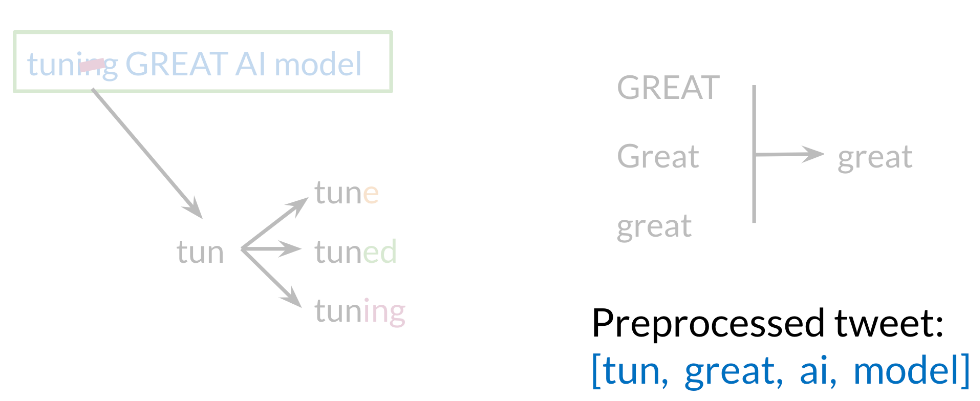
https://aman.ai/coursera-nlp/

Preprocessing

When preprocessing, you have to perform the following:

1. Eliminate handles and URLs
2. Tokenize the string into words.
3. Remove stop words like "and, is, a, on, etc."
4. Stemming- or convert every word to its stem. Like dancer, dancing, danced, becomes 'danc'. You can use porter stemmer to take care of this.
5. Convert all your words to lower case.

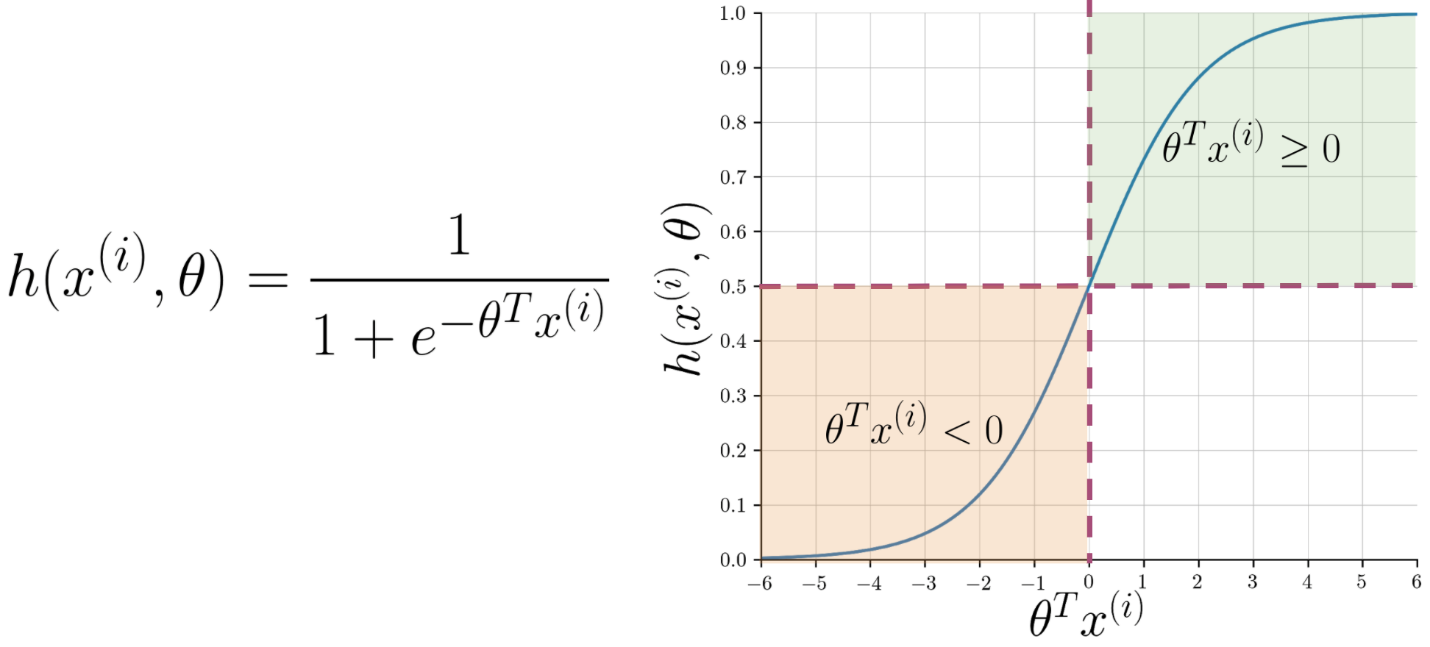
For example the following tweet "@YMourri and @AndrewYNg are tuning a GREAT AI model at https://deeplearning.ai!!!" after preprocessing become



