Assignment 1, Theoretical Part

Charles Huard

February 17, 2019

Question 1

1. The Heaviside function is a piecewise function, with three different value. We therefore need to show that for a particular value of the heaviside function, the derivative of the Relu over the same domain is equal to the Heavyside fonction. Two domains are of interest, $x < 0$ and $x > 0$. Since Relu isn't differientiable at X = 0, we don't need to prove the equality at this value.

$$Case : x < 0$$

$$Relu = 0(\frac{\partial 0}{\partial x} = 0)$$

$$Heavyside = 0$$

$$Heavyside = \frac{\partial Relu}{\partial x}$$

$$Case : x > 0$$

$$Relu = x, \frac{\partial X}{\partial x} = 1$$

$$Heavyside = 1$$

$$Heavyside = \frac{\partial Relu}{\partial x}$$

■

2. $g(x) = \int H(x)$, by definition since we established that $\frac{\partial g(x)}{\partial x} = H(x)$

   $g(x) = xH(x)$, makes the positive part of H(x) linear, which perfectly mirroirs g(x)

3. H(x) is a 3 piece function, so if we show each piece can be approximate by the sigmoid with a large k, we would show that H(x) can be approximated by a sigmoid with a large k.

   Let N be a large interger.

$$Case : x < 0$$

$$e^{-kx} + 1 = N$$

$$\frac{1}{e^{-kx} + 1} = \frac{1}{N} \approx 0 = H(x)$$

$$Case : x = 0$$

$$\frac{1}{e^{0} + 1} = \frac{1}{2} = H(x)$$

$$Case : x > 0$$

$$-kx = -N$$

$$e^{-N} \approx 0$$

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

■

4. By the definition that is provided, $F[\phi] = \int_R F(x)\phi(x)dx$
   Using integration by parts, we express the derivative to be

$$F'[\phi] = F(x)\phi(x)\Big|_{-\infty}^{\infty} - \int_R F(x)\phi'(x)dx$$

   By the definition provided, $\phi(x) = 0$ at ∞ and $-\infty$. We can simply the expression to be

$$F'[\phi] = -\int_R F(x)\phi'(x)dx$$

   Which is the desired result.

   We then use this definition to express $H'(x) = -\int_{-\infty}^{\infty} H(x)\phi'(x)dx$
   By definition H(x) = 0 over $x < 0$. Using this we can reduce the integral to be

$$H'(x) = -\int_0^{\infty} H(x)\phi'(x)dx$$

   By definition H(x) = 1 over $x > 0$. Using this we can reduce the integral to be

$$H'(x) = -\int_0^{\infty} \phi'(x)dx = -\Big|_0^{\infty} \phi(x) = -(\phi(\infty) - \phi(0))$$

   By definition $\phi(\infty) = 0$

$$H'(x) = -(0 - \phi(0)) = \phi(0)$$

■

Question 2

1. By definition the softmax is

$$\frac{\partial S(x)_i}{\partial x} = \frac{\partial}{\partial x}\frac{exp(x_i)}{\sum_k exp(x_k)}$$

Applying Quotient Rule

$$= \frac{\frac{\partial exp(x)_i}{\partial x_j}(\sum_k exp(x_k)) - \frac{\partial \sum_k exp(x_k)}{\partial x_j} exp(x_i)}{(\sum_k exp(x_k))^2}$$

Refactors to

$$= \frac{\frac{\partial exp(x_i)}{\partial x_j}}{\sum_k exp(x_k)} - \frac{exp(x_i)}{(\sum_k exp(x_k))^2}\frac{\partial}{\partial x_j}(\sum_k exp(x_k))$$

Since $\frac{\partial exp(x_i)}{\partial x_j}$ is 1 if i = j and else, we can express that derivative as being $\delta_{ij} exp(x_i)$. Simlilarly, $\frac{\partial}{\partial x_j}(\sum_k exp(x_k)) = exp(x_j)$
Then

$$= \frac{\delta_{ij} exp(x_i)}{\sum_k exp(x_k)} - \frac{exp(x_i)}{\sum_k exp(x_k)}\frac{exp(x_j)}{\sum_k exp(x_k)}$$

$$= \frac{exp(x_i)}{\sum_k exp(x_k)}(\delta_{ij} - \frac{exp(x_j)}{\sum_k exp(x_k)})$$

