

Feedback — IV. Linear Regression with Multiple Variables

[Help Center](#)

You submitted this quiz on **Mon 26 Jan 2015 10:00 AM CET**. You got a score of **5.00** out of **5.00**.

Question 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 (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 $x_2^{(4)}$? (Hint: midterm = 89, final = 96 is training example 1.) Please enter your answer in the text box below. If applicable, please provide at least two digits after the decimal place.

You entered:

-0.46982

Your Answer		Score	Explanation
-0.46982	✓	1.00	
Total		1.00 / 1.00	

Question Explanation

The mean of x_2 is 6675.5 and the range is $8836 - 4761 = 4075$. So $x_1^{(4)}$ is $\frac{4761 - 6675.5}{4075} = -0.47$.

Question 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** quickly then levels off. Based on this, which of the following conclusions seems most plausible?

Your Answer	Score	Explanation
<input type="radio"/> Rather than use the current value of α , it'd be more promising to try a smaller value of α (say $\alpha = 0.1$).		
<input type="radio"/> Rather than use the current value of α , it'd be more promising to try a larger value of α (say $\alpha = 1.0$).		
<input checked="" type="radio"/> $\alpha = 0.3$ is an effective choice of learning rate.	✓ 1.00	We want gradient descent to quickly converge to the minimum, so the current setting of α seems to be good.
Total	1.00 / 1.00	

Question 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?

Your Answer	Score	Explanation
<input checked="" type="radio"/> X is 14×4 , y is 14×1 , θ is 4×1	✓ 1.00	
<input type="radio"/> X is 14×3 , y is 14×1 , θ is 3×3		
<input type="radio"/> X is 14×4 , y is 14×4 , θ is 4×4		
<input type="radio"/> X is 14×3 , y is 14×1 , θ is 3×1		
Total	1.00 / 1.00	

Question Explanation

X has m rows and $n + 1$ columns (+1 because of the $x_0 = 1$ term). y is an m -vector. θ is an $(n + 1)$ -vector.

Question 4

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 θ to our data. Should you prefer gradient descent or the normal equation?

Your Answer	Score	Explanation
-------------	-------	-------------

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

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

✓ 1.00

With $n = 200000$ features, you will have to invert a 200001×200001 matrix to compute the normal equation. Inverting such a large matrix is computationally expensive, so gradient descent is a good choice.

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

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

Total	1.00 /
	1.00

Question 5

Which of the following are reasons for using feature scaling?

Your Answer	Score	Explanation
-------------	-------	-------------

<input checked="" type="checkbox"/> It speeds up gradient descent by making it require fewer iterations to get to a good solution.	✓	0.25	Feature scaling speeds up gradient descent by avoiding many extra iterations that are required when one or more features take on much larger values than the rest.
<input type="checkbox"/> It prevents the matrix $X^T X$ (used in the normal equation) from being non-invertable (singular/degenerate).	✓	0.25	$X^T X$ can be singular when features are redundant or there are too few examples. Feature scaling does not solve these problems.
<input type="checkbox"/> It speeds up gradient descent by making each iteration of gradient descent less expensive to compute.	✓	0.25	The magnitude of the feature values are insignificant in terms of computational cost.
<input type="checkbox"/> It speeds up solving for θ using the normal equation.	✓	0.25	The magnitude of the feature values are insignificant in terms of computational cost.
Total		1.00 / 1.00	