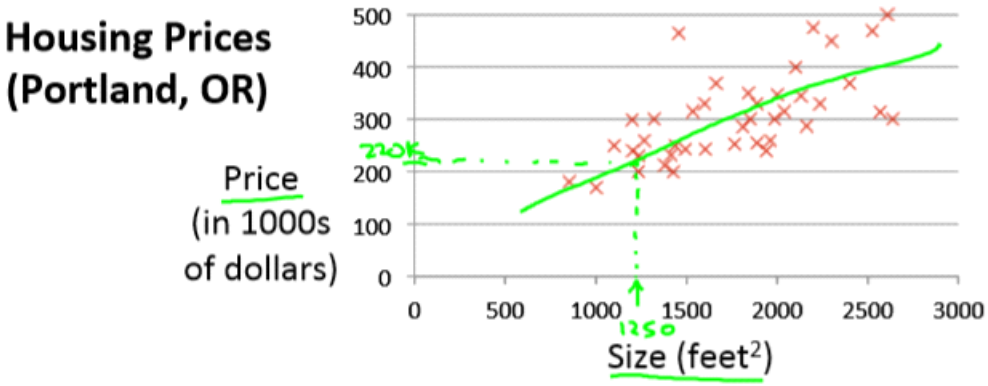
Linear Regression

2018년 5월 21일 월요일 오후 7:12

https://blog.naver.com/joyfull1
Jaegul Joo - 고려대 교수

집의 크기와 가격의 상관 관계, 단순히 2가지 요소로
여태 쌓인 데이터를 가지고.



Supervised Learning

Given the "right answer" for each example in the data.

Regression Problem

Predict real-valued output

Classification: Discrete-valued output

Regression : Y or N
이번에는 그냥 실수 값 기준으로.

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 = 47

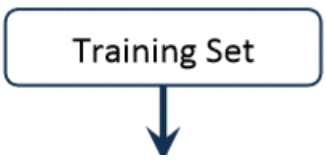
Notation:

- m = Number of training examples
- x's = "input" variable / features
- y's = "output" variable / "target" variable

$x^{(1)} = 2104$
$x^{(2)} = 1416$
$y^{(1)} = 460$

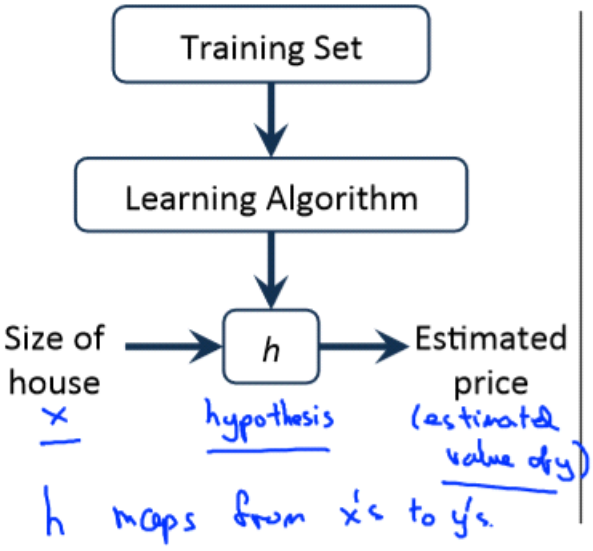
(x, y) - one training example
$(x^{(i)}, y^{(i)})$ - i^{th} training example

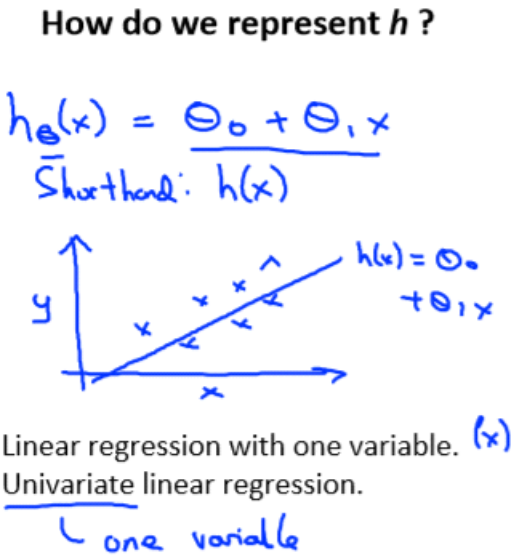
Feature x, y



How do we represent *h* ?

