Regression

Index

KU-BIG

1. Introduction

회귀분석(Regression Analysis)

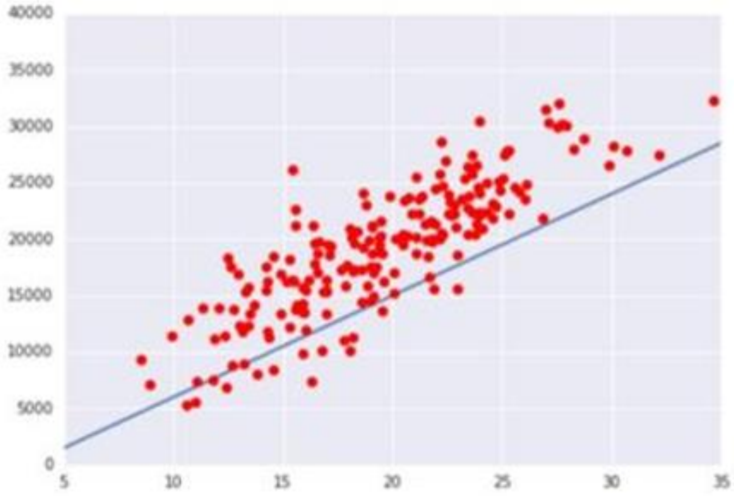
: 어떤 결과값이 존재할 때, 그 결과값을 결정할 것이라고 추정되는

입력값과 결과값의 연관관계를 찾아 결과값을 예측하는 기법

회귀모델(Regression model)

$$y = h(x_1, x_2,\dots,x_k; \beta_1, \beta_2 \dots,\beta_k) + e$$

2.1 선형회귀모델 (Linear regression)



$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

- x와 y는 선형관계이다
- 오차항은 $N(0, \sigma^2)$을 따른다
- 독립변수⊥오차항

2.1 선형회귀모델 (Linear regression)

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$, $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$를 β에 대해서 Minimize

KU-BIG

2.2 GLM (Generalized Linear Model)

: 오차항이 정규분포를 따르지 않는 경우를 포함하는 선형모형의 확장

ex) 이항변수

- Random Component : response variable Y
- Systematic Component : X
- Link function : a function of E(Y) related to X

  $\mu = E(y)$ 를 $g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$처럼

  linear predictor와 연결시키는 함수

2.2 GLM (Generalized Linear Model)

- $g(\mu) = \mu$: identity link function, $-\infty < \mu < \infty$

  $\mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ => Ordinary Regression

- $g(\mu) = log(\mu)$: log link function, $0 \leq \mu < \infty$

  $log(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ => Poisson Regression

- $g(\mu) = log(\frac{\mu}{1-\mu})$: logit link function, $0 \leq \mu < \infty$

- $log(\frac{\mu}{1-\mu}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ => Logistic Regression

2.2 GLM (Generalized Linear Model)

$$f(y_i;\theta_i,\phi) = exp[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)]$$

$$\sum_{i=1}^{n} logf(y_i;\theta_i,\phi) = \sum_{i=1}^{n}[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)]$$

만약 $Y \sim f(y;\theta)$ 라면 아래의 식을 유도할 수 있다.

$$E\{\frac{\partial}{\partial\theta}logf(Y;\theta)\} = 0 \quad (3)$$

$$-E\{\frac{\partial^2}{\partial\theta^2}logf(Y;\theta)\} = E[\{\frac{\partial}{\partial\theta}logf(Y;\theta)\}] \quad (4)$$

GLM density에 대해서 (3)으로부터 $\mu_i = E(Y_i) = b'(\theta)$ 를 유도할 수 있다.

또한, (4)로부터 $Var(Y_i) = b''(\theta_i)a(\phi)$ 식을 유도할 수 있다.

2.2 GLM (Generalized Linear Model)

또한, (4)로부터 $Var(Y_i) = b''(\theta_i)a(\phi)$ 식을 유도할 수 있다.

GLM은 $\eta_i = g(\mu_i) = \boldsymbol{\beta}\mathbf{x}_i$와 같이, link function인 $g(\cdot)$를 사용해서 $\eta_i = \boldsymbol{\beta}\mathbf{x}_i$와 $\mu_i = E(Y_i)$를 연결한다. $\mu_i = b'(\theta_i)$, $hence$ $g = b'(\theta_i)^{-1}$이고, 이때 g를 canonical link라 부른다.

$$\sum_{i=1}^{n} logf(y_i;\theta_i,\phi) = \sum_{i=1}^{n}[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)] \qquad (5)$$

likelihood가 (5)와 같기 때문에, 이를 베타에 대해 미분한 결과를 0으로 두고 계산한다.

