Model and Cost Function

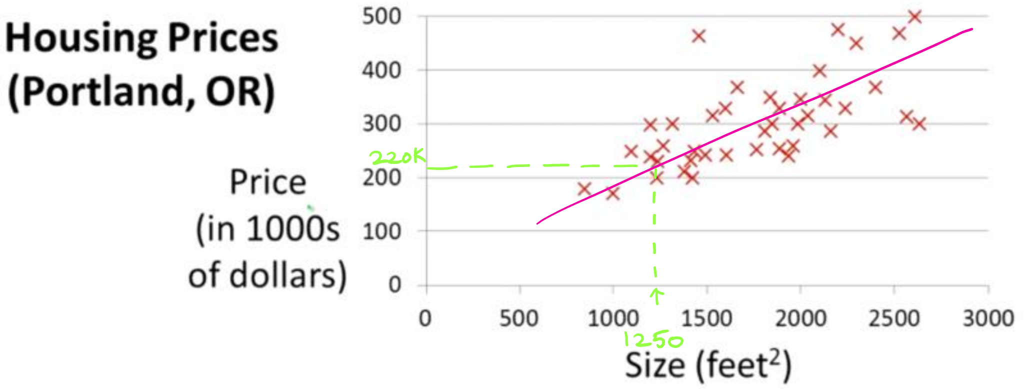
26 November 2020 01:24 PM

MODEL REPRESENTATION

→ Here we will go through an example of a Supervised Machine Learning problem of Regression Analysis

EXAMPLE

We are going to use the dataset for housing prices which we used earlier.



Q Using Machine Learning we have to tell what the price might be for a house which has a size of 1250 sq. ft.

↳ We could fit a straight line through the data and say that the price might be 220K

↳ This is an example of supervised machine learning as we have given the "right answers" — the actual prices — and that too a regression problem as the variable we are predicting (again, the price) is actually a continuous variable

→ More formally, the dataset that we have here can be called a Training set of housing prices

Training set of housing prices (Portland, OR)

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...
...	- - -

m rows

- Notation

m — Number of training examples

x — "input" variable / features

y — "output" / "target" variable

(x, y) — One training example representation

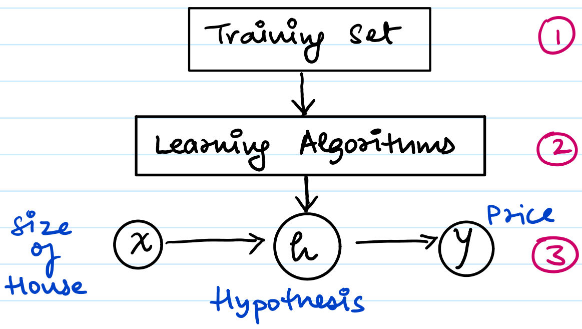
$(x^{(i)}, y^{(i)})$ — i^{th} training example → (i) is the row-indexer

↳ (x', y') ⇒ (2104, 460)

→ Here is how a supervised learning algorithm works:

Training Set

① We collect data to form our training set. That training set to fed to the learning



① We collect data to form our training set. That training set to fed to the learning algorithm.

② It is the job of the learning algorithm to then output a function called the hypothesis function

③ Hypothesis is a function that takes the features as input and then tries to output the estimated value of target variable

→ The next thing we need to decide when designing a learning algorithm is how do we represent this hypothesis h ?
↳ Let's start with something simple like this :

⟨We can represent $h_\theta(x)$ as $h(x)$⟩

LINEAR REGRESSION with one variable

$$\boxed{h_\theta(x) = \theta_0 + \theta_1 x}$$

→ Univariate Linear Regression

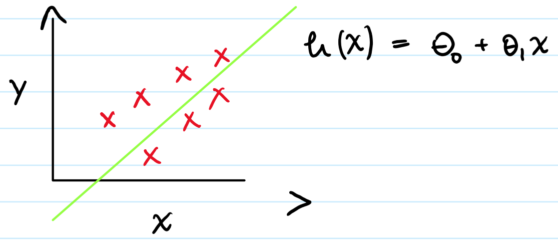
θ_i → Parameters

→ If you're familiar with Algebra, you'll recognize this equation instantly. This is nothing but the equation of a straight line → $y = mx + c$

→ What the hypothesis function is saying is that y is some straight line function of x that we are predicting.