[*tun*,*great*,*ai*,*model*]. Hence you can see how we eliminated handles, tokenized it into words, removed stop words, performed stemming, and converted everything to lower case.

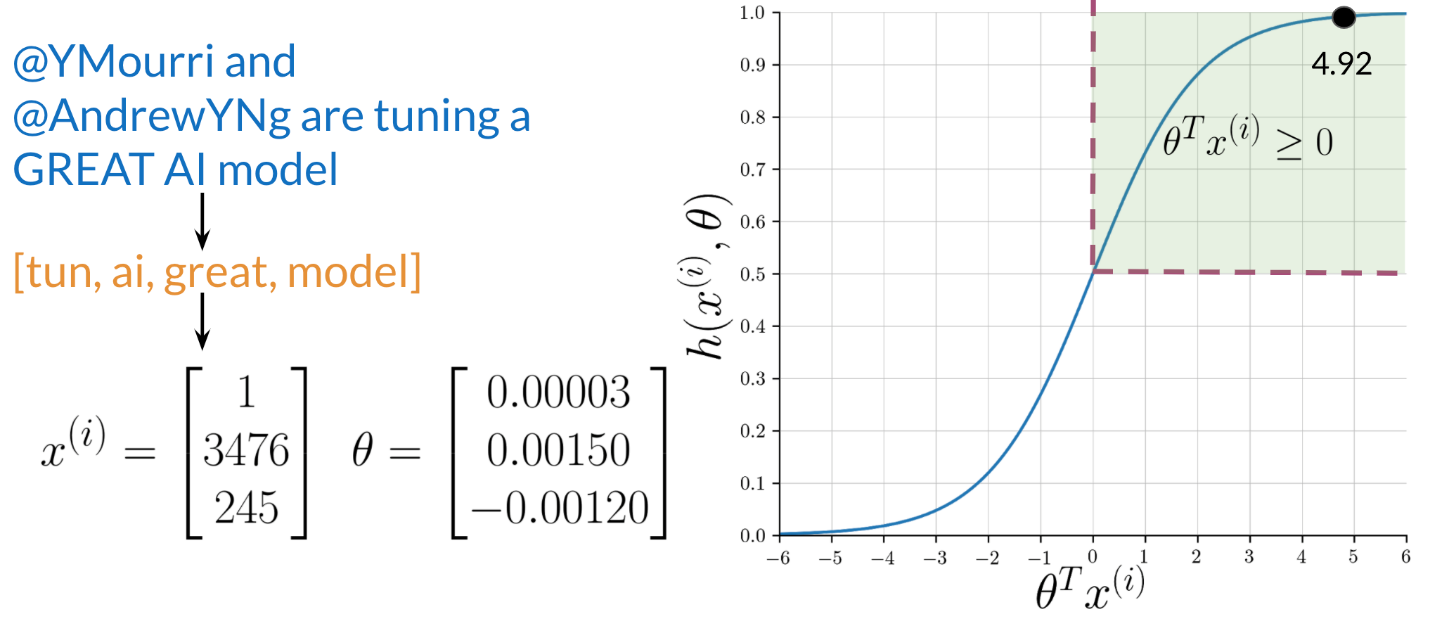
Logistic Regression Overview

Logistic regression makes use of the sigmoid function which outputs a probability between 0 and 1. The sigmoid function with some weight parameter \theta*θ* and some input x^{(i)}*x*(*i*) is defined as follows.



