Iowa State University

Department of Electrical and Computer Engineering

Deep Machine Learning: Theory and Practice

EE 526X

Homework 2

Author:
Vishal Deep

Instructor:
Dr. Zhengdao Wang

October 7, 2019

1 Problem 1

We have a softplus function,

$$f(z) = \log(1 + e^z) \tag{1}$$

Taking first derivative of the function:

$$f'(z) = \frac{1}{(1 + e^z)} e^z \tag{2}$$

multiplying and dividing by e^{-z},

$$f'(z) = \frac{e^z . e^{-z}}{(1 . e^{-z} + e^z . e^{-z})} = \frac{1}{(1 + e^{-z})} \tag{3}$$

Now taking second derivative of softplus function,

$$f''(z) = \frac{(1 + e^{-z}).0 + e^{-z}}{(1 + e^{-z})^2} \tag{4}$$

$$f''(z) = \frac{e^{-z}}{(1 + e^{-z})^2} \tag{5}$$

From eq. (5), numerator is always positive because exponential is a positive number and any power to a positive is number is always a positive number. The denominator is always positive because is squared. Therefore we can say that second derivative of a softplus function is always positive for every value of z. This implies that a softplus function is convex in z.

2 Problem 2

We have a vector $z = [z1, z2, ..., zn]^T$ and $p = [p1, p2, ..., pn]^T$, where p is output after z is applied to a softmax function.

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{6}$$

The jacobian matrix will have two types of elements diagonal and off-diagonal, we will calculate both and then generalize the result for all elements.

Diagonal row elements (i=j):

$$\frac{\partial p_i}{\partial z_i} = \frac{\sum_{j=1}^{n} e^{z_j} . e^{z_i} - e^{z_i} . e^{z_i}}{(\sum_{j=1}^{n} e^{z_j})^2}$$

$$\frac{\partial p_i}{\partial z_i} = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} - \left(\frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \right)^2$$

$$\frac{\partial p_i}{\partial z_i} = p_i - p_i^2 \tag{7}$$

1

off-diagonal row elements (i≠j):

$$\frac{\partial p_i}{\partial z_j} = \frac{\sum_{j=1}^n e^{z_j}.0 - e^{z_i}.e^{z_j}}{(\sum_{j=1}^n e^{z_j})^2}$$

$$\frac{\partial p_i}{\partial z_j} = -\frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}} \times \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}$$

$$\frac{\partial p_i}{\partial z_j} = -p_i p_j \tag{8}$$

The Jacobian matrix is given by

$$\frac{\partial p}{\partial z} = \begin{bmatrix} \frac{\partial p_0}{\partial z_0} & \frac{\partial p_0}{\partial z_1} & \cdots & \frac{\partial p_0}{\partial z_n} \\ \frac{\partial p_1}{\partial z_0} & \frac{\partial p_1}{\partial z_1} & \cdots & \frac{\partial p_1}{\partial z_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial p_n}{\partial z_0} & \frac{\partial p_n}{\partial z_1} & \cdots & \frac{\partial p_n}{\partial z_n} \end{bmatrix} \tag{9}$$

using eq. (7) and eq. (8), we get

$$\frac{\partial p}{\partial z} = \begin{bmatrix} p_0 - p_0^2 & -p_0 p_1 & \cdots & -p_0 p_n \\ -p_1 p_0 & p_1 - p_1^2 & \cdots & -p_1 p_n \\ \vdots & \vdots & \ddots & \vdots \\ -p_n p_0 & -p_n p_1 & \cdots & p_n - p_n^2 \end{bmatrix} \tag{10}$$

3 Problem 3

We have $y = [y1, y2, ..., yn]^T$, a correct probability vector. And the cross entropy is given by

$$J(z) = -\sum_{i=1}^n y_i \log p_i \tag{11}$$

eq. (11) is a dot product or element-wise product of y_i and $\log p_i$. Taking derivative of eq. (11) w.r.t. p_i,

$$\frac{\partial J}{\partial p_i} = -\frac{\partial}{\partial p_i} y_i \log p_i = -y_i \frac{\partial \log p_i}{\partial p_i}$$

$$\frac{\partial J}{\partial p_i} = -\frac{y_i}{p_i} \tag{12}$$

Also from eqs. (7) and (8) we know $\frac{\partial p_i}{\partial z_j}$,

$$\frac{\partial p_i}{\partial z_j} = \begin{cases} p_i - p_i^2 & i = j \\ -p_i p_j & i \neq j \end{cases} \tag{13}$$

from eqs. (12) and (13), we can find $\frac{\partial J}{\partial z_i}$

$$\frac{\partial J}{\partial z_i} = \sum_{j=1}^{n} \frac{\partial J}{\partial p_j} \frac{\partial p_j}{\partial z_i}$$

$$= \frac{\partial J}{\partial p_i} \frac{\partial p_i}{\partial z_i} + \sum_{i \neq j} \frac{\partial J}{\partial p_j} \frac{\partial p_j}{\partial z_i}$$

$$= -\frac{y_i}{p_i}(p_i - p_i^2) + \sum_{i \neq j}(-\frac{y_j}{p_j})(-p_i p_j)$$

$$= -y_i(1 - p_i) + \sum_{i \neq j} y_j p_i$$

$$= -y_i + y_i p_i + p_i \sum_{j} y_j$$

$$= p_i \left(y_i + \sum_{j} y_j \right) - y_i$$

$$= p_i - y_i$$

Because y is one hot encoded and $\sum_j y_j = 1$

$$\frac{\partial J}{\partial z} = \begin{bmatrix} p_0 - y_0 \\ p_1 - y_1 \\ \vdots \\ p_n - y_n \end{bmatrix} \tag{14}$$

