Day2 : Introduction to Bayesian Regression and MCMC

[FastCampus] AI센터 베이지안 통계과정
강사: 전인수 (isjeon@vision.snu.ac.kr)
MAY 29, 2019

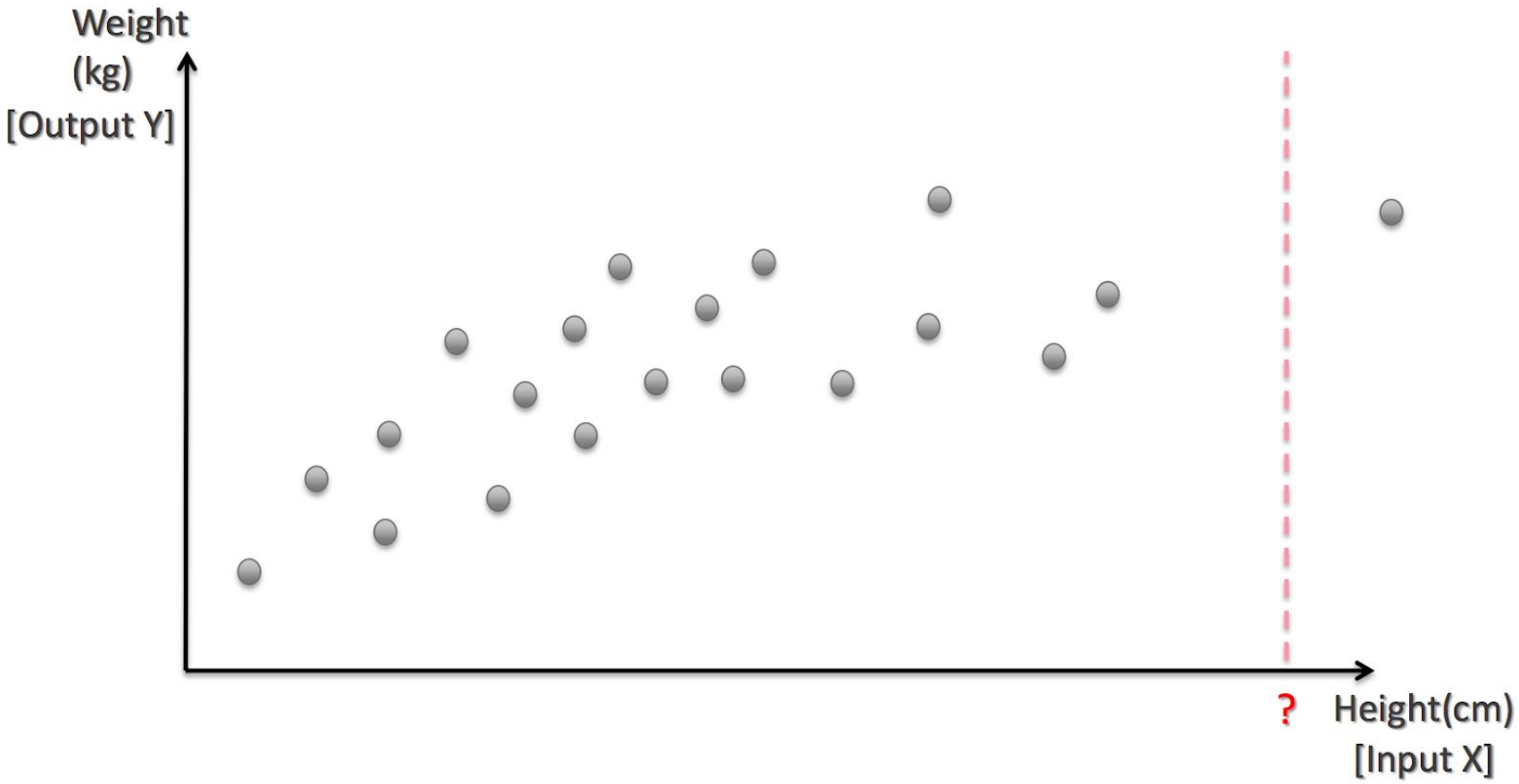
목차

- Review of Regression Models
  - Linear Regression
  - Logistic Regression
- Bayesian Linear Regression
- Bayesian Logistic Regression
- MCMC approximation examples (with PyMC3)

