Solution for Assignment 1

1. Softmax

(a). $softmax(x_i + c) = \frac{e^{x_i+c}}{\Sigma_j e^{x_j+c}} = \frac{e^c * e^{x_i}}{e^c * \Sigma_j e^{x_j}} = \frac{e^{x_i}}{\Sigma_j e^{x_j}} = softmax(x_i)$

2. Neural Network Basics

(a). $\sigma(x) = \frac{1}{1+e^{-x}}$ $\sigma'(x) = \frac{1}{(1+e^{-x})^2} e^{-x} = \sigma^2(x)(\frac{1}{\sigma(x)} - 1) = \sigma(x)(1 - \sigma(x))$

(b). $CE(y, \hat{y}) = -\Sigma_i y_i log(\hat{y}_i)$

Assume: $y_i = 1, i = k$; $y_i = 0, i \neq k$

Thus: $CE(y, \hat{y}_i) = -log(softmax(\theta_k))$

$$CE(y, \hat{y}) = -\frac{\alpha y_k log \frac{e^{\theta_k}}{\Sigma_j e^{\theta_j}}}{\alpha \theta} = \hat{y} - y$$

(c). Assume: $\theta_i^{(1)} = \Sigma_j x_j W_{ji}^{(1)} + b_1$

$h_i = \sigma(\theta_i^{(1)})$

$\theta_i^{(2)} = \Sigma_j h_j W_{ji}^{(2)}$

$\hat{y}_i = softmax(\theta_i^{(2)})$

Then:

$\frac{\alpha}{\alpha h_i} J = \Sigma_j \frac{\alpha J}{\alpha \theta_j^{(2)}} \frac{\alpha \theta_j^{(2)}}{\alpha h_i} = \Sigma_j (\hat{y}_j - y_j) * W_{ij}^{(2)} = (\hat{y} - y) W_{i*}^{(2)T}$

$\frac{\alpha}{\alpha \theta_i^{(1)}} J = \Sigma_j \frac{\alpha J}{\alpha h_j} \frac{\alpha h_j}{\alpha \theta_i^{(1)}} = \Sigma_j (\hat{y} - y) W^{(2)T} \frac{\alpha \sigma(\theta_j^{(1)})}{\alpha \theta_i^{(1)}} = (\hat{y} - y) W_{i*}^{(2)T} \sigma'(\theta_i^{(1)})$

Thus:

$\frac{\alpha}{\alpha x_i} J = \Sigma_j \frac{\alpha J}{\alpha \theta_j^{(1)}} \frac{\alpha \theta_j^{(1)}}{\alpha x_i} = \Sigma_j (\hat{y} - y) W_{j*}^{(2)T} \frac{\alpha \Sigma_i W_{ij}^{(1)} + b_i}{\alpha x_i} = \Sigma_j (\hat{y} - y) W_{j*}^{(2)T} \sigma'(\theta_j) W_{ij} = [(\hat{y} - y)(W^{(2)T} \bigotimes \sigma'(\theta^{(1)}))] W_{i*}^{(1)T}$

, Where $\bigotimes$ denotes broadcasting and element-wise multiply.

(d). The first Parameter Matrix is of $(D_x + 1) * H$ size, The second Parameter Matrix is of $(H + 1) * D_y$ size (1 denotes bias); Thus total parameter number is
$(D_x + 1) * H + (H + 1) * D_y = H(D_x + D_y) + H + D_y$

(g). For $W^{(2)}$:

$\frac{\alpha}{\alpha W_{ij}^{(2)}} J = \Sigma_j \frac{\alpha J}{\alpha \theta_j^{(2)}} \frac{\alpha \theta_j^{(2)}}{\alpha W_{ij}^{(2)}} = (\hat{y}_j - y_j) h_i$

Thus: $\frac{\alpha}{\alpha W^{(2)}} J = h^T (\hat{y} - y)$

For $W^{(1)}$:

$\frac{\alpha}{\alpha W_{ij}^{(1)}} J = \Sigma_j \frac{\alpha J}{\alpha \theta_j^{(1)}} \frac{\alpha \theta_j^{(1)}}{\alpha W_{ij}^{(1)}} = (\hat{y} - y) W_{i*}^{(2)T} \sigma'(\theta_i^{(1)}) x_i$

Thus: $\frac{\alpha}{\alpha W^{(1)}} J = X^T \frac{\alpha J}{\alpha \theta^{(1)}}$

3. word2vec

(a). assume: all vectors below are coloumn vectors. Let $\theta = U^T v_c$,

Then $\frac{\alpha CE(y,\hat{y_o})}{\alpha v_c} = \frac{\alpha CE(y,\hat{y}_o)}{\alpha \theta} \frac{\alpha \theta}{\alpha v_c} = U(\hat{y} - o)$