$h_\theta(x) = \theta_0 + \theta_1 x$

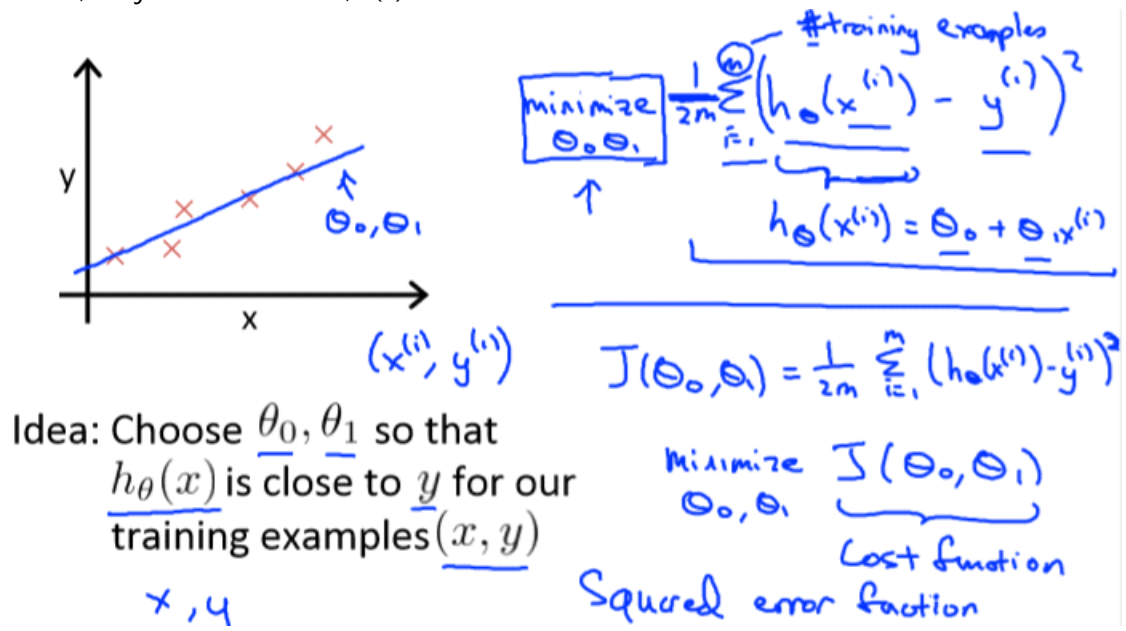




Linear model 이라고 가정

Cost function

=Loss , Objective Function, J(x)



Train data 로 도출된 그래프와

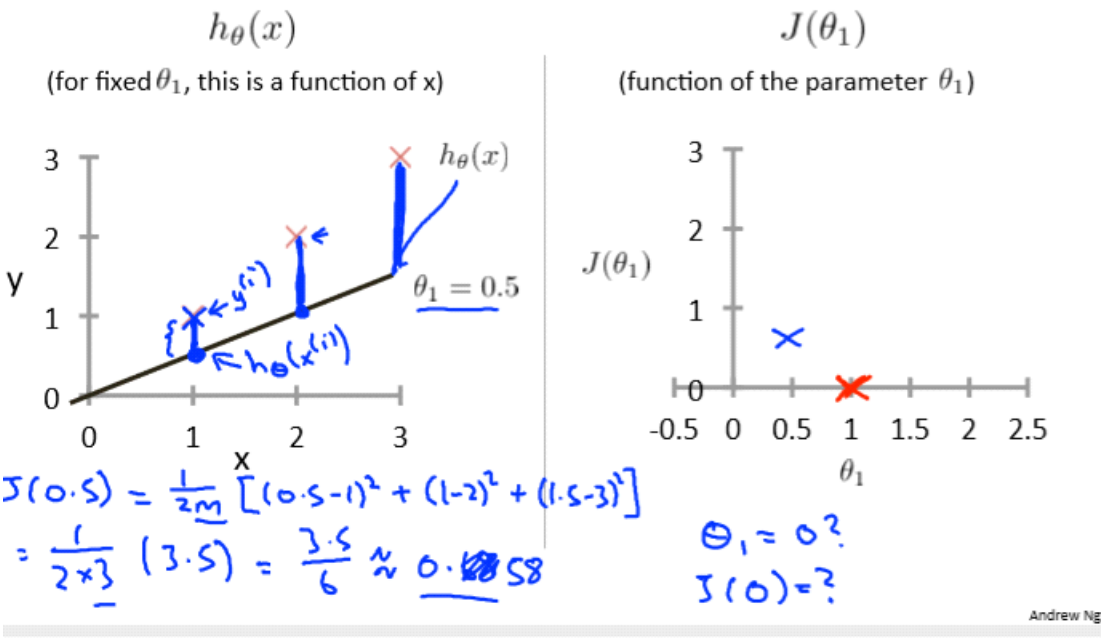
Test data 를 통하여 cost function 식을 통해 나온 값이 최소가 되도록 변수를 조정한다.