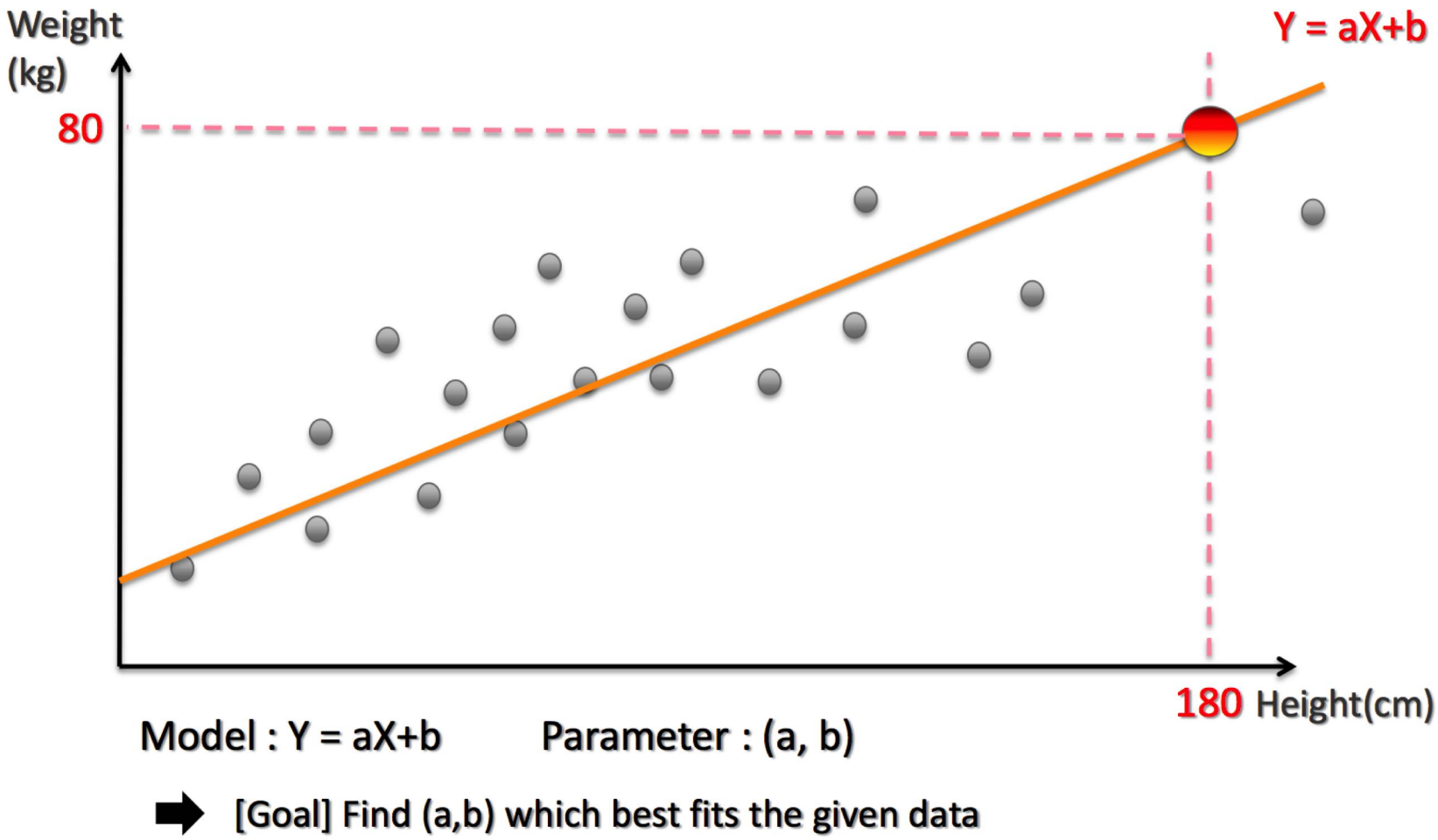
Ch 1. Review of Regression Models

기본적인 Linear Regression 과 Logistic Regression을 Review한다

Linear Regression

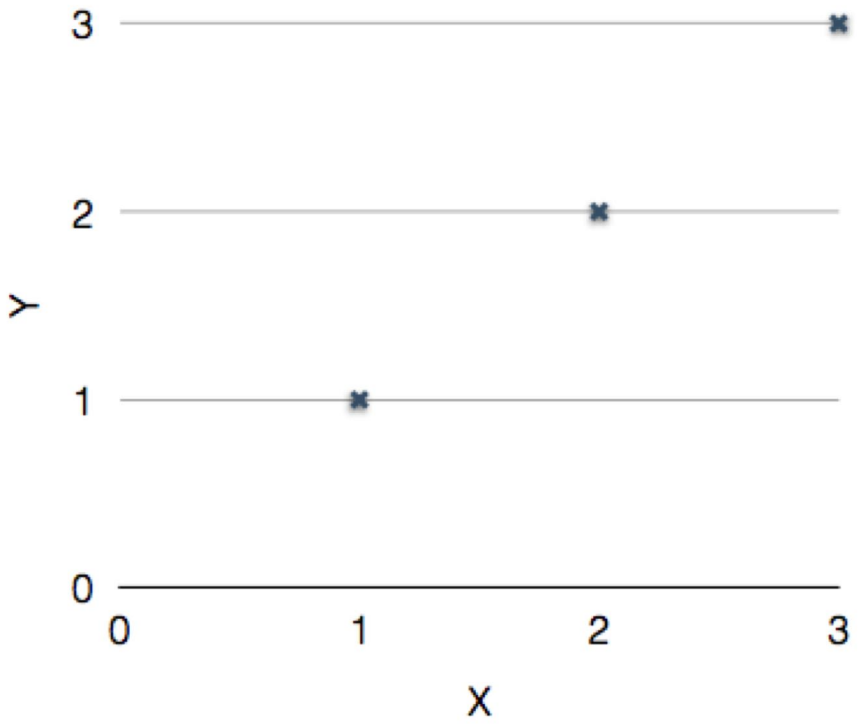


Linear Regression



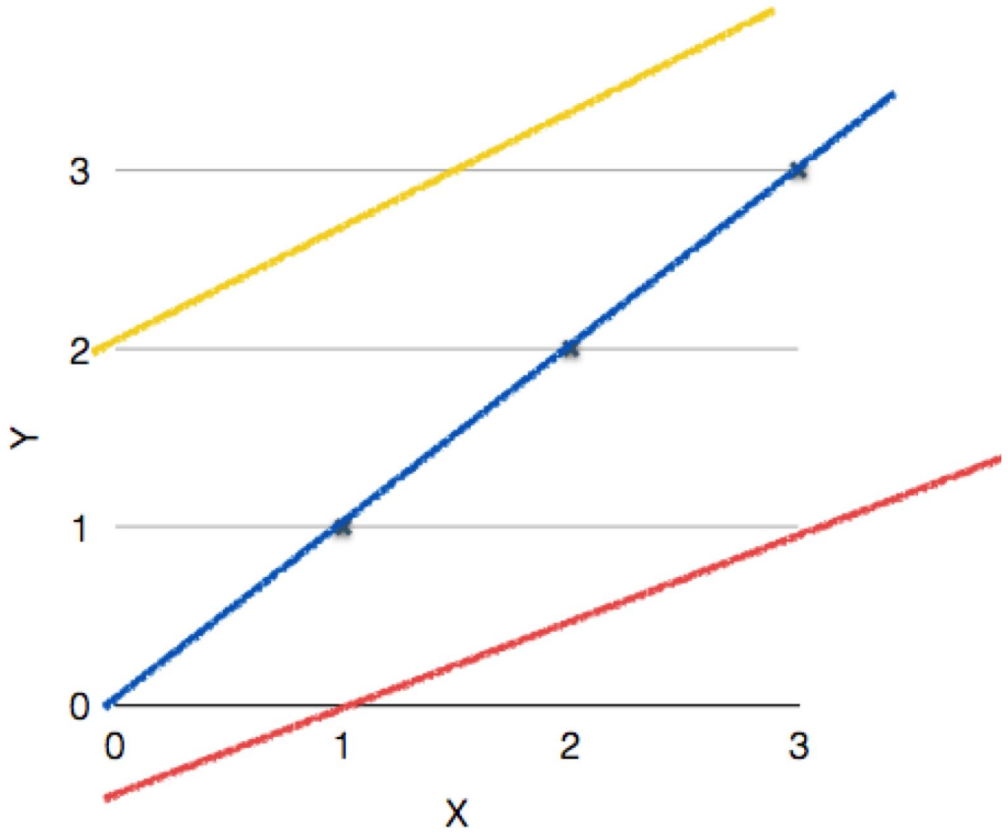
Linear Regression - Example

x	Y
1	1
2	2
3	3

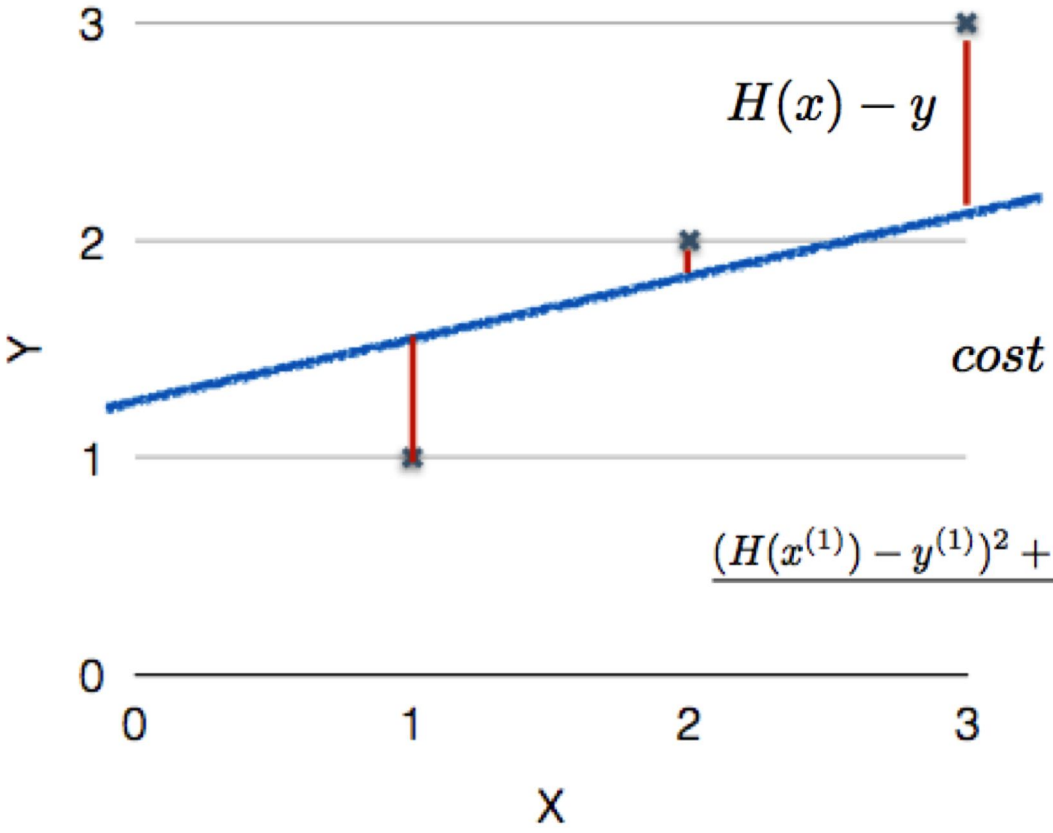


Which model is the best?

x	Y
1	1
2	2
3	3



Cost Function - Measure Error of Model



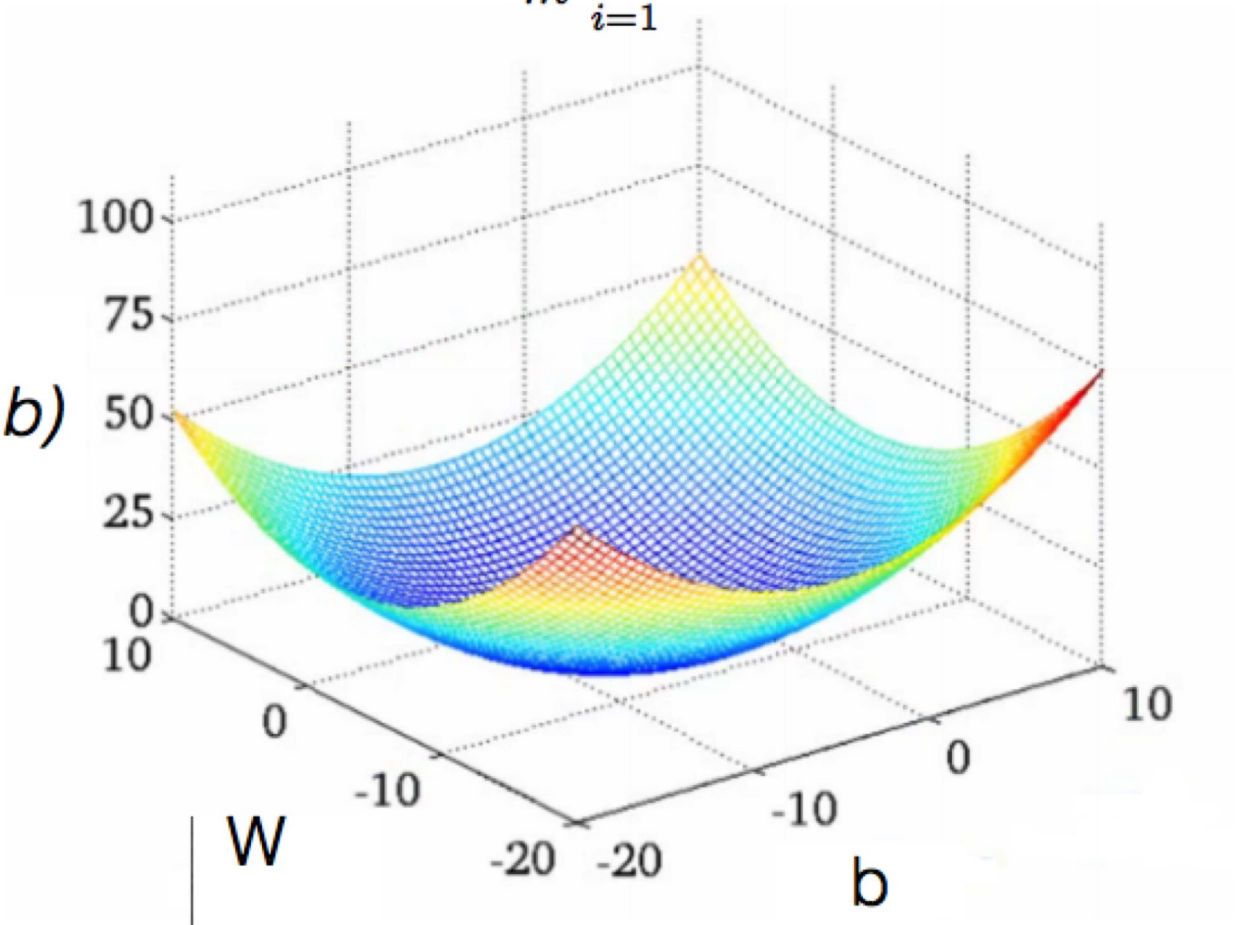
$$H(x) = Wx + b$$

$$cost = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$

$$\frac{(H(x^{(1)}) - y^{(1)})^2 + (H(x^{(2)}) - y^{(2)})^2 + (H(x^{(3)}) - y^{(3)})^2}{3}$$

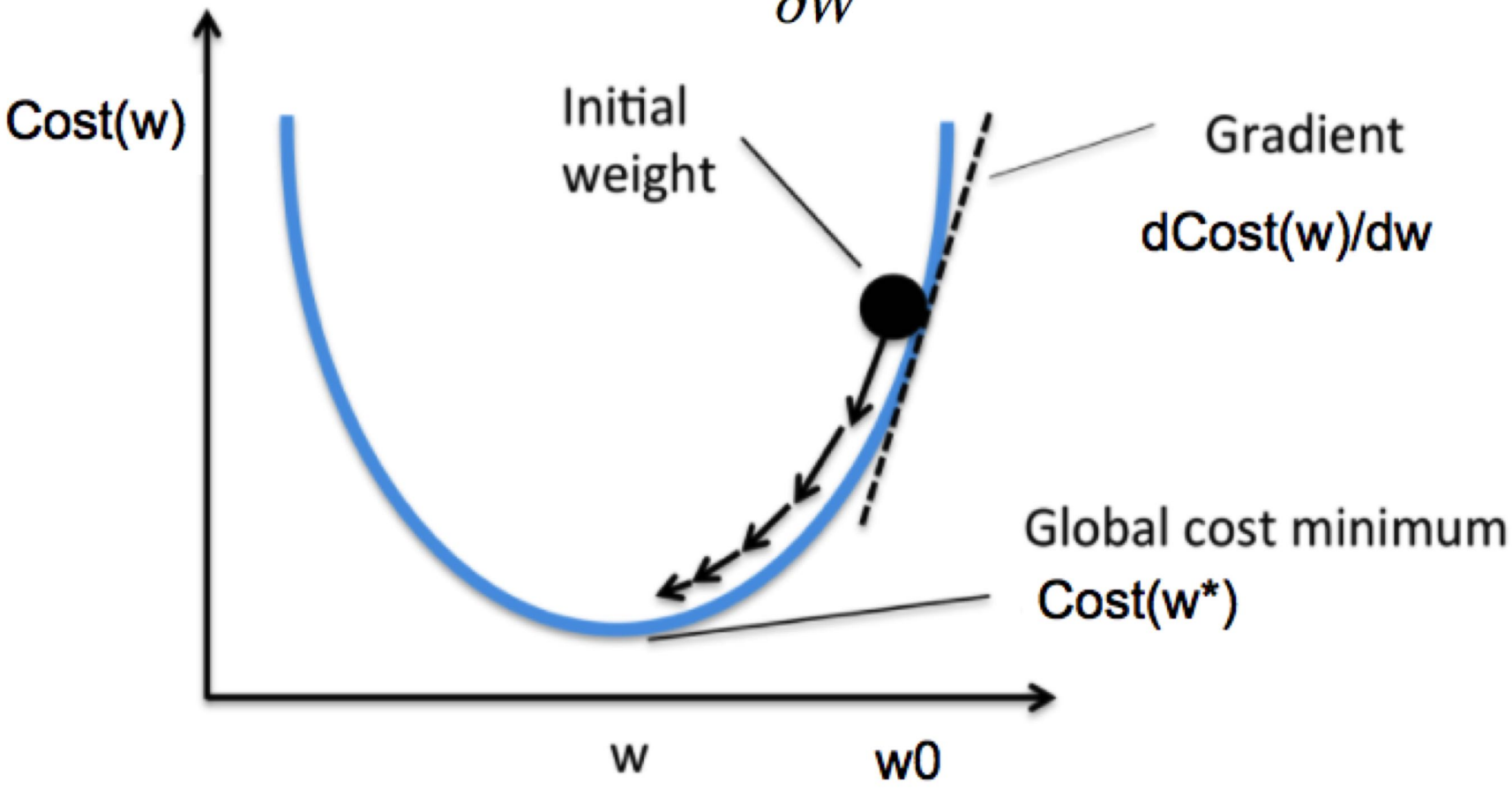
Cost Function Space

$$H(x) = Wx + b \qquad cost(W, b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$



