1.EXERCISES IN BACKPROPAGATION

1. Assume we have a n-dimensional input $x = [x_1, x_2, \ldots, x_n]^T$. And we have m units in the hidden layers and the activation function of each unit is sigmoid. And we have q-dimensional output $\hat{y} = [y_1, y_2, \ldots, y_q]^T$.
   The neural network looks like this:

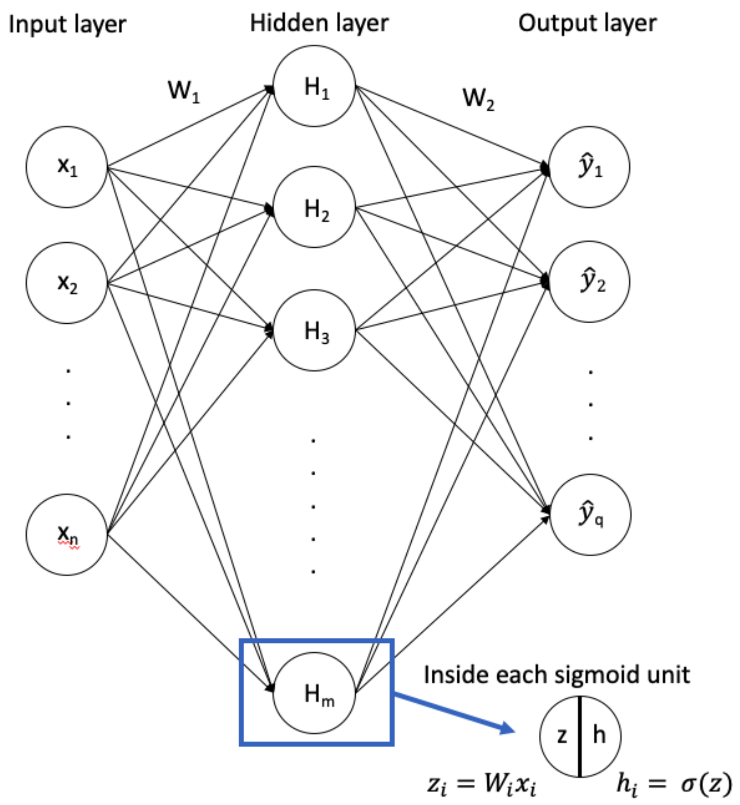

Figure 1. Network Layout

In each unit of hidden layer, there are two values: one is z, the input value of this unit; another one is h, the output value of this unit.
Here is computation graph:

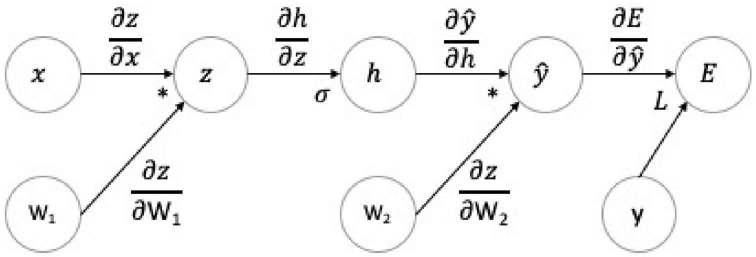
$\frac{\partial z}{\partial x}$
$\frac{\partial h}{\partial z}$
$\frac{\partial \hat{y}}{\partial h}$
$\frac{\partial E}{\partial \hat{y}}$
x
z
h
$\hat{y}$
E
*
σ
*
L
W_1
$\frac{\partial z}{\partial W_1}$
W_2
$\frac{\partial z}{\partial W_2}$
y

Figure 2. Computation graph

In the forward passes, we can calculate the forward values.

$$z = W_1 x$$

$$h = \sigma(z) = \sigma(W_1 x), \sigma(x) = \frac{1}{1+e^{-x}}$$

$$\hat{y} = W_2 h = W_2 \sigma(w_1 x)$$

$$E = ||\hat{y} - y||_2^2$$

In the backward passes, we need to calculate these partial derivatives first: $\frac{\partial z}{\partial x}$, $\frac{\partial z}{\partial W_1}$, $\frac{\partial h}{\partial z}$, $\frac{\partial \hat{y}}{\partial h}$, $\frac{\partial \hat{y}}{\partial W_2}$, $\frac{\partial E}{\partial \hat{y}}$

Here is how we calculate $\frac{\partial E}{\partial \hat{y}}$:

$$E = \sqrt{(\hat{y_1} - y_1)^2 + (\hat{y_2} - y_2)^2 + \cdots + (\hat{y_q} - y_q)^2}^2 = (\hat{y_1} - y_1)^2 + (\hat{y_2} - y_2)^2 + \cdots + (\hat{y_q} - y_q)^2$$

$$\frac{\partial E}{\partial \hat{y}} = [\frac{\partial E}{\partial \hat{y_1}}, \frac{\partial E}{\partial \hat{y_2}}, \ldots, \frac{\partial E}{\partial \hat{y_q}}]$$