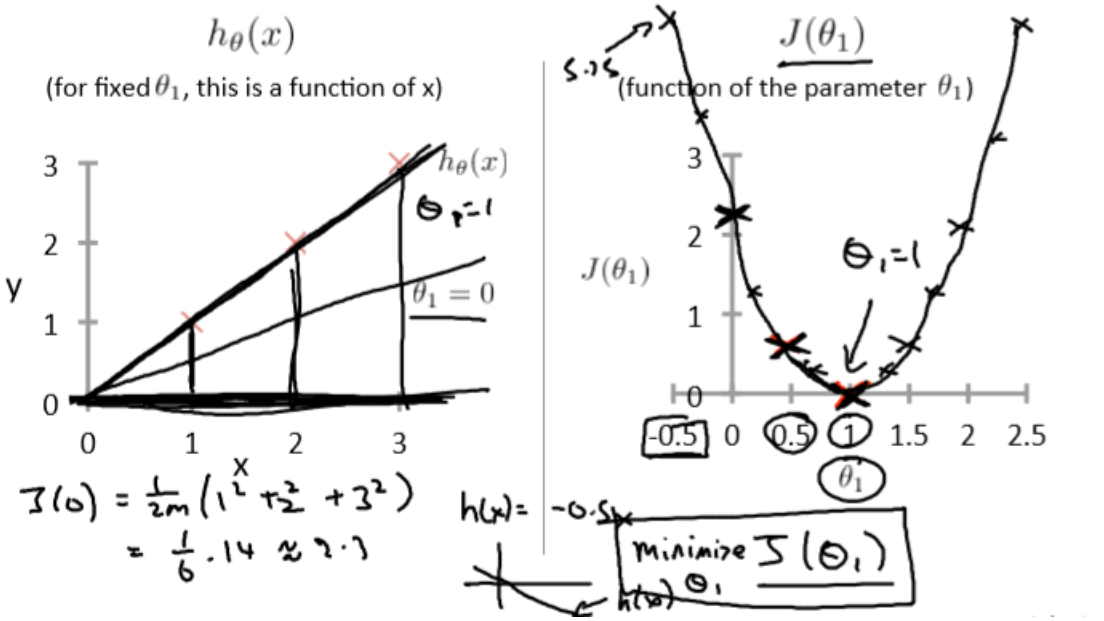
(통계학)절대값 대신 제곱을 이용하는 이유는 값이 차이가 큰 상태의 가중치를 많이 주기 위해서

식에 대한 구간의 범위를 나눌 필요도 없고 계산이 편리해서 제곱을 사용한다.

M은 평균을 구하기 위해서 2는 미분 할 때 발생하는 값을 제거하기위해.

(2차식 미분 ->2*)





