Chain rule for one layer neural network

Tang Xin (tangxint@gmail.com)

December 22, 2018

Abstract

This is a draft for understanding how to use chain rule to compute the derivative of cost with respect to weights and bias.

1 Problem

Assume that we have a very simple neural network. It can be described by the following Equation:

$$y = \sigma(XW + b), \tag{1}$$

where $W \in R^{m\times 1}, X \in R^{1\times m}$ and b is a scalar.

We also assume that we have n pairs of training samples $(x_i, y_i)_{i=1}^{n}$.

In order to train the neural network, we have used mean square error as loss function. So, we have

$$\begin{aligned} L &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \sigma(x_i W + b))^2, \end{aligned} \tag{2}$$

where x_i is the i-th training sample and $\hat{y}_i$ is the prediction of the one layer neural network.

So, we need to compute the derivative of L with respect to W and b.

2 The decomposition of cost

In Equation 2, we can see the cost is the average cost for all training samples. Hence let's denote $L_i = (y_i - \hat{y}_i)^2$. We can rewrite the cost L as:

$$L = \frac{1}{n}\sum_{i=1}^{n} L_i$$

Obviously, the derivatives of L with respect to W and b are as following:

$$\frac{\partial L}{\partial W} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial L_i}{\partial W} \tag{3}$$

$$\frac{\partial L}{\partial b} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial L_i}{\partial b} \tag{4}$$

$$\tag{5}$$

Therefore, we just need to compute the derivative of cost for each sample L_i.

3 The derivative of L_i with respect to W

As we known,

$$\begin{aligned} L_i &= (y_i - \hat{y}_i)^2 && (6) \\ &= (y_i - \sigma(z_i))^2 && (7) \\ &= (y_i - \sigma(x_i W + b))^2 && (8) \end{aligned}$$

Therefore, according to chain rule, the derivative of L_i respect to W is

$$\begin{aligned} \frac{\partial L_i}{\partial W} &= \frac{\partial L_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial W} && (9) \\ &= \frac{\partial L_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_i} \frac{\partial z_i}{\partial W} && (10) \end{aligned}$$

At the same time, from Equation (11), we have

$$\begin{aligned} \frac{\partial L_i}{\partial \hat{y}_i} &= 2(\hat{y}_i - y_i) && (11) \\ \frac{\partial \hat{y}_i}{\partial z_i} &= \frac{\partial \sigma(z_i))}{\partial z_i} = \sigma z_i(1 - \sigma z_i) && (12) \\ \frac{\partial z_i}{\partial W} &= x_i^T && (13) \end{aligned}$$

Therefore, we get

$$\frac{\partial L_i}{\partial W} = 2(\hat{y}_i - y_i)\sigma z_i(1 - \sigma z_i)x_i^T \quad (14)$$

References