Ⓠ Why only this Linear Function ?
→ This case is a simple building block, and the hypothesis function can be more complex as we go on

LINEAR REGRESSION



→ To define the supervised learning problem slightly more formally, our goal is, given a training set, to learn a function $h: X \to Y$ so that $h(x)$ is a good predictor for the corresponding value of y.

⊡ COST FUNCTION

→ Suppose we are working on the same Linear Regression problem

Size in feet^2 (x)	Price (\$) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...
...	...

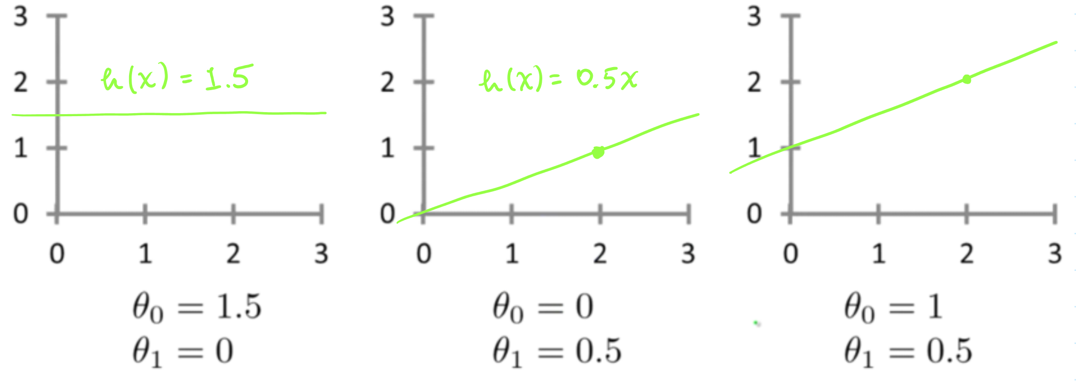
} m rows

θ_i — Parameters

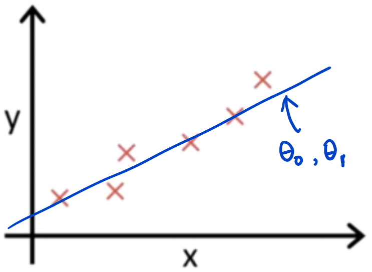
$$\boxed{\text{Hypothesis}: \quad h_\theta(x) = \theta_0 + \theta_1 x}$$

Ⓠ To get the final equation for $h_\theta(x)$ we must figure out the values for the parameters θ_0 and θ_1. So how do we do that?

parameters θ_0 and θ_1. So how do we do that ?



→ With different values for θ_0 and θ_1, we get different hypothesis functions



Now, we want to find the values for the parameters so that we are able to fit a straight line through the data

IDEA

Choose θ_0, θ_1 so that $h(x)$ is close to y for our training examples

→ Let's formalize this approach for Linear Regression :

$$\boxed{\underset{\theta_0, \theta_1}{\text{minimize}} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2}$$

Here we are trying to minimize the difference between the estimated target variable value and the actual target variable value and squaring it, for each training record, and summing them up to find the total amount of squared difference

To ease calculations later, we will now average out this quantity and take half of it.

↳

$$\boxed{\underset{\theta_0, \theta_1}{\text{minimize}} \; \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2}$$

This implies that we are finding the values θ_0 and θ_1 which are causing the hypothesis to be minimized

$h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)}$

→ Notation-wise, this is called the Cost Function and is written like this :

COST FUNCTION

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

OBJECTIVE

$$\underset{\theta_0, \theta_1}{\text{minimize}} \; J(\theta_0, \theta_1)$$

This particular cost function is the most widely used cost function for Regression problems

OBJECTIVE minimize $J(\theta_0, \theta_1)$
θ_0, θ_1

for Regression problems

→ This cost function is also called the Squared Error Function

⊡ COST FUNCTION : INTUITION

→ So far we have built up the following equations :

Hypothesis	$h_\theta(x) = \theta_0 + \theta_1 x$
Parameters	θ_0 , θ_1
Cost Function	$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$
Goal	$\underset{\theta_0, \theta_1}{\text{minimize}} \; J(\theta_0, \theta_1)$

→ To get a better understanding of the cost Function let's take a step back and start again, but with a simplified cost function.
↳ Let's start with just 1 parameter

