# Week 2-Linear Regression with multiple Variable

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 *hθ*​(*x*)=*θ*0​+*θ*1​*x*1​+*θ*2​*x*2​, where *x*1​ is the midterm score and *x*2​ is (midterm score)^2. 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*1(1)​? (Hint: midterm = 89, final = 96 is training example 1.) Please round off your answer to two decimal places and enter in the text box below.



1. You run gradient descent for 15 iterations with *α*=0.3 and compute  *J*(*θ*) after each iteration. You find that the value of  *J*(*θ*) **increases** over time. Based on this, which of the following conclusions seems most plausible?

Rather than use the current value of *α*, it'd be more promising to try a larger value of *α* (say *α*=1.0).

Rather than use the current value of *α*, it'd be more promising to try a smaller value of *α* (say *α*=0.1).

*α*=0.3 is an effective choice of learning rate.

1. Suppose you have *m*=23 training examples with *n*=5 features (excluding the additional all-ones feature for the intercept term, which you should add). The normal equation is *θ*=(*XT X*)-1 *XT y*. For the given values of m and n, what are the dimensions of *θ*, X, and y in this equation?

Xis 23×6, y is 23×6, *θ* is 6×6

X *is* 23×5, y is 23×1, *θ* is 5×5

Xis 23×5, y is 23×1, *θ* is 5×1

X *is* 23×6, yis 23×1, *θ* is 6×1

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

The normal equation, since it provides an efficient way to directly find the solution.

The normal equation, since gradient descent might be unable to find the optimal *θ*.

Gradient descent, since it will always converge to the optimal *θ*.

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

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

It prevents the matrix X^TX (used in the normal equation) from being non-invertable (singular/degenerate).

It speeds up gradient descent by making it require fewer iterations to get to a good solution.

It speeds up solving for *θ* using the normal equation.

It is necessary to prevent gradient descent from getting stuck in local optima.