Y=x 를 따르는 데이터라 할 때 결과 값, 그래프

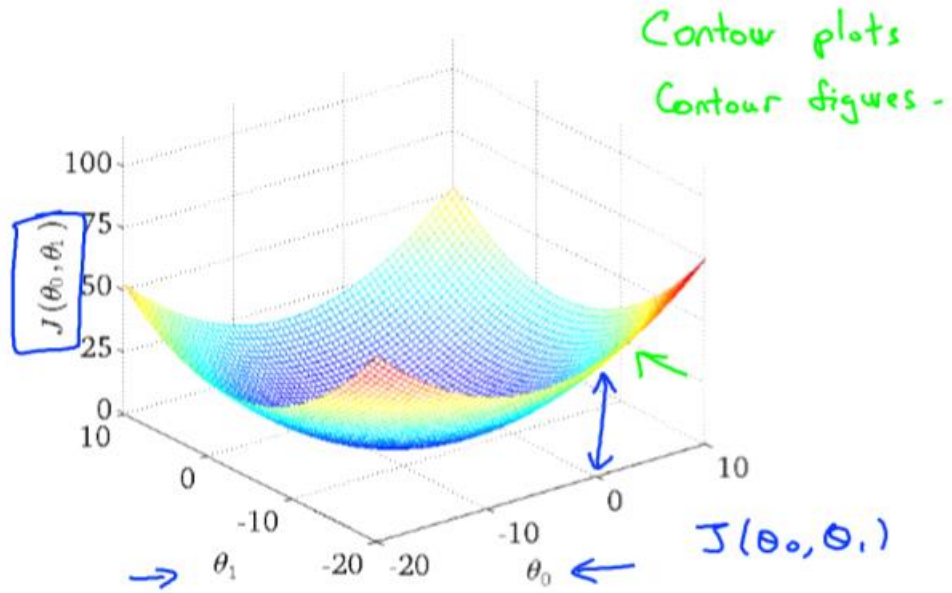
Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

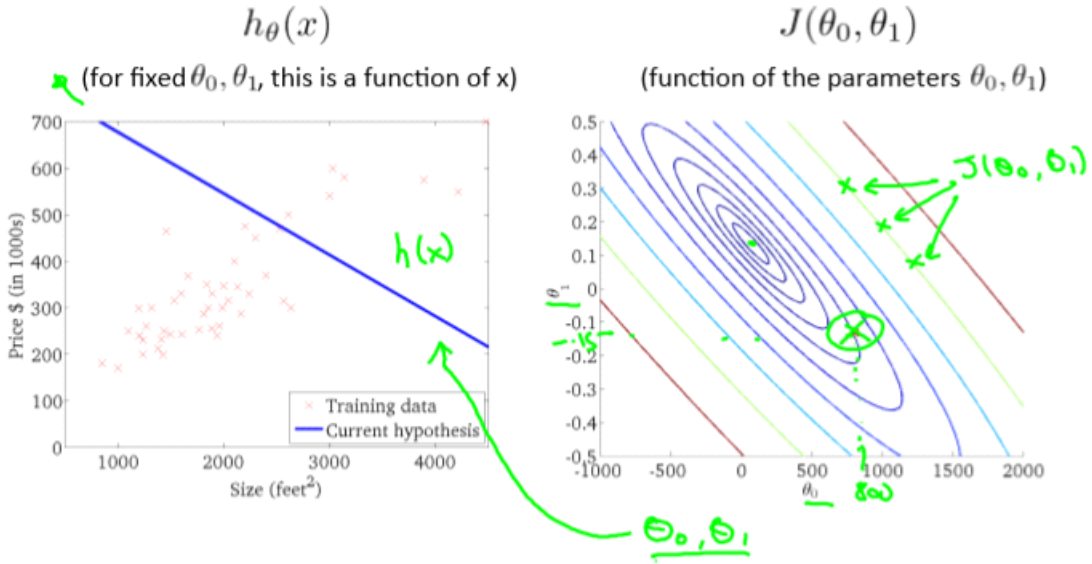
Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

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

Bias 추가, 입력 ->vector(조합 a,b)



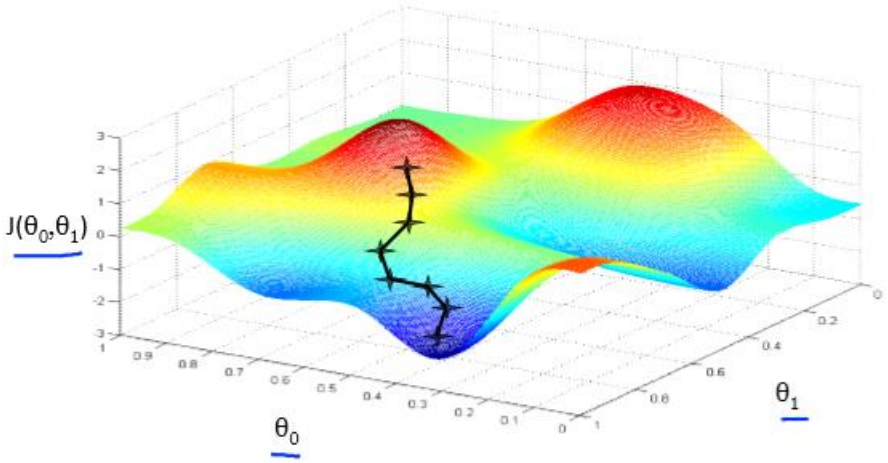
A, B 조합의 에러 Z 축에 찍어서 나타난 모양.



