Gradient Descent Homework

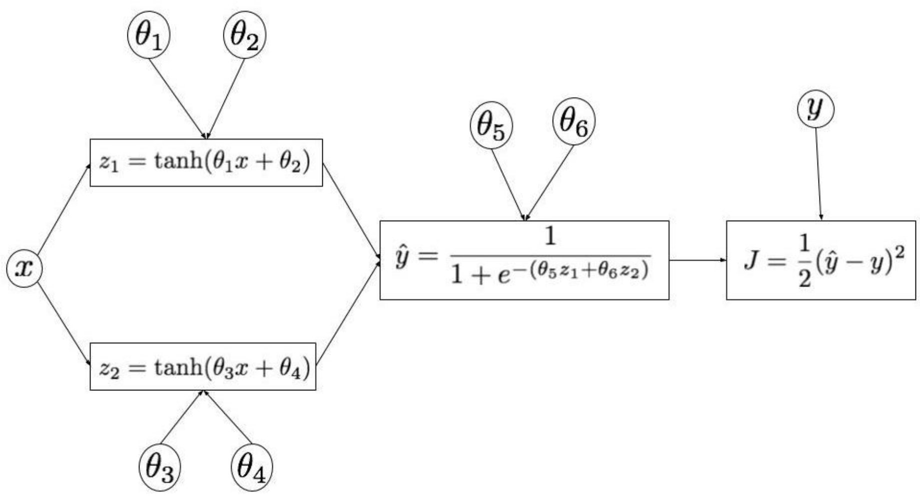
Optimization Methods for Analytics, MSA 8100

Fall 2018

Problem

For the computational graph below, perform one step of gradient descent of $J(x, y; \theta)$ with respect to the parameters $\theta = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6]$ using a learning rate of $\alpha = 2$. Let

$$x = 3,\ \ y = 1,\ \ \text{and}\ \ \theta = \left[\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}, \theta_5^{(0)}, \theta_6^{(0)}\right] = [1, -2, -0.5, 1, -2, 3]. \quad (1)$$



**Recall that* $\theta_j^{(k+1)} = \theta_j^{(k)} - \alpha \left.\frac{\partial J}{\partial \theta_j}\right|_{\substack{(x,y)\\ \theta^{(k)}}}.$

Solution for one parameter:

In order to calculate one update in the direction of, say, θ_1, first we need to find $\frac{\partial J}{\partial \theta_1}$ using the chain rule (and using the definitions in the graph):

$$\begin{aligned}\frac{\partial J}{\partial \theta_1} &= \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_1} \frac{\partial z_1}{\partial \theta_1} \\ &= (\hat{y} - y) \cdot \left(\frac{\theta_5 e^{-(\theta_5 z_1 + \theta_6 z_2)}}{\left(1 + e^{-(\theta_5 z_1 + \theta_6 z_2)}\right)^2} \right) \cdot \left(x \left(1 - \tanh^2(\theta_1 x + \theta_2) \right) \right).\end{aligned} \quad (2)$$

Now we evaluate this derivative using the values in (1). We can find first z_1, z_2, and $\hat{y}$ by plugging in these values, in order to simplify the numerical calculations (we use the expressions in the graph for z_1, z_2, $\hat{y}$):

$$z_1 = 0.76159416, \ \ z_2 = -0.4621176, \ \ \hat{y} = 0.05168399\,. \tag{3}$$

Next, we use (1) and (3) to plug in in (2). We get that

$$\left.\frac{\partial J}{\partial \theta_j}\right|_{\substack{(x=3,y=1)\\ \theta^{(0)}}} = 0.11712138\,. \tag{4}$$

Finally, we are ready to perform the update on θ_1:

$$\theta_1^{(1)} = \theta_1^{(0)} - \alpha \left.\frac{\partial J}{\partial \theta_1}\right|_{\substack{(x=3,y=1)\\ \theta^{(0)}}} = 1 - (2)(0.11712138) = \underline{0.76575724}\,.$$

Doing this for each θ_j will constitute one step of gradient descent.

General strategy:

To sum up, the process of completing one update for θ_j consists of the following steps.

1. Find the derivative of J with respect to θ_j (you will need the chain rule since the functions are nested).
2. Evaluate this derivative using the values in (1).
3. Multiply this derivative by α and subtract it from the initial value $\theta_j^{(0)}$.

You will see that point $\theta = \left[\theta_1^{(1)}, \theta_2^{(1)}, \theta_3^{(1)}, \theta_4^{(1)}, \theta_5^{(1)}, \theta_6^{(1)}\right]$ is an improvement over $\theta^{(0)}$ in the search for a minimizer. You can verify this by comparing the values of J at $\theta^{(0)}$ and at $\theta^{(1)}$ (in fact, you will see that the function decreases from 0.44965163 to 0.00380892).

Observation

This computation graph is actually a neural network with one hidden layer of two nodes. It can perform binary classification given a training set of several examples x, y.