Coursera Machine Learning Notes

Lectured by Andrew Ng

Typed by Aris Zhu Yi Qing

July 21, 2020

Contents

Chapter 1

Neural Networks

1.1 Notations

- $a_k^{(i)}$: activation energy at layer i of element k.
- $a^{(L)}$: the activation energy of the last layer, i.e. the values in the output layer.
- $\Theta_{ij}^{(l)}$: the weight between the element i in layer $l+1$ and the element j in layer l.
- y (or Y): target output(s).
- $\hat{y}$ (or $\hat{Y}$): derived output(s), equivalent to $a^{(L)}$.

1.2 Backward Propagation Derivation/Proof

As seen in the previous chapters, after deriving $J(\Theta)$, i.e.

$$J(\Theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}[-y_k^{(i)}\log\left((h_\theta(x^{(i)}))_k\right) - (1-y_k^{(i)})\log\left(1-(h_\theta(x^{(i)}))_k\right)],$$

or

$$J(\Theta) = \frac{1}{m} * \mathscr{S}(-Y \; .* \log(\hat{Y}) - (1-Y) \; .* \log(1-\hat{Y})),$$

where $\mathscr{S}$ refers to the sum of all the elements in the matrix, and $.*$ refers to element-wise multiplication. The next challenge is, as with previous chapters, to find out what represents

$$\frac{\partial J}{\partial \Theta_{ij}^{(l)}}.$$

For the sake of illustration, we are ignoring the regularization terms and dimension incompatibility.

Now the following is where the magic begins. It turns out that

$$\begin{aligned}\Delta^{(l)} := \frac{\partial J}{\partial \Theta^{(l)}} &= \frac{\partial z^{(l+1)}}{\partial \Theta^{(l)}} \frac{\partial J}{\partial z^{(l+1)}} \\ &= \frac{\partial(\Theta^{(l)} * a^{(l)})}{\partial \Theta^{(l)}} \delta^{(l+1)} \\ &= a^{(l)} \delta^{(l+1)},\end{aligned}$$

where

$$\begin{aligned}\delta^{(l)} := \frac{\partial J}{\partial z^{(l)}} &= \frac{\partial z^{(l+1)}}{\partial z^{(l)}} \frac{\partial J}{\partial z^{(l+1)}} \\ &= \frac{\partial(\Theta^{(l)} * g(z^{(l)}))}{\partial z^{(l)}} \delta^{(l+1)} \\ &= \delta^{(l+1)} \Theta^{(l)} g'(z^{(l)})\end{aligned}$$

and

$$\begin{aligned}\delta^{(L)} &= \frac{\partial J}{\partial z^{(L)}} \\ &= \frac{\partial \hat{y}}{\partial z^{(L)}} \frac{\partial J}{\partial \hat{y}},\end{aligned}$$

so we need to figure out the value of the two fractions to obtain the "terminating value" of the δs. On the other hand, we can also derive this by figuring out

$$\delta_k^{(L)} = \frac{\partial \hat{y}_k}{\partial z_k^{(L)}} \frac{\partial J}{\partial \hat{y}_k} \tag{1.1}$$

and line them up to return to vector form, $\delta^{(L)}$. For the sake of clarity of proof, we will use equation (1.1).

Let's tackle the first fraction. Since $\hat{y}_k = \sigma(z_k^{(L)})$, and

$$\begin{aligned}\sigma'(z) &= (-1)(1+e^{-z})^{-2} \frac{\partial(1+e^{-z})}{\partial z} \\ &= \frac{1}{1+e^{-z}} \cdot \frac{1+e^{-z}-1}{1+e^{-z}} \\ &= \sigma(z)(1-\sigma(z)),\end{aligned}$$

we can derive that

$$\frac{\partial \hat{y}_k}{\partial z_k^{(L)}} = \hat{y}_k(1 - \hat{y}_k). \tag{1.2}$$

Now let's tackle the second fraction. From the provided $J(\Theta)$, we look at the k-th output, and calculate the derivative of the loss with respect to its activation energy: (In the following equation, y_k is abbreviated to y, $\hat{y}_k$ is abbreviated to $\hat{y}$ for clarity)

$$\begin{aligned}
\frac{\partial J}{\partial \hat{y}} &= -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} \\
&= \frac{\hat{y}(1-y) - (1-\hat{y})y}{\hat{y}(1-\hat{y})} \\
&= \frac{\hat{y} - y}{\hat{y}(1-\hat{y})}.
\end{aligned}$$

i.e.

$$\frac{\partial J}{\partial \hat{y}_k} = \frac{\hat{y}_k - y_k}{\hat{y}_k(1 - \hat{y}_k)} \tag{1.3}$$

Substituting equations (1.2) and (1.3) into (1.1), and convert into vectorized form, we can get that

$$\delta^{(L)} = \hat{y} - y.$$