Cost Function (Logistic Regression Model)

An example of the supervised learning problem of fitting logistic regression model. We have a training set of m training examples and as usual, each of our examples is represented by a that's $n + 1$ dimensional. Given this training set, how do we choose, or how do we fit the parameter's theta?

Back when we were developing the linear regression model, we used the following cost function.

Linear regression: $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$

and instead of writing the squared error term, we can write:

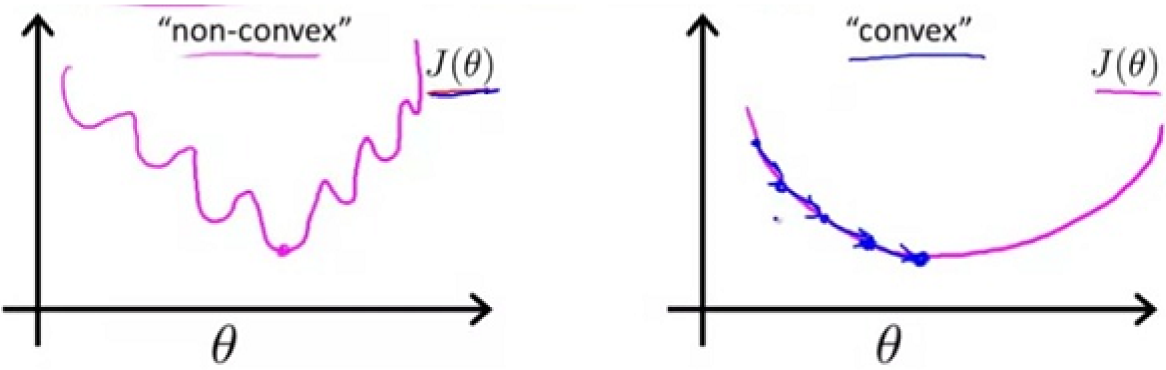
$J(\theta) = \text{cost}(h_\theta(x^{(i)}) - y^{(i)})$

and we define the cost function like:

$\text{cost}(h_\theta(x^{(i)}) - y^{(i)}) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$

The cost function worked fine for linear regression. But here, we're interested in logistic regression. If we could minimize the cost function, that will work okay.

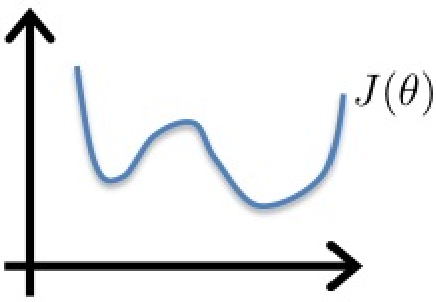
But it turns out that if we use this particular cost function for logistic regression, this would be a non-convex function of the parameter's data and if we were to run gradient descent on this sort of function It is not guaranteed to converge to the global minimum.

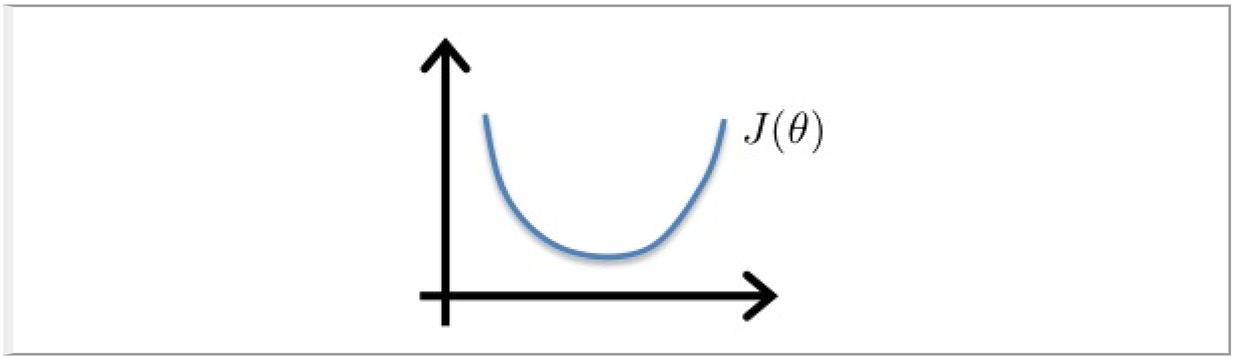


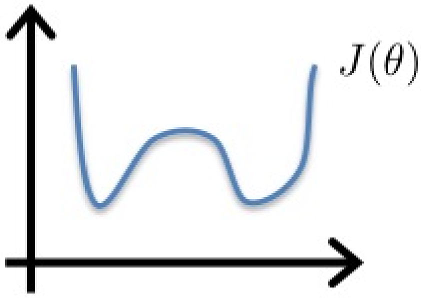
Whereas in contrast what we would like is to have a cost function $J(\theta)$ that is convex, that is a single bow-shaped function, so that if you run theta in the we would be guaranteed that would converge to the global minimum.

