Answer: Given the cost function

$$J = -\log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) - \sum_{k=1}^{K} \log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \tag{12}$$

we use the chain rule

$$\begin{aligned} \frac{\partial J}{\partial \boldsymbol{v}_c} &= -\frac{\boldsymbol{u}_o^T \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)(1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)} - \sum_{k=1}^{K} \frac{\boldsymbol{u}_k^T \sigma(\boldsymbol{u}_k^T \boldsymbol{v}_c)(1 - \sigma(\boldsymbol{u}_k^T \boldsymbol{v}_c))}{\sigma(\boldsymbol{u}_k^T \boldsymbol{v}_c)} \\ &= -\boldsymbol{u}_o^T (1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) - \sum_{k=1}^{K} \boldsymbol{u}_k^T (1 - \sigma(\boldsymbol{u}_k^T \boldsymbol{v}_c)) \end{aligned} \tag{13}$$

Similarly

$$\frac{\partial J}{\partial \boldsymbol{u}_o} = (\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c) - 1) \boldsymbol{v}_c^T \tag{14}$$

and

$$\frac{\partial J}{\partial \boldsymbol{u}_k} = (1 - \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \boldsymbol{v}_c^T \tag{15}$$

for $o \neq k$