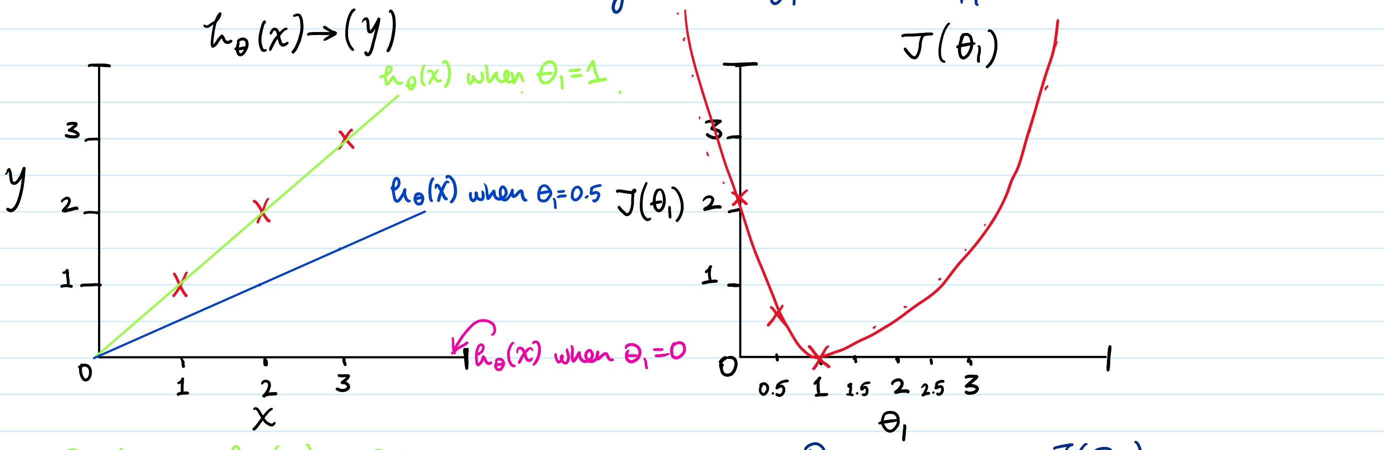
$$\boxed{h_\theta(x) = \theta_1 x}$$

→ Cost function remains the same but the hypothesis contains only 1 variable

COST FUNCTION
$$\boxed{J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(\underbrace{h_\theta\left(x^{(i)}\right)}_{\theta_1 x^{(i)}} - y^{(i)} \right)^2}$$

OBJECTIVE
$$\underset{\theta_1}{\text{minimize}} \; J(\theta_1)$$

→ Let's now look at how the value of the Hypothesis affects the Cost Function



When $\theta_1 = 1$, $h_\theta(x) = \theta_1 x$
$\Rightarrow x$
$\Rightarrow y = x$

$$\Rightarrow J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$
$= $ ZERO

θ_1	$J(\theta_1)$
1	0
0.5	0.58
0	2.33

When $\theta_1 = 0.5$, $h_\theta(x) = \theta_1 x$
$\Rightarrow \frac{x}{2}$
$\Rightarrow y = \underline{x}$

We can keep on computing these J values and plot them on the $J(\theta)$ graph above, and eventually

$$\Rightarrow \frac{x}{2}$$

$$\Rightarrow y = \frac{x}{2}$$

$$\Rightarrow J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$= \frac{1}{2m} \times \left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right]$$

$$= \frac{1}{2\times3} \times \frac{7}{2} = \frac{7}{12} \approx 0.58$$

When $\theta_1 = 0$, $h_\theta(x) = 0$

$$\Rightarrow y = 0$$

$$\Rightarrow J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$= \frac{1}{2\times3} \times \left[1^2 + 2^2 + 3^2\right]$$

$$= \frac{7}{3} \approx 2.33$$

these J values and plot them on the $J(\theta)$ graph above, and eventually what we will get is the parabolic line somewhat similar to the one shown above

→ Each value of θ_1 gives a different value of $h_\theta(x)$

→ Each value of $h_\theta(x)$ gives a different value of $J(\theta)$

→ Now, remember that the optimization objective of the learning algorithm is to choose a value of θ_1 such that it minimizes the value of $J(\theta_1)$.

