1. Intro

1.1 What is machine learning?

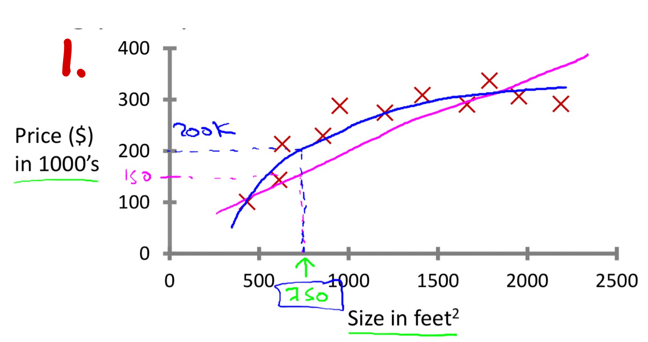
- Field of study that gives computers the ability to learn without being explicitly programmed.

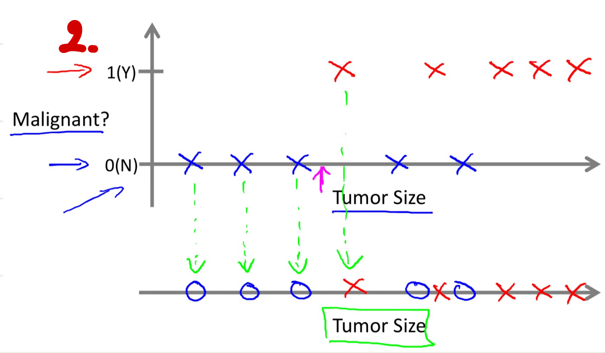
- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

1.2. Supervised Learning

: "right answers" given

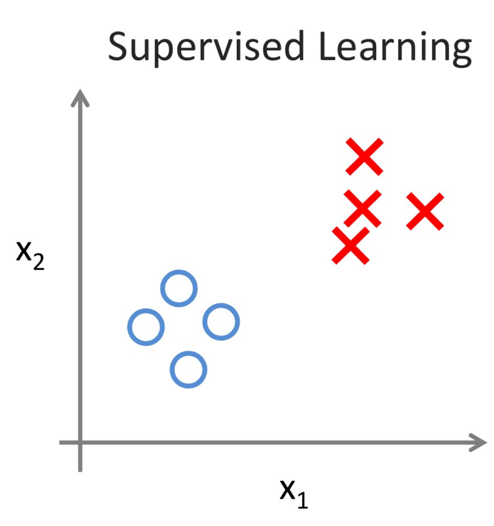
1. regression : continuous valued output ex) price
2. classification : discrete valued output ex) 0 or 1

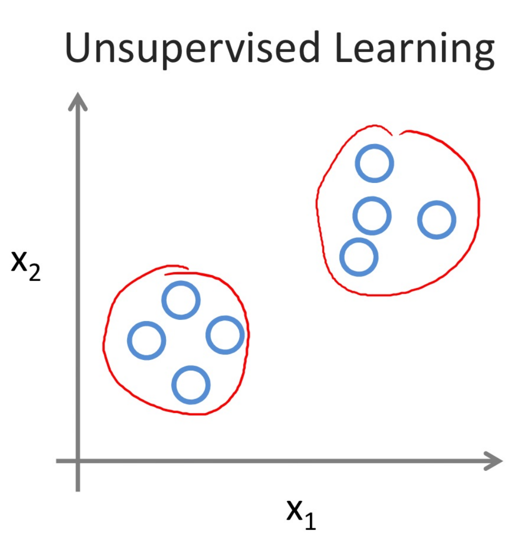




1.3 Unsupervised Learning

little or no idea what our results should look like.
ex) clustering





2. Linear regression

2.1 Cost function

choose θ_0, θ_1 so that $h_\theta(x)$ is close to y for our training examples (x, y)

$$\boxed{\underset{\theta_0 \theta_1}{\text{minimize}} \; \frac{1}{2m} \sum_{i=1}^{m} \big(\underbrace{h_\theta(x^{(i)})}_{= (\theta_0 + \theta_1 x^{(i)})} - y^{(i)}\big)^2}$$

m : # training examples

$\underset{\theta_0 \theta_1}{\text{minimize}} \; J(\theta_0, \theta_1)$: Cost function (squared error fn)