Gradient Descent Algorithm

$$W := W - \alpha \frac{\partial}{\partial W} cost(W)$$



Linear Regression Model

Model: $$H(x) = Wx + b$$

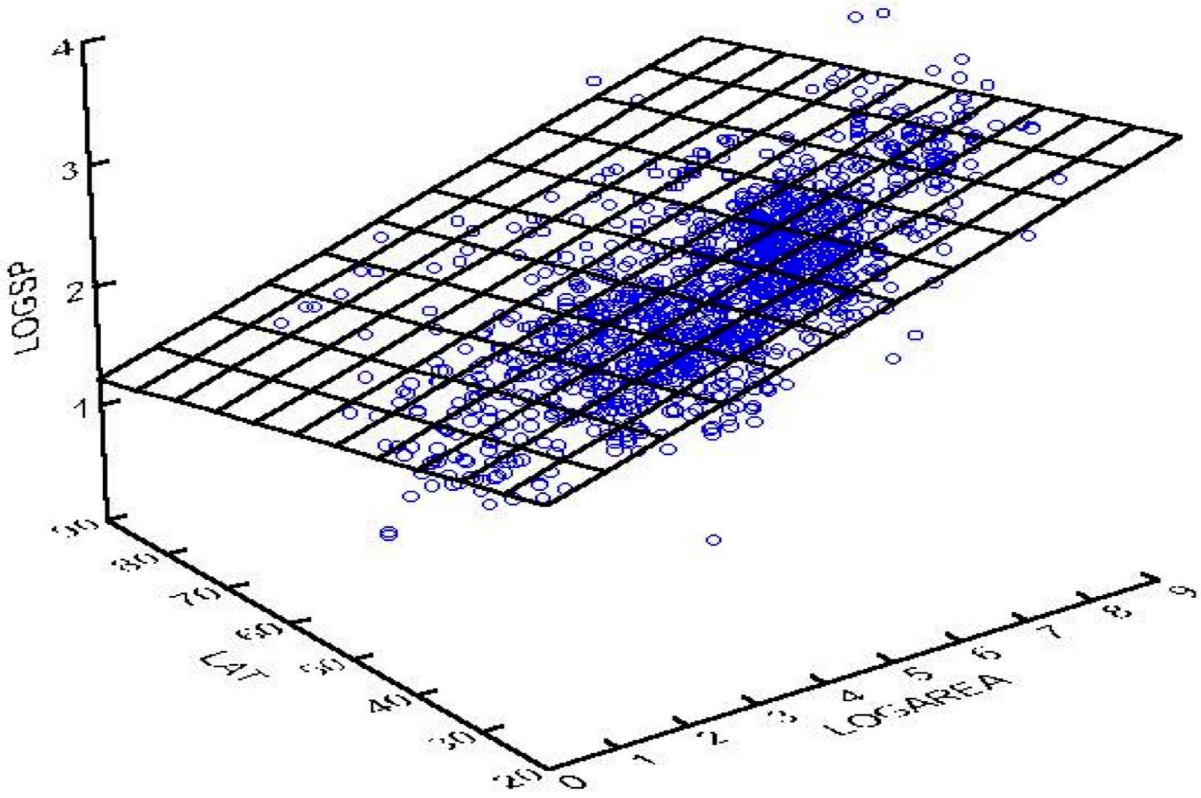
Cost Function: $$cost(W) = \frac{1}{2m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

Minimize Cost:
W, b $$\underset{W,b}{\text{minimize}}\, cost(W, b)$$

Gradient Descent: $$W := W - \alpha \frac{\partial}{\partial W} cost(W)$$

Multivariable Linear Regression

$$H(x_1, x_2, x_3) = w_1 x_1 + w_2 x_2 + w_3 x_3 + b$$



Logistic Regression (Classification)

Linear Regression 이 연속적(continuous)인 값을 다룬다면
Classification 문제는 0/1 이산적(discrete)값을 다룬다

이메일이 스펨인지(1)/아닌지(0)

온라인 거래가 사기인지(1) / 아닌지 (0)

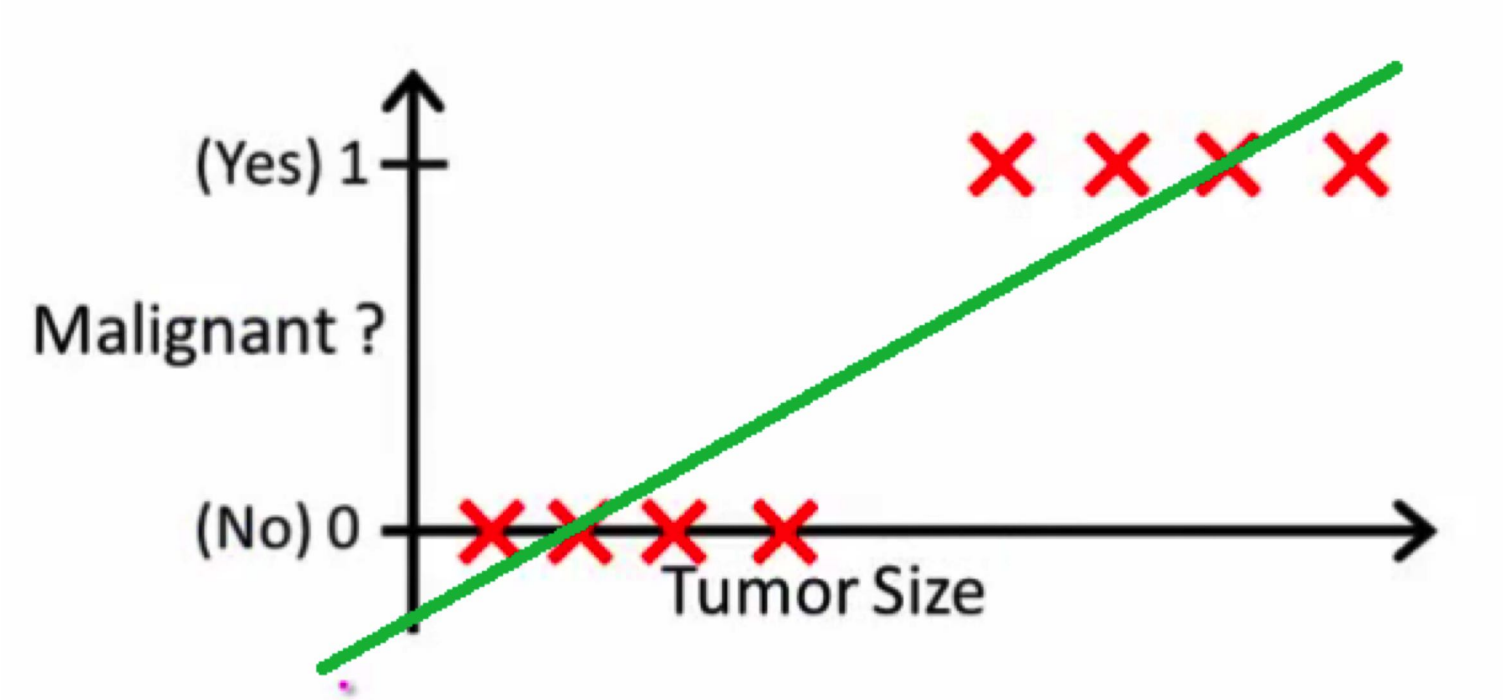
악성 종양인지(1) / 아닌지(0)

개인지(1) / 고양이인지(0)

…

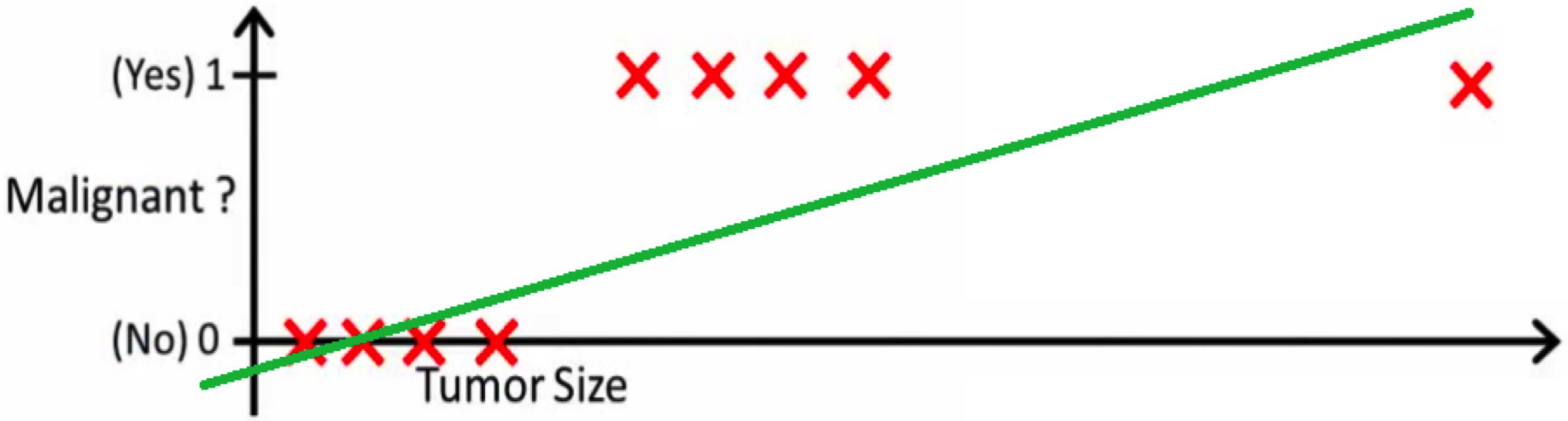
Logistic Regression

- 종양의 크기에 따라, 악성 종양(1)인지 아닌지(0)를 Label해둔 Data가 있다고 하자



Logistic Regression

- 이와 같은 경우 Linear Regression으로 악성 종양을 예측 할 수 있다.



- 그러나, 특별히 크기가 큰 종양이 관측 되는 경우, Linear Regression 모형이 올바르지 않을 수 있다

Logistic Regression (Classification)

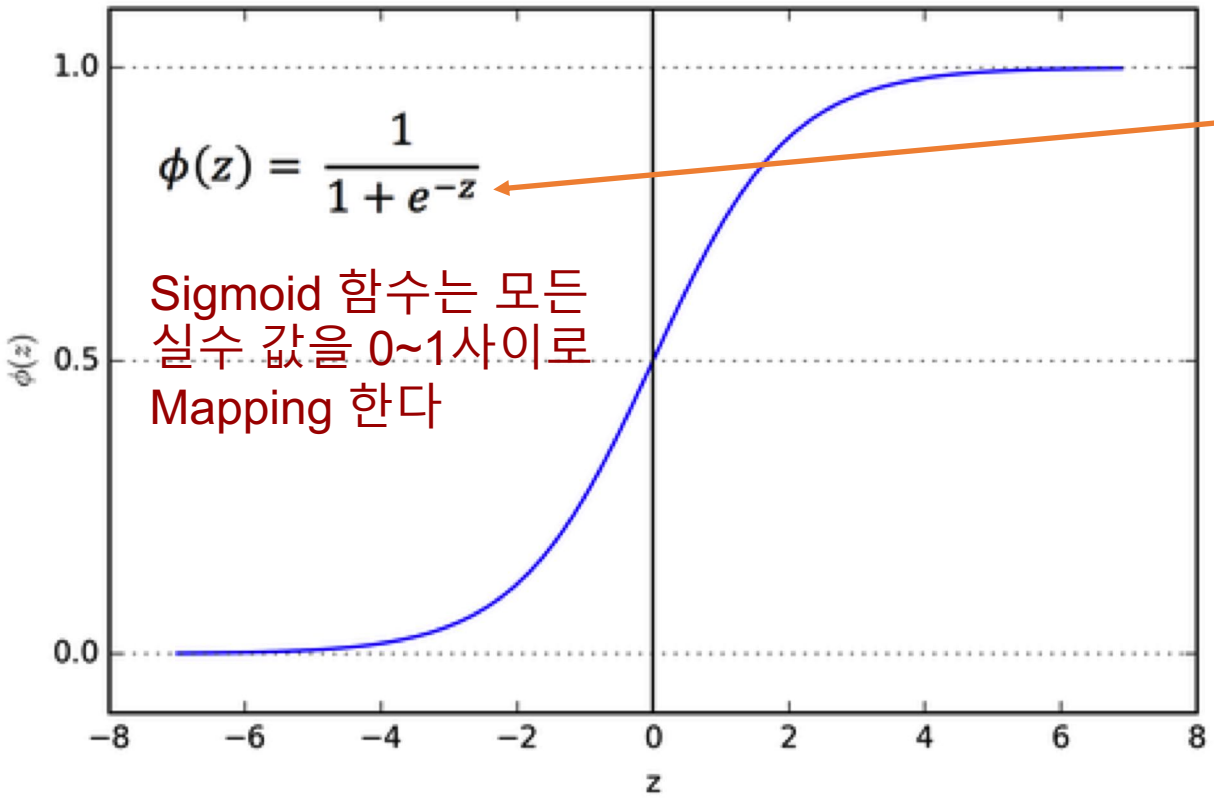
- Logistic Regression에 서다루는 Label Y값은 0 또는 1.
- Linear Regression 모형을 사용하는 경우 1보다 큰 값이나, 0보다 작은 값이 발생할 수 있다.

