Week 2-1: Linear Regression
with Multiple Variables

1. Suppose $m = 4$ students have taken some class, and the class had a midterm exam and a final exam. You have collected a dataset of their scores on the two exams, which is as follows:

Midterm Exam	(Midterm Exam)^2	Final Exam
89	7921	96
72	5184	74
94	8836	87
69	4761	78

You'd like to use polynomial regression to predict a student's final exam score from their midterm exam score. Concretely, suppose you want to fit a model of the form $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$, where x_1 is the midterm score and x_2 is the square of the midterm score. Further, you plan to use both feature-scaling (dividing by the max-min, or range, of a feature) and mean normalization. What is the normalized feature $x_2^{(4)}$? (Hint: midterm = 69, final = 78 is training example 4.) Please round off your answer to two decimal places and enter in the text box below.

> We have our score vector as: $\begin{bmatrix} 7921 \\ 5184 \\ 8836 \\ 4761 \end{bmatrix}$. The range of these values is $8836 - 4761 = 4075$. We divide the vector by the range to get: $\begin{bmatrix} 1.94 \\ 1.27 \\ 2.17 \\ 1.17 \end{bmatrix}$. The next step is to find the mean of the values of x_2, which is about 1.64. Finally, we replace each x_i in the vector with $x_i - \mu$ to get: $\begin{bmatrix} 0.31 \\ -0.37 \\ 0.53 \\ -0.47 \end{bmatrix}$.
>
> Thus, the value of the normalized feature $x_2^{(4)}$ is $\boxed{-0.47}$.

2. You run gradient descent for 15 iterations with $\alpha = 0.3$ and compute $J(\theta)$ after each iteration. You find that the value of $J(\theta)$ decreases slowly and is still decreasing after 15 iterations. Based on this, which of the following conclusions seems most plausible?

> Rather than using the current value of α, it'd be more promising to try a larger value of α, say 1.0.

3. Suppose you have $m = 14$ training examples with $n = 3$ features (excluding the additional all-ones feature for the intercept term, which you should add). The normal equation is $\theta = (X^T X)^{-1} X^T y$. For the given values of m and n, what are the dimensions of θ, X, and y in this equation?

> X is 14×4, y is 14×1, θ is 4×1.

4. Suppose you have a dataset with $m = 1000000$ examples and $n = 200000$ features for each example. You want to use multivariable linear regression to fit the parameters θ to our data. Should you prefer gradient descent or the normal equation?

Gradient descent, since $(X^T X)^{-1}$ will be very slow to compute in the normal equation.

5. Which of the following are reasons for using feature scaling?

(a) It speeds up gradient descent by making each iteration of gradient descent less expensive to compute.