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i \frac{\partial L_i}{\partial \beta_j} = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)}(\frac{\partial \mu_i}{\partial \eta_i}) = 0 \quad , \quad \mu_i = g^{-1}(\sum_j \beta_j x_{ij})$$

2.2.1 로지스틱 회귀 (Logistic Regression)

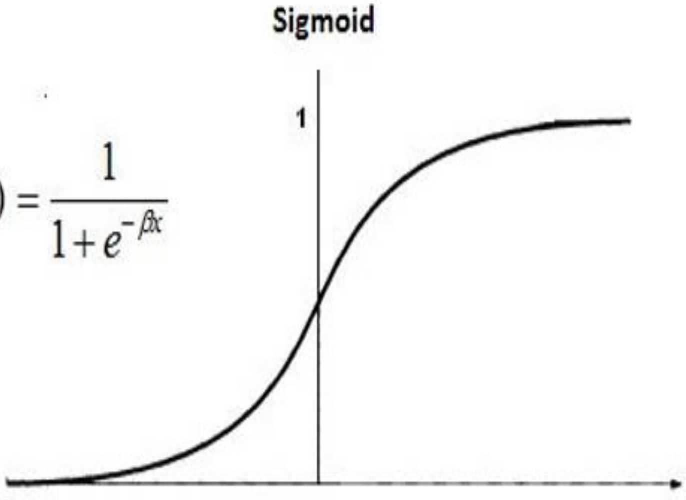
: 종속변수가 0과 1의 값을 갖는 이항변수일 때 사용하는 모형 ($\hat{y} = P(y = 1)$)

$$y = p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \quad 0 < p(\mathbf{x}) < 1$$

$$ln \frac{\mathrm{p}(\mathbf{x})}{1 - \mathrm{p}(\mathbf{x})} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \quad 0 < p(\mathbf{x}) < 1$$

$$p(\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}} = \frac{1}{1 + e^{-\beta\mathbf{x}}}, \quad 0 < p(\mathbf{x}) < 1$$

$$f(x) = \frac{1}{1 + e^{-\beta x}}$$



2.2.1 로지스틱 회귀 (Logistic Regression)

만약 $n_i y_i \sim^{iid} Binomial(n_i, p_i)$ 이라면,

$$f(y_i; p_i, n_i) = \binom{n_i}{n_i y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - n_i y_i} = exp[\frac{y_i\theta_i - log\{1+exp(\theta_i)\}}{1/n_i} + log\binom{n_i}{n_i y_i}]$$

where $\theta_i = log\{\frac{\pi_i}{1-\pi_i}\}$, $b(\theta_i) = log\{1+exp(\theta_i)\}$, $\quad and \quad a(\phi) = 1/n_i$

$$E(Y_i) = b'(\theta_i) = \frac{\partial}{\partial\theta_i} log\{1+exp(\theta_i)\} = \frac{exp(\theta_i)}{1+exp(\theta_i)} = \pi_i$$

$$Var(Y_i) = b''(\theta_i)a(\phi) = \frac{exp(\theta_i)}{\{1+exp(\theta_i)\}^2 n_i} = \pi_i(1-\pi_i)/n_i$$

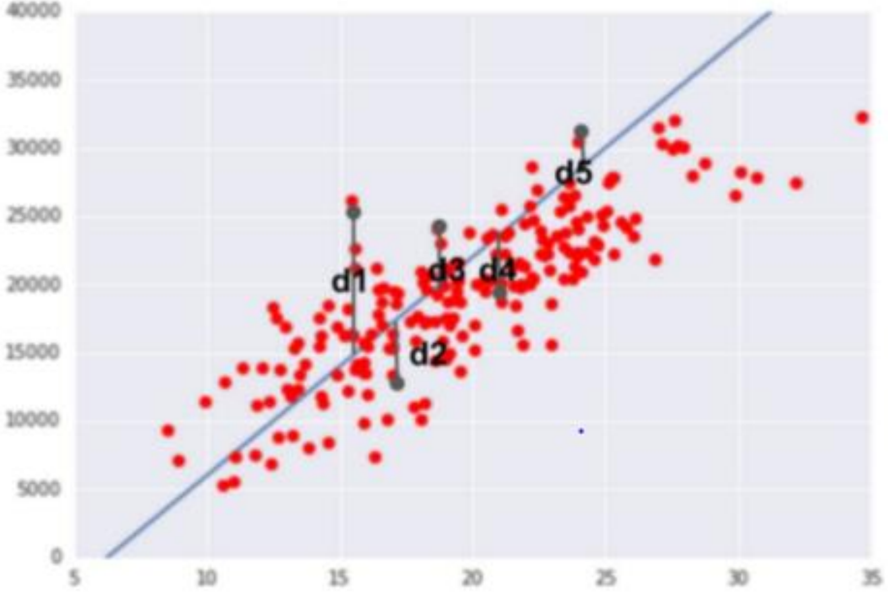
$$\eta_i = g(\mu_i) = \theta_i = (b')^{-1}(\mu_i) = log\frac{\pi_i}{1-\pi_i} = \boldsymbol{\beta}\mathbf{x}_i$$ (link function)

