Elements of Scientific Computing with Julia

March 5, 2015

Fitting a Line to Data

Linear regression is a *statistical learning* approach for quantitatively predicting an outcome $\mathbf{y}$ on the basis of predictor variables $\mathbf{x}_i$, such that the relationship between the outcome and the variables is linear. Mathematically we write:

Fitting a Line to Data

*L*inear regression is a *statistical learning* approach for quantitatively predicting an outcome $\mathbf{y}$ on the basis of predictor variables $\mathbf{x}_i$, such that the relationship between the outcome and the variables is linear. Mathematically we write:

$$\mathbf{y} \approx \theta_0 + \theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + \ldots + \theta_n \mathbf{x}_n$$

Fitting a Line to Data

Linear regression is a *statistical learning* approach for quantitatively predicting an outcome $\mathbf{y}$ on the basis of predictor variables $\mathbf{x}_i$, such that the relationship between the outcome and the variables is linear. Mathematically we write:

$$\mathbf{y} \approx \theta_0 + \theta_1 \mathbf{x}_1 + \theta_2 \mathbf{x}_2 + ... + \theta_n \mathbf{x}_n$$

that is $\mathbf{y}$ can be approximated by a linear combination of n predictor variables.

Fitting a Line to Data

In this lecture we will focus our attention on *single* variable regression and leave the multivariable case for the next lecture. In the single variable case we have that the outcome $\mathbf{y}$ can be approximated by a linear relationship involving only one input variable $\mathbf{x}$, or mathematically:

Fitting a Line to Data

In this lecture we will focus our attention on *single* variable regression and leave the multivariable case for the next lecture. In the single variable case we have that the outcome $\mathbf{y}$ can be approximated by a linear relationship involving only one input variable $\mathbf{x}$, or mathematically:

$$\mathbf{y} \approx \theta_0 + \theta_1 \mathbf{x}$$

Supervised Learning

There are different contexts in which we may discuss the topic of linear regression. In this lecture, we will study it in the light of machine learning and thus use vocabulary pertinent to this field.

Supervised Learning

There are different contexts in which we may discuss the topic of linear regression. In this lecture, we will study it in the light of machine learning and thus use vocabulary pertinent to this field.

In machine learning there are two main sorts of problems:

Supervised Learning

There are different contexts in which we may discuss the topic of linear regression. In this lecture, we will study it in the light of machine learning and thus use vocabulary pertinent to this field.

In machine learning there are two main sorts of problems:

- *supervised learning* and
- *unsupervised learning*

Supervised Learning

There are different contexts in which we may discuss the topic of linear regression. In this lecture, we will study it in the light of machine learning and thus use vocabulary pertinent to this field.

In machine learning there are two main sorts of problems:

- *supervised learning* and
- *unsupervised learning*

In this course we will only address supervised learning problems.

Supervised Learning

In supervised learning, we wish to infer a function that classifies or fits future data, based on a *training data set:*

Supervised Learning

In supervised learning, we wish to infer a function that classifies or fits future data, based on a *training data set:*

data that we already have collected and know the "right" output for each input

Supervised Learning

In supervised learning, we wish to infer a function that classifies or fits future data, based on a *training data set:*

data that we already have collected and know the "right" output for each input

Let us motivate our discussion with an example. Consider the following training data set on the next slide, where we have on the left column the number of years for which a person went to school and on the right column the income they now make, in thousands of dollars per year.

Example: Education v. Income

"Education"	"Income"
10.0	26.6588
10.4013	27.3064
10.8428	22.1324
11.2441	21.1698
11.6455	15.1926
12.087	26.399
12.4883	17.4353
12.8896	25.5079
13.291	36.8846
13.7324	39.6661
14.1338	34.3963
14.5351	41.498
...	

Example: Education v. Income

- "x" or input is the education level of an individual
- "y" or output is the income level of the same individual

Example: Education v. Income

- "x" or input is the education level of an individual
- "y" or output is the income level of the same individual

Since we have many inputs for different observations (say m of them), then we put them all in a $m \times 1$ vector $\mathbf{x}$.