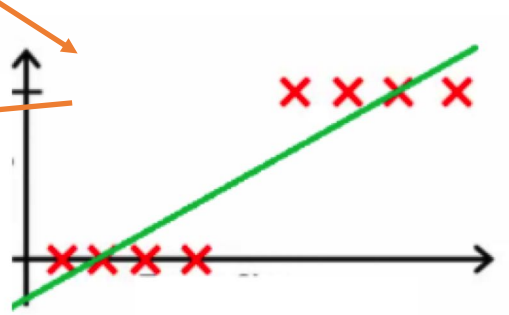
$$H(x) = Wx + b$$

- 이 경우 Threshold에 기반을 두어 H(x)가 일정 값 이상이면 Y=1로 예측하는 편이 정확도가 높다. (일정 값 이하 Y=0)
- 즉, Y의 범위가 0<Y<1인 모형을 원한다

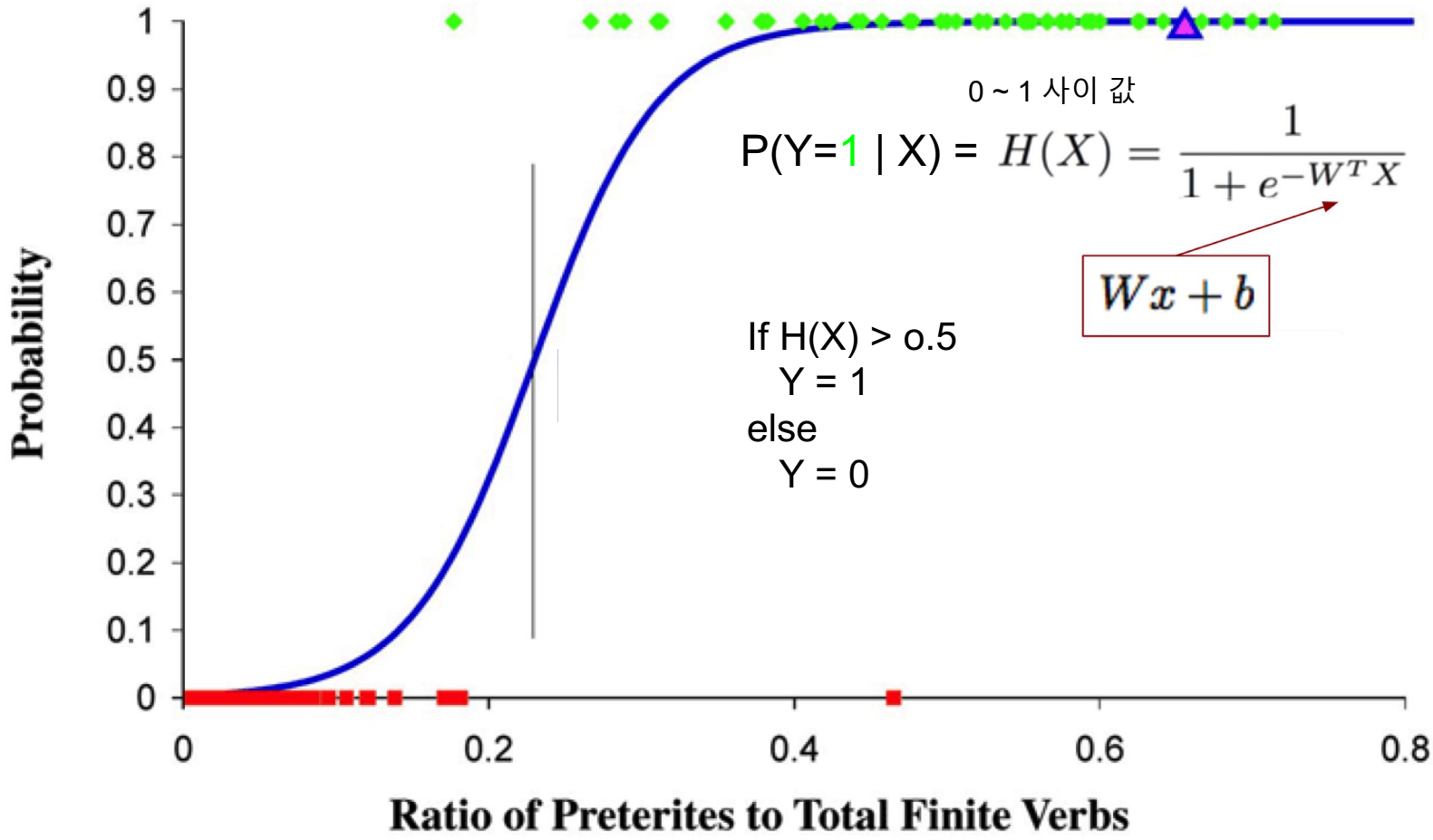
Logistic Regression - Sigmoid 함수

$$H(X) = \frac{1}{1 + e^{-W^T X}}$$





Logistic Regression - Sigmoid 함수



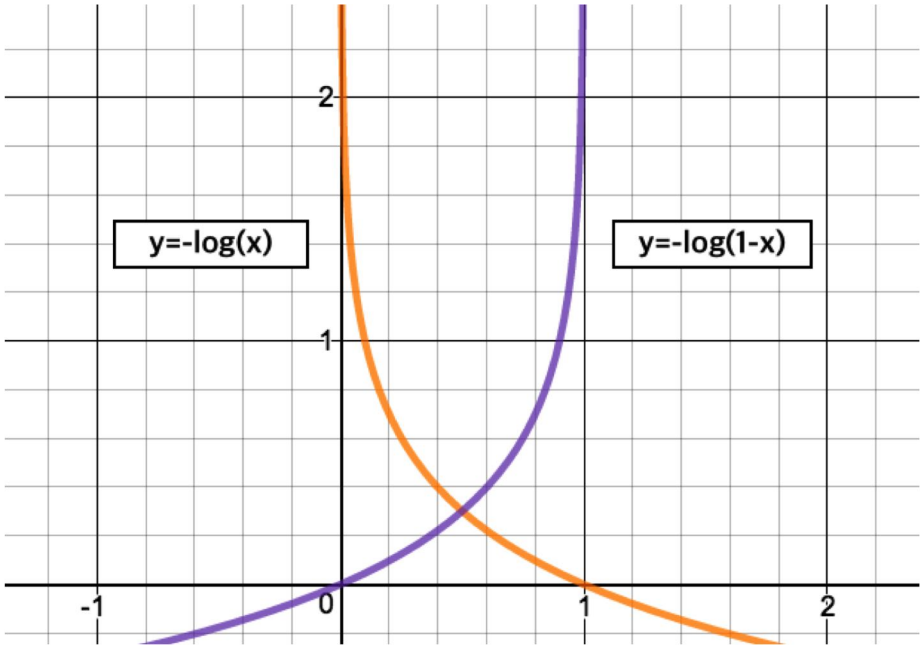
Logistic Regression - Cost Function

$$cost(W) = \frac{1}{m}\sum \quad c(H(x), y)$$

0 ~ 1 사이 값

$$H(X) = \frac{1}{1 + e^{-W^T X}}$$

$$c(H(x), y) = \begin{cases} -log(H(x)) & : y = 1 \\ -log(1 - H(x)) & : y = 0 \end{cases}$$



(y=1)

H(x) -> 1 cost -> 0

H(x) -> 0 cost -> ∞

(y=0)

H(x) -> 0 cost -> 0

H(x) -> 1 cost -> ∞

Logistic Regression - Cost Function

$$cost(W) = \frac{1}{m}\sum \quad c(H(x), y)$$

0 ~ 1 사이 값

$$H(X) = \frac{1}{1 + e^{-W^T X}}$$

$$c(H(x), y) = \begin{cases} -log(H(x)) & : y = 1 \\ -log(1 - H(x)) & : y = 0 \end{cases}$$

