HW 2

Andrei A Simion

January 26, 2023

1 Negative Sampling for CBOW

In class we looked at the Skip-Gram and CBOW models and we looked at Negative Sampling. In the Skip-Gram model, we want to predict the the outside words from the center word. Negative Sampling removed the softmax dependency, which is expensive. The upshot is for a (w_c, w_o) pair we have

$$P(w_o|w_c) = \frac{\exp b_{w_o}^{\intercal} a_{w_c}}{\sum_{j=1}^{|V|} \exp b_{w_j}^{\intercal} a_{w_c}}$$

and replace this by

$$P(w_o|w_c) = (\frac{1}{1 + \exp -b_{w_o}^{\intercal} a_{w_c}}) E_{w_k \sim P(w)}[\prod_{k=1}^{K} \frac{1}{1 + \exp b_{w_k}^{\intercal} a_{w_c}}]$$

You can consider the expectation by: ”Draw K random samples from the set V, with probability $P(w)$”. For CBOW, we want to predict the inner word from the words around it. Thus, if $m = 1$, for example, we have

$$P(w_c|w_{c-1}, w_{c+1}) = \frac{\exp b_{w_c}^{\intercal} a_{avg}}{\sum_{j=1}^{|V|} \exp b_{w_j}^{\intercal} a_{avg}}$$

In this case, a_{avg} is the average a vector of the words w_{c-1}, w_{c+1}. The first goal is to submit what the objective for Negative Sampling would look like for CBOW. I.e., for the above example, what would it look like? Please submit a formula with justification. Your next goal is to take the notebook I give you and, using the hints and the notebook for Skip-Gram in class, implement the Negative Sampling Approach for CBOW. Can you print out the associated vectors for the validation words? Are they related, in turn, to each validation word.

2 Mathematical Problems

Below are some mathematical drills.

Problem 1 Consider the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. What is the derivate of $\sigma(x)$ in terms of $\sigma(x)$. You need to get the derivative and then simplify a bit. Do the same for hyperbolic tangent, $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Problem 2 Assume you do CBOW and Skip-Gram with negative sampling. Assume $m = 1$. Which method, on average, will get more training samples? Suppose there are 10 training samples with 7, 8, and 11 tokens. How many training sampling (positive training samples), will each method get. Draw a picture of a sentence with token counts and think about the number of samples each method gives. This is why Skip-Gram is used more often. It is more "sample efficient": you get more training data.

Problem 3 Assume you have input $a_0 = x$, and you set $z_0 = w^{[1]}a_0 + b^{[1]}$, then $a_1 = \sigma(z_0)$, then $z_1 = w^{[2]}a_1 + b^{[1]}$ and finally $a_2 = \sigma(z_1)$. Assume that the loss is $l = -log(a_2)$. What is the derivative of l with respect to each of the 4 parameters $w^{[1],[2]}$ and $b^{[1],[2]}$ (4 derivatives - express in terms of a and z)? What happens if z_0 is very large to the derivative $\frac{da_1}{dz_0}$? How would this affect learning for $w^{[1]}$ and $b^{[1]}$. Everything here is a scalar, 1 dimensional. It's like you have 1 training sample and you are doing *Stochastic* gradient descent: batch size = 1.

Problem 1 Consider the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. What is the derivate of $\sigma(x)$ in terms of $\sigma(x)$. You need to get the derivative and then simplify a bit. Do the same for hyperbolic tangent, $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

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

$$\frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

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

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

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

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

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

b. $tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$

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

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

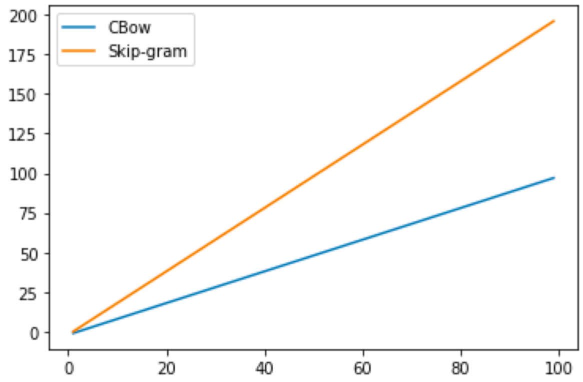
$$= 1 - \frac{(e^x-e^{-x})^2}{(e^x+e^{-x})^2}$$