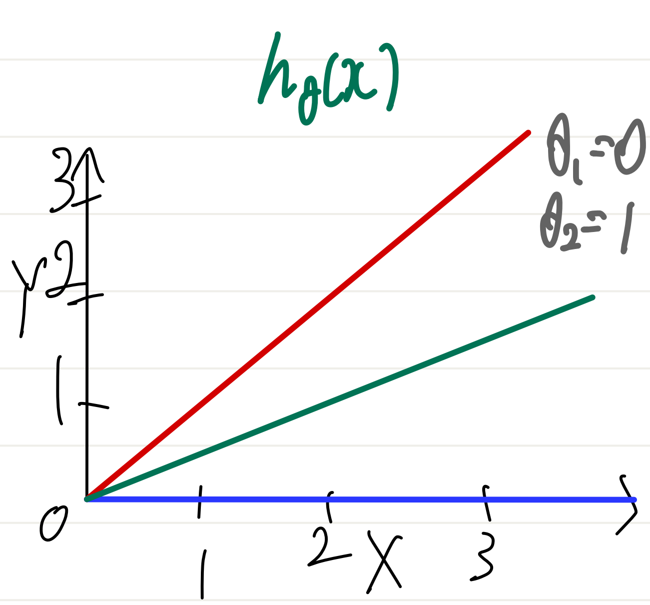
2.2 Cost function institution

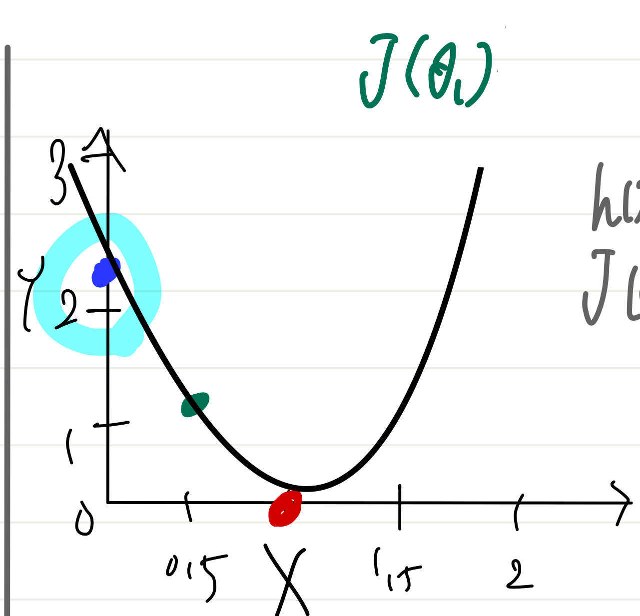
Hypothesis : $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters : θ_0, θ_1

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

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





$h(x) = x$

$J(0) = \frac{1}{2m}(1^2 + 2^2 + 3^2)$

$= \frac{1}{2 \cdot 3} \cdot 14 \approx 2.3$

2. Linear regression

2.3 Gradient descent

repeat until convergence

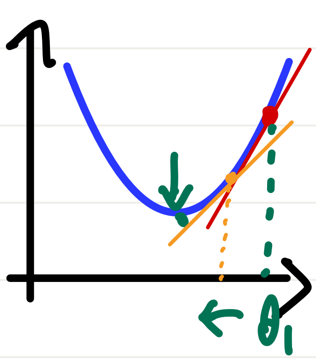
$$\boxed{\theta_j := \theta_j - \underbrace{\alpha}_{\text{learning rate}} \frac{d}{d\theta_j} J(\theta_0, \theta_1)}$$

(for $j=0$ and $j=1$)

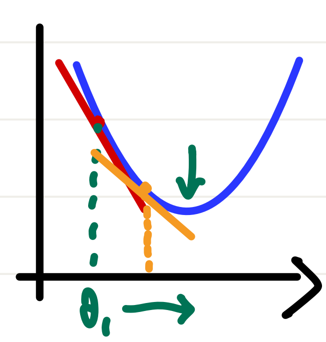
$temp0 := \theta_0 - \alpha \frac{d}{d\theta_0} J(\theta_0, \theta_1)$
$temp1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta_0, \theta_1)$
$\theta_0 := temp0$
$\theta_1 := temp1$

simultaneously update θ_0 & θ_1

2.4 Gradient descent intuition

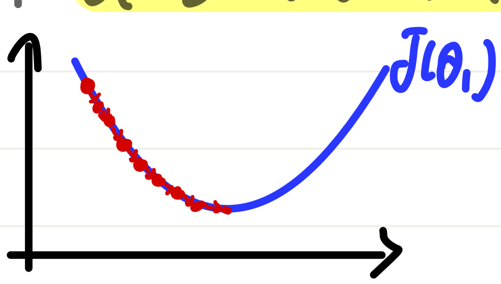


$\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta_1)$
≥ 0 (positive number)