For each partial derivatives,

$$\frac{\partial E}{\partial \hat{y_i}} = 2(\hat{y_i} - y_i)$$

Thus,

$$\frac{\partial E}{\partial \hat{y}} = [2(\hat{y_1} - y_1), 2(\hat{y_2} - y_2), \ldots, 2(\hat{y_q} - y_q)] = 2(\hat{y} - y)$$

Here is how we calculate $\frac{\partial h}{\partial z}$:

$$h = \sigma(z)$$

$$h = \begin{bmatrix} h_1 \\ h_2 \\ \cdots \\ h_i \\ \cdots \\ h_m \end{bmatrix} = \sigma \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_i \\ \cdots \\ z_m \end{bmatrix}$$

We can get the vector Jacobian

$$J = \frac{\partial h}{\partial z} = \begin{bmatrix} \frac{\partial h_1}{\partial z_1} & \frac{\partial h_1}{\partial z_2} & \cdots & \frac{\partial h_1}{\partial z_m} \\ \frac{\partial h_2}{\partial z_1} & \frac{\partial h_2}{\partial z_2} & \cdots & \frac{\partial h_2}{\partial z_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial z_1} & \frac{\partial h_n}{\partial z_2} & \cdots & \frac{\partial h_n}{\partial z_m} \end{bmatrix}$$

Because $h_i = \frac{1}{1+e^{-z_i}}$
Thus, When $i \neq j$, $\frac{\partial h_i}{\partial z_j} = 0$
Thus, J is a diagonal matrix. When $i = j$, $J_{ij} = \frac{\partial h_i}{\partial z_j} = \frac{1}{1+e^{-z_j}}(1 - \frac{1}{1+e^{-z_j}})$

$$J = \frac{\partial h}{\partial z} = \begin{bmatrix} \frac{1}{1+e^{-z_1}}(1 - \frac{1}{1+e^{-z_1}}) & & \\ & \ddots & \\ & & \frac{1}{1+e^{-z_m}}(1 - \frac{1}{1+e^{-z_m}}) \end{bmatrix}$$

Here is how we calculate $\frac{\partial Z}{\partial W_1}$:

$$Z = W^1 x$$

, here we use W^1 to represent W_1

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_i \\ \dots \\ z_m \end{bmatrix} = \begin{bmatrix} W^1_{11} & W^1_{12} & \cdots & W^1_{1n} \\ W^1_{21} & W^1_{22} & \cdots & W^1_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W^1_{m1} & W^1_{12} & \cdots & W^1_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

We can get the vector Jacobian

$$J = \frac{\partial z}{\partial W_1} = \begin{bmatrix} \frac{\partial z_1}{\partial W^1_{1k}} & \frac{\partial z_1}{\partial W^1_{2k}} & \cdots & \frac{\partial z_1}{\partial \partial W^1_{mk}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_n}{\partial W^1_{1k}} & \frac{\partial z_n}{\partial W^1_{2k}} & \cdots & \frac{\partial z_n}{\partial W^1_{mk}} \end{bmatrix}$$

And what we need to address is that the size of J is $m * (m * n)$, and

$$\frac{\partial z_i}{\partial W^1_j} = [\frac{\partial z_i}{\partial W^1_{j1}}, \frac{\partial z_i}{\partial W^1_{j2}}, \dots, \frac{\partial z_i}{\partial W^1_{jn}}]$$

Because $z_i = W^1_{i1}x_1 + W^1_{i2}x_2 + \cdots + W^1_{in}x_n$
Thus, When $i \neq j$, $\frac{\partial z_i}{\partial W^1_{jk}} = 0$
When $i = j$, $\frac{\partial z_i}{\partial W^1_{jk}} = x_k, k = 1, 2, \dots, n$

$$\frac{\partial z_i}{\partial W^1_j} = [x_1, x_2, ..x_n] = x^T$$

Thus, J is a diagonal matrix. When $i = j$, $J_{ij} = x^T$

$$J = \frac{\partial z}{\partial W_1} = \begin{bmatrix} [x_1, \dots, x_n] & & \\ & \ddots & \\ & & [x_1, \dots, x_n] \end{bmatrix}$$

We can get $\frac{\partial \hat{y}}{\partial W_2}$ in the same way.

$$\frac{\partial \hat{y}}{\partial W_2} = \begin{bmatrix} [h_1, \dots, h_m] & & \\ & \ddots & \\ & & [h_1, \dots, h_m] \end{bmatrix}$$