Example: Education v. Income

- "x" or input is the education level of an individual
- "y" or output is the income level of the same individual

Since we have many inputs for different observations (say m of them), then we put them all in a $m \times 1$ vector $\mathbf{x}$.

Similarly, $\mathbf{y}$ is a $m \times 1$ vector of outputs.

Example: Education v. Income

- "x" or input is the education level of an individual
- "y" or output is the income level of the same individual

Since we have many inputs for different observations (say m of them), then we put them all in a $m \times 1$ vector $\mathbf{x}$.

Similarly, $\mathbf{y}$ is a $m \times 1$ vector of outputs.

In linear regression, we wish to infer the function of the line that best fits data in a certain input-output relationship, that is, we wish to come up with a linear function

Example: Education v. Income

- "x" or input is the education level of an individual
- "y" or output is the income level of the same individual

Since we have many inputs for different observations (say m of them), then we put them all in a $m \times 1$ vector $\mathbf{x}$.

Similarly, $\mathbf{y}$ is a $m \times 1$ vector of outputs.

In linear regression, we wish to infer the function of the line that best fits data in a certain input-output relationship, that is, we wish to come up with a linear function

$$\mathbf{h}_\theta(\mathbf{x}) = \theta_0 + \theta_1 \mathbf{x}$$

Example: Education v. Income

- "x" or input is the education level of an individual
- "y" or output is the income level of the same individual

Since we have many inputs for different observations (say m of them), then we put them all in a $m \times 1$ vector $\mathbf{x}$.

Similarly, $\mathbf{y}$ is a $m \times 1$ vector of outputs.

In linear regression, we wish to infer the function of the line that best fits data in a certain input-output relationship, that is, we wish to come up with a linear function

$$\mathbf{h}_\theta(\mathbf{x}) = \theta_0 + \theta_1 \mathbf{x}$$

that best approximates our output vector $\mathbf{y}$.

Example: Education v. Income

In the formula above we have that θ_i's are the *parameters* or the *weights* of the linear map from $\mathbf{x}$ to $\mathbf{y}$.

Example: Education v. Income

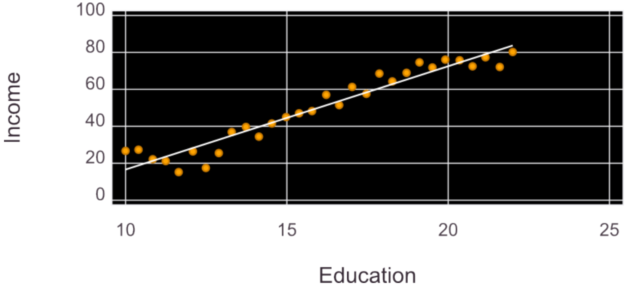
In the formula above we have that θ_i's are the *parameters* or the *weights* of the linear map from $\mathbf{x}$ to $\mathbf{y}$.

So we want to find the parameters (in this case they are the slope and y-intercept of our line) that give us a line of best fit as in the picture below:

Example: Education v. Income

In the formula above we have that θ_i's are the *parameters* or the *weights* of the linear map from $\mathbf{x}$ to $\mathbf{y}$.

So we want to find the parameters (in this case they are the slope and y-intercept of our line) that give us a line of best fit as in the picture below:



How to Choose the θ's?

So how do we choose the parameters θ_i?

How to Choose the θ's?

So how do we choose the parameters θ_i?

In this lecture we will do that via the *Gradient Descent Method*.

How to Choose the θ's?

So how do we choose the parameters θ_i?

In this lecture we will do that via the *Gradient Descent Method*.

In the next lecture we will generalize linear regression and gradient descent for multiple variables (that is, multiple input vectors $\mathbf{x}_i$), and also learn the *Normal Method* for solving for the parameters θ_i.

How to Choose the θ's?

