Stanford CS224n Assignment 1.

1. Softmax.

a) Q. proof that $\forall x \in \mathbb{R}^n, \forall c \in \mathbb{R}$: softmax$(x+c)$ = softmax(c)

A. Softmax$(x+c)$ = softmax$(\langle x_1+c; x_2+c; \cdots; x_i+c; \cdots x_n+c \rangle)$

Let softmax$(x+c)_i$ be the i^{th} component of softmax$(x+c)$

$$\text{softmax}(x+c)_i = \frac{e^{x_i+c}}{\sum_{j=1}^{n} e^{x_j+c}}$$

$$= \frac{e^{c} \cdot e^{x_i}}{\sum_j e^{c} \cdot e^{x_j}}$$

$$= \frac{\cancel{e^{c}} \cdot e^{x_i}}{\cancel{e^{c}} \cdot \sum_j e^{x_j}}$$ (can simplify because e^{c} is never $= 0 \; \forall c$)

$$= \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$= \text{softmax}(x)_i$$

$$\Rightarrow \boxed{\text{Softmax}(x+c) = \text{Softmax}(x)}$$

1

2) Neural Network Basics

a) $$\sigma(x) = \frac{1}{1+\exp(-x)}$$

$$\Rightarrow \frac{d\sigma}{dx}(x) = \frac{0\,(1+e^{-x}) - 1\cdot(-e^{-x})}{(1+e^{-x})^2}$$

$$= \frac{e^{-x} + (1-1)}{(1+e^{-x})(1+e^{-x})}$$

$$= \frac{\cancel{1+e^{-x}}}{\cancel{(1+e^{-x})}(1+e^{-x})} - \frac{1}{(1+e^{-x})(1+e^{-x})}$$

$$= \frac{1}{(1+e^{-x})} - \frac{1}{(1+e^{-x})^2}$$

$$= \sigma(x) - \sigma^2(x)$$

$$\boxed{\frac{d\sigma}{dx}(x) = \sigma(x)\,(1-\sigma(x))}$$

b) $$\nabla_\theta CE(y,\hat{y}) = \nabla_\theta\left(-\sum_i y_i \log(\hat{y}_i)\right)$$

$$= \nabla_\theta\left(-\log(\hat{y}_k)\right)$$ | where k is the correct class

$$= -\frac{1}{\hat{y}_k}\cdot\nabla_\theta \hat{y}_k \qquad (1)$$

$$\cancel{\nabla_\theta \hat{y}_k = \nabla_\theta \frac{e^{\theta_k}}{\sum_i e^{\theta_i}} = \frac{e^{\theta_k}\cdot\Sigma - e^{\theta_k}\cdot e^{\theta_k}}{\Sigma^2}}$$

$$\cancel{= \left(\frac{1}{\Sigma}\cdot\hat{y}_k\right) - \left(\hat{y}_k\right)^2}$$

$$\nabla_\theta \hat{y}_k = \,?$$

Let's distinguish two case $\left.\frac{\partial \hat{y}_k}{\partial \theta_i}\right|_{i=k}$ (a) and $\left.\frac{\partial \hat{y}_k}{\partial \theta_i}\right|_{i\neq k}$ (b)

(a) $$\frac{\partial \hat{y}_k}{\partial \theta_k} = \frac{\partial}{\partial \theta_k}\frac{e^{\theta_k}}{\sum_j e^{\theta_j}} = \frac{e^{\theta_k}\cdot\Sigma_j - e^{\theta_k}\cdot e^{\theta_i}}{\Sigma_j^2} = \frac{e^{\theta_k}}{\Sigma_j}\cdot\left(1 - \frac{e^{\theta_j}}{\Sigma_j}\right)$$

$$= \hat{y}_k\,(1-\hat{y}_j)$$

2)

b) $CE(y,\hat{y}) = -\sum_j y_j \log(\hat{y}_j)$ $\qquad \hat{y}_j = softmax(z)_j$