$$cost(W) = -\frac{1}{m}\sum ylog(H(x)) + (1 - y)log(1 - H(x))$$

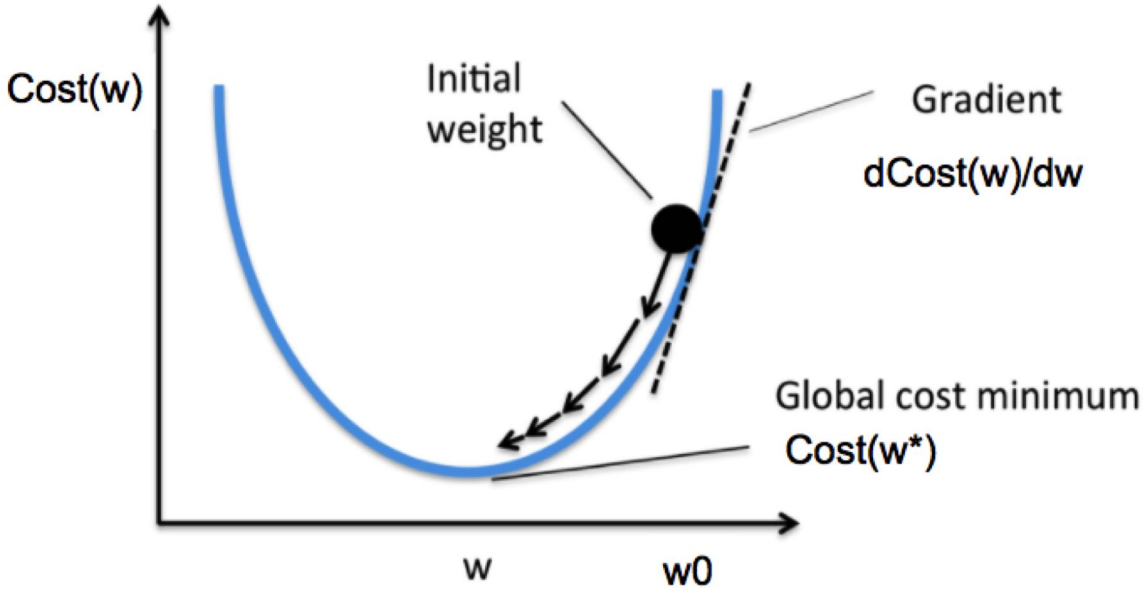
Cross Entropy Cost Function

Logistic Regression Model

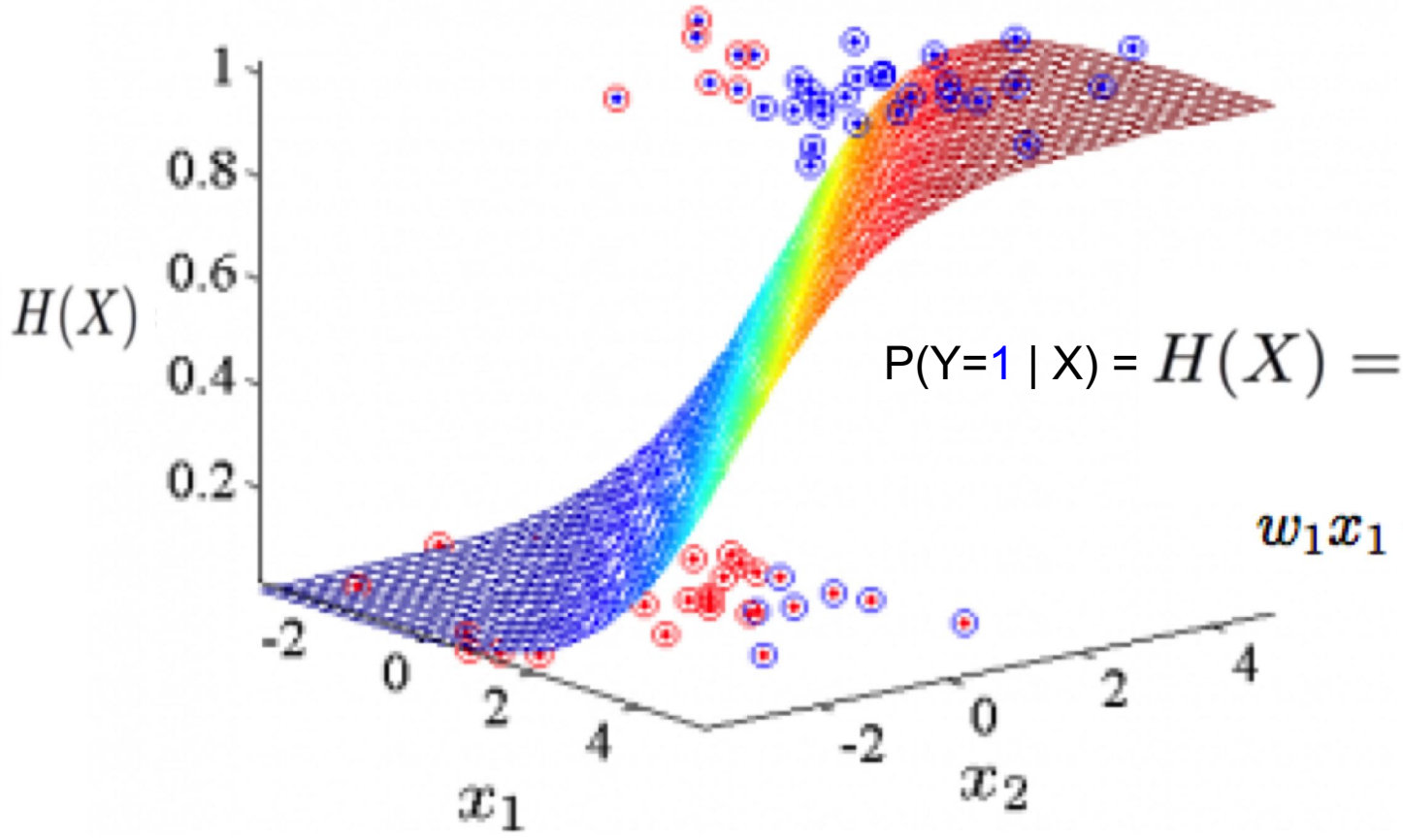
Model: $$H(X) = \frac{1}{1 + e^{-W^T X}}$$

Cost Function: $$cost(W) = -\frac{1}{m} \sum ylog(H(x)) + (1 - y)log(1 - H(x))$$

Gradient Descent: $$W := W - \alpha \frac{\partial}{\partial W} cost(W)$$



Multivariable Logistic Regression



P(Y=1 | X) = $H(X) = \frac{1}{1+e^{-W^T X}}$

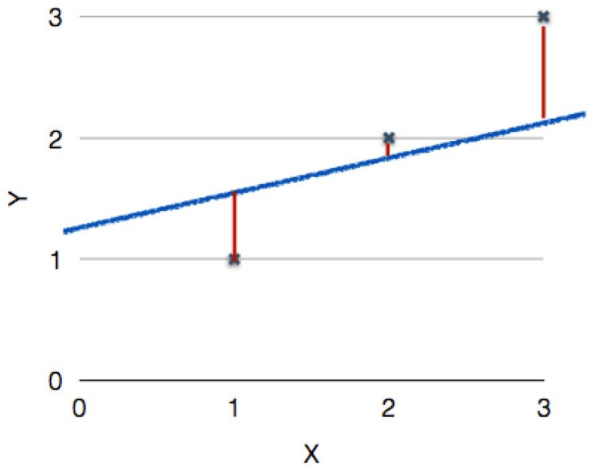
$w_1 x_1 + w_2 x_2 + b$

Ch 2. Bayesian Linear Regression

(Probabilistic) Linear Regression

- Linear Regression Model

$$y = W^T x$$



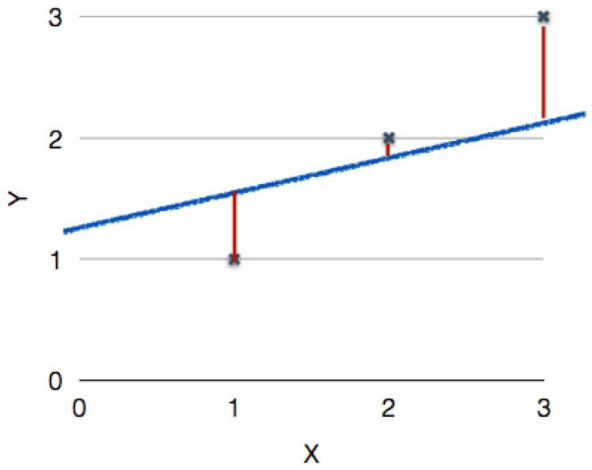
- Cost Function (MSE)

$$L = \frac{1}{2}\|\hat{y} - y\|_2^2$$

(Probabilistic) Linear Regression

- Linear Regression Model

$$y = W^T x$$



- Cost Function (MSE)

$$L = \frac{1}{2}\|\hat{y} - y\|_2^2$$

= Error 및 Cost function에 대한 가정 역시 모델링의 일부분.

(Probabilistic) Linear Regression

- Target y 와 Input x의 관계를 다음과 같이 noise ϵ 를 이용해서 표현 후,

$$\hat{y} = y + \epsilon = W^T x + \epsilon = \sum_{j=1}^{D} w_j x_j + \epsilon$$

- 노이즈 ϵ 가 Zero Mean Gaussian분포를 따른다고 가정하면 (i.e. $\epsilon \sim N(0,1)$),

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right) \qquad \epsilon = \hat{y} - y$$

- 다음과 같은 Probability Model P(Y|X;θ)를 정의할 수 있다.

$$P(\hat{y}|x, W) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(\hat{y} - W^T x)^2}{2\sigma^2}\right)$$

$$P(\hat{y}|x, W) = N(\hat{y}|W^T x, \sigma^2)$$

(Probabilistic) Linear Regression - MLE

- Maximum Likelihood Estimation (MLE) 식을 살펴 보면

$$W_{MLE} = \arg\max_{W} N(\hat{y}|W^T x, \sigma^2)$$

- Log-Likelihood 식으로 대체 하면

$$W_{MLE} = \arg\max_{W} \log\left(\exp\left(-\frac{(\hat{y} - W^T x)^2}{2\sigma^2}\right)\right)$$

$$= \arg\max_{W} -\frac{1}{2\sigma^2}(\hat{y} - W^T x)^2$$

$$= \arg\min_{W} \frac{1}{2\sigma^2}(\hat{y} - W^T x)^2$$

(Probabilistic) Linear Regression - MLE

- $\sigma^2 = 1$ 이라고 가정하면

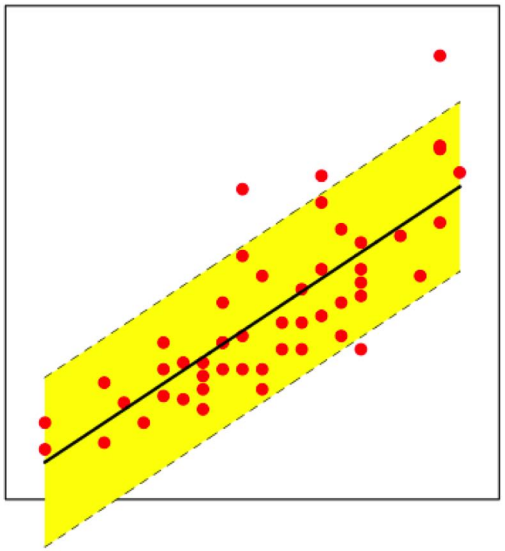
$$W_{MLE} = \arg\min_W \frac{1}{2}(\hat{y} - W^T x)^2$$

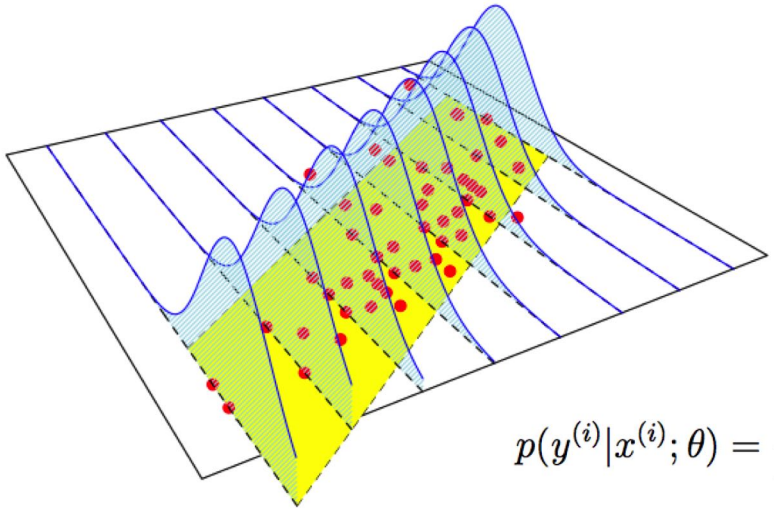
$$= \arg\min_W \frac{1}{2}\sum_i (\hat{y}_i - W_i x_i)^2$$

$$= \arg\min_W \frac{1}{2}\|\hat{y} - W^T x\|_2^2$$

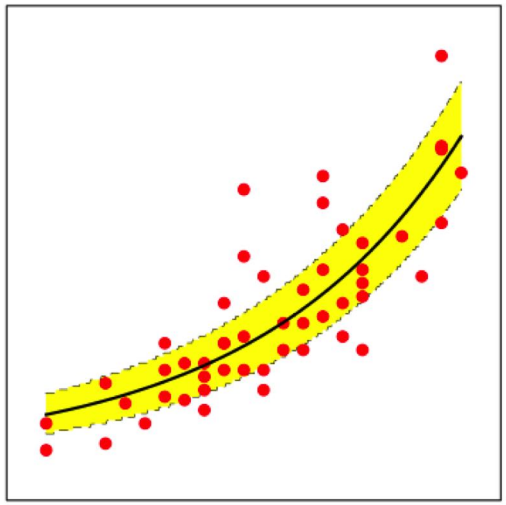
- 즉, Gaussian Likelihood 사용해 전개한 MLE 식이 원래 Linear Regression Cost function과 같은 MSE임을 알 수 있다.

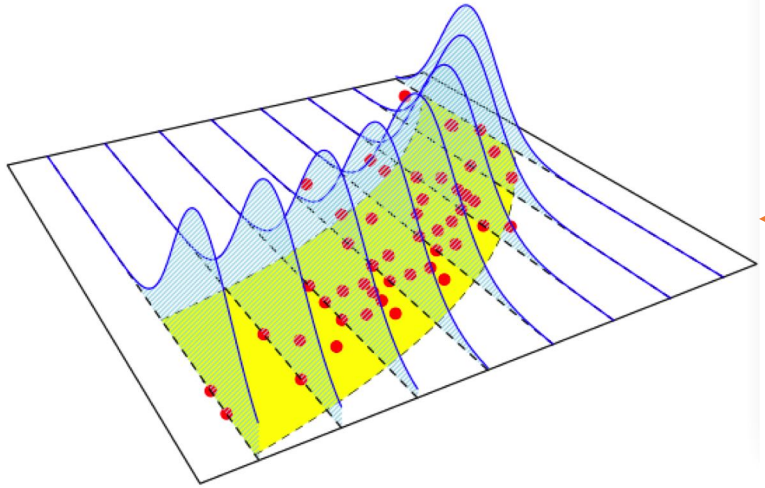
(Probabilistic) Linear Regression





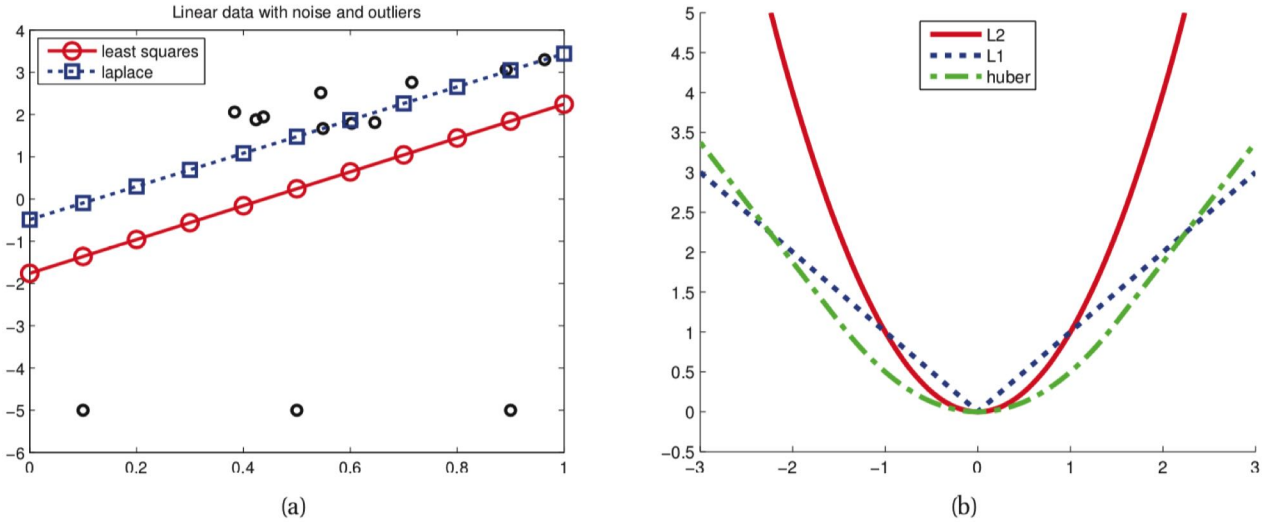
$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$





