Practical Class 05

Gradient Descent

1. In this problem,you will solve a linear regression problem using two different approaches. Here is a small training set:

x	y
1	3
-1	-2
2	4

   Assume that $w_0 = 0$, so that the line is without bias as:

   $$h_{w_1} = w_1 x,$$

   and the cost function is

   $$J(w_1) = \sum_{i=1}^{N} (w_1 x_i - y_i)^2.$$

   (a) First, substitute the values of x_i and y_i from the training data into $J(w_1)$, and write the derivative. Set it equal to zero and solve for w_1.

   (b) Second, use the general formula we derived in the lecture for the value of w_1 that minimizes $J(w_1)$.

2. In this problem you will execute a few steps of gradient descent using the training data from Problem 1. However, this time you will not assume that $w_0 = 0$, so you will have to update both w_0 and w_1 within the gradient descent algorithm. Below are two different initialization for w_0 and w_1. In each case, execute one step of gradient descent (update both w_0 and w_1) using a step size of $\eta = 0.1$, and write the new values w_0' and w_1'.

   (a) $w_0 = 0$, $w_1 = 0$

   (b) $w_0 = 0$, $w_1 = 2$

   Now, for each of the two cases, measure the distance between the original point (w_0, w_1) in parameter space and the new point (w_0', w_1'):

   $$d((w_0, w_1), (w_0', w_1')) = \sqrt{(w_0' - w_0)^2 + (w_1' - w_1)^2}.$$

   (c) Write the distance moved by the two updates in (a) and (b).

   (d) Which update, (a) or (b), do you think finished closer to the minimum? Why?

   (e) The gradient descent algorithm says “repeat until convergence”. What would be a reasonable test for convergence?

3. Another reasonable cost function for linear regression would be

   $$J(w_0, w_1) = \sum_{i=1}^{N} \left| h_w(x_i) - y_i \right|.$$

   This penalizes the absolute value of the difference between the predicted value and the true value, instead of the square of the difference. However, Carl Friedrich Gauss (1777–1855) chose to use squared loss instead, and squared loss is still much more widely used today. Why do you think squared loss might be more popular? Briefly explain.