등고선 모형으로 위에서 값을 바라 봤을 때 높이를 표현.

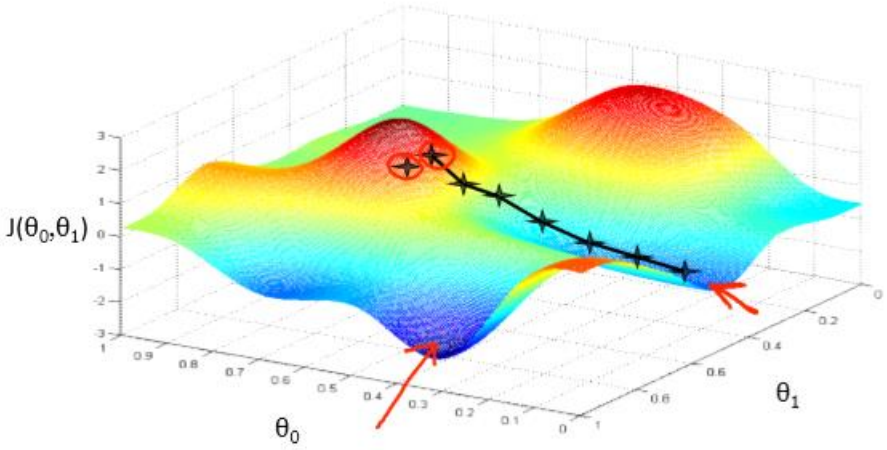
Linear regression with one variable

Gradient descent



가장 낮은 값에 도달 하는 방법.

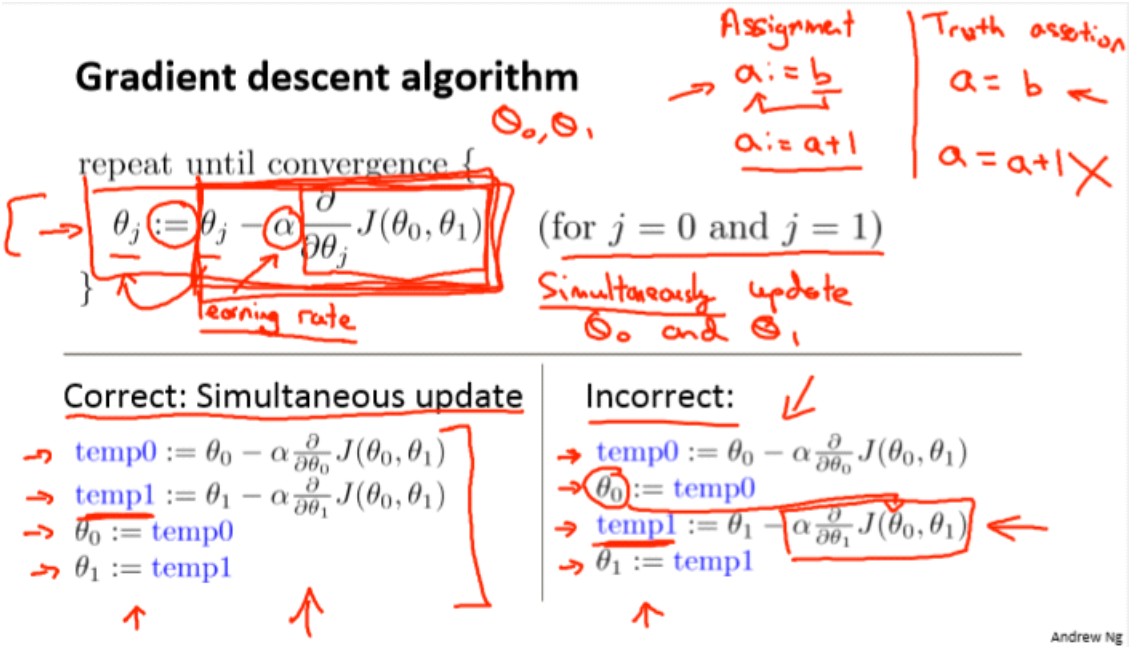
-> 초기값에서 부터 주변 미분 값 중에 가장 큰 지점을 찾아 간다.