$f(x^{(i)};\theta)$

Robust Linear Regression w/ Laplacian



- Noise ϵ에 Laplacian Distribution을 가정하는 경우 Negative Loglikelihood (NLL) cost은 l1 function을 사용하는 형태가 된다.

$$p(y|\mathbf{x}, \mathbf{w}, b) = \text{Lap}(y|\mathbf{w}^T\mathbf{x}, b) \propto \exp(-\frac{1}{b}|y - \mathbf{w}^T\mathbf{x}|) \qquad |y - \mathbf{w}^T\mathbf{x}|$$

- Outlier에 강건한 Linear Model Fitting이 가능하다

Maximum Likelihood Estimation (MLE)

- MLE : Probability Model(Likelihood)를 가정한 후 주어진 data에 적합한 모형을 학습하는 방법

$$\theta_{MLE} = \arg\max_{\theta} \log P(X|\theta)$$

$$= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta)$$

$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)$$

Maximum A Posterior (MAP) Estimation

- MAP : MLE와 마찬가지로 주어진 data에 적합한 모형을 학습하는 방법, 그러나 Likelihood 뿐만 아닌 parameter에 대한 Prior를 함께 가정

$$\theta_{MAP} = \arg\max_{\theta} P(X|\theta)P(\theta)$$

$$= \arg\max_{\theta} \log P(X|\theta)P(\theta)$$

$$= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta)P(\theta)$$

$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)P(\theta)$$

Bayes' Rule

$$p(\theta|x) = \frac{p(\theta)f(x|\theta)}{p(x)}$$

MAP vs MLE

- P(θ)를 Uniform Distribution 가정하는 경우 1/c으로 고정.

$$\theta_{MAP} = \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)P(\theta)$$

$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)\, const$$

$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)$$

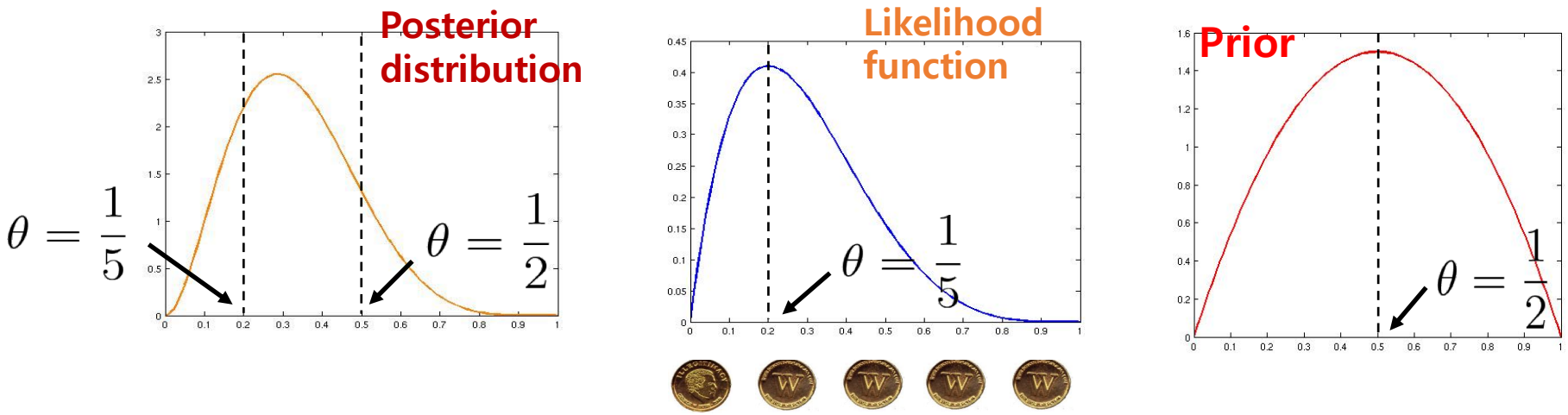
$$= \theta_{MLE}$$

- 즉, MLE는 MAP의 Special Case임을 알 수 있다

베이즈 갱신(Bayes Update)

Posterior Likelihood Prior

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int d\theta P(x|\theta)P(\theta)}$$



Posterior

$$P(\theta|x) = \frac{\theta^x(1-\theta)^{1-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{P(x)B(\alpha,\beta)}$$

$$\propto \theta^{\alpha+x-1}(1-\theta)^{\beta+(1-x)-1} = Beta(\hat{\alpha},\hat{\beta})$$

updated Prior with observed data

Bayesian Linear Regression

- Linear Regression 모형의 parameter W에 Prior를 가정하면

$$P(W|\hat{y}, x) = P(\hat{y}|x, W)P(W|\mu_0, \sigma_0^2)$$

- Likelihood는 Gaussian, Prior는? zero mean Gaussian again!

$$P(W|\mu_0, \sigma_0^2) = N(0, \sigma_0^2)$$

- 즉,

$$P(W|\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\sigma_0^2\pi}} \exp\left(-\frac{(W-\mu_0)^2}{2\sigma_0^2}\right) \qquad \mu_0 = 0,$$

$$\propto \exp\left(-\frac{W^2}{2\sigma_0^2}\right)$$

Bayesian Linear Regression w/ Gaussian Prior (=Ridge Regression)

- Posterior? $P(W|\hat{y},x) = P(\hat{y}|x,W)P(W|\mu_0,\sigma_0^2)$

$$\propto \exp\left(-\frac{(\hat{y}-W^Tx)^2}{2\sigma^2}\right)\exp\left(-\frac{W^2}{2\sigma_0^2}\right)$$

- MAP estimation w/ log-likelihood

$$\log P(W|\hat{y},x) \propto -\frac{1}{2\sigma^2}(\hat{y}-W^Tx)^2 - \frac{1}{2\sigma_0^2}W^2$$

$$= -\frac{1}{2\sigma^2}\|\hat{y}-W^Tx\|_2^2 - \frac{1}{2\sigma_0^2}\|W\|_2^2$$

- 즉, $\log P(W|\hat{y},x) \propto -\frac{1}{2}\|\hat{y}-W^Tx\|_2^2 - \frac{\lambda}{2}\|W\|_2^2$

Bayesian View of Linear Regression

Likelihood	Prior	Name
Gaussian	Uniform	Least squares
Gaussian	Gaussian	Ridge
Gaussian	Laplace	Lasso
Laplace	Uniform	Robust regression
Student	Uniform	Robust regression

$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2$$

Ridge Regression
(l2 penalty)

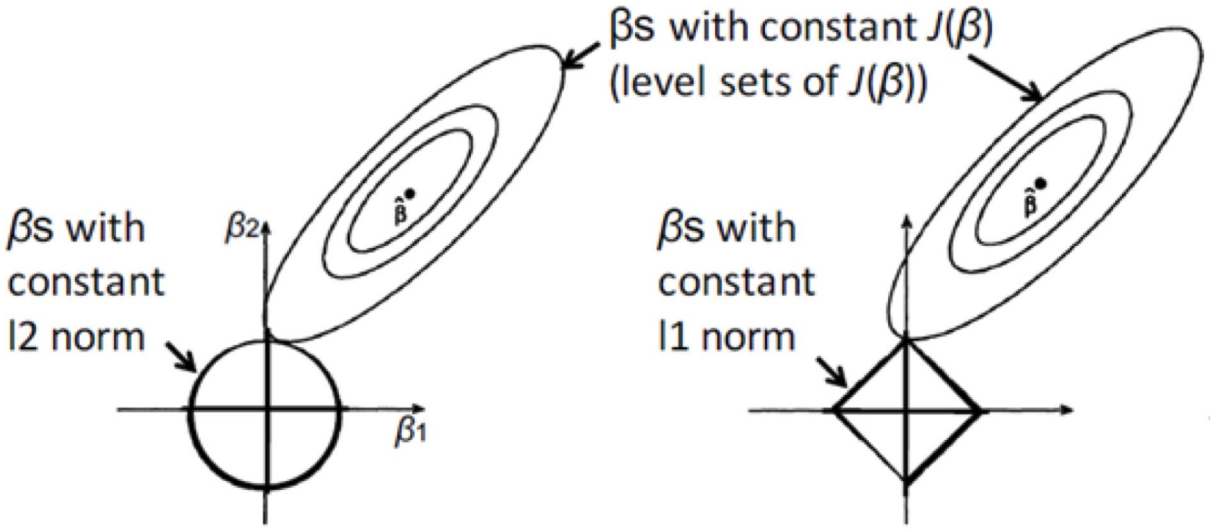
$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_1$$

Lasso
(l1 penalty)

$\lambda \geq 0$

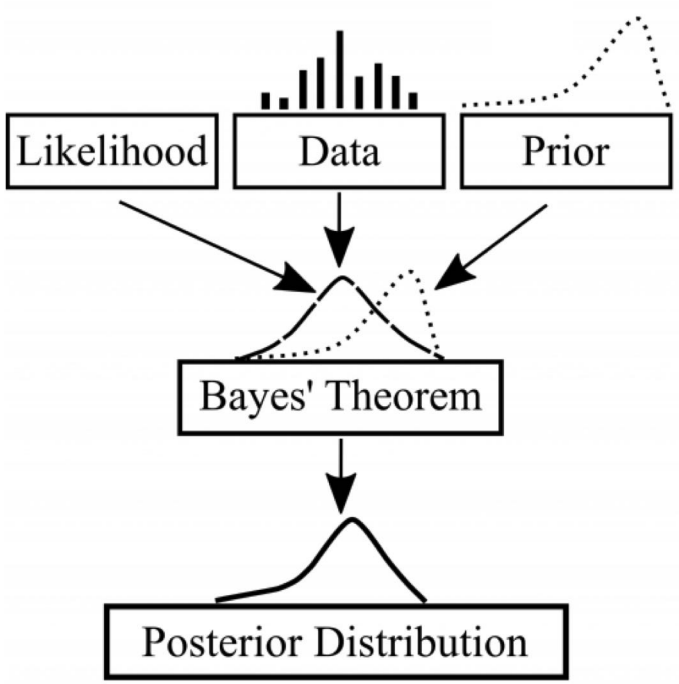
Ridge Regression:
$\text{pen}(\beta) = \|\beta\|_2^2$

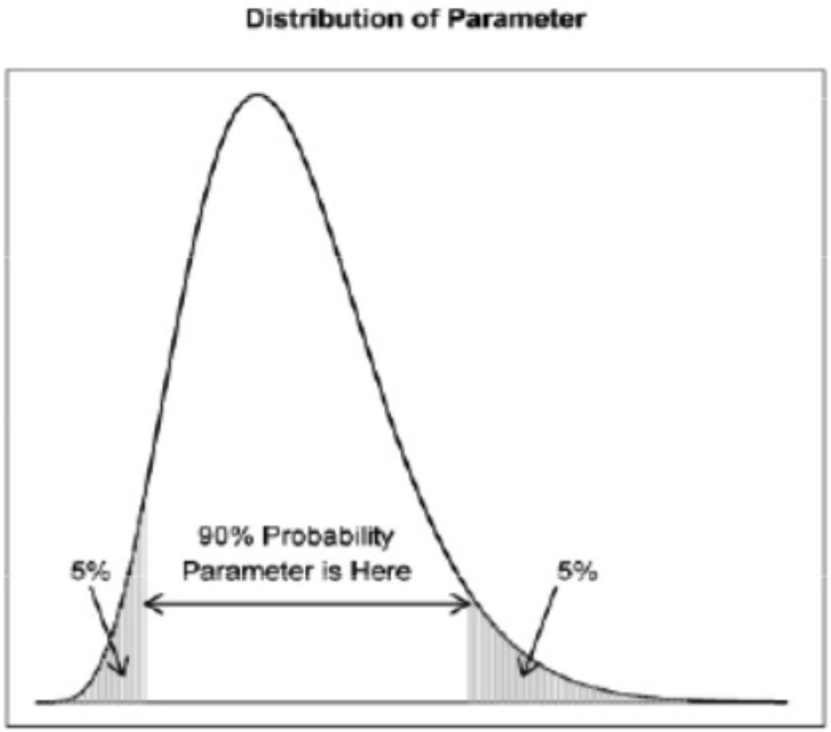
Lasso:
$\text{pen}(\beta) = \|\beta\|_1$