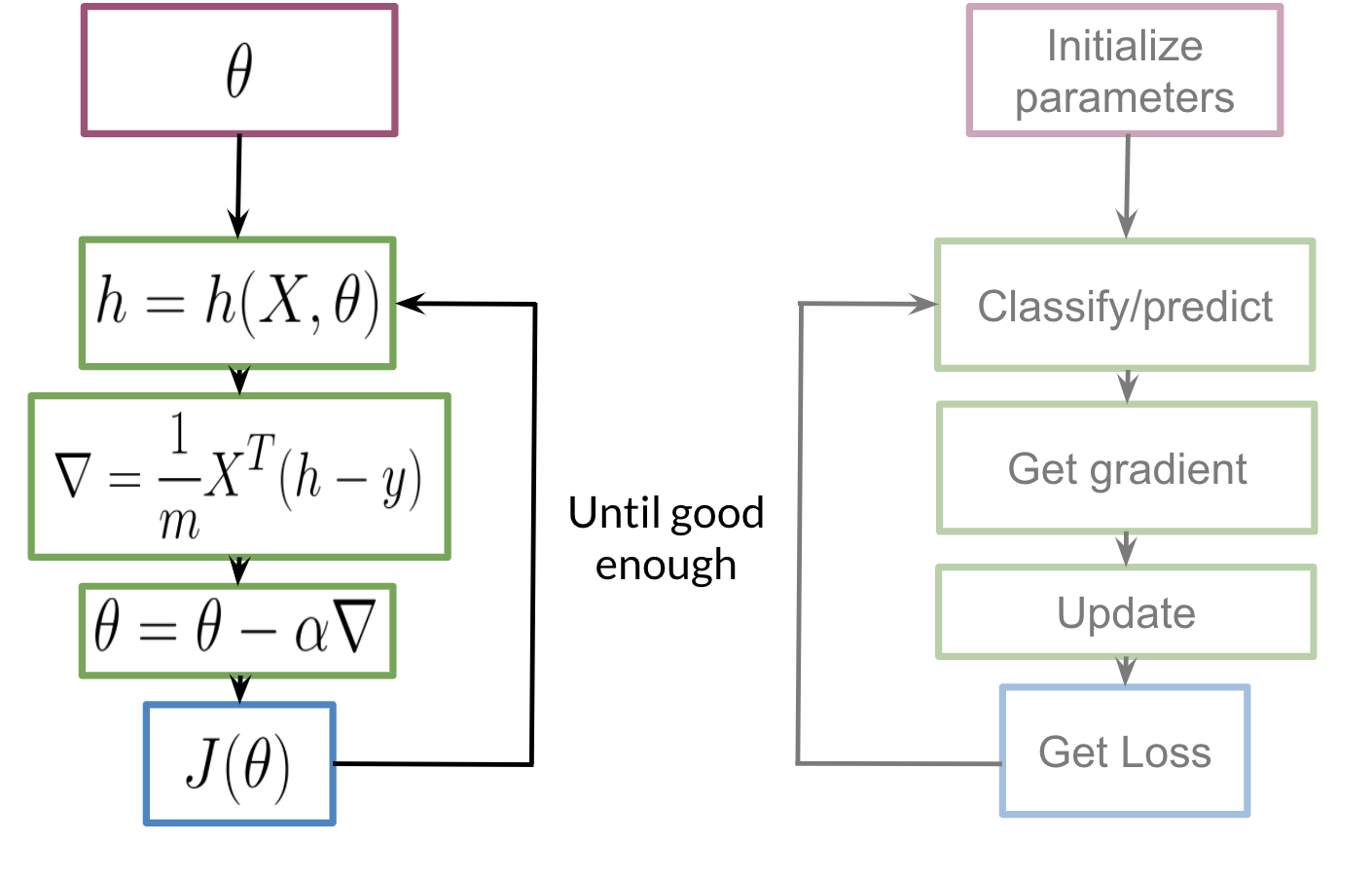
Note that as \theta^Tx^{(i)}*θTx*(*i*) gets closer and closer to -\infty−∞ the denominator of the sigmoid function gets larger and larger and as a result, the sigmoid gets closer to 00. On the other hand, as \theta^Tx^{(i)}*θTx*(*i*) gets closer and closer to \infty∞ the denominator of the sigmoid function gets closer to 1 and as a result the sigmoid also gets closer to 11.