$$= 1 - (tanh(x))^2$$

Problem 2 Assume you do CBOW and Skip-Gram with negative sampling. Assume $m = 1$. Which method, on average, will get more training samples? Suppose there are 10 training samples with 7, 8, and 11 tokens. How many training sampling (positive training samples), will each method get. Draw a picture of a sentence with token counts and think about the number of samples each method gives. This is why Skip-Gram is used more often. It is more ”sample efficient”: you get more training data.

a. The Skip-gram will have more training Samples,

# of Token	CBow	Skip-gram
7	5	12
8	6	14
11	9	20
n	n-2	2n-2



Problem 3 Assume you have input $a_0 = x$, and you set $z_0 = w^{[1]}a_0 + b^{[1]}$, then $a_1 = \sigma(z_0)$, then $z_1 = w^{[2]}a_1 + b^{[1]}$ and finally $a_2 = \sigma(z_1)$. Assume that the loss is $l = -log(a_2)$. What is the derivative of l with respect to each of the 4 parameters $w^{[1],[2]}$ and $b^{[1],[2]}$ (4 derivatives - express in terms of a and z)? What happens if z_0 is very large to the derivative $\frac{da_1}{dz_0}$? How would this affect learning for $w^{[1]}$ and $b^{[1]}$. Everything here is a scalar, 1 dimensional. It's like you have 1 training sample and you are doing *Stochastic* gradient descent: batch size = 1.

$$\frac{dl}{dw_1} = -\frac{1}{a_2} \cdot \frac{da_2}{dw_1}$$

$$= -\frac{1}{a_2} \cdot \frac{da_2}{dz_1} \cdot \frac{dz_1}{dw_1}$$

$$= -\frac{1}{a_2} \cdot \sigma(z_1) \cdot (1 - \sigma(z_1)) \cdot \frac{dz_1}{da_1} \cdot \frac{da_1}{dw_1}$$

$$= -\frac{1}{a_2} \cdot \sigma(z_1) \cdot (1 - \sigma(z_1)) \cdot w_2 \cdot \frac{da_1}{dz_0} \cdot \frac{dz_0}{dw_1}$$

$$= -\frac{1}{a_2} \cdot \sigma(z_1) \cdot (1 - \sigma(z_1)) \cdot w_2 \cdot \sigma(z_0) \cdot (1 - \sigma(z_1)) \cdot a_0$$

$$\frac{dl}{dw_2} = -\frac{1}{a_2} \cdot \frac{da_2}{dw_2}$$

$$= -\frac{1}{a_2} \cdot \frac{da_2}{dz_1} \cdot \frac{dz_1}{dw_2}$$

$$= -\frac{1}{a_2} \cdot \sigma(z_1) \cdot (1 - \sigma(z_1)) \cdot a_1$$

$$\frac{dl}{db_1} = -\frac{1}{a_2} \cdot \frac{da_2}{dw_2}$$

$$= -\frac{1}{a_2} \cdot \frac{da_2}{dz_1} \cdot \frac{dz_1}{da_1} \cdot \frac{da_1}{dz_0} \cdot \frac{dz_0}{db_1}$$

$$= -\frac{1}{a_2} \cdot \sigma(z_1) \cdot (1 - \sigma(z_1)) \cdot w_2 \cdot \sigma(z_0) \cdot (1 - \sigma(z_1))$$

$$\frac{dL}{db_2} = -\frac{1}{a_2} \cdot \frac{da_2}{db_2}$$

$$= -\frac{1}{a_2} \cdot \frac{da_2}{dz_1} \cdot \frac{dz_1}{db_2}$$

$$= -\frac{1}{a_2} \cdot \sigma(z_1) \cdot (1 - \sigma(z_1))$$

When z_0 gets really large, $\frac{da_0}{dz_0}$ will get really close to 0

w_1, b_1 will stop learning because of low gradient.