By definition of the softmax

$$= S(x_i)(\delta_{ij} - S(x_j))$$

■

2. Knowing $\frac{\partial S(x_i)}{\partial x_j} = S(x_i)(\delta_{ij} - S(x_j))$, distributing we get

$$S(x_i)\delta_{ij} - S(x_i)S(x_j)$$

the left part can be expressed as $diag(S(x_i))$ and the right part can now be expressed as Softmax of a single indice.
We then express the Jacobian matrix as

$$J(S(x)) = diag(S(x)) - S(x)S(x)^T$$

3. First case: i != j

$$\frac{\partial \sigma(x_i)}{\partial x_j} = 0$$

Then if i = j, we need to solve

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

Using the quotient rule

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

Then using simple algebra

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

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

By definition of the sigmoid

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

With both cases we express

$$J(\sigma(x)) = diag(\sigma(x)(1-\sigma(x)))$$

4. We need to show O(n) for the Softmax and the Sigmoid

For the Sigmoid, since $\frac{\partial}{\partial x}\sigma(x)$ is a diagonal matrix, this become a vector multiplication between the diagonal of both matrices, since all other results yields 0. Knowing that the diagonal of an nxn matrix as n elements, we only need to do n multiplication, therefore the multiplication is O(n)

For the Softmax : We showed previously that we can represent the Jacobian matrix of the softmax as follow :

$$= \begin{bmatrix} \sigma(1)(\delta_{ij} - \sigma(1)) & \dots & \sigma(1)(\delta_{ij} - \sigma(k)) \\ \vdots & \ddots & \vdots \\ \sigma(k)(\delta_{ij} - \sigma(1)) & \dots & \sigma(k)(\delta_{ij} - \sigma(k)) \end{bmatrix}$$

Which can be simplified as follow, knowning kroenecker delta definition :

$$= \begin{bmatrix} \sigma(1) - \sigma(1)\sigma(1) & \dots & -\sigma(1)\sigma(k) \\ \vdots & \ddots & \vdots \\ -\sigma(1)\sigma(k) & \dots & \sigma(k) - \sigma(k)\sigma(k) \end{bmatrix}$$

We see that the Jacobian of the softmax is a symmetric matrix. We know from linear algebra that a symmetric matrix is always diagonalizable. By preprocessing the jacobian first to get a diagonal matrix, the matrix vector operation becomes a diagonal matrix vector multiplication. This operation can be simplified as an element product between the diagonal of the matrix and the vector. This elementwise product is O(n), which makes the multiplication O(n).

Question 3

1.

$$S(x+c) = \frac{e^{x+c}}{\sum_k e^{x+c}}$$
$$= \frac{e^x e^c}{\sum_k e^x e^c}$$
$$= \frac{e^x e^c}{e^c \sum_k e^x}$$
$$= \frac{e^x}{\sum_k e^x}$$
$$\blacksquare$$

2. Proof by contradiction, let S(x) be invariant of scalar multiplication. Then S(x) = S(xc) should hold true.$Let x_1 = 2, x_2 = 4 and c = 2$. Then

$$S(x) = [\frac{e^2}{e^2+e^4}, \frac{e^4}{e^2+e^4}] = [0.1192, 0.88] = S(xc) = [\frac{e^4}{e^4+e^8}, \frac{e^8}{e^4+e^8}] = [0.0179, 0.98]$$

Since the equation is false, we conclude that S(x) is not invariant under scalar multiplication.

We also observe that if $c > 0$, the multiplication by a scalar c raise the value of the most probable class and lowers the other.
If c = 0,$S(x) = \frac{e^0}{\sum_k e^0} = \frac{1}{\sum_j 1}$, which means all class are equally probable and we get an uniform distribution.

3. We must show $\sigma(z) = S(x_2)$ and $1 - \sigma(z) = S(x_1)$ for z being a scalar function of x.

Let $z = x_2 - x_1$

$$\sigma(z) = \frac{1}{1+exp(-z)} = \frac{1}{1+exp(x_1)exp(-x_2)}$$
$$= \frac{exp(x_2)}{exp(x_2)(1+exp(x_1)exp(-x_2)} = \frac{exp(x_2)}{exp(x_2)+exp(x_1)}$$