Now given a tweet, you can transform it into a vector and run it through your sigmoid function to get a prediction as follows:



Logistic Regression: Training

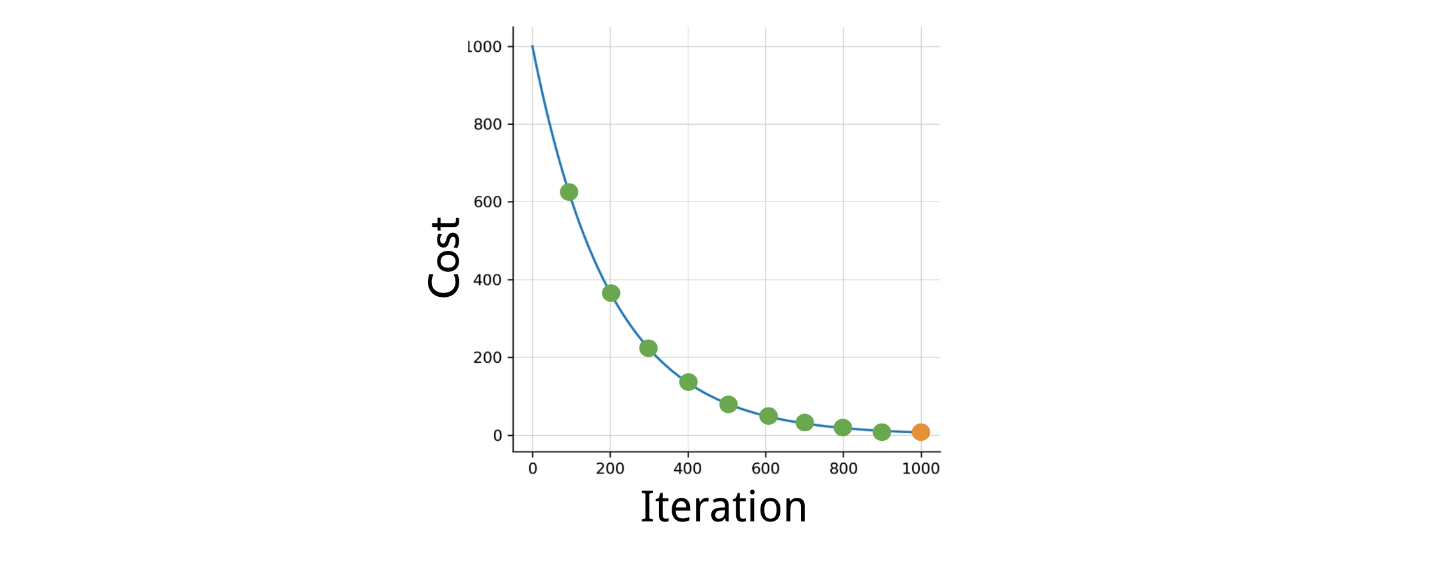
To train your logistic regression function, you will do the following:



You initialize your parameter \theta*θ*, that you can use in your sigmoid, you then compute the gradient that you will use to update \theta*θ*, and then calculate the cost. You keep doing so until good enough.

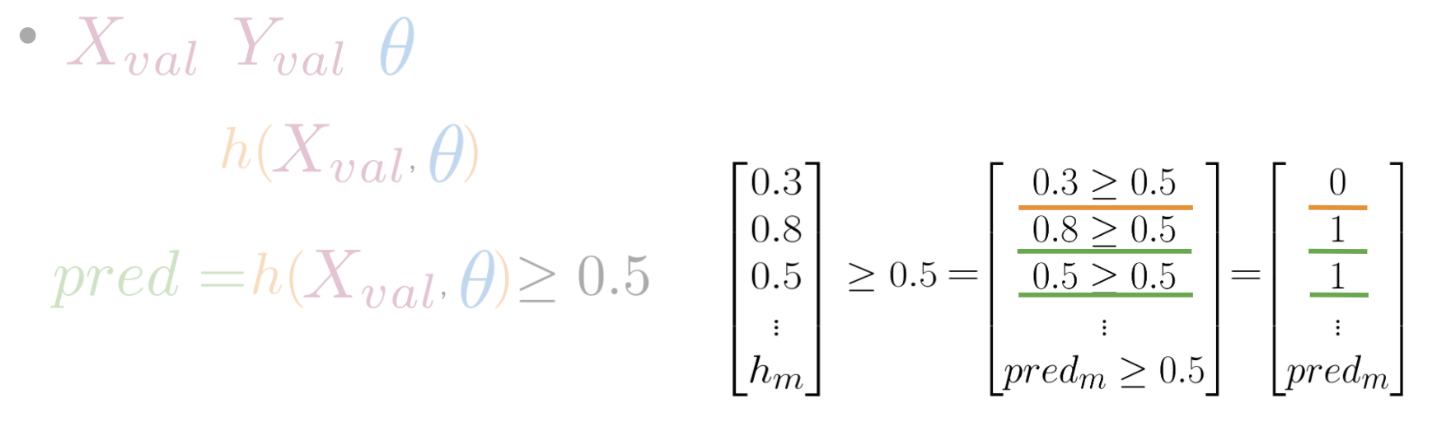
**Note:** If you do not know what a gradient is, don't worry about it. I will show you what it is at then end of this week in an optional reading. In a nutshell, the gradient allows you to learn what \theta*θ* is so that you can predict your tweet sentiment accurately.

