

20기 정규세션

ToBig's 19기 강의자

오유진

Regression Analysis

회귀분석

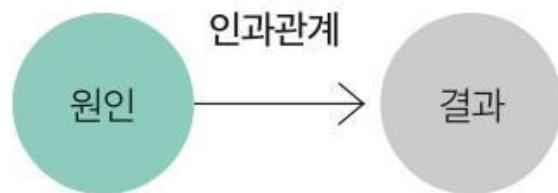
Intro

1. 머신러닝 알고리즘

지도학습 (Supervised Learning)	비지도학습 (Unsupervised Learning)	강화학습 (Reinforcement Learning)
<ul style="list-style-type: none">- 입력과 결과값(Label) 이용한 학습- 회귀(Regression)- 분류(Classification)	<ul style="list-style-type: none">- 입력만을 이용한 학습- 군집화(Clustering)	<ul style="list-style-type: none">- Agent가 주어진 State에서 Action을 취했을 때, 이로부터 얻는 Reward를 최대화하는 방향으로 학습
Ex) 선형회귀, 로지스틱 회귀, KNN, SVM, Decision Tree	Ex) K-Means Clustering	

Intro

2. 인과관계 VS 상관관계



원인이 있었기 **때문에** 결과가 생겨났다.



원인과 결과의 관계가 아니다.

인과관계(Causality)

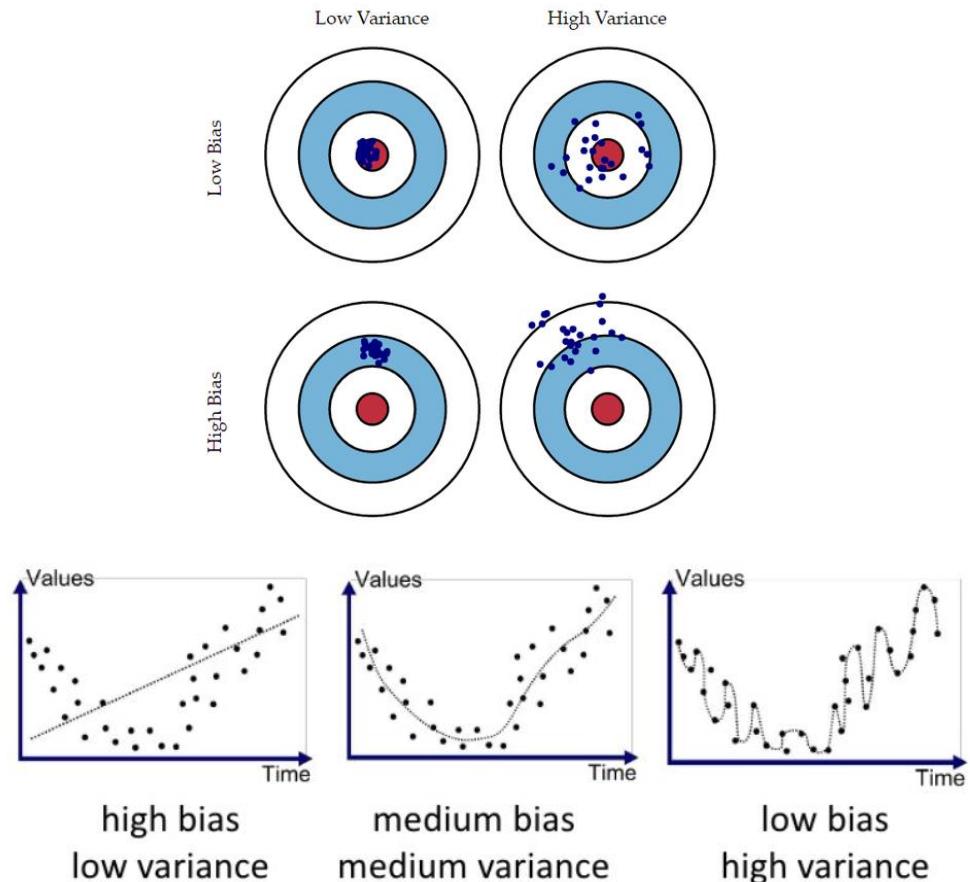
- 어떤 사실과 다른 사실 사이의 원인과 결과 관계

상관관계 (Association, Correlation)

- 두 변량 중 한쪽이 증가함에 따라, 다른 한쪽이 증가하거나 감소하는 관계 (선형적 관계)
- 상관관계가 존재할 때, **필연적으로** 인과관계가 존재하는 것은 **아님**

Intro

3. 편향(Bias) VS 분산(Variance)



Bias(편향)

