CS 341: Machine Learning
Homework 1

September 12, 2012

Instructions

Complete all required problems and turn in by the beginning of class on Wed. Sep 19. You may submit hard copy solutions in class, or electronic solutions through ella. Work on optional problems for an extra challenge and to gain a deeper insight into the material. I will grade these for extra credit, which may have a small effect on your final course grade.

You may work together with other students, but *I highly encourage you to attempt the problems first on your own.* Please make sure to:

- Write your name on your submission
- Write the name of all students with whom you collaborated
- Cite any sources you used other than the textbook, course notes, or Coursera

Problems

1. (Originally Exercise 2 from Lecture 2). 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

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

   (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. (The only derivative you need to know is $\frac{d}{dw_1}(w_1 x - y)^2 = 2(w_1 x^2 - xy)$; alternately, you may want to use $\frac{d}{dw_1}(aw_1^2 + bw_1 + c) = 2aw_1 + b$ for any constants a, b, c.)

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

2. On page 719 of your textbook, Equation (18.3) gives a closed form solution to the problem of finding w_0 and w_1 to minimize $J(w_0, w_1)$. The steps to derive this solution are not shown—they are similar

to the steps we followed to derive the formula for the value of w_1 that minimizes $J(w_1)$, but they are more difficult because there are two unknowns.

Imagine computing w_0 and w_1 using Equation (18.3) in a language like Java. Write pseudocode for a function that takes as input two arrays `x` and `y`, each of length N, containing the training data. The function should compute w_0 and w_1.

(a) How many passes through the data are required? (One pass = one for-loop to read the entries of each array).

(b) Write the pseudocode.

(c) (Optional). Derive Equation (18.3).

3. 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 initializations 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 $\alpha = 0.1$, 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?

4. Page 720 of the textbook describes a very common variant of gradient descent called *stochastic gradient descent*. Read the description in the text and write pseudocode for stochastic gradient descent. Make sure to clearly define the inputs to the procedure.

5. (Optional). Another reasonable cost function for linear regression would be

$$J(w_0, w_1) = \sum_{i=1}^{N} |h_{\mathbf{w}}(x) - y_i|.$$

   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.

6. (Optional). A continuous twice-differentiable function is *convex* if its second derivative is always positive. Show that $J(w_1)$ from Problem 1 is convex.

7. How much time did you spend on the required problems for this homework?