3. Cost function

: 실제값과 예측값의 차이에 대한 함수

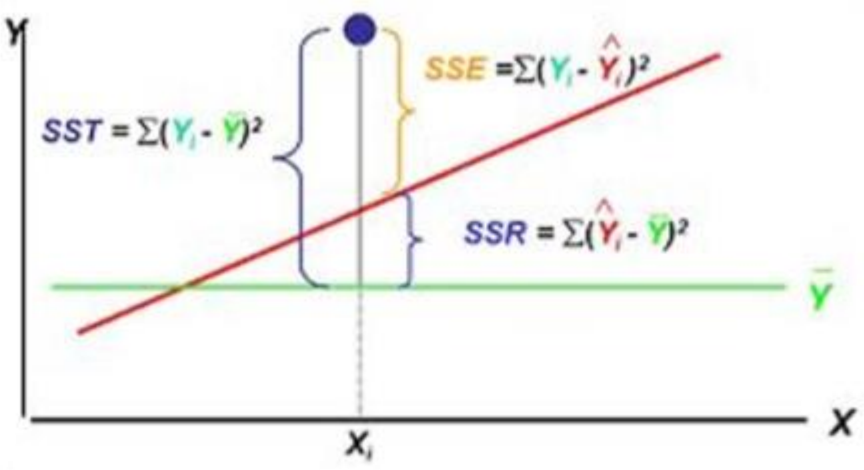
3.1 MSE (Mean Square Error)

$$\text{residual} = y - \hat{y}$$



$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

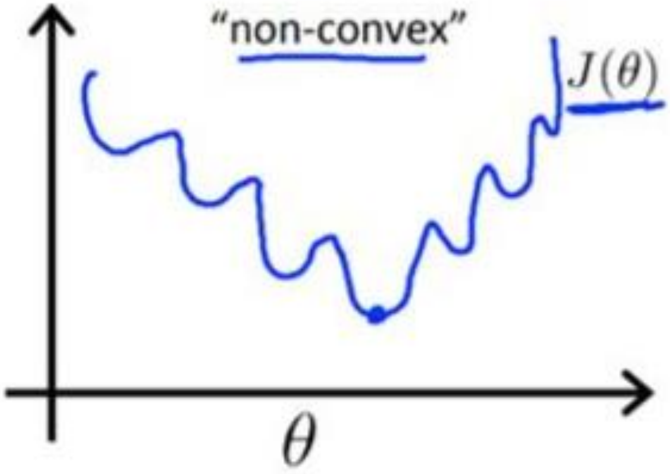
3.1 MSE (Mean Square Error)