$$\nabla_z CE(y,\hat{y}) = \nabla_z \left(-\sum y_j \log(\hat{y}_j)\right)$$

$$= -\nabla_z \log(\hat{y}_k) \text{ ; where } k \text{ is the correct label}$$

$$= -\nabla_z \log\left(\frac{e^{z_k}}{\sum_j e^{z_j}}\right)$$

$$= -\nabla_z \left(\log(e^{z_k})\right) + \nabla_z \log\left(\sum_j e^{z_j}\right)$$

$$\cancel{= -\nabla_z e^{z_k}}$$

$$= \underbrace{-\nabla_z (z_k)}_{(1)} + \underbrace{\nabla_s \log(s) \otimes \nabla_z (s) \Big|_{s=\sum_j e^{z_j}}}_{(2)}$$

- let's simplify the first part (1)

$$\nabla_z (z_k) = \begin{bmatrix} \frac{\partial z_k}{\partial z_1} \\ \vdots \\ \frac{\partial z_k}{\partial z_k} \\ \vdots \\ \frac{\partial z_n}{\partial z_k} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 1 \}\text{row}=k\} \\ \vdots \\ 0 \end{bmatrix} = y \text{ (one\_hot)} \quad (3)$$

- now let's handle the second part (2)

* $\nabla_s \log(s) = \frac{1}{s} = \frac{1}{\sum_j e^{z_j}}$ (4)

* $\nabla_z (s) = \nabla_z \left(\sum_j e^{z_j}\right) = \begin{bmatrix} \frac{\partial}{\partial z_1}(e^{z_1} + \underbrace{e^{z_2} + \dots e^{z_n}}_{0 \dots 0}) \\ \vdots \\ \frac{\partial}{\partial z_i}(\underbrace{e^{z_1} + e^{z_2} + \dots}_{0 \dots 0} e^{z_i} + \underbrace{\dots e^{z_n}}_{0 \dots 0}) \\ \vdots \\ \frac{\partial}{\partial z_n}(\underbrace{e^{z_1} + e^{z_2} + \dots}_{0 \dots 0} + e^{z_n}) \end{bmatrix}$

$$= \begin{bmatrix} \vdots \\ e^{z_i} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ e^{z_i} \\ \vdots \end{bmatrix}$$

$\sigma'(z) = \sigma(z)(1-\sigma(z))$

$\frac{\partial e^{x}}{\partial x} = \frac{\partial e^{z}}{\partial z} \cdot \frac{\partial z}{\partial x}$

(3), (4), (5) $\Rightarrow \nabla_z CE(y, \hat{y})$

$$= -y + \underbrace{\left(\begin{bmatrix} \vdots \\ e^{z_i} \\ \vdots \end{bmatrix} \cdot \frac{1}{\sum_j e^{z_j}}\right)}$$

$$= -y + \hat{y}$$

$$\Rightarrow \boxed{\nabla_z CE(y, \hat{y}) = \hat{y} - y}$$

c) $\nabla_x J = \nabla_x CE(y, \hat{y}) = ?$

The flow is the following

$$x \longrightarrow h \longrightarrow \hat{y} \longrightarrow J = CE(y, \hat{y})$$

$h = \text{sigmoid}(xW_1 + b_1)$, $\hat{y} = \text{softmax}(hW_2 + b_2)$, $J = -\sum_j y_j \log(\hat{y}_j)$

in the previous question we found $\frac{\partial J}{\partial h} = \hat{y} - y$

$$\Rightarrow \frac{\partial J}{\partial x} = \frac{\partial J}{\partial h} \cdot \frac{\partial h}{\partial x} = (\hat{y} - y) \cdot \frac{\partial h}{\partial x}$$

so we need to find $\frac{\partial h}{\partial x}$

$$\frac{\partial h}{\partial x} = \frac{\partial}{\partial x} \text{sigmoid}(xW_1 + b_1) = \sigma'(xW_1 + b_1) \cdot \frac{\partial}{\partial x}(xW_1 + b_1)$$

$$= \sigma'(xW_1 + b_1)$$

2)

c) $\frac{\partial J}{\partial x} = \frac{\partial CE(y,\hat{y})}{\partial x} = ?$