Usually you keep training until the cost converges. If you were to plot the number of iterations versus the cost, you should see something like this:



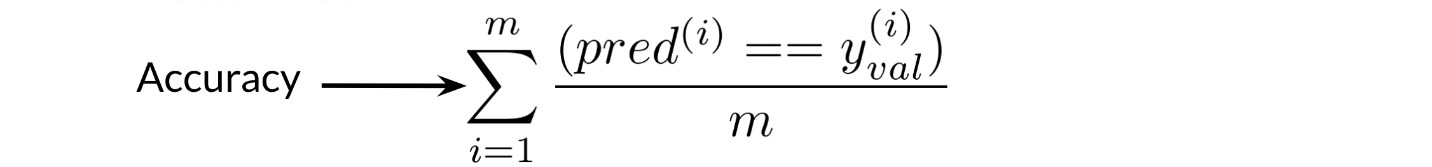
Logistic Regression: Testing

To test your model, you would run a subset of your data, known as the validation set, on your model to get predictions. The predictions are the outputs of the sigmoid function. If the output is \geq =0.5 ≥=0.5, you would assign it to a positive class. Otherwise, you would assign it to a negative class.



In the video, I briefly mentioned X*X* validation. In reality, given your X*X* data you would usually split it into three components. X\_{train}, X\_{val}, X\_{test}*Xtrain*​,*Xval*​,*Xtest*​. The distribution usually varies depending on the size of your data set. However, an 80,10,1080,10,10 split usually works fine.

To compute accuracy, you solve the following equation:



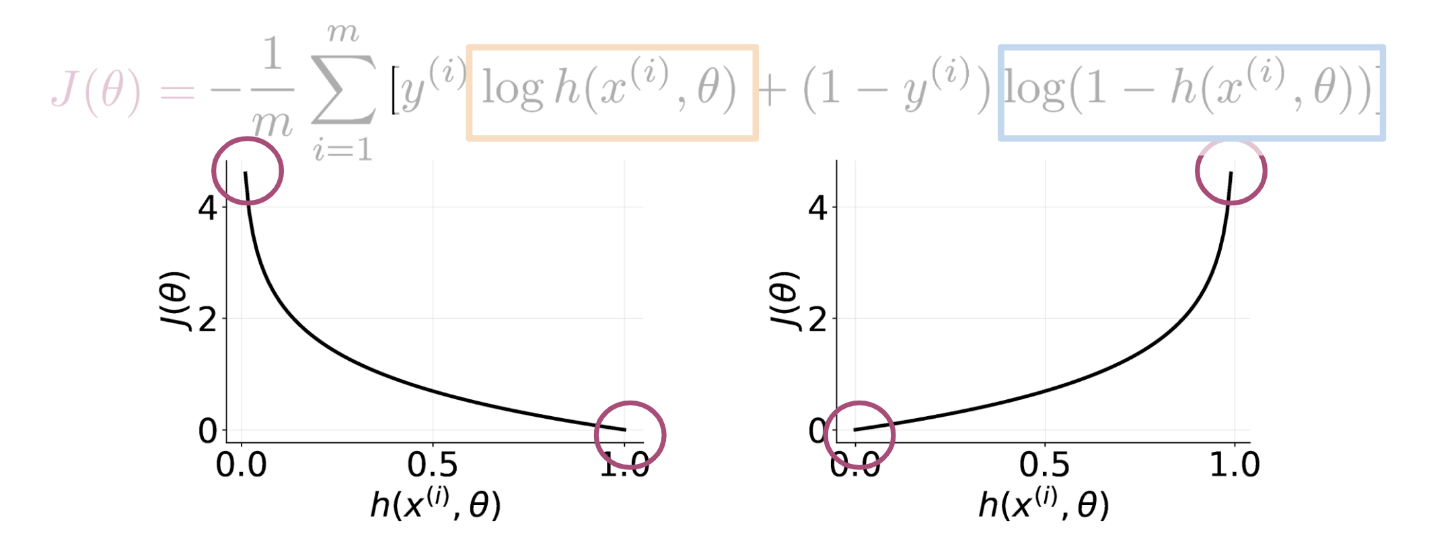
In other words, you go over all your training examples, m*m* of them, and then for every prediction, if it was wright you add a one. You then divide by m*m*.