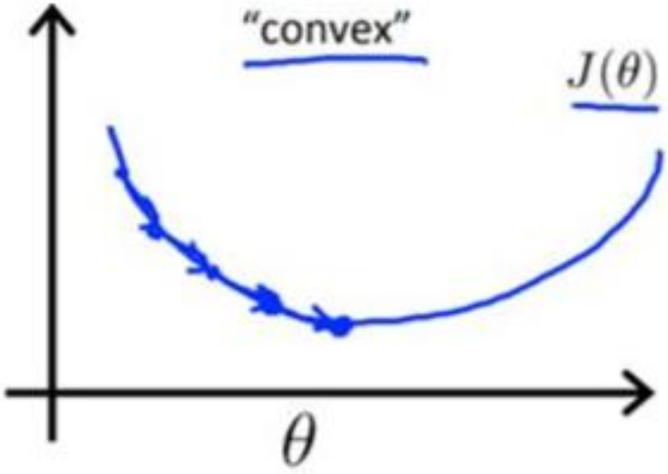


$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y - \hat{y})^2$$

3.2 Cross Entropy

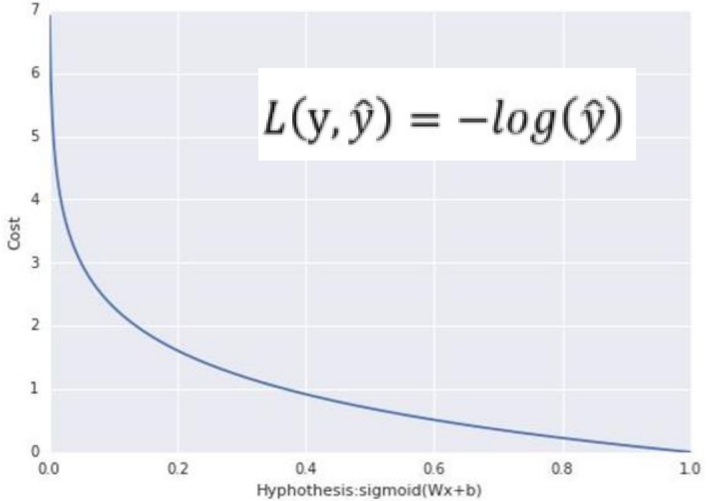


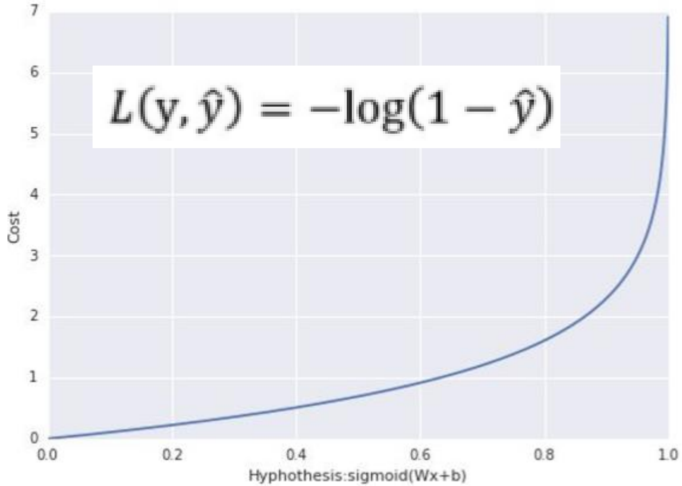


3.2 Cross Entropy

$$Cross\ Entropy = -\frac{1}{m}\left[\sum_{i=1}^{m} y^i \cdot log\widehat{y^i} + (1 - y^i) \cdot lo\,g(1 - \widehat{y^i})\right]$$

$$L(y, \hat{y}) = -[y\log(\hat{y}) + (1 - y)\log(1 - \hat{y})]$$

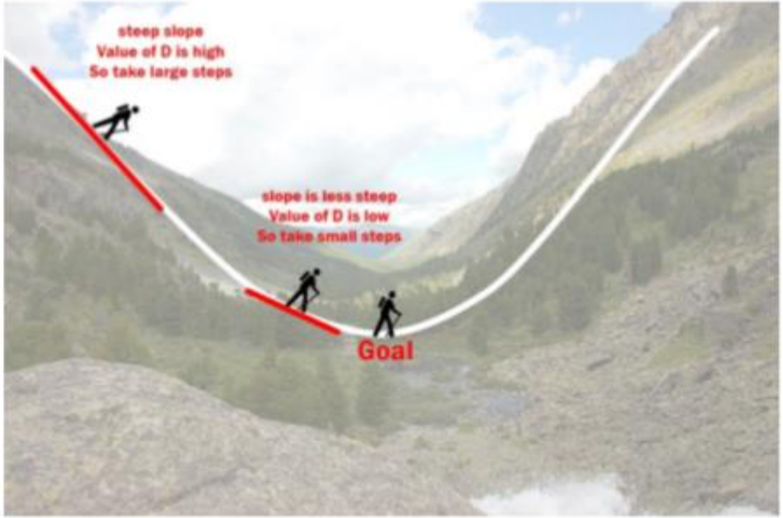




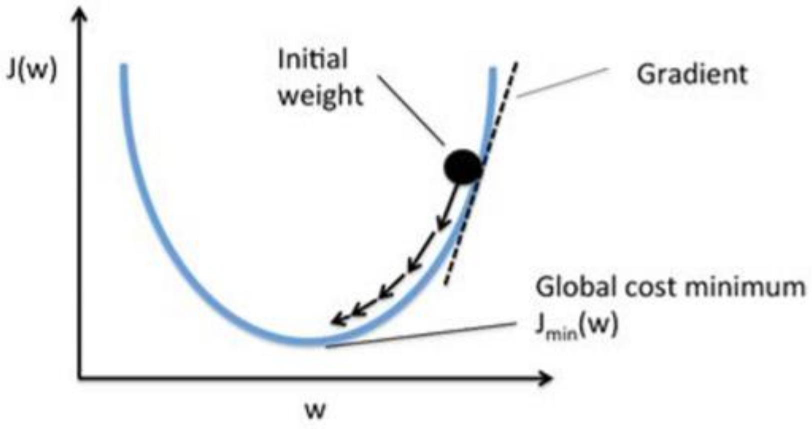
4. Optimization: Gradient Descent

Optimization : 어떤 목적함수의 함숫값을 최적화하는 모수를 찾는 문제
Gradient Descent : 기울기가 0일 부분을 찾아가는 알고리즘

$$
\begin{aligned}
&\textit{Repeat } \{ \\
&\quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\
&\quad (\textit{simultaneously update all } \theta_J) \\
&\}
\end{aligned}
$$



4. Optimization: Gradient Descent



KU-BIG