But before we get into all that, let us get a better intuition for these parameters, simply by experimenting with a few different combinations of θ_0 and θ_1:

How to Choose the θ's?

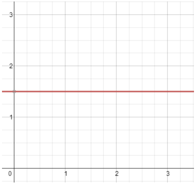
But before we get into all that, let us get a better intuition for these parameters, simply by experimenting with a few different combinations of θ_0 and θ_1:

$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 1.5, \theta_1 = 0$:

How to Choose the θ's?

But before we get into all that, let us get a better intuition for these parameters, simply by experimenting with a few different combinations of θ_0 and θ_1:

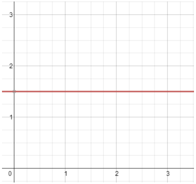
$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 1.5, \theta_1 = 0$:



How to Choose the θ's?

But before we get into all that, let us get a better intuition for these parameters, simply by experimenting with a few different combinations of θ_0 and θ_1:

$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 1.5, \theta_1 = 0$:

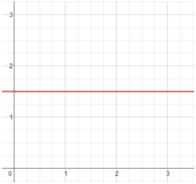


$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 0, \theta_1 = 0.5$:

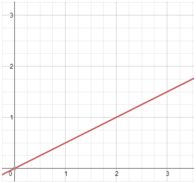
How to Choose the θ's?

But before we get into all that, let us get a better intuition for these parameters, simply by experimenting with a few different combinations of θ_0 and θ_1:

$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 1.5, \theta_1 = 0$:



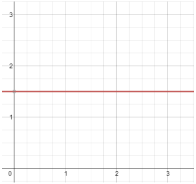
$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 0, \theta_1 = 0.5$:



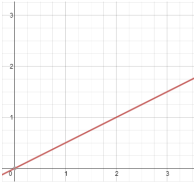
How to Choose the θ's?

But before we get into all that, let us get a better intuition for these parameters, simply by experimenting with a few different combinations of θ_0 and θ_1:

$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 1.5, \theta_1 = 0$:



$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 0, \theta_1 = 0.5$:

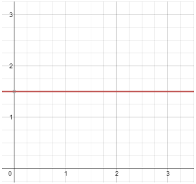


$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 1, \theta_1 = 0.5$:

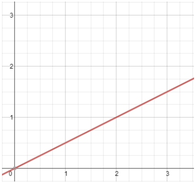
How to Choose the θ's?

But before we get into all that, let us get a better intuition for these parameters, simply by experimenting with a few different combinations of θ_0 and θ_1:

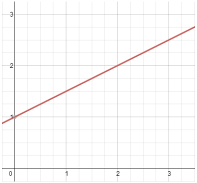
$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 1.5, \theta_1 = 0$:



$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 0, \theta_1 = 0.5$:



$h(x) = \theta_0 + \theta_1 x$,
$\theta_0 = 1, \theta_1 = 0.5$:



How to Choose the θ's?

Note how changing the parameters θ_i changes the slope and position of our line in the plane.

How to Choose the θ's?

Note how changing the parameters θ_i changes the slope and position of our line in the plane.

As an exercise, I encourage you to read different datasets in Julia (with one input variable, where the output is quantitative), scatter plot the data and play around with manually picking the slope and intercept parameters for the line that best fits the data visually.

The Cost Function

In order for us to be able to apply the gradient descent algorithm for solving our liner regression problem, we will need a function to minimize (as we will see later gradient descent is an algorithm for finding the input that minimizes the output value of a function).

The Cost Function

In order for us to be able to apply the gradient descent algorithm for solving our liner regression problem, we will need a function to minimize (as we will see later gradient descent is an algorithm for finding the input that minimizes the output value of a function).

This function is called the *cost function* and it is defined as follows:

The Cost Function

In order for us to be able to apply the gradient descent algorithm for solving our liner regression problem, we will need a function to minimize (as we will see later gradient descent is an algorithm for finding the input that minimizes the output value of a function).

This function is called the *cost function* and it is defined as follows:

$$C(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

The Cost Function

This function outputs the sum of the squared errors between the predicted outputs and the actual outputs, given inputs θ_0 and θ_1.