- 데이터 내 모든 정보를 고려하지 않기에 알고리즘이 지속적으로 잘못된 내용을 학습하는 경향성
- Underfitting과 관련

Variance(분산)

- Highly flexible model에 데이터를 fit함으로써, 실제 현상과 관계 없는 random한 것들까지 학습하는 알고리즘의 경향성
- Overfitting과 관련

Contests

Unit 01 | 선형 회귀분석

Unit 02 | 회귀 진단

Unit 03 | 로지스틱 회귀분석

Unit 04 | 최대우도추정 & 평가지표

Unit 01 | 선형 회귀분석

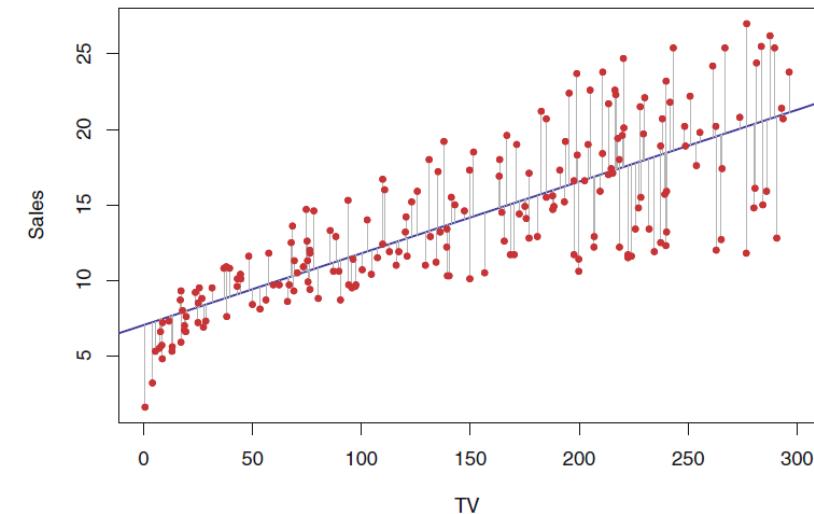
선형 회귀분석(Linear Regression)

“회귀분석”

- 설명변수 (X)에 대응하는 반응변수 (Y)와 가장 비슷한 값 (\hat{Y})을 출력하는 함수를 찾는 과정
- 변수들의 관계를 기술하고 형태를 파악하는 통계적인 기법

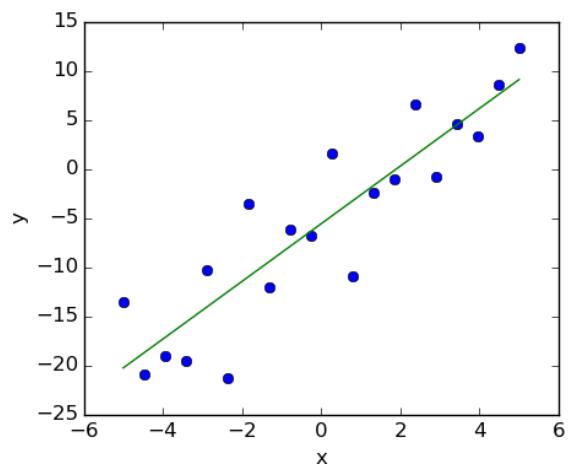
“선형회귀분석”

- 반응변수와 한 개 이상의 설명변수와의 선형 상관관계를 모델링하는 회귀분석 기법
- Ex) 해당 연도 수확량(X)에 따른 열매 개수(Y)