4 Problem 4

We have

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{15}$$

$$\mathbf{y} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{16}$$

$$\mathbf{z} = \mathbf{Wx} + \mathbf{b} \tag{17}$$

1st Iteration:

$$z^{[0]} = W^{[0]}x + b^{[0]}.1$$

$$= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix}$$

$$z^{[0]} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{18}$$

We are using a softmax function, therefore p matrix is given by

$$p^{[0]} = \frac{e^{z_i}}{\sum_{i=1}^{4} e^{z_i}}$$

$$p^{[0]} = \begin{bmatrix} \frac{e^0}{e^0+e^0+e^0+e^0} & \frac{e^0}{e^0+e^0+e^0+e^0} \\ \frac{e^0}{e^0+e^0+e^0+e^0} & \frac{e^0}{e^0+e^0+e^0+e^0} \end{bmatrix} = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix} \tag{19}$$

and cost J is given by cross entropy function,

$$J^{[0]} = -\sum_{i=1}^{4} y_i \log p_i = -\boldsymbol{y}(\log \boldsymbol{p})^T$$

$$J^{[0]} = -\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \log 0.25 & \log 0.25 \\ \log 0.25 & \log 0.25 \end{bmatrix} = -\begin{bmatrix} \log 0.25 & \log 0.25 \\ \log 0.25 & \log 0.25 \end{bmatrix} \tag{20}$$

Now doing back propagation, from eq. (14)

$$\boldsymbol{dz}^{[0]} = \boldsymbol{p} - \boldsymbol{y} = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -0.75 & 0.25 \\ 0.25 & -0.75 \end{bmatrix} \tag{21}$$

$$\boldsymbol{dW}^{[0]} = \boldsymbol{dz}^{[0]} \boldsymbol{X}^T = \begin{bmatrix} -0.75 & 0.25 \\ 0.25 & -0.75 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -0.75 & 0.25 \\ 0.25 & -0.75 \end{bmatrix} \tag{22}$$

$$\boldsymbol{db}^{[0]} = \boldsymbol{dz}^{[0]} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.75 & 0.25 \\ 0.25 & -0.75 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix} \tag{23}$$

Now update weights and bias matrices with learning rate = 1,

$$\boldsymbol{W}^{[1]} = \boldsymbol{W}^{[0]} - 1.\boldsymbol{dW}^{[0]} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} -0.75 & 0.25 \\ 0.25 & -0.75 \end{bmatrix} = \begin{bmatrix} 0.75 & -0.25 \\ -0.25 & 0.75 \end{bmatrix} \tag{24}$$

$$\boldsymbol{b}^{[1]} = \boldsymbol{b}^{[0]} - 1.\boldsymbol{db}^{[0]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \tag{25}$$

2nd Iteration:

$$z^{[1]} = W^{[1]}x + b^{[1]}.1$$

$$= \begin{bmatrix} 0.75 & -0.25 \\ -0.25 & 0.75 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix}$$

$$z^{[1]} = \begin{bmatrix} 0.25 & -0.75 \\ -0.75 & 0.25 \end{bmatrix} \tag{26}$$

Softmax function with $z^{[1]}$,

$$p^{[1]} = \frac{e^{z_i}}{\sum_{i=1}^{4} e^{z_i}} = \begin{bmatrix} 0.365 & 0.134 \\ 0.134 & 0.365 \end{bmatrix} \tag{27}$$

$$J^{[1]} = -\sum_{i=1}^{4} y_i \log p_i = -\boldsymbol{y}(\log \boldsymbol{p})^T$$

$$J^{[1]} = -\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \log 0.365 & \log 0.134 \\ \log 0.134 & \log 0.365 \end{bmatrix} = -\begin{bmatrix} \log 0.365 & \log 0.134 \\ \log 0.134 & \log 0.365 \end{bmatrix} \tag{28}$$

Now doing back propagation, from eq. (14)

$$\boldsymbol{dz}^{[1]} = \boldsymbol{p} - \boldsymbol{y} = \begin{bmatrix} 0.365 & 0.134 \\ 0.134 & 0.365 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -0.634 & 0.134 \\ 0.134 & -0.634 \end{bmatrix} \tag{29}$$

$$\boldsymbol{dW}^{[1]} = \boldsymbol{dz}^{[1]} \boldsymbol{X}^T = \begin{bmatrix} -0.634 & 0.134 \\ 0.134 & -0.634 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -0.634 & 0.134 \\ 0.134 & -0.634 \end{bmatrix} \tag{30}$$

$$\boldsymbol{db}^{[1]} = \boldsymbol{dz}^{[1]} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.634 & 0.134 \\ 0.134 & -0.634 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix} \tag{31}$$

Now update weights and bias matrices with learning rate = 1,

$$\boldsymbol{W}^{[2]} = \boldsymbol{W}^{[1]} - 1.\boldsymbol{dW}^{[1]} = \begin{bmatrix} 0.75 & -0.25 \\ -0.25 & 0.75 \end{bmatrix} - \begin{bmatrix} -0.634 & 0.134 \\ 0.134 & -0.634 \end{bmatrix} = \begin{bmatrix} 1.384 & -0.384 \\ -0.384 & 1.384 \end{bmatrix} \tag{32}$$

$$\boldsymbol{b}^{[2]} = \boldsymbol{b}^{[1]} - 1.\boldsymbol{db}^{[1]} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} - \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{33}$$