Full Bayesian Approach

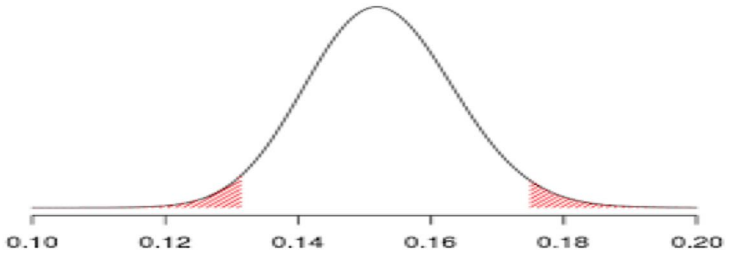
- 사실 Bayesian Approach에서 핵심적으로 구하고자 하는 것은 Posterior Distribution P(w|y,x) 이다.
  - MAP estimation은 Bayesian Estimation의 Shortcut 일뿐





Full Bayesian Approach

- Parameter w 에 대한 Posterior Distribution P(w|y,x)을 구하게 되면, w에 대한 credibility interval 및 Optimal 값 추정은 물론,



- 새로운 Data Point에 대한 Predictive Distribution을 구할 수 있다.

$$P(y'|\hat{y},x) = \int_W P(y'|x',W)P(W|\hat{y},x) \qquad (x',y')$$

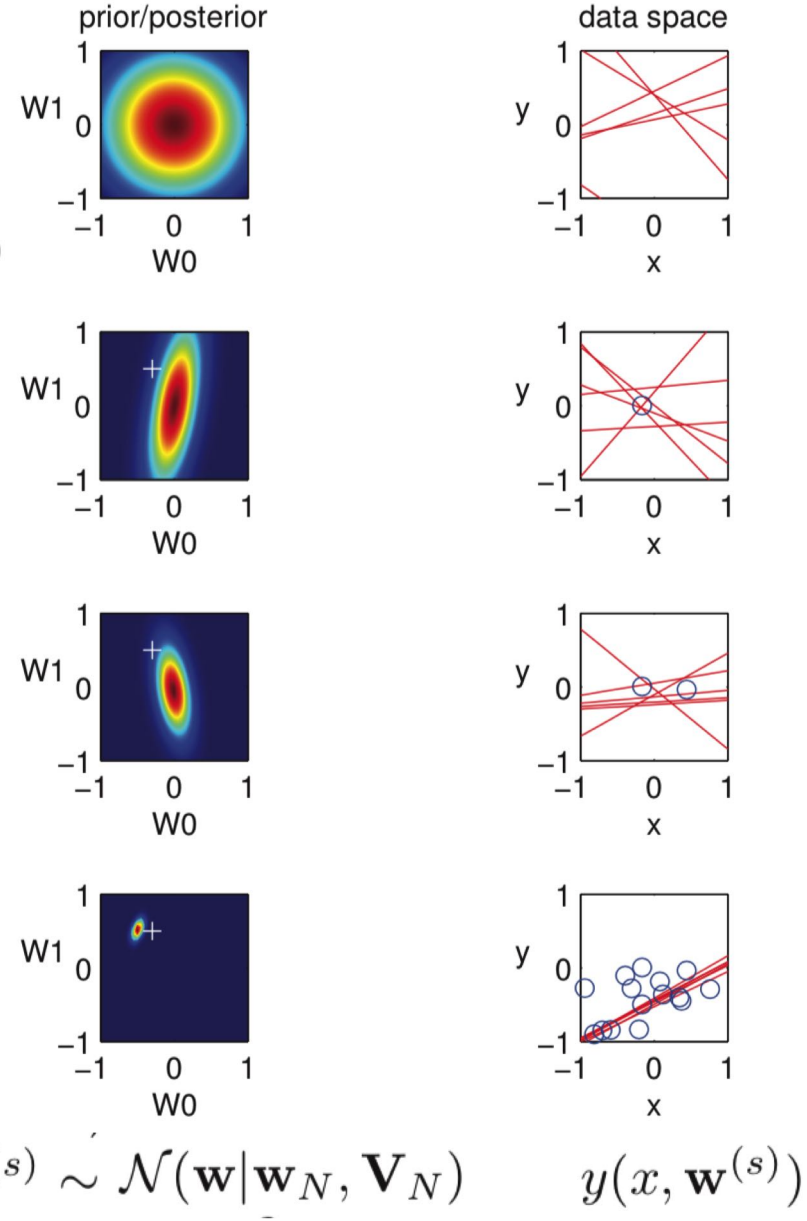
$$= \mathbb{E}_W\left[P(y'|x',W)\right]$$

Posterior Distribution

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{X},\mathbf{y},\sigma^2) &\propto \mathcal{N}(\mathbf{w}|\mathbf{w}_0,\mathbf{V}_0)\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I}_N) \\
&= \mathcal{N}(\mathbf{w}|\mathbf{w}_N,\mathbf{V}_N) \\
\mathbf{w}_N &= \mathbf{V}_N\mathbf{V}_0^{-1}\mathbf{w}_0 + \frac{1}{\sigma^2}\mathbf{V}_N\mathbf{X}^T\mathbf{y} \\
\mathbf{V}_N^{-1} &= \mathbf{V}_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} \\
\mathbf{V}_N &= \sigma^2(\sigma^2\mathbf{V}_0^{-1} + \mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}
$$

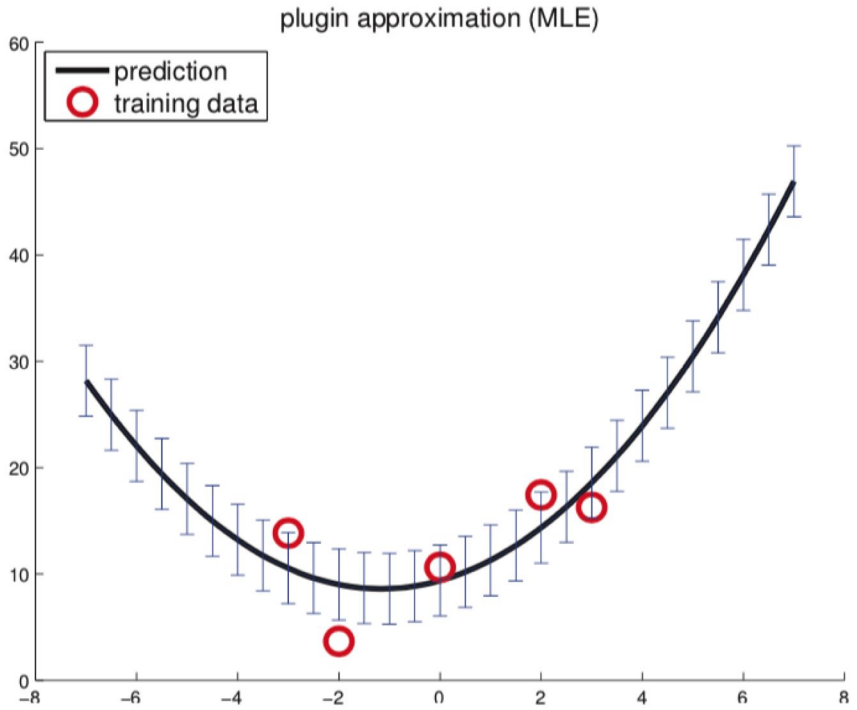
If $\mathbf{w}_0 = \mathbf{0}$ and $\mathbf{V}_0 = \tau^2\mathbf{I}$, and let $\lambda = \frac{\sigma^2}{\tau^2}$, it is "ridge regression setting"

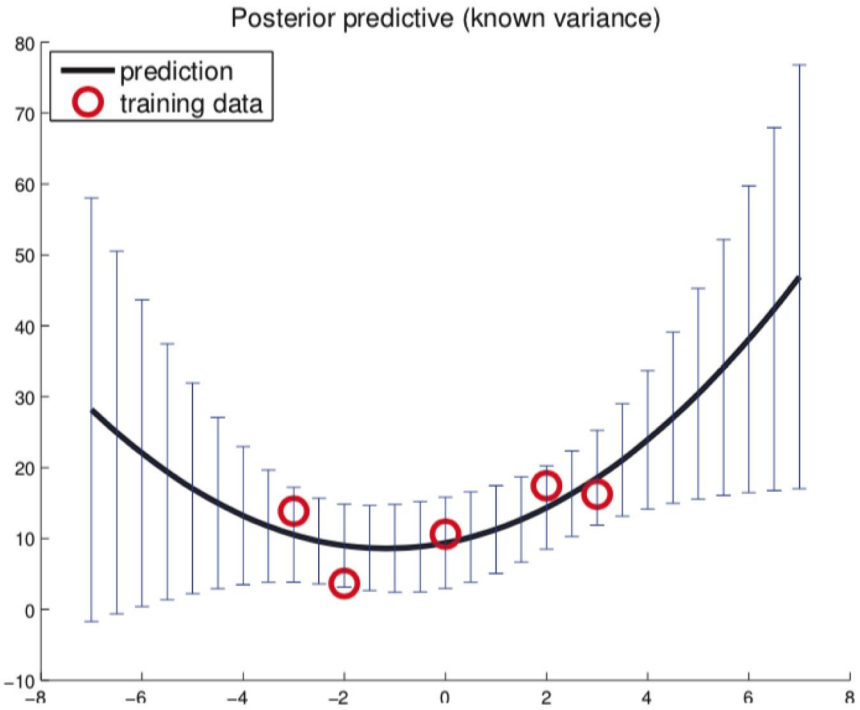
$$
\begin{aligned}
y(x,\mathbf{w}) &= w_0 + w_1 x + \epsilon \\
w_0 &= -0.3 \\
w_1 &= 0.5
\end{aligned}
$$



$\mathbf{w}^{(s)} \sim \mathcal{N}(\mathbf{w}|\mathbf{w}_N,\mathbf{V}_N)$ $y(x,\mathbf{w}^{(s)})$