$$= S(x_2)$$

$$1 - \sigma(z) = 1 - \frac{1}{1+exp(-z)} = 1 - \frac{1}{1+exp(x_1)exp(-x_2)}$$

$$= 1 - \frac{exp(x_2)}{exp(x_2)(1 + exp(x1)exp(-x_2))} = 1 - \frac{exp(x_2)}{exp(x_2) + exp(x_1)}$$

$$= \frac{exp(x_1) + exp(x_2)}{exp(x_1) + exp(x_2)} - \frac{exp(x_2)}{exp(x_1)exp(x_2)} = \frac{exp(x_1)}{exp(x_1)exp(x_2)}$$

$$= S(x_1)$$

■

4. Let $y_i = x_i - x_1$. We then propose a S(x) with K - 1 parameters to be of this form

$$f(y2, y3, ..., y_k)_i = \begin{cases} \frac{exp(y_i)}{(\sum_{j=2}^{k} exp(y_k))+1} & i \neq 1 \\ 1 - \sum_{j=2}^{k} f(y_j) & i = 1 \end{cases}$$

We then show that $f(y2, y3, ..., y_k)_i = S(x1, x2, ..., x_k)_i$ for all i

For $i \neq j$ We have

$$f(y_2, y3,..., y_k)_i = \frac{exp(x_i - x_1)}{(\sum_{j=2}^{k} exp(x_j - x_1)) + 1}$$

$$= \frac{exp(x_i)}{(\sum_{j=2}^{k} exp(x_j)) + exp(x_1)}$$

$$= S(x_1, ..., x_k)_i$$

For $i = j$

$$S(x_1, ..., x_k)_i = 1 - \sum_{j=2}^{k} S(x1, ..., x_k)_j = 1 - \sum_{j=2}^{k} f(y_2, ..., y_k)_j$$

■

Question 4

1. If we show a linear relationship between the sigmoid and tanh, we can transform the sigmoid form into the tanh form, given the correct parameters.

   We show the sigmoid/tanh relation to be $tanh(x) = 2\sigma(2x) - 1$

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

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

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

$$= tanh(x)$$

   Then we transform the sigmoid activated equation. Let $\Theta' = (2\omega^{(1)}, 2\omega^{(2)}), \omega_{j0}^{(1)} = 0, \omega_{j0}^{(2)} = \omega_{k0}^{(2)} + 1$ Reparametrizing with Θ' and biases

$$y(x,\Theta,\sigma)_k = \sum_{j=1}^{M} 2\omega_{kj}^{(2)}\sigma(\sum_{i=1}^{D} 2\omega_{ji}^{(2)}x_i) + \omega_{j0}^{(2)} - 1$$

$$= 2\sum_{j=1}^{M} \omega_{kj}^{(2)}\sigma(2\sum_{i=1}^{D} \omega_{ji}^{(2)}x_i) + \omega_{j0}^{(2)} - 1$$

   Using the above tanh/sigmoid relation

$$= \sum_{j=1}^{M} \omega_{kj}^{(2)} tanh(\sum_{i=1}^{D} \omega_{ji}^{(2)}x_i) + \omega_{j0}^{(2)}$$

$$= y(x,\Theta',tanh)_k$$

$$\blacksquare$$

   Therefore, the relation $\Theta' = 2\Theta$, with $\sigma_0^{'1} = 0$ and $\sigma_0^{'2} = \sigma_0^2 + 1$ gives an equivalent 2 layers NN with sigmoid and tanh activations.

Question 5

1.

$$y(x; W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}, \phi) = \sum W_j^{(2)}\phi(\sum W_i^{(1)}x_i + b^{(1)}) + b^{(2)}$$

2.

$$= \sum W_j^{(2)} \phi(W_i^{(1)} \sum x_i + b^{(1)}) + b^{(2)}$$

$$= \begin{bmatrix} \phi(W_i^{(1)} \sum x_i + b^{(1)}) & \dots & \phi(W_i^{(1)} \sum x_i + b^{(1)}) & 1 \\ \vdots & & & \\ \phi(W_N^{(1)} \sum x_i + b^{(1)}) & \dots & \phi(W_N^{(1)} \sum x_i + b^{(1)}) & 1 \end{bmatrix} \begin{bmatrix} W_1^{(2)} \\ W_2^{(2)} \\ \vdots \\ W_{N-1}^{(2)} \\ b^{(2)} \end{bmatrix} = \begin{bmatrix} f(x^{(1)}), & f(x^{(2)}), & \dots & , f(x^{(N)}) \end{bmatrix}$$

3. Let's re-order the above matrix so that the values on each value of each row are stricly greater than the values of the rows under them. With the new ordering, let's then take $b_j^{(1)} > \epsilon > 0$. Since all values above the diagonal are bigger then $b_j^{(1)}$, they all get positive. The diagonal is also positive, since $\epsilon > 0$. The values under the diagonal get negative, since $wx + \epsilon < b_j^{(1)}$. Since ϕ is Relu, all negative values = 0. So we get an upper triangular matrix, with a diagonal > 0.