The network look like this:

$$\underset{\in \mathbb{R}^{D_x}}{x} \longrightarrow \underset{\in \mathbb{R}^{H}}{z_1} \longrightarrow \underset{\in \mathbb{R}^{H}}{h} \longrightarrow \underset{\in \mathbb{R}^{D_y}}{z_2} \longrightarrow \underset{\in \mathbb{R}^{D_y}}{\hat{y}} \longrightarrow \underset{\in \mathbb{R}}{CE}$$

$z_1 = xW_1 + b_1 \qquad h = \text{Sigmoid}(z_1) \qquad z_2 = hW_2 + b_2 \qquad \hat{y} = \text{softmax}(z_2)$

in the previous question we found $\frac{\partial J}{\partial z_2} = \hat{y} - y$

- $\frac{\partial J}{\partial x} = \frac{\partial J}{\partial z_2} \cdot \frac{\partial z_2}{\partial x} = \boxed{(\hat{y} - y)} \frac{\partial z_2}{\partial x}$ (1)

- $\frac{\partial z_2}{\partial x} = \frac{\partial z_2}{\partial h} \cdot \frac{\partial h}{\partial x}$ | $\frac{\partial z_2}{\partial h} = \frac{\partial (hW_2 + b_2)}{\partial h} = W_2^T$

  $= \boxed{W_2^T} \cdot \frac{\partial h}{\partial x}$ (2)

- $\frac{\partial h}{\partial x} = \frac{\partial h}{\partial z_1} \cdot \frac{\partial z_1}{\partial x} = \boxed{\sigma'(z_1)} \cdot \frac{\partial z_1}{\partial x}$ (3)

- $\frac{\partial z_1}{\partial x} = \frac{\partial (xW_1 + b_1)}{\partial x} = \boxed{W_1^T}$ (4)

(1), (2), (3), (4) $\Rightarrow \underset{\in \mathbb{R}^{D_x}}{\frac{\partial J}{\partial x}} = (\hat{y} - y) \cdot W_2^T \cdot \sigma'(z_1) \cdot W_1^T$

$(1 \times D_y) \cdot (D_y \times H) \cdot (H \times H) \cdot (H \times D_x) = (1 \times D_x)$

d) $W_1 : D_x \times H \qquad b_1 : 1 \times H$

W_2: ~~$D_y \times H$~~ $H \times D_y \qquad b_2 : 1 \times D_y$

$$\boxed{\text{total number of parmameters} = (D_x \times H) + (H \times D_y) + H + D_y}$$

3) Word2Vec

a) skip-gram model $\hat{y} = P(o|c) = \dfrac{\exp(u_o^T v_c)}{\sum_w^{|W|} \exp(u_w^T \cdot v_c)}$

$J = J_{softmax-CE}(o, v_c, U) = CE(y, \hat{y})$ | Where o is the idx of the predicted word

$\dfrac{\partial J}{\partial v_c} = ?$

$CE(y, \hat{y}) = -\sum_w^W y_w \log(\hat{y}_w) = -\log(\hat{y}_o) = -u_o^T v_c + \log \sum_w^W e^{u_w^T \cdot v_c}$

$\Rightarrow \dfrac{\partial J}{\partial v_c} = \dfrac{\partial}{\partial v_c}\left(-u_o^T v_c + \log \sum_w^W \exp(u_w^T v_c)\right)$

$= -u_o + \dfrac{1}{\sum_w^W \exp(u_w^T \cdot v_c)} \otimes \dfrac{\partial}{\partial v_c}\left(\sum_w^W \exp(u_w^T v_c)\right) \quad (1)$

(element-wise multiplication)

$\dfrac{\partial}{\partial v_c}\left(\sum_w^W \exp(u_w^T v_c)\right) = \sum_w^W \dfrac{\partial}{\partial v_c}\left(e^{u_w^T v_c}\right)$

$= \sum_w^W u_w e^{u_w^T v_c} \quad (2)$

(1) and (2) $\Rightarrow$

$\dfrac{\partial J}{\partial v_c} = -u_o + \dfrac{\sum_w^W u_w e^{u_w^T \cdot v_c}}{\sum_{w'}^W e^{u_{w'}^T v_c}}$