(b). $\frac{\alpha CE(y,\hat{y_o})}{\alpha U_{ij}} = \Sigma_k \frac{\alpha CE(y,\hat{y_o})}{\alpha \theta_k} \frac{\alpha \theta_k}{\alpha U_{ij}} = \Sigma_k (\hat{y_k} - o_k) \frac{\alpha \Sigma_i U_{ki}^T (V_c)_i}{\alpha U_{ij}} = (\hat{y_j} - o_j)(v_c)_i$

Thus: $\frac{\alpha CE(y,\hat{y_o})}{\alpha U} = v_c (\hat{y} - o)^T$

(c). $\frac{\alpha J_{neg}}{\alpha v_c} = -\frac{\alpha log(\sigma(u_o^T v_c))}{\alpha v_c} - \Sigma_{k=1}^K \frac{\alpha log(\sigma(-u_k^T v_c))}{\alpha v_c}$

$= -\frac{1}{\sigma(u_o^T v_c)} \sigma(u_o^T v_c)(1 - \sigma(u_o^T v_c)) u_o - \Sigma_{k=1}^K \frac{1}{\sigma(-u_k^T v_c)} \sigma(-u_k^T v_c)(1 - \sigma(-u_k^T v_c))(-u_k)$

$= (\sigma(u_o^T v_c) - 1) u_o - \Sigma_{k=1}^K (\sigma(-u_k^T v_c) - 1) u_k$

$\frac{\alpha J_{neg}}{\alpha u_w} = -\frac{\alpha log(\sigma(u_o^T v_c))}{\alpha u_w} - \Sigma_{k=1}^K \frac{\alpha log(\sigma(1 - u_k^T v_c))}{\alpha u_w}$,

When $w = 0$, $\frac{\alpha J_{neg}}{\alpha u_w} = [\sigma(u_o^T v_c) - 1] v_c$,

When $w \in S$, $\frac{\alpha J_{neg}}{\alpha u_w} = [1 - \sigma(-u_w^T v_c)] v_c$

When $w \notin S$, $\frac{\alpha J_{neg}}{\alpha u_w} = 0$

Reason for why this cost function can be more effienct: $CE(y, \hat{y})$ contains lots of exponents calculation while this cost function only contains K+1 exponents calculation. And exponents calculation requires lots of calculating time.

(d). **For Skip-Gram**:

$\frac{\alpha}{\alpha v_k} J_{skip-gram} = \Sigma_{-m \le j \le m, j \ne 0} \frac{\alpha F(w_{c+j}, v_c)}{\alpha v_k}$,

$\frac{\alpha}{\alpha u_k} J_{skip-gram} = \Sigma_{-m \le j \le m, j \ne 0} \frac{\alpha F(w_{c+j}, v_c)}{\alpha u_k}$

For CBOW:

$\frac{\alpha}{\alpha v_k} J_{CBOW} = \frac{\alpha F(w_{c+w}, \hat{v})}{\alpha v_k} = \frac{\alpha F(w_{c+w}, \hat{v})}{\alpha \hat{v}} \frac{\alpha \hat{v}}{\alpha v_k}$

when $k \in [c-m, c+m], k \ne c$, $\frac{\alpha}{\alpha v_k} J_{CBOW} = \frac{\alpha F(w_{c+w}, \hat{v})}{\alpha \hat{v}}$

else, $\frac{\alpha}{\alpha v_k} J_{CBOW} = 0$

$\frac{\alpha}{\alpha u_k} J_{CBOW} = \frac{\alpha F(w_c, \hat{v})}{\alpha u_k}$

4. Sentiment Analysis

(b). Reason: avoid parameters being too large; also this can avoid overfitting.

(d). Reason:

The glove model is trained by large contexts, so the pretrained vectors can better describe a word in vector.

The pretrained vectors have more dimensions(50-dimension) than what we trained(10-dimension).

The glove model is trained by statistical methods, which is different from skip-gram methods. The glove model has a better model for word vector training.

(e). As regualrization value grows larger, training accuracy is droping while dev accuracy is first increasing then decreasing.

We can see the regularization here plays a role for avoid overfitting.

But when the regularization growing too large, the cost function seems will be smaller when the parameter is small, so in this case, too large regularization value will cause parameter becoming very small.

(f). The model tends to predict sentences as '+' or '-'. This may because the sentiment analysis criteria is obsure and the labelled '+' or '-' training sets can cross over the whole vector space. And the vector summing up for average seems not a good idea, because when "not good" with "bad" vs. "not bad" with "good" will sometimes generate the same average vector, but these are two different situations.

(g). Sentence1(line71): it 's refreshing to see a girl-power movie that does n't feel it has to prove anything .

Analyze: the most important attitude word here is "refreshing", however, the whole words after "see" have words like "doesn't feel", "anything" can lead the sentence to a negative space. The reason is that all vectors weights the same in senctence average vector.