   Knowing this we can now try to evaluate $N = \infty$. With the triangular matrix, we can now see that the difference betweenx^i and x^{i+1} is now any definite value. Since x^i is a sum of linear functions, which can give any value, and the difference between x^i and x^{i+1} is also linear and any definite value, we can conclude that the resulting function can be any continous function f.

4. We order the matrix the same way we did in the previous question. For the upper triangle matrix, we get $wx > b$ and the difference $= +\infty$. For the diagonal the difference is 0, which $\phi(0)$ is positive. For the lower matrix, we get $wx < b$, which $\phi(-\infty) = 0$. We then get an upper triangular matrix, with a positive diagonal.

   As we concluded previously, this upper triangular matrix can be use to approximate any continous fonction f.

Question 6

Kernel flipped = [2,0,1]

Full Convolution: [1,2,3,4] * [2,0,1] = [(2x0 + 0 +1x1), (2x0 + 0 +1x2), (2x1 + 0 +1x3),(2x2 + 0 +1x4),(2x3 + 0 +1x0), (2x4 + 0 +1x0)] = [1,2,5,8,6,8]

Same Convolution: [1,2,3,4] * [2,0,1] = [(2x0 + 0 +1x2), (2x1 + 0 +1x3),(2x2 + 0 +1x4),(2x3 + 0 +1x0)] = [2,5,8,6]

Valid Convolution: [1,2,3,4] * [2,0,1] = [(2x1 + 0 +1x3),(2x2 + 0 +1x4)] = [5,8]

Question 7

Using formula $o = (\frac{i+2p-k}{s}) + 1$ from class notes, we calculate dimension at each layer.

First layer: i = 256, k = 8, s=2, p=0 and we have 3 channels. $o = (\frac{256+2(0)-8}{2}) + 1 = 125$ Since there is 64 kernels, output dimension is 125x125x64.

Second layer: Maxpooling preserves number of channels, but divide square area. 125/5 = 25, output dimension is 25x25x64
Last layer: : i = 25, k = 4, s=1, p=1 and we have 64 channels. $o = (\frac{25+2(1)-4}{1}) + 1 = 24$ Since there is 128 kernels, output dimension is 24x 24x128.

1. Output of the last layer is 24x24x128 = 73 728 dimensions.

2. Parameters of the function is given by O x O x number of inputs channel. We have 24 x 24 x 64 = 36864 parameters.

Question 8

1. (a) k = 8, then $32 = (\frac{64+2(p)-8}{s}) + 1$. p = 3 , s = 2 satisfy the equation.

   Correct configuration is: k=8, s=2, p=3,d=1.

   (b) d = 7, s = 2. Then $32 = (\frac{64+2(p)-\hat{k}}{s}) + 1$. Let p = 3, then $\hat{k}$ need to be 8. $\hat{k} = k + (k-1)(7-1)$. k = 2 statisfies the equation.

   Correct configuration is: k=2, s=2, p=3,d=7.

2. (a) k = 4 and s = 2
   (b) 4 x 4 x 64

3. (a) p = 0, d=1 . $4 = (\frac{8+2(0)-k}{s}) + 1$. s = 2 , k = 2 satisfy the equation.

   Correct configuration is: k=2, s=2, p=0,d=1.

   (b) d = 2, p = 2. Then $4 = (\frac{8+2(2)-\hat{k}}{s}) + 1$. Let s = 1, then $\hat{k}$ need to be 9. $\hat{k} = k + (k-1)(2-1)$. k = 5 statisfies the equation.

   Correct configuration is: k=5, s=1, p=2,d=2.

   (c) d = 1, p = 1. Then $4 = (\frac{8+2(1)-k}{s}) + 1$ .s = 3 , k = 3 satisfy the equation statisfy the equation.

   Correct configuration is: k=3, s=3, p=1,d=1.