In the Week 3 Discussion Forum, take a moment to discuss the questions below. Post your answers briefly and precisely and contribute to the discussion by responding to another post.

Suppose the features in your training set have very different scales. Which algorithms might suffer from this, and how? What can you do about it?

Why is gradient descent introduced rather than just leveraging normal equations to estimate the parameters in linear regression?

**Post your response by Friday of Week 3 (11:59 p.m.), then respond to at least one peer’s post by Sunday of Week 3 (11:59 p.m.). Please list reference pages that helped you answer the question so fellow students can learn from these as well.**

1. Suppose the features in your training set have very different scales. Which algorithms might suffer from this, and how? What can you do about it?
2. The normal equations method does not require normalizing the features, so it remains unaffected by features in the training set having very different scales.
3. Feature scaling is required for the various gradient descent algorithms. Feature scaling will help gradient descent converge quicker.

The cost function will have the shape of an elongated bowl, so the Gradient Descent algorithms will take a long time to converge.  
To solve this, you should scale the data before training the model.  
Note that the Normal Equation will work just fine without scaling

Gradient Descent techniques will take a very long time to converge since the cost function will be shaped like an extended bowl.

You should scale the data before training the model to address this.

Noting that scaling is not necessary for the Normal Equation to function,

1. Why is gradient descent introduced rather than just leveraging normal equations to estimate the parameters in linear regression?

In the normal equation estimate, no need to iterate while in gradient descent need for many more iterations. So, that could make it slower. But in the normal equation, it is more Computationally expensive when large no of features, because of the need to take the inverse of an N x N matrix in order to solve the parameters. And gradient descent will be reasonably efficient and will do something acceptable when you have a very large number of features. So for that why gradient descent is introduced to estimate the large no of parameters in linear regression.

Hello Rumana,

Thanks for sharing, it’s a good explanation of why gradient is decent over a normal equation for the estimate of the linear regression. How gradient descent reduces the cost function by modifying the model’s.