$= -u_o + \sum_w^W u_w \cdot \dfrac{e^{u_w^T v_c}}{\sum_{w'}^W e^{u_{w'}^T v_c}}$

$\boxed{\dfrac{\partial J}{\partial v_c} = -u_o + \sum_w^W u_w \cdot \hat{y}_w} \rightarrow (F1)$

3)

b) $\frac{\partial J}{\partial u_w} = \frac{\partial}{\partial u_w}\left(-u_o^T \cdot v_c + \log \sum_{w}^{W} \exp(u_w^T \cdot v_c)\right)$

$$= \frac{\partial}{\partial u_w}(-u_o^T v_c) + \frac{\partial}{\partial u_w}\left[\log \sum_{w}^{W} \exp(u_w^T v_c)\right]$$

$$= \begin{cases} 0 & \text{if } o \neq w \\ -v_c & \text{if } o = w \end{cases} + \frac{1}{\sum_{w'}^{W} \exp(u_{w'}^T \cdot v_c)} \otimes \frac{\partial}{\partial u_w} \sum_{w}^{W} \exp(u_w^T \cdot v_c)$$

$$\cancel{= \begin{cases} 0 & \text{if } o \neq w \\ -v_c & \text{if } o = w \end{cases} + \frac{1}{\sum_{w'}^{W} \exp(u_{w'}^T v_c)} \otimes \sum_{w}^{W} v_c \exp(u_w^T v_c)}$$

$$\cancel{= \begin{cases} 0 & \text{if } o \neq w \\ -v_c & \text{if } o = w \end{cases} + W \cdot v_c \cdot \frac{\sum}{\sum} = 1}$$

$$\cancel{\boxed{\frac{\partial J}{\partial u_w} = \begin{cases} v_c & \text{if } o \neq w \\ 0 & \text{if } o = w \end{cases}}}$$

$$= \begin{cases} 0 & \text{if } o \neq w \\ -v_c & \text{if } o = w \end{cases} + \frac{1}{\sum_{w'}^{W} \exp(u_{w'}^T \cdot v_c)} \otimes v_c \left[\frac{\partial e^{u_1^T v_c}}{\partial v} ; \ldots ; \frac{\partial e^{u_W^T v_c}}{\partial v}\right] \quad \partial e^{u_w^T v_c}$$

$$= \begin{cases} 0 & \text{if } o \neq w \\ -v_c & \text{if } o = w \end{cases} + v_c \cdot \frac{e^{u_w^T v_c}}{\sum_{w'}^{W} \exp(u_{w'}^T v_c)} \leftarrow \text{softmax}_w$$

$$\boxed{\frac{\partial J}{\partial u_w} = \begin{cases} v_c \cdot \hat{y}_w & \text{if } o \neq w \\ v_c(1-\hat{y}_w) & \text{if } o = w \end{cases}} \rightarrow (F2)$$

c) $J_{neg\,sample}(o, v_c, U) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$

c-1) $\frac{\partial J}{\partial v_c}$

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

$$= -\frac{1}{\sigma(u_o^T v_c)} \otimes \left(\sigma'(u_o^T v_c) \cdot \frac{\partial}{\partial v_c} u_o^T v_c\right) - \sum_{k=1}^{K} \frac{1}{\sigma(-u_k^T v_c)} \otimes \left(\sigma'(-u_k^T v_c) \frac{\partial(\cdot)}{\partial v_c}\right.$$

$$= -\frac{1}{\sigma(u_o^T v_c)} \otimes \left(\sigma'(u_o^T v_c) \cdot u_o\right) - \sum_{k=1}^{K} \frac{1}{\sigma(-u_k^T v_c)} \otimes \left(\sigma'(-u_k^T v_c) \cdot (-1) u_k\right)$$

$$= -\frac{\cancel{\sigma(u_o^T v_c)}(1-\sigma(u_o^T v_c))}{\cancel{\sigma(u_o^T v_c)}} \otimes u_o + \sum_{k=1}^{K} \frac{\cancel{\sigma(-u_k^T v_c)}(1-\sigma(-u_k^T v_c))}{\cancel{\sigma(-u_k^T v_c)}} \otimes u_k$$