The Cost Function

This function outputs the sum of the squared errors between the predicted outputs and the actual outputs, given inputs θ_0 and θ_1.

Thus, minimizing this function means getting the θ_0 and θ_1 that causes C to be smallest.

The Cost Function

This function outputs the sum of the squared errors between the predicted outputs and the actual outputs, given inputs θ_0 and θ_1.

Thus, minimizing this function means getting the θ_0 and θ_1 that causes C to be smallest.

It seems intuitive to minimize the magnitude of the errors, but you may be wondering why minimize the squared errors versus just the absolute value of the errors.

The Cost Function

This function outputs the sum of the squared errors between the predicted outputs and the actual outputs, given inputs θ_0 and θ_1.

Thus, minimizing this function means getting the θ_0 and θ_1 that causes C to be smallest.

It seems intuitive to minimize the magnitude of the errors, but you may be wondering why minimize the squared errors versus just the absolute value of the errors.

For our purposes, the explanation is simple - minimization means taking derivatives, and therefore the cost function must be continuous and differentiable.

The Cost Function

Note that the cost function is a *multivariate quadratic function*.

The Cost Function

Note that the cost function is a *multivariate quadratic function*.

If C were a quadratic function of one variable, then in a two dimensional space its graph would look like a concave-up parabola. Such a parabola would have a global minimum at the vertex.

The Cost Function

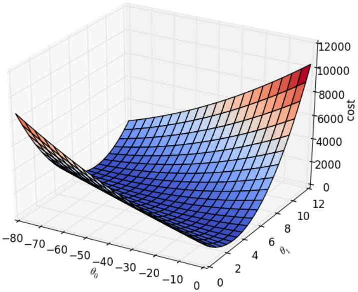
Note that the cost function is a *multivariate quadratic function*.

If C were a quadratic function of one variable, then in a two dimensional space its graph would look like a concave-up parabola. Such a parabola would have a global minimum at the vertex.

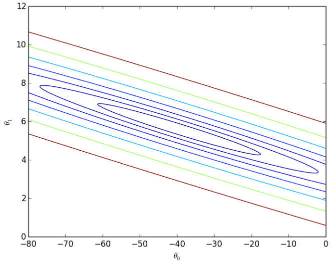
Analogously, in our case C's graph lives in a three dimensional space and it is a "bowl-up" shaped surface, with a global minimum at the base of the bowl.

The Cost Function

The surface plot:



The contour plot:



Gradient Descent

In this section we will explore the gradient descent algorithm for optimizing the cost function defined earlier.

Gradient Descent

In this section we will explore the gradient descent algorithm for optimizing the cost function defined earlier.

The gradient descent algorithm is a type of search algorithm, that is we start with an initial guess and then we update our guess iteratively by some rule.

Gradient Descent

*I*n this section we will explore the gradient descent algorithm for optimizing the cost function defined earlier.

The gradient descent algorithm is a type of search algorithm, that is we start with an initial guess and then we update our guess iteratively by some rule.

The rule for gradient descent is that from our previous guess we should step down (with some step size) toward the direction of steepest descent.

Gradient Descent

In this section we will explore the gradient descent algorithm for optimizing the cost function defined earlier.

The gradient descent algorithm is a type of search algorithm, that is we start with an initial guess and then we update our guess iteratively by some rule.

The rule for gradient descent is that from our previous guess we should step down (with some step size) toward the direction of steepest descent.

If we iterate this rule in a “bowl-up” shaped surface with a not-to-large step size, we shall converge to a minimum.

The Gradient Descent Algorithm

Gradient Descent Algorithm for Bivariate Functions

repeat until convergence $\{$

for $j = 0$ and $j = 1$, simultaneously update:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} C(\theta_0, \theta_1)$$

$\}$

The Gradient Descent Algorithm

So we take a guess on each θ, then update each θ *simultaneously* by taking a step of size α in the respective direction of steepest descent as dictated by the partial derivative of the cost function with respect to the θ that is being updated.