# Optional Logistic Regression: Cost Function

This is an advanced optional reading where we delve into the details.. If you do not get the math, do not worry about it - you will be just fine by moving onto the next component. In this part, I will tell you about the intuition behind why the cost function is designed the way it is. I will then show you how to take the derivative of the logistic regression cost function to get the gradients.

The logistic regression cost function is defined as

J(\theta)=-\frac{1}{m} \sum\_{i=1}^{m}\left[y^{(i)} \log h\left(x^{(i)}, \theta\right)+\left(1-y^{(i)}\right) \log \left(1-h\left(x^{(i)}, \theta\right)\right)\right]*J*(*θ*)=−*m*1​∑*i*=1*m*​[*y*(*i*)log*h*(*x*(*i*),*θ*)+(1−*y*(*i*))log(1−*h*(*x*(*i*),*θ*))]



As you can see in the picture above, if y = 1*y*=1 and you predict something close to 00, you get a cost close to \infty∞. The same applies for then y=0*y*=0 and you predict something close to 11. On the other hand if you get a prediction equal to the label, you get a cost of 0. In either, case you are trying to minimize J(\theta)*J*(*θ*).

## Math Derivation

To show you why the cost function is designed that way, let us take a step back and write up a function that compresses the two cases into one case.

P(y | x^{(i)}, \theta ) = h(x^{(i)}, \theta)^{y^{(i)}} (1-h(x^{(i)}, \theta)^{(1-y^{(i)})}*P*(*y*∣*x*(*i*),*θ*)=*h*(*x*(*i*),*θ*)*y*(*i*)(1−*h*(*x*(*i*),*θ*)(1−*y*(*i*))

From the above, you can see that when y = 1*y*=1, you get h( x^{(i)}, \theta)*h*(*x*(*i*),*θ*), and when y = 0*y*=0, you get (1-h( x^{(i)}, \theta))(1−*h*(*x*(*i*),*θ*)), which makes sense, since the two probabilities equal to 1. When y=0*y*=0, you want (1-h( x^{(i)}, \theta))(1−*h*(*x*(*i*),*θ*)) to be close to 11, which means that h( x^{(i)})*h*(*x*(*i*)) needs to be close to 00, and therefore h( x^{(i)}, \theta)*h*(*x*(*i*),*θ*) close to 11. When y=1*y*=1, you want h( x^{(i)}, \theta)=1*h*(*x*(*i*),*θ*)=1.

Now we want to find a way to model the entire data set and not just one example. To do so, we will define the likelihood as follows:

L(\theta) = \prod\_{i=1}^m h(\theta, x^{(i)})^{y^{(i)}} (1-h(\theta, x^{(i)}))^{(1-y^{(i)})}*L*(*θ*)=∏*i*=1*m*​*h*(*θ*,*x*(*i*))*y*(*i*)(1−*h*(*θ*,*x*(*i*)))(1−*y*(*i*))

The \prod∏ symbol tells you that you are multiplying the terms together and not adding them. Note that if we mess up the classification of one example, we end up messing up the overall likelihood score, which is exactly what we intended. We want to fit a model to the entire dataset where all data points are related. One issue is that as m*m* gets larger, what happens to L(\theta)*L*(*θ*)? It goes close to zero, because both numbers h(x^{(i)}, \theta)*h*(*x*(*i*),*θ*) and (1-h(x^{(i)}, \theta))(1−*h*(*x*(*i*),*θ*)) are bounded between 00 and 11. Since we are trying to maximize h(\theta, x^{(i)})*h*(*θ*,*x*(*i*)) if y =1*y*=1, and minimize h(\theta, x^{(i)})*h*(*θ*,*x*(*i*)) if y =0*y*=0 in L(\theta)*L*(*θ*), we can introduce the log and just maximize the log of the function. (We are maximizing the same function just in a different space). Introducing the \loglog, allows us to write the \loglog of a product as the sum of each \loglog. Here are two identities that will come in handy:

\log a\*b\*c = \log a + \log b + \log c log*a*∗*b*∗*c*=log*a*+log*b*+log*c*

\log a^b = b \log a log*ab*=*b*log*a*

Given the two identities above, we can rewrite the equation as follows:

\max\_{h(x^{(i)}, \theta )} \log L(\theta) = \log \prod\_{i=1}^m h(x^{(i)}, \theta)^{y^{(i)}} (1-h(x^{(i)}, \theta)^{(1-y^{(i)})} max*h*(*x*(*i*),*θ*)​log*L*(*θ*)=log∏*i*=1*m*​*h*(*x*(*i*),*θ*)*y*(*i*)(1−*h*(*x*(*i*),*θ*)(1−*y*(*i*))

= \sum\_{i=1}^m \log h(x^{(i)}, \theta)^{y^{(i)}} (1-h(x^{(i)}, \theta)^{(1-y^{(i)})} =∑*i*=1*m*​log*h*(*x*(*i*),*θ*)*y*(*i*)(1−*h*(*x*(*i*),*θ*)(1−*y*(*i*))

= \sum\_{i=1}^m \log h(x^{(i)}, \theta)^{y^{(i)}} + \log (1-h(x^{(i)}, \theta)^{(1-y^{(i)})} =∑*i*=1*m*​log*h*(*x*(*i*),*θ*)*y*(*i*)+log(1−*h*(*x*(*i*),*θ*)(1−*y*(*i*))

= \sum\_{i=1}^m y^{(i)} \log h(x^{(i)}, \theta)+ (1-y^{(i)}) \log (1-h(x^{(i)}, \theta))=∑*i*=1*m*​*y*(*i*)log*h*(*x*(*i*),*θ*)+(1−*y*(*i*))log(1−*h*(*x*(*i*),*θ*))

Hence, we now divide by m*m*, because we want to see the average cost.

\frac {1}{m} \sum\_{i=1}^m y^{(i)} \log h(x^{(i)}, \theta) + (1-y^{(i)}) \log (1- h(x^{(i)}, \theta))*m*1​∑*i*=1*m*​*y*(*i*)log*h*(*x*(*i*),*θ*)+(1−*y*(*i*))log(1−*h*(*x*(*i*),*θ*))

Remember that we were maximizing h(\theta, x^{(i)})*h*(*θ*,*x*(*i*)) in the equation above. It turns out that maximizing an equation is the same as minimizing its negative. Think of x^2*x*2, feel free to plot it to see that for you yourself. Hence we add a negative sign and we end up minimizing the cost function as follows.

J(\theta)=-\frac{1}{m} \sum\_{i=1}^{m}\left[y^{(i)} \log h\left(x^{(i)}, \theta\right)+\left(1-y^{(i)}\right) \log \left(1-h\left(x^{(i)}, \theta\right)\right)\right]*J*(*θ*)=−*m*1​∑*i*=1*m*​[*y*(*i*)log*h*(*x*(*i*),*θ*)+(1−*y*(*i*))log(1−*h*(*x*(*i*),*θ*))]

A vectorized implementation is:

|  |
| --- |
| *h*=*g*(*Xθ*)  *J*(*θ*)=1*m*⋅(−*yT*log(*h*)−(1−*y*)*T*log(1−*h*)) |

# Optional Logistic Regression: Gradient

### This is an optional reading where I explain gradient descent in more detail. Remember, previously I gave you the gradient update step, but did not explicitly explain what is going on behind the scenes.

The general form of gradient descent is defined as:

|  |
| --- |
| *Repeat*{*θj*:=*θj*−*α*∂∂*θjJ*(*θ*)} |

For all j*j*. We can work out the derivative part using calculus to get:

|  |
| --- |
| *Repeat*{*θj*:=*θj*−*αm*∑*i*=1*m*(*h*(*x*(*i*),*θ*)−*y*(*i*))*x*(*i*)*j*} |

A vectorized implementation is:

\theta := \theta - \frac{\alpha}{m} X^{T} (H(X , \theta ) - Y)*θ*:=*θ*−*mα*​*XT*(*H*(*X*,*θ*)−*Y*)

### **Partial derivative of J(θ)**

|  |
| --- |
| *h*(*x*)′=(11+*e*−*x*)′=−(1+*e*−*x*)′  (1+*e*−*x*)2=−1′−(*e*−*x*)′  (1+*e*−*x*)2=0−(−*x*)′  (*e*−*x*)(1+*e*−*x*)2=−(−1)(*e*−*x*)(1+*e*−*x*)2=*e*−*x*(1+*e*−*x*)2=(11+*e*−*x*)(*e*−*x*1+*e*−*x*)=*h*(*x*)  (+1−1+*e*−*x*1+*e*−*x*)=*h*(*x*)(1+*e*−*x*1+*e*−*x*−11+*e*−*x*)=*h*(*x*)(1−*h*(*x*)) |