초기값이 다른 점에서 시작하면 다른 값에 수렴 할 수도 있다.

but 다른 값이 더 낮은 rate를 가지고 있다. -> Local minimal 문제 발생

근본적인 문제. (그렇게 큰 문제는 아니다.)



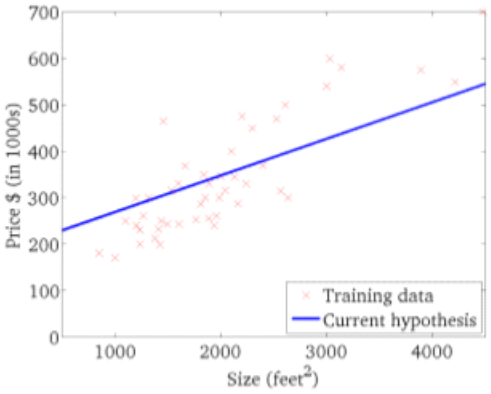
편미분을 통해서 x,y 축 방향 각각의 방향을 정한다.

기울기 값이 큼에 따라 이동하는 방향으로 얼마나 갈지 정해진다.

기울기가 가파를 수록 많이 간다.(극점이 멀다), 완만할 수록 (극점이 가깝다).
기울기 음수 -> + 방향, 양수 -> -방향 , so '-'를 붙여 준다.
Learning rate (알파) 값이 너무 낮으면 연산 수가 많아진다.

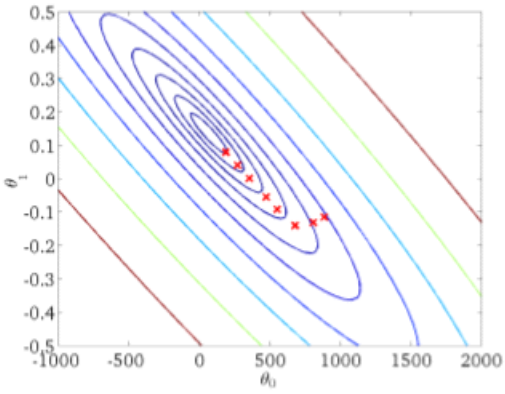
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

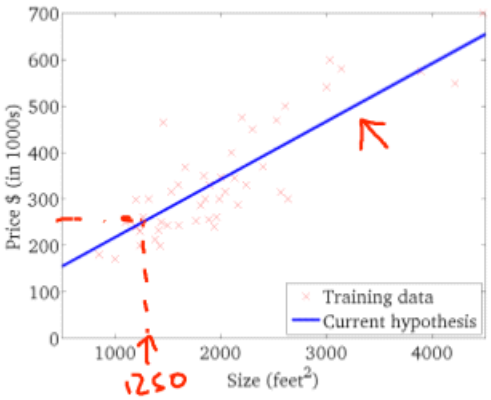
(function of the parameters θ_0, θ_1)



Andrew

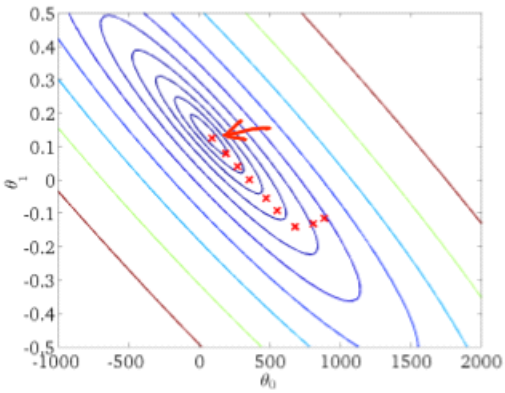
$h_\theta(x)$

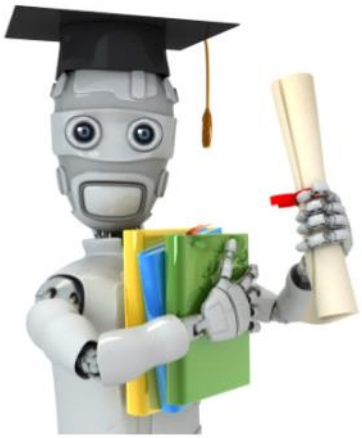
(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)





Linear Regression with multiple variables

Multiple features

Machine Learning

Multiple features (variables).