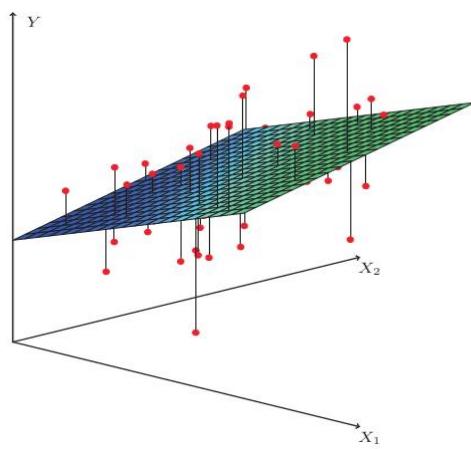
Unit 01 | 선형 회귀분석

단순 선형 회귀



$$y = \beta_0 + \beta_1 x + \varepsilon$$

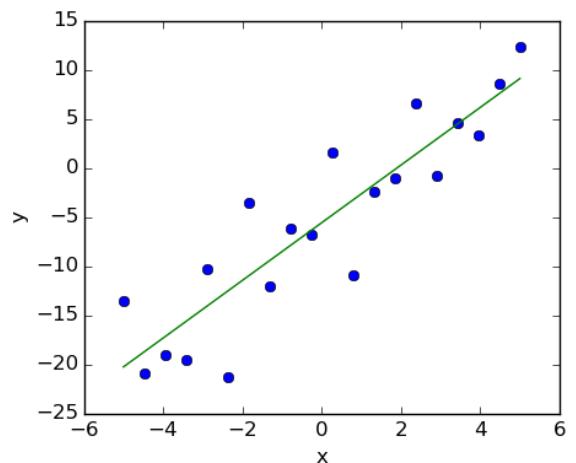
다중 선형 회귀



$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

Unit 01 | 선형 회귀분석

단순 선형 회귀



$$y = \beta_0 + \beta_1 x + \varepsilon$$

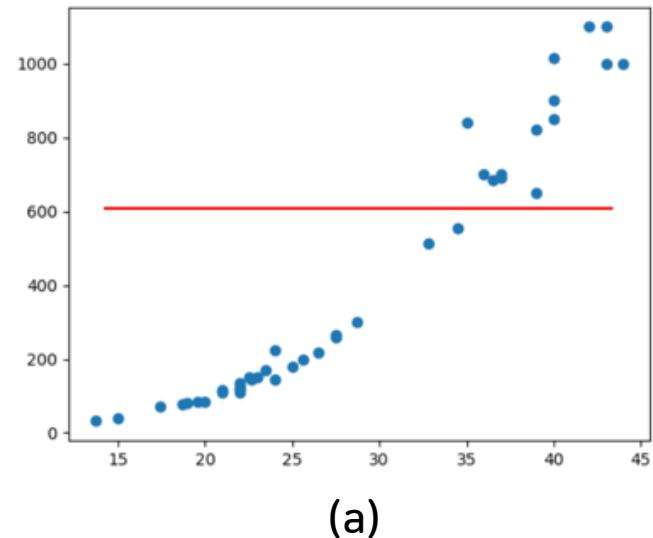
Formulation : $\hat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

- β_0, β_1 : 회귀계수

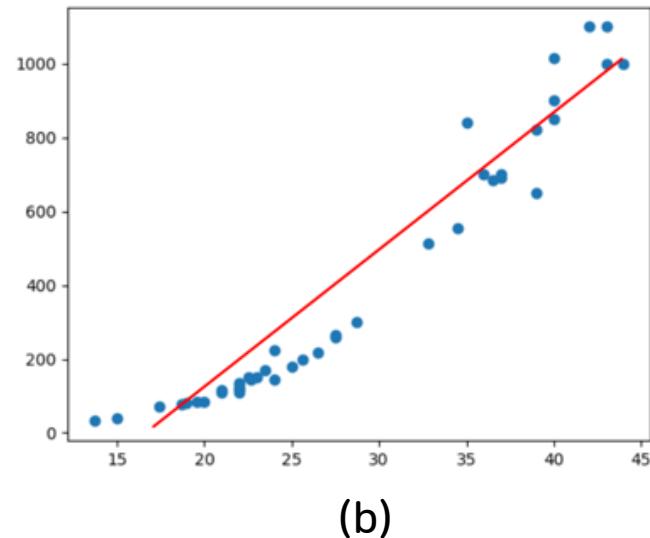
- $\widehat{\beta}_0, \widehat{\beta}_1$: 예측된 회귀계수

Unit 01 | 선형 회귀분석

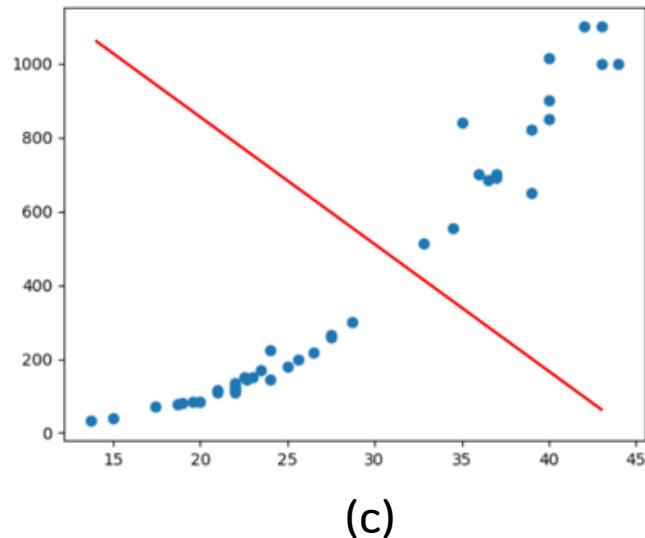
Optimal한 선형관계를 어떻게 