Feedback — IV. Linear Regression with Multiple Variables

Help

You submitted this quiz on **Sat 9 Nov 2013 8:03 PM PST**. 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 $x_1^{(1)}$? (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:

0.32

Your Answer		Score	Explanation
0.32	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

The mean of x_1 is 81 and the range is $94 - 69 = 25$ So $x_1^{(1)}$ is $\frac{89-81}{25} = 0.32$.

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 $J(\theta)$ **increases** over time. Based on this, which of the following conclusions seems most plausible?

Your Answer		Score	Explanation
Rather than use the current value of α, it'd be more promising to try a smaller value of α (say $\alpha = 0.1$).	✔	1.00	Since the cost function is increasing, we know that gradient descent is diverging, so we need a lower learning rate.
$\alpha = 0.3$ is an effective choice of learning rate.			
Rather than use the current value of α, it'd be more promising to try a larger value of α (say $\alpha = 1.0$).			
Total		1.00 / 1.00	

Question 3

Suppose you have $m = 28$ training examples with $n = 4$ 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
X is 28×4, y is 28×1, θ is 4×4		

◉ X is 28×5, y is 28×1, θ is 5×1	✔	1.00	
○ X is 28×4, y is 28×1, θ is4×1			
○ X is 28×5, y is 28×5, θ is 5×5			
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 $n = 15$ 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 $(X^T X)^{-1}$ will be very slow to compute in the normal equation.			
○ 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.	✔	1.00	With $n = 15$ features, you will have to invert a 15×15 matrix to compute the normal equation. This is a simple inversion, so the normal equation is efficient.
○ Gradient descent, since it will always converge to the optimal θ.			
Total		1.00 / 1.00	

Question 5

Which of the following are reasons for using feature scaling?

Your Answer		Score	Explanation
☐ It is necessary to prevent the normal equation from getting stuck in local optima.	✔	0.25	The cost function $J(\theta)$ for linear regression has no local optima.
☑ 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.
☐ It prevents the matrix X^TX (used in the normal equation) from being non-invertable (singular/degenerate).	✔	0.25	X^TX can be singular when features are redundant or there are too few examples. Feature scaling does not solve these problems.
☐ It is necessary to prevent gradient descent from getting stuck in local optima.	✔	0.25	The cost function $J(\theta)$ for linear regression has no local optima.
Total		1.00 / 1.00	