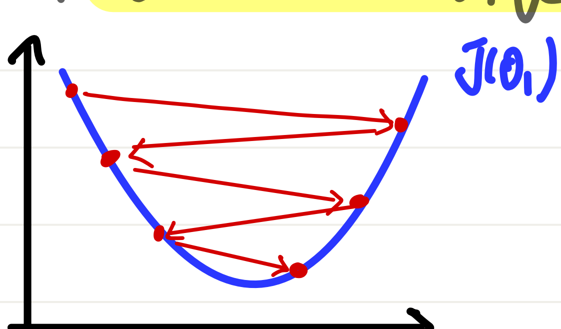


$\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} (J(\theta_1))$
≤ 0 (negative number)

1) If α is too small → gradient descent can be slow



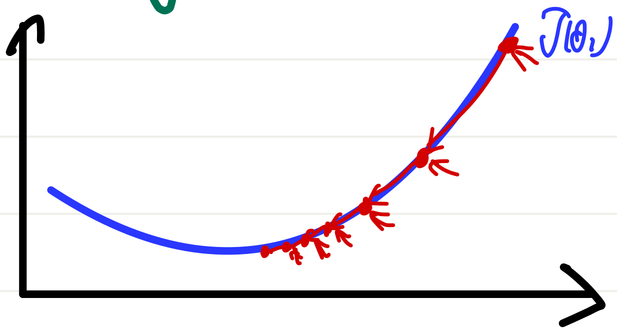
2) If α is too large → gradient descent can overshoot the minimum. (fail to converge or even diverge)



2. Linear regression

- gradient descent $=0$ (minimum)
  $\Rightarrow$ 움직이지 X

- learning rate α fixed $\Rightarrow$ gradient descent can converge to a local minimum
  $\Rightarrow$ No need to decrease α
  (automatically take smaller steps)



2.5 Gradient descent for linear regression

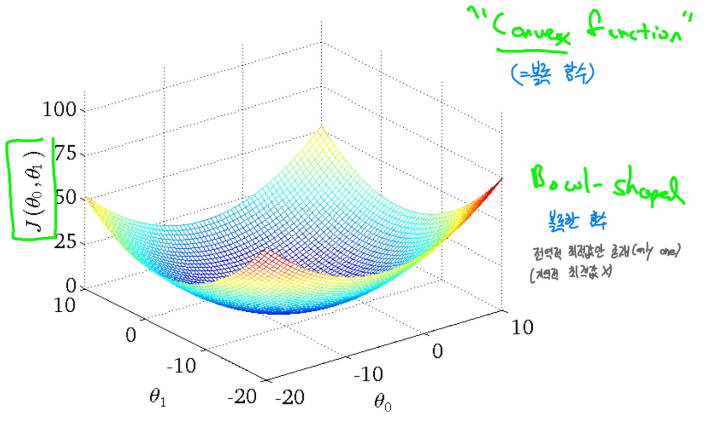
$$j=0: \frac{\partial}{\partial\theta_0} J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

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

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

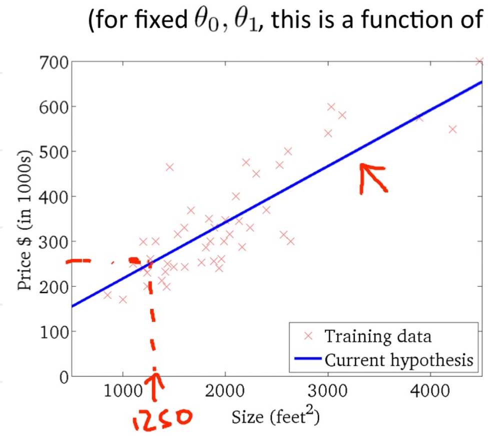
$$\theta_1 := \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)\cdot x^{(i)}$$