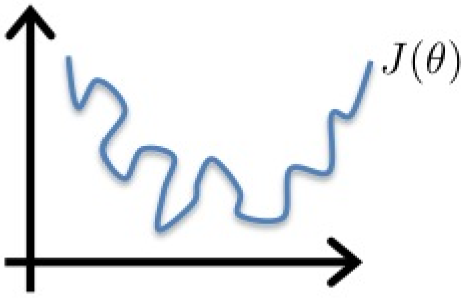
So what we'd like to do is, instead of come up with a different cost function, that is convex, and so that we can apply a great algorithm, like gradient descent and be guaranteed to find the global minimum.

Video Question: Consider minimizing a cost function $J(\theta)$. Which one of these functions is convex?

- 



- 

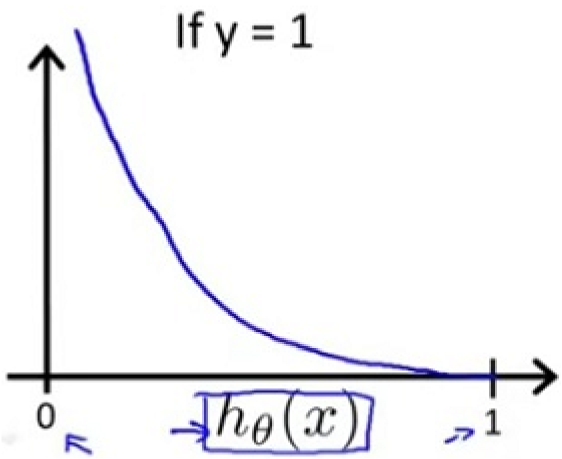
- 

Logistic Regression cost function

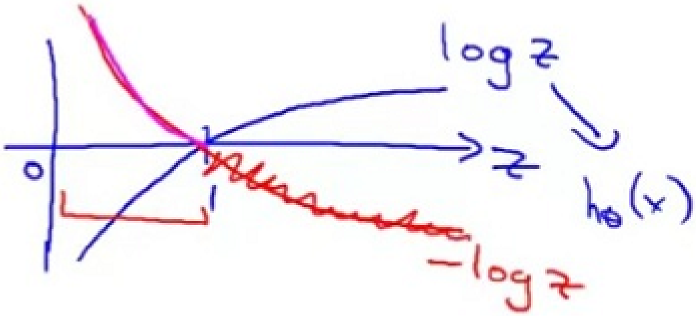
The cost function that we're going to use for logistic regression is:

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

If $y = 1$, then the cost function is $-\log(h_\theta(x))$, and we plot that, we find that it looks like this:



One way to see why the plot looks like this is because if we were to plot $\log z$ and $-\log z$ (z is playing the role of $h_\theta(x)$) will look like this:



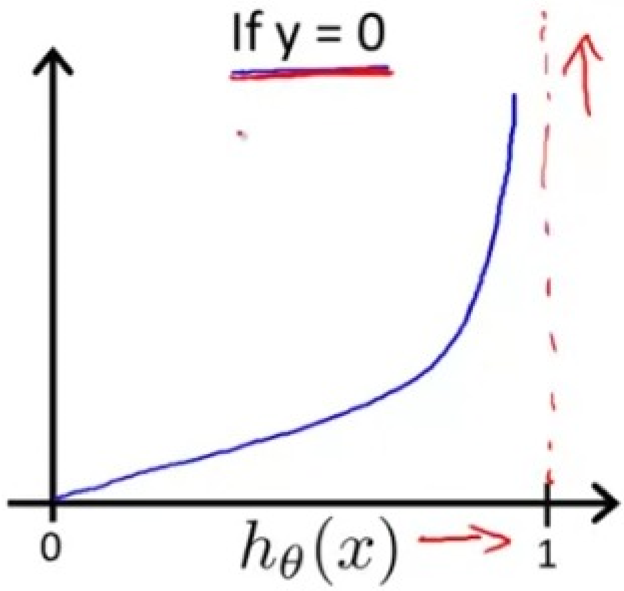
This cost function has a few interesting and desirable properties:

$$\text{Cost} = 0 \text{ if } y = 1, h_\theta(x) = 1$$
$$\text{But as} \quad h_\theta(x) \to 0$$
$$Cost \to \infty$$

Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1|x;\theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

If $h_\theta(x) = 1$, if that hypothesis predicts that $y = 1$ and if indeed $y = 1$ then the $cost = 0$. And that's where we'd like it to be because if we correctly predict the output y, then the cost is 0.