↳ In simpler terms, we want to minimize our COST

→ Looking at the graph for $J(\theta_1)$, we find that the value that does so is $\boxed{\theta_1 = 1}$

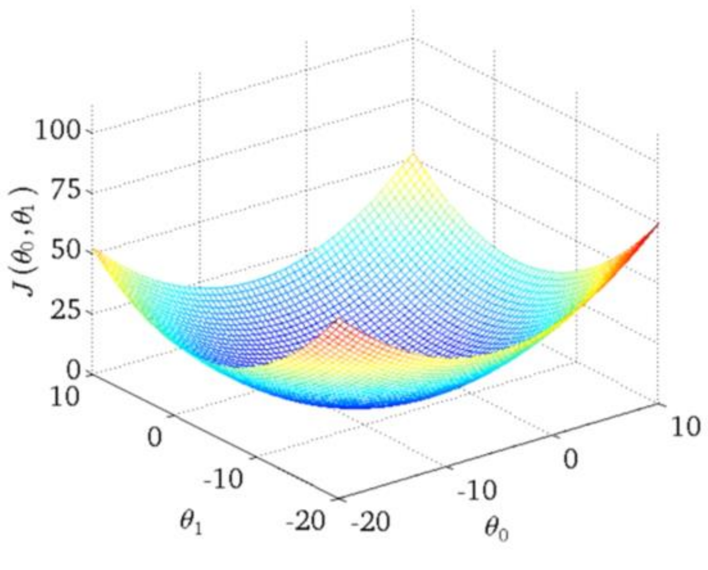
→ Also note that the hypothesis for $\theta_1 = 1$ is also the best possible fit to our data as it covers all the points perfectly

⊡ COST FUNCTION : INTUITION WITH MORE PARAMETERS

→ As you would expect, when we have more than one parameter, the situation gets a little more complex

→ The visualization also gets complex as now we need to have more planar axis to plot the cost function

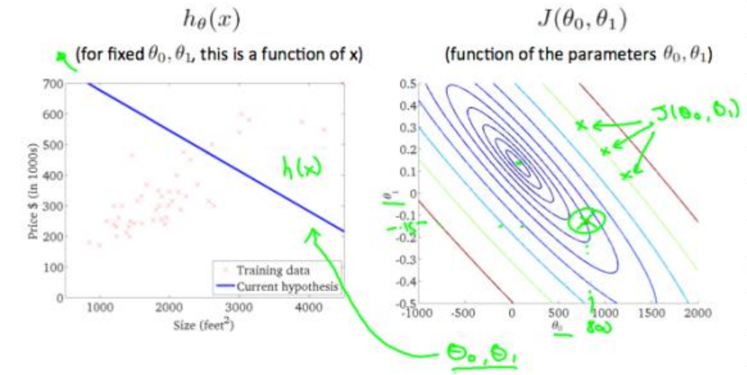
↳ For instance, a cost function plot with 2 parameters θ_0 and θ_1 will look something like this :



→ As we can see, $J(\theta_0, \theta_1)$ can be found out by considering θ_0 and θ_1 as bases and $J(\theta_0, \theta_1)$ to be the height

→ There is another representation as well to plot the cost function → CONTOUR PLOTS

→ A contour plot is a graph that contains many contour lines - A contour line of a two variable function has a constant value at all points of the same line

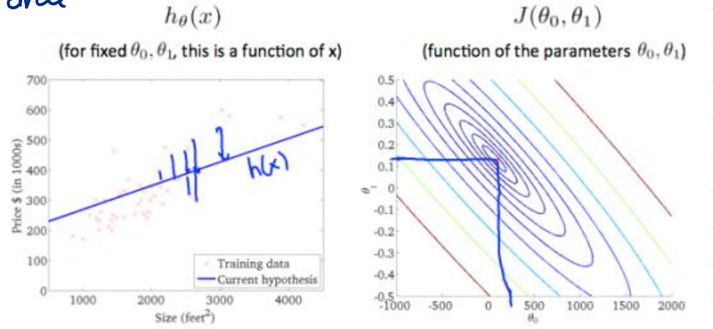


→ Taking any color and going along the 'circle' one would expect to get the same value of the cost function

→ The three green points found on the green circle (above, right) have the same value for $J(\theta_0, \theta_1)$ <and as a result they are found on the same line>

→ The graph below minimizes the cost function as much as possible and consequently the result of θ_0 and θ_1 tend to be around 0.12 and 250 respectively.

→ Plotting those values on the graph to the right seems to put our point in the center of the 'inner most' circle


$h_\theta(x)$
(for fixed θ_0, θ_1, this is a function of x)
$J(\theta_0, \theta_1)$
(function of the parameters θ_0, θ_1)
Price $ (in 1000s)
Size (feet^2)
Training data
Current hypothesis
h(x)
θ_1
θ_0