update θ_0 & θ_1 simultaneously



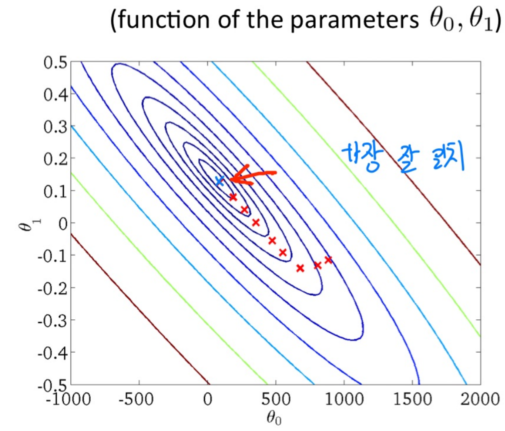
$h_\theta(x)$

(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)



"Batch" Gradient Descent

집단 : each step of gradient descent uses all the training examples.

3. Linear Algebra review

3.1 Matrices and vectors

Matrix Elements

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

row

column

$A_{11} = 1$

$A_{32} = 6$

3×2 $\mathbb{R}^{3 \times 2}$

Vector : n×1 matrix

$$y = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$y_1 = 1$

$y_2 = 2$

$y_3 = 3$

$\mathbb{R}^3$ 3-dimensional vector

3.2 Addition and multiplication

ex) $\begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix} + \begin{bmatrix} 4 & 0.5 \\ 2 & 5 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0.5 \\ 4 & 10 \\ 3 & 2 \end{bmatrix}$ *같은 차원의 행렬끼리만 계산 가능

3×2 3×2 3×2

ex) $3 \times \begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix} \times 3 = \begin{bmatrix} 3 & 0 \\ 6 & 15 \\ 9 & 3 \end{bmatrix}$

순서 바뀌어도 상관 X

ex) $\begin{bmatrix} 1 & 2 & 1 & 5 \\ 0 & 3 & 0 & 4 \\ -1 & -2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 14 \\ 13 \\ -7 \end{bmatrix}$

$1 \times 1 + 2 \times 3 + 1 \times 2 + 5 \times 1 = 14$

$0 \times 1 + 3 \times 3 + 0 \times 2 + 4 \times 1 = 13$

$-1 \times 1 + (-2) \times 3 + 0 \times 2 + 0 \times 1 = -7$

ex) $\begin{bmatrix} 1 & 3 & 2 \\ 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 0 & 1 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 11 & 10 \\ 9 & 14 \end{bmatrix}$

$\begin{bmatrix} 1 \times 1 & 3 \times 0 & 2 \times 5 \\ 4 \times 1 & 0 \times 0 & 1 \times 5 \end{bmatrix} = \begin{bmatrix} 11 \\ 9 \end{bmatrix}$

$\begin{bmatrix} 1 \times 3 & 3 \times 1 & 2 \times 2 \\ 4 \times 3 & 0 \times 1 & 1 \times 2 \end{bmatrix} = \begin{bmatrix} 10 \\ 14 \end{bmatrix}$

2×3 3×2 2×2

같아야 함

3. Linear Algebra review

- 교환법칙 (commutative) 성립 X $\Rightarrow$ $A \times B \neq B \times A$
- 결합법칙 (Associative) 성립 O $\Rightarrow$ $A \times B \times C \rightarrow \left.\begin{matrix} A \times (B \times C) \\ (A \times B) \times C \end{matrix}\right)$ same!
- Identity Matrix $= I \, (I_{n \times n}) \Rightarrow A \cdot I = I \cdot A = A$

$$\begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}$$

3.3 Inverse transpose

- $A(A^{-1}) = A^{-1} \cdot A = I$
  └ inverse

A is an $m \times m$ matrix, if it has an inverse.
If A don't have an inverse $\Rightarrow$ singular
or degenerate

- $A = \begin{bmatrix} 1 & 2 & 0 \\ 3 & 5 & 9 \end{bmatrix}$ 2×3

$A^T = \begin{bmatrix} 1 & 3 \\ 2 & 5 \\ 0 & 9 \end{bmatrix}$ 3×2