Here is how we calculate $\frac{\partial Z}{\partial x}$:
We can get the vector Jacobian

$$J = \frac{\partial z}{\partial x} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} & \cdots & \frac{\partial z_1}{\partial \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_n}{\partial x_1} & \frac{\partial z_n}{\partial x_2} & \cdots & \frac{\partial z_n}{\partial x_n} \end{bmatrix}$$

Because $z_i = W^1_{i1}x_1 + W^1_{i2}x_2 + \cdots + W^1_{in}x_n$
Thus,

$$\frac{\partial z_i}{\partial x_j} = W^1_{ij}$$

Thus,

$$J = \frac{\partial z}{\partial x} = W_1$$

We can get $\frac{\partial \hat{y}}{\partial h}$ in the same way.

$$\frac{\partial \hat{y}}{\partial h} = W_2$$

Finally, Our goal is to update W_1 and W_2, so we need to get $\frac{\partial E}{\partial W_1}$ and $\frac{\partial E}{\partial W_2}$.

$$\frac{\partial E}{\partial W_2} = \frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial W_2}$$

$$\frac{\partial E}{\partial W_1} = \frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial h} * \frac{\partial h}{\partial z} * \frac{\partial z}{\partial W_1}$$

And other derivatives we can get are:

$$\frac{\partial E}{\partial \hat{y}}$$

$$\frac{\partial E}{\partial h} = \frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial h}$$

$$\frac{\partial E}{\partial z} = \frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial h} * \frac{\partial h}{\partial z}$$

$$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial h} * \frac{\partial h}{\partial z} * \frac{\partial z}{\partial x}$$

And all the partial derivatives were calculated above. We can multiply them one by one to get these results.
As we can see, the most expensive pass is to get $\frac{\partial E}{\partial W_1}$. Let's compare $\frac{\partial E}{\partial W_1}$ to $\frac{\partial E}{\partial W_2}$, since the computation of $\frac{\partial z}{\partial W_1}$ and $\frac{\partial \hat{y}}{\partial W_2}$ is very similar. Thus computing $\frac{\partial E}{\partial W_1}$ has two more operations($\frac{\partial \hat{y}}{\partial h}$ and $\frac{\partial h}{\partial z}$) than computing $\frac{\partial E}{\partial W_2}$. And the computing complexity of $\frac{\partial E}{\partial \hat{y}}$ and $\frac{\partial \hat{y}}{\partial h}$ are similar and the computing complexity of $\frac{\partial h}{\partial z}$, $\frac{\partial z}{\partial w_1}$ and $\frac{\partial \hat{y}}{\partial w_2}$ are similar .Thus, computing $\frac{\partial E}{\partial W_1}$ is about 2 times more expensive than computing $\frac{\partial E}{\partial W_2}$.

2.SLOW RATE OF DESCENT

1. Here is the gradient.

$$L(w_1, w_2) = 0.5(aw_1^2 + bw_2^2)$$

$$\nabla L(w_1, w_2) = [\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}]$$

$$\frac{\partial L}{\partial w_1} = 2 * 0.5 * aw_1 = aw_1$$

$$\frac{\partial L}{\partial w_2} = 2 * 0.5 * bw_2 = bw_2$$

$$\nabla L(w_1, w_2) = [aw_1, bw_2]$$

We can assume that a and b are positive.
When $\nabla L(w_1, w_2) = [aw_1, bw_2] = 0$, which means that $w_1 = w_2 = 0$, then we can achieve the minimum value of L.

2.

$$w_1(t+1) = w_1(t) - \eta \nabla L(w_1(t))$$

$$w_1(t+1) = w_1(t) - \eta a w_1(t) = (1 - a\eta) w_1(t) = \rho_1 w_1(t)$$

$$\rho_1 = 1 - a\eta$$

In the same way, We can get $\rho_1 = 1 - b\eta$

3. If we want the gradient descent converge, we need to meet this requirement:

$$\lim_{t\to\infty} = \frac{w_1(t+1)-w_1^*}{w_1(t)-w^*} = \frac{(1-a\eta)w_1(t)-w_1^*}{w_1(t)-w_1^*} = \mu, 0 < \mu < 1$$

Thus,

$$1-a\eta \to \mu$$

Then

$$0 < 1-a\eta < 1$$

$$0 < \eta < \frac{1}{a}$$

In the same way we can get that

$$0 < \eta < \frac{1}{b}$$

Thus,

$$0 < \eta < min(\frac{1}{a}, \frac{1}{b})$$

4. If $\frac{a}{b}$ is a very large ratio, for example, if $\frac{a}{b} = h, h \to \infty$ Then

$$a = hb$$

Then

$$0 < \eta < min(\frac{1}{hb}, \frac{1}{b}) = min\frac{1}{hb}$$

$$\because \frac{1}{hb} \to 0, \therefore \eta \to 0$$

Thus, the learning rate is very small and the convergence rate of gradient descent is very slow.