Size (feet^2) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

$m = 47$

$x^{(2)} = \begin{bmatrix} 1416 \\ 3 \\ 2 \\ 40 \end{bmatrix}$

$x_3^{(2)} = 2$

Notation:

- n = number of features ($n = 4$)
- $x^{(i)}$ = input (features) of i^{th} training example.
- $x_j^{(i)}$ = value of feature j in i^{th} training example.

Hypothesis:

Previously: $h_\theta(x) = \theta_0 + \theta_1 x$

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$

e.g. $h_\theta(x) = 80 + 0.1 x_1 + 0.01 x_2 + 3 x_3 - 2 x_4$ (x_4: age)

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

For convenience of notation, define $x_0 = 1$. ($x_0^{(i)} = 1$)

$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1}$ $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1}$

$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n$ ($x_0 = 1$)

$= \theta^T x$

$[\theta_0\ \theta_1 \cdots \theta_n]$ (θ^T, $(n+1) \times 1$ matrix) $\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$ (x)

$\theta^T x$

Multivariate linear regression.

Col Vector 로 정의하고 필요할때 Transpose 한다.

Gradient descent for multiple variables

Y = aX1 +bX2+c;

F(a,b,c) = 1/2m ((a+2b+c-3)^2 + (2a+5b+c)^2)

df/da = 1/2m (2(a+2b+c-3)*1 +..... (편미분 하는 법)

X1	X2	Y
1	2	3
2	5	4

예측치와 실측치와의 차이가 미분 값에 들어간다.

미분하면 (미분 상수)*(예측치-실측치)*(편미분 인자의 값), 들의 합의 평균으로 구하게 된다.
C로 미분한 값은 (예측치-실측치)들의 합이 양수면 C를 낮추게 된다, 음수면 C의 값을 증가시킨다.
(편미분 인자의 값)은 배수 값에 해당 된다.

내적 계산을 통해서 한방에 값을 구할 수 있다.

Hypothesis: $h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

Parameters: $\theta_0, \theta_1, \ldots, \theta_n$

Cost function:

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

Gradient descent:

Repeat {

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ (for $j = 0$ and $j = 1$)

}

Gradient Descent

Previously (n=1):

Repeat {

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$$

(simultaneously update θ_0, θ_1)

}

New algorithm ($n \geq 1$):

Repeat {

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update θ_j for $j = 0, \ldots, n$)

}

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_1^{(i)}$$

}

$$\theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_2^{(i)}$$

상황에 따라서 위와 같이 여러개의 미분 값을 구해야 할 수도 있다.

Linear Regression with multiple variables

Gradient descent in practice I: Feature Scaling

Examples: $m = 4$.

x_0	Size (feet^2) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
1	2104	5	1	45	460
1	1416	3	2	40	232
1	1534	3	2	30	315
1	852	2	1	36	178

$$X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix} \quad y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

m x (n+1)

m-dimensional vector

$$\theta = (X^T X)^{-1} X^T y$$

75	152
93	185
90	180
100	196
70	142

Hypothesis using matrix

$$w_1x_1 + w_2x_2 + w_3x_3 + ... + w_nx_n$$

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} x_{11}w_1 + x_{12}w_2 + x_{13}w_3 \\ x_{21}w_1 + x_{22}w_2 + x_{23}w_3 \\ x_{31}w_1 + x_{32}w_2 + x_{33}w_3 \\ x_{41}w_1 + x_{42}w_2 + x_{43}w_3 \\ x_{51}w_1 + x_{52}w_2 + x_{53}w_3 \end{pmatrix}$$

$$H(X) = XW$$

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} x_{11}w_1 + x_{12}w_2 + x_{13}w_3 \\ x_{21}w_1 + x_{22}w_2 + x_{23}w_3 \\ x_{31}w_1 + x_{32}w_2 + x_{33}w_3 \\ x_{41}w_1 + x_{42}w_2 + x_{43}w_3 \\ x_{51}w_1 + x_{52}w_2 + x_{53}w_3 \end{pmatrix}$$

[5, 3] [3, 1] [5, 1]

$$H(X) = XW$$