Assignment 2 CS224n

mhmd.sl.elhady

June 2019

1 Written Part

1.1 1.a

Required: Show that

$$-\Sigma_w y_w log(\hat{y}_w) = -log(\hat{y}_o)$$

Solution for the cross entropy loss, y_w only is 1 when w is the predicted word and 0 other wise.

Therefore: $-\Sigma_w y_w log(\hat{y}_w)$ becomes $log(\hat{y}_w)$

1.2 1.b

Required: Get partial derivative of $J_{n.softmax}(v_c, o, u)$ in terms of $y, \hat{y}, and U$

we know this deravatives:

$$J = CE(y, \hat{y})\ \hat{y} = softmax(\theta)\ \frac{\partial J}{\partial \theta} = (\hat{y} - y)$$

where $\theta = U^T v_c$

We can use chain rules to compute the derivative:

$$\frac{\partial J}{\partial v_c} = \frac{\partial J}{\partial \theta}\frac{\partial \theta}{\partial v_c}\ = (\hat{y} - y)\frac{\partial U^T v_c}{\partial v_c}\ = U^T(\hat{y} - y)$$

1.3 1.c

Recall part 1.b

$$\frac{\partial J}{\partial v_c} = \frac{\partial J}{\partial \theta}\frac{\partial \theta}{\partial U}\ = (\hat{y} - y)\frac{\partial U^T v_c}{\partial U}\ = v_c(\hat{y} - y)^T$$

1.4 1.d

Required: derivative of softmax w.r.t x

Solution:

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

That leads to the following eqn:

$$\frac{e^{-x}}{1+e^{-x}} * \frac{1}{1+e^{-x}}$$

Add +- 1 to the left term

$$\frac{e^{-x}+1-1}{1+e^{-x}} * \frac{1}{1+e^{-x}}$$

$$\frac{e^{-x}+1-1}{1+e^{-x}} * \frac{1}{1+e^{-x}}$$

The right term is the actual $\sigma(x)$, by a bit of simple algebra to the left term it will give us

$$(1-\sigma(x))\sigma(x)$$

1.5 1.e

a

Required:

$$\frac{\partial J_{neg\_sample}}{\partial v_c}$$

Solution:

$$\frac{\partial J_{neg\_sample}}{\partial v_c} = \frac{\partial - log(\sigma(u_o^T v_c))}{\partial vc} - \Sigma_{k=1}^{K} \frac{\partial log(\sigma(-u_k^T v_c))}{\partial vc}$$

We know that $\frac{\partial \sigma(x)}{x} = \sigma(x)(1-\sigma(x))$

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

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

b

Required:

$$\frac{\partial J_{neg\_sample}}{\partial u_o}$$

Solution:

$$\frac{\partial J_{neg\_sample}}{\partial u_o} = \frac{-log(\sigma(u_o^T v_c))}{\partial u_o} - \Sigma_K^{k=1} \frac{\partial log(\sigma(-u_k^T v_c))}{\partial u_o}$$

The right term is indep. of u_o so it will be 0, therefore we are concerned only with the derivative of the left hand term.

$$(\sigma(u_o^T v_c) - 1)v_c$$

c
Required:

$$\frac{\partial J_{neg\_sample}}{\partial u_k}$$

Solution:

$$\frac{\partial J_{neg\_sample}}{\partial u_o} = \frac{-log(\sigma(u_o^T v_c))}{\partial u_k} - \Sigma_K^{k=1} \frac{\partial log(\sigma(-u_k^T v_c))}{\partial u_k}$$

The left term is indep. of u_k so it will be 0, therefore we are concerned only with the derivative of the right term. That gives us.

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

This Solution is better than the previous one as the summation is done only on the k negative samples instead of the whole vocabulary as in softmax loss

2 1.6

a

$$\frac{\partial J_skip - gram}{\partial U} = \Sigma_{-m<j<m} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

$$\frac{\partial J_skip - gram}{\partial U} = \Sigma_{-m<j<m} v_c(\hat{y} - y)^T$$

b

$$\frac{\partial J_skip - gram}{\partial v_c} = \Sigma_{-m<j<m} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

$$\frac{\partial J_skip - gram}{\partial v_c} = \Sigma_{-m<j<m} U^T(\hat{y} - y)^T$$

c

$$\frac{\partial J_skip - gram}{\partial v_w} wherew \neq c = 0$$

We do not update any other word vector other than center word