If we plot the function when If $y = 0$ (then the cost function is $-\log(1 - h_\theta(x))$) the plot look like this:



If $h_\theta(x) = 0$ and $y = 0$, then the hypothesis elted. The predicted y of z is equal to 0, and it turns out y is equal to 0, so at this point, the cost function is going to be 0.

Video Question: In logistic regression, the cost function for our hypothesis outputting (predicting) $h_\theta(x)$ on a training example that has label $y \in \{0, 1\}$ is:

$$\text{cost}(h_\theta(x), y) = \begin{cases} -\log h_\theta(x) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Which of the following are true? Check all that apply.

If $h_\theta(x) = y$, then $\text{cost}(h_\theta(x), y) = 0$ (for $y = 0$ and $y = 1$).

> if $y = 0$, then $\mathrm{cost}(h_\theta(x), y) \to \infty$ as $h_\theta(x) \to 1$.

- if $y = 0$, then $\mathrm{cost}(h_\theta(x), y) \to \infty$ as $h_\theta(x) \to 1$.

> Regardless of whether $y = 0$ or $y = 1$, if $h_\theta(x) = 0.5$, then $\mathrm{cost}(h_\theta(x), y) > 0$.

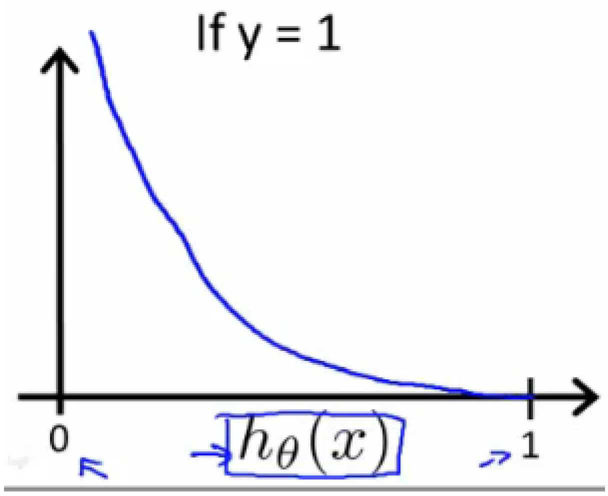
Summary

We cannot use the same cost function that we use for linear regression because the Logistic Function will cause the output to be wavy, causing many local optima. In other words, it will not be a convex function.

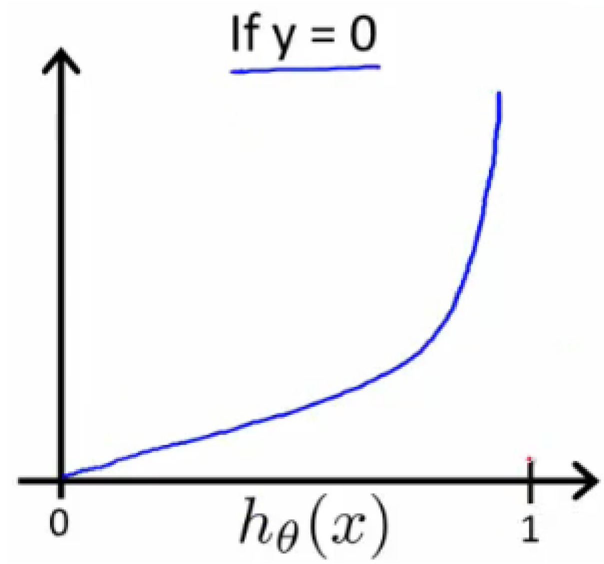
Instead, our cost function for logistic regression looks like:

$$
\begin{aligned}
&J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) \\
&\mathrm{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad && \text{if y = 1} \\
&\mathrm{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad && \text{if y = 0}
\end{aligned}
$$

When $y = 1$, we get the following plot for $J(\theta)$ vs $h_\theta(x)$:



Similarly, when $y = 1$, we get the following plot for $J(\theta)$ vs $h_\theta(x)$:



$$
\begin{aligned}
&\mathrm{Cost}(h_\theta(x), y) = 0 \text{ if } h_\theta(x) = y \\
&\mathrm{Cost}(h_\theta(x), y) \to \infty \text{ if } y = 0 \text{ and } h_\theta(x) \to 1 \\
&\mathrm{Cost}(h_\theta(x), y) \to \infty \text{ if } y = 1 \text{ and } h_\theta(x) \to 0
\end{aligned}
$$

If our correct answer 'y' is 0, then the cost function will be 0 if our hypothesis function also outputs 0. If our hypothesis approaches 1, then the cost function will approach infinity.

If our correct answer 'y' is 1, then the cost function will be 0 if our hypothesis function outputs 1. If our hypothesis approaches 0, then the cost function will approach infinity.

Note that writing the cost function in this way guarantees that $J(\theta)$ is convex for logistic regression.