$$\boxed{\frac{\partial J}{\partial v_c} = -(1-\sigma(u_o^T v_c))\, u_o + \sum_{k=1}^{K} (1-\sigma(-u_k^T v_c))\, u_k} \rightarrow (F3)$$

c-2) $$\frac{\partial J}{\partial u_w} = \begin{cases} 0 & \text{if } o \neq w \\ -(1-\sigma(u_o^T v_c))\, v_c & \text{if } o = w \end{cases} + \begin{cases} (1-\sigma(-u_k^T v_c))\, v_c & \text{if } w = k \\ 0 & \text{if } w \neq k \end{cases}$$

$$\Rightarrow \boxed{\frac{\partial J}{\partial u_w} = \begin{cases} 0\ ; \text{if } w \neq o \text{ and } w \neq \text{all negative samples } k \\ -(1-\sigma(u_o^T v_c))\, v_c\ ; \text{if } w = o, \text{the output word} \\ (1-\sigma(-u_k^T v_c))\, v_c\ ; \text{if } w = k, \text{one of the negative samples} \end{cases}}$$

(F4)

3)

c) The cost function with the negative-sample loss is much more efficient to compute because we don't have to look at all the words in the vocabulary to compute the gradient at each step.

d) d-1) SkipGram $v \rightarrow \langle u$

$$\frac{\partial J_{SG}}{\partial v_k} = \sum_{-m, m}^{j} \frac{\partial}{\partial v_k} F(u_{c+j}, v_c) = \begin{cases} 0 \;; \text{ if } k \neq c \\ \sum_{-m,m}^{j} \frac{\partial}{\partial v_c} F(u_{c+j}, v_c) \end{cases}$$

$$\frac{\partial J_{SG}}{\partial v_k} = \begin{cases} 0 \;; \text{ if } k \neq c \\ \sum_{-m,m}^{j} \left[-u_{c+j} + \sum_{w}^{W} u_w \hat{y}_w \right] \text{ if } k = c \text{ and if } \underline{\text{softmax-loss}} \\ \sum_{-m,m}^{j} \left[(\sigma(u_{c+j}^T v_c) - 1) u_{c+j}^T + \sum_{n=1}^{N} (1 - \sigma(-u_n^T v_c)) u_n \right] \text{ if } k = c \text{ and } \underline{\text{neg-loss}} \end{cases}$$

Here we change k previously denoting negative samples to $n \in (1, -N)$ to break confusion with outer k

$$\frac{\partial J_{SG}}{\partial u_k} = \sum_{-m,m}^{j} \frac{\partial}{\partial u_k} F(u_{c+j}, v_c) = \begin{cases} 0 \text{ if } k \notin \underbrace{[c-m, c+m]}_{E} \\ \sum_{-m,m}^{j} \frac{\partial}{\partial u_k} F(u_{c+j}, v_c) \text{ if } k \in E = [c-m, c+1, c+m] \end{cases}$$

$$\frac{\partial J_{SG}}{\partial u_k} = \begin{cases} 0 \;; \text{ if } k \notin E = [c-m, c+m] \text{ and softmax} \\ \sum_{-m,m}^{j} v_c y_{c+j} \text{ if } k \in E \text{ and } k \neq c \text{ and softmax} \\ \sum_{-m,m}^{j} v_c (1 - y_{c+j}) \text{ if } k \in E \text{ and } k = c \text{ and softmax} \\ \sum_{-m,m}^{j} (\sigma(-u_n^T v_c)) v_c \text{ if } k \notin E \text{ and } k = n = \text{one of the neg-samples and neg-sample} \\ \sum_{-m,m}^{j} (\sigma(- \end{cases}$$

(this k is different to the k used previously to denote neg-samples)

$$\frac{\partial J_{SG}}{\partial u_k} = \sum_{-m,m}^{j} F(u_{c+j}, v_c)$$

~~we~~ ~~we~~

- the ~~output word~~ is neither ~~the predicted or~~ center
- ~~the only~~

- For the softmax-loss we have 2 cases
  - ~~u_k is neither the center nor context word (C1)~~
  - u_k is a Context word (C2)
  - u_k is the Center (C3)