First calculate derivative of sigmoid function (it will be useful while finding partial derivative of J(θ)):

Note that we computed the partial derivative of the sigmoid function. If we were to derive h(x^{(i)}, \theta)*h*(*x*(*i*),*θ*) with respect to \theta\_j*θj*​, you would get h(x^{(i)}, \theta)(1-h(x^{(i)}, \theta))x^{(i)}\_j*h*(*x*(*i*),*θ*)(1−*h*(*x*(*i*),*θ*))*xj*(*i*)​. Note that we used the chain rule there, because we multiply by the derivative of \theta^Tx^{(i)} *θTx*(*i*) with respect to \theta\_j*θj*​. Now we are ready to find out resulting partial derivative:

|  |
| --- |
| ∂∂*θjJ*(*θ*)=∂∂*θj*−1*m*∑*i*=1*m*[*y*(*i*)*log*(*h*(*x*(*i*),*θ*))+(1−*y*(*i*))*log*(1−*h*(*x*(*i*),*θ*))]=−1*m*∑*i*=1*m*[*y*(*i*)∂∂*θjlog*(*h*(*x*(*i*),*θ*))+(1−*y*(*i*))∂∂*θjlog*(1−*h*(*x*(*i*),*θ*))]=−1*m*∑*i*=1*m*⎡⎣*y*(*i*)∂∂*θjh*(*x*(*i*),*θ*)*h*(*x*(*i*),*θ*)+(1−*y*(*i*))∂∂*θj*(1−*h*(*x*(*i*),*θ*))1−*h*(*x*(*i*),*θ*)⎤⎦=−1*m*∑*i*=1*m*⎡⎣*y*(*i*)∂∂*θjh*(*x*(*i*),*θ*)*h*(*x*(*i*),*θ*)+(1−*y*(*i*))∂∂*θj*(1−*h*(*x*(*i*),*θ*))1−*h*(*x*(*i*),*θ*)⎤⎦=−1*m*∑*i*=1*m*⎡⎣*y*(*i*)*h*(*x*(*i*),*θ*)(1−*h*(*x*(*i*),*θ*))∂∂*θjθTx*(*i*)*h*(*x*(*i*),*θ*)+−(1−*y*(*i*))*h*(*x*(*i*),*θ*)(1−*h*(*x*(*i*),*θ*))∂∂*θjθTx*(*i*)1−*h*(*x*(*i*),*θ*)⎤⎦=−1*m*∑*i*=1*m*⎡⎣*y*(*i*)*h*(*x*(*i*),*θ*)(1−*h*(*x*(*i*),*θ*))∂∂*θjθTx*(*i*)*h*(*x*(*i*),*θ*)−(1−*y*(*i*))*h*(*x*(*i*),*θ*)(1−*h*(*x*(*i*),*θ*))∂∂*θjθTx*(*i*)1−*h*(*x*(*i*),*θ*))⎤⎦=−1*m*∑*i*=1*m*[*y*(*i*)(1−*h*(*x*(*i*),*θ*))*x*(*i*)*j*−(1−*y*(*i*))*h*(*x*(*i*),*θ*)*x*(*i*)*j*]=−1*m*∑*i*=1*m*[*y*(*i*)(1−*h*(*x*(*i*),*θ*))−(1−*y*(*i*))*h*(*x*(*i*),*θ*)]*x*(*i*)*j*=−1*m*∑*i*=1*m*[*y*(*i*)−*y*(*i*)*h*(*x*(*i*),*θ*)−*h*(*x*(*i*),*θ*)+*y*(*i*)*h*(*x*(*i*),*θ*)]*x*(*i*)*j*=−1*m*∑*i*=1*m*[*y*(*i*)−*h*(*x*(*i*),*θ*)]*x*(*i*)*j*=1*m*∑*i*=1*m*[*h*(*x*(*i*),*θ*)−*y*(*i*)]*x*(*i*)*j* |

The vectorized version:

\nabla J(\theta) = \frac{1}{m} \cdot X^T \cdot \left(H\left(X, \theta\right) - Y\right)∇*J*(*θ*)=*m*1​⋅*XT*⋅(*H*(*X*,*θ*)−*Y*)

Congratulations, you now know the full derivation of logistic regression :) !