Predictive Distribution





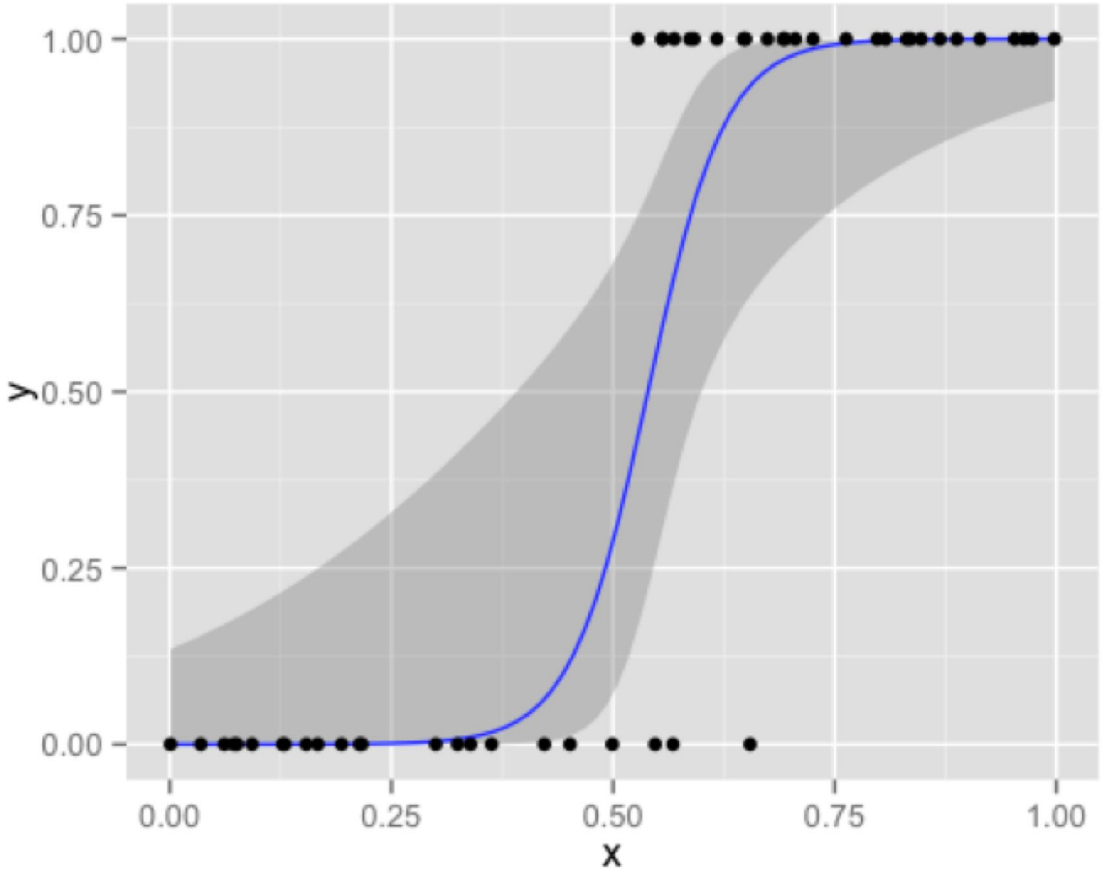
$$
\begin{aligned}
p(y|\mathbf{x}, \mathcal{D}, \sigma^2) &= \int \mathcal{N}(y|\mathbf{x}^T\mathbf{w}, \sigma^2)\mathcal{N}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N)d\mathbf{w} \\
&= \mathcal{N}(y|\mathbf{w}_N^T\mathbf{x}, \sigma_N^2(\mathbf{x})) \\
\sigma_N^2(\mathbf{x}) &= \sigma^2 + \mathbf{x}^T\mathbf{V}_N\mathbf{x}
\end{aligned}
$$

Bayesian Linear Regression

- 변수 X, Y들의 관계를 확률 모형 $\mathrm{P}(Y|X)$로 정의하고, Likelihood Function을 정의하였다.
- MLE (Maximum Likelihood Estimation)을 통하여 최적의 Parameter를 찾는 과정이 Least-square cost를 최소화하는 것과 같았다.
- Bayesian Linear Regression에서 Parameter에 대한 Prior를 (zero mean) Gaussian으로 정의하였다.
- MAP (Maximum A Posterior Estimation)을 통하여 최적의 Parameter를 찾는 과정은 l2 regularization을 적용한 Ridge Regression과 같았다.
- Gaussian-Gaussian은 Conjugate 관계임으로 Posterior 및 Predictive Distribution 등을 수식으로 구할 수 있다.

Ch 2. Bayesian Logistic Regression (and MCMC with Code)

Logistic Regression (Classification Model)

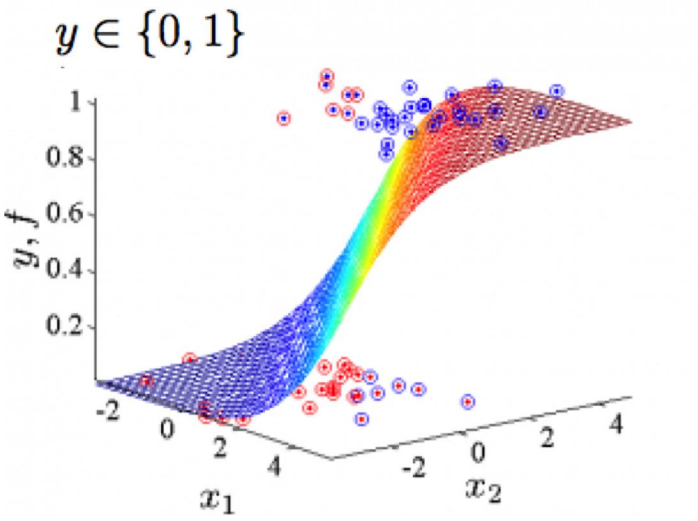


Logistic Regression

- Logistic Regression Model

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$\theta^T x = \theta_0 + \sum_{j=1}^{n} \theta_j x_j$$



- Cost Function (Cross Entropy)

$$J(\theta) = \frac{1}{m} \sum c(h_\theta(x), y)$$

$$c(h_\theta(x), y) = \begin{cases} -\log\big(h_\theta(x)\big) & : \quad y = 1 \\ -\log\big(1 - h_\theta(x)\big) & : \quad y = 0 \end{cases}$$

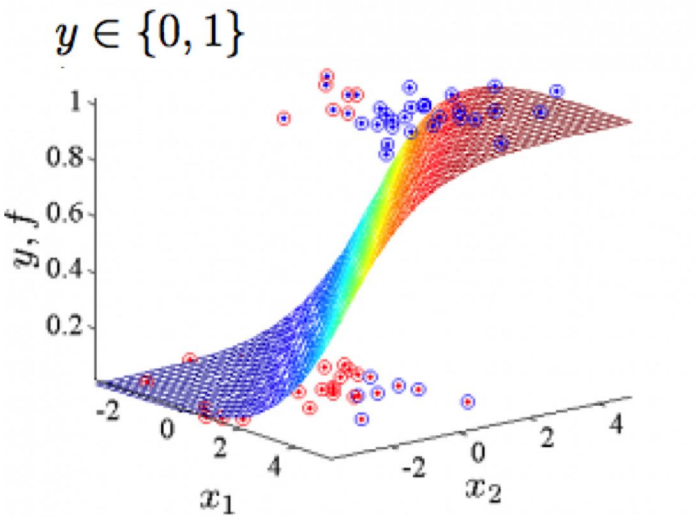
(OR) $\quad c(h_\theta(x), y) = -y \log\big(h_\theta(x)\big) - (1 - y)\log(1 - h_\theta(x))$

Logistic Regression's Cost Function?

- Logistic Regression Model

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$\theta^T x = \theta_0 + \sum_{j=1}^{n} \theta_j x_j$$



- Cost Function (Cross Entropy)

$$J(\theta) = \frac{1}{m} \sum c(h_\theta(x), y)$$

Cross Entropy 사용 근거는?
= 미분이 가능하다?
= Non-Negativity?
= Convexity?

$$c(h_\theta(x), y) = \begin{cases} -\log\big(h_\theta(x)\big) & : \quad y = 1 \\ -\log\big(1 - h_\theta(x)\big) & : \quad y = 0 \end{cases}$$

(OR) $\quad c(h_\theta(x), y) = -y \log\big(h_\theta(x)\big) - (1 - y)\log(1 - h_\theta(x))$

Logistic Regression의 확률적 표현

- Hypothesis를 인풋 변수 x가 주어졌을 때 아웃풋 변수 y에 대한 확률 모형 P(Y|X)로 가정한다.

$$
\begin{aligned}
P(y = 1 \mid x; \theta) &= h_\theta(x) \\
P(y = 0 \mid x; \theta) &= 1 - h_\theta(x)
\end{aligned}
$$

- Logistic Regression 모형을 Bernoulli Distribution을 활용한 모형으로 해석할 수 있다.

$$
p(y \mid x; \theta) = (h_\theta(x))^y \, (1 - h_\theta(x))^{1-y}
$$

$$
y|x; \theta \sim \text{Bernoulli}(\phi)
$$

Logistic Regression의 확률적 표현

- 관측 데이터들이 독립시행으로 얻어졌다고 가정, θ 에 대한 likelihood 함수를 다음과 같이 적을 수 있다.

$$
\begin{aligned}
L(\theta) &= p(\vec{y} \mid X; \theta) \\
&= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}
\end{aligned}
$$

- Log-likelihood

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))
\end{aligned}
$$

Logistic Regression의 확률적 표현

- Log likelihood의 gradient를 계산,

$$\begin{aligned}
\frac{\partial}{\partial \theta_j}\ell(\theta) &= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)\frac{\partial}{\partial \theta_j}g(\theta^T x) \\
&= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)g(\theta^T x)(1-g(\theta^T x)\frac{\partial}{\partial \theta_j}\theta^T x \\
&= \left(y(1-g(\theta^T x)) - (1-y)g(\theta^T x)\right)x_j \\
&= \left(y - h_\theta(x)\right)x_j
\end{aligned}$$

- Gradient Ascent를 이용해 MLE Optimization을 수행 할 수 있다.

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}$$

Bayesian Logistic Regression?

- Logistic Regression의 경우 Linear Regression와 다르게 Conjugate 조건을 만족하는 parameter의 prior distribution이 존재하지 않는다.

- 즉, Analytical 하게 수식으로 구해지는 Posterior Distribution존재하지 않는다.

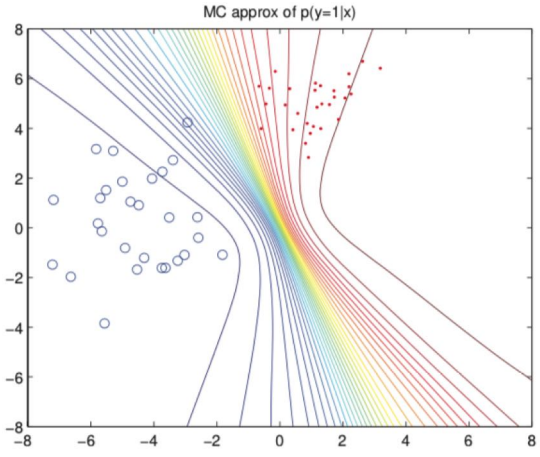
- 하지만, 여전히 Parameter에 Prior를 지정 할 수 있으며, MCMC sampling기법을 통해서 Posterior Distribution의 Approximation을 구할 수 있다.

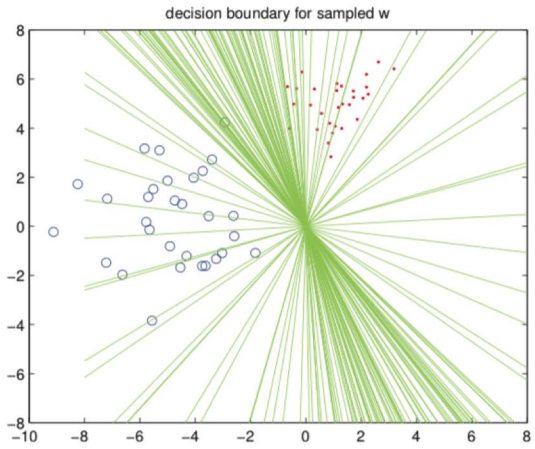
Markov Chain Monte Carlo (MCMC)

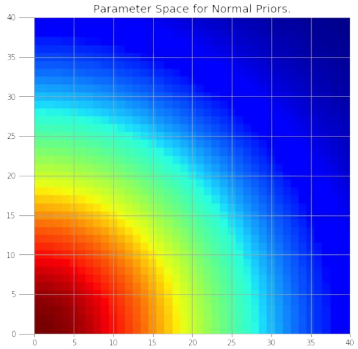
- MCMC 기법이란? 타겟 확률 분포(Target Probability Distribution)으로 부터 랜덤 샘플을 얻는 방법.
  - Target분포를 Stationary Distribution으로 가지는 Markov Chain을 만들어 Sample을 얻는 방법이다.
- 예) Logistic Regression

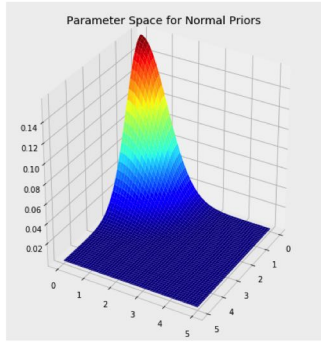
$$p(y=1|\mathbf{x},\mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S}\text{sigm}((\mathbf{w}^s)^T\mathbf{x})$$

$$\mathbf{w}^s \sim p(\mathbf{w}|\mathcal{D})$$









Let's See MCMC approximation Examples (with PyMC3)