The Gradient Descent Algorithm

So we take a guess on each θ, then update each θ *simultaneously* by taking a step of size α in the respective direction of steepest descent as dictated by the partial derivative of the cost function with respect to the θ that is being updated.

The intuition here is that the slope of the tangent line to our current point is the steepest slope we can get on a line that would still include the point.

Choosing α

In machine learning, we call the α parameter "the learning rate".
Note:

Choosing α

In machine learning, we call the α parameter “the learning rate”. Note:

- if α is too large our algorithm may diverge, as our steps will be so large as to “climb the walls” of our surface

Choosing α

In machine learning, we call the α parameter "the learning rate". Note:

- if α is too large our algorithm may diverge, as our steps will be so large as to "climb the walls" of our surface
- if α is too small, our algorithm may take a long time to converge

Choosing α

In machine learning, we call the α parameter “the learning rate”. Note:

- if α is too large our algorithm may diverge, as our steps will be so large as to “climb the walls” of our surface
- if α is too small, our algorithm may take a long time to converge

Tweaking α can be seen as more of an art than a science and it will depend on your particular problem.

Choosing α

In machine learning, we call the α parameter "the learning rate". Note:

- if α is too large our algorithm may diverge, as our steps will be so large as to "climb the walls" of our surface
- if α is too small, our algorithm may take a long time to converge

Tweaking α can be seen as more of an art than a science and it will depend on your particular problem.

The best way to check that you are picking a good α is to graph (or otherwise output) your *cost* at each iteration and verify that it is indeed decreasing and the rate of decrease is not too slow.

Let me take that derivative for you...

Using basic Calculus you can easily verify that:

$$j = 0 \Rightarrow \frac{\partial}{\partial \theta_0} C(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$$

$$j = 1 \Rightarrow \frac{\partial}{\partial \theta_1} C(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

Back to our Example

So now we have a powerful tool for determining the θ parameters in your linear regression model.

Back to our Example

So now we have a powerful tool for determining the θ parameters in your linear regression model.

Let's go over the accompanying IJulia notebook for this lecture which illustrates linear regression applied to the education versus income data.

Homework 6

The Advertising data set (found in "Advertising.csv" on Canvas or GitHub) has data pertaining to the sales in thousands of units versus the advertising budget in thousands of dollars of different advertising media such as TV, radio and newspaper.

Homework 6

The Advertising data set (found in “Advertising.csv” on Canvas or GitHub) has data pertaining to the sales in thousands of units versus the advertising budget in thousands of dollars of different advertising media such as TV, radio and newspaper.

1. Using linear regression with least squares and gradient descent, as outlined in lecture 6, determine the line that best fits the relationship between sales and advertising budget for each medium.

Homework 6

The Advertising data set (found in “Advertising.csv” on Canvas or GitHub) has data pertaining to the sales in thousands of units versus the advertising budget in thousands of dollars of different advertising media such as TV, radio and newspaper.

1. Using linear regression with least squares and gradient descent, as outlined in lecture 6, determine the line that best fits the relationship between sales and advertising budget for each medium.
2. Plot your solution line against the data for each feature considered (that is, sales versus TV budget, radio budget and newspaper budget).

Homework 6

3. Plot the surface of your cost function for each feature considered.

Homework 6

3. Plot the surface of your cost function for each feature considered.
4. Compare your results. Is there a medium for advertising that you would claim to be best above all others? Which one and why? Write a short report of your findings using LaTeX. Make sure to include your formulas and figures obtained in the process of solving this problem.

Homework 6

3. Plot the surface of your cost function for each feature considered.
4. Compare your results. Is there a medium for advertising that you would claim to be best above all others? Which one and why? Write a short report of your findings using LaTeX. Make sure to include your formulas and figures obtained in the process of solving this problem.

Commit and push your .ipynb, .tex, and .pdf files for this homework into your GitHub private repository by 5/19 at 11:59 PM.