- For the negsampling-loss we have 3 cases
  - u_k is neither neg-sample, nor Context, nor center word (C1)
  - u_k is neg-sample word (C2)
  - u_k is context word (C3)
  - ~~u_k is center word (C4)~~

So for softmax-loss:

$$\frac{\partial J_{SG}}{\partial u_k} = \begin{cases} \cancel{0 \; ; \; \text{if } k \notin [c-m; c+m] = E} & \cancel{(C1)} \\ \sum_{-m,m}^{j} v_c \cdot \hat{y}_{c+j} \; ; \; \text{if } k \in E \text{ and } k \neq c & (2) \\ \sum_{-m,m}^{j} v_c (1 - \hat{y}_{c+j}) \; ; \; \text{if } k \in E \text{ and } k = c & (3) \end{cases}$$

for negative-sampling loss:

$$\frac{\partial J_{SG}}{\partial u_k} = \begin{cases} 0 \; ; \; \text{if } k \notin [c-m; c+m] = E \text{ and } u_k \text{ is not neg-sample} & (C1) \\ \sum_{-m,m}^{j} v_c (1 - \sigma(-u_k^T v_c)) \; ; \; \text{if } k \notin E \text{ and } u_k \text{ is neg-sample} & (C2) \\ \sum_{-m,m}^{j} v_c (\sigma(u_{c+j}^T v_c) - 1) \; ; \; \text{if } k \in E & (C3) \end{cases}$$

3)

d) d-2) CBOW $v \rightarrow \hat{v} \rightarrow u$

$$\frac{\partial J}{\partial v_k} = \frac{\partial}{\partial v_k} F(u_c, \hat{v})$$

$$\boxed{\frac{\partial J}{\partial v_k} = \begin{cases} 0; & \text{if } k \notin E = [c-m, c+m] \\ -u_c + \sum_{w}^{W} u_w \hat{y}_w; & \text{if } k \in E \text{ and softmax} \\ (\sigma(u_c^T \cdot \hat{v}) - 1) u_c + \sum_{n}^{N} (1 - \sigma(-u_n^T \hat{v})) u_n & \text{if } k \in E \text{ and neg-loss} \end{cases}}$$

$$\frac{\partial J}{\partial u_k} = \frac{\partial}{\partial u_k} F(u_c, \hat{v})$$

$$\frac{\partial J}{\partial u_k} = \begin{cases} \hat{v} \cdot \hat{y}_k & \text{if } k \neq c \\ \hat{v}(1 - \hat{y}_k) & \text{if } k = c \end{cases} \quad \text{Softmax}$$

$$\begin{cases} 0; \text{ if } k \neq c \text{ and } w_k \neq \text{all negative samples } n \\ (1 - \sigma(-u_n^T \hat{v}))\hat{v}; \text{ if } k \neq c, \text{ and } k = n \text{ a negative sample} \\ (\sigma(u_k^T \hat{v}) - 1)\hat{v}; \text{ if } k = c \end{cases} \quad \text{Neg-sampling}$$

3) g) What i can see on the q3-word-vectors.png:

1) articles (a, the) are more close to each other than they are to other words

2) semantically {antonym / similar} words (wonderful, great, amazing) (boring, waste, worth) are cluster next to each other ⇒ they appear in similar context

4) b) We want to add reguralization to our training procedure to keep the weights of our model low and hence fight overfitting; which ultimately will lead to more generalization power when confronted to unseen data

d) the pretrained works better because:
- trained on more data ⇒ more generalization power
- hyperparameters were chosed more carefully for pretrained vectors
- Maybe Glove ~~may capture~~ is better for this task.

e) - After a threshold (= 10^{-2} here) decreasing the regularization factor doesn't increase the accuracy that much
- train accuracy is lower than dev accuracy once the model has "converged" which is strange
- the previous point may suggest we may be underfitting.

f) - globally the model tend to avoid very extreme reviews (++, --) hence those classes are poorly classified
- the model favored (-, +) classes a bit too much

g)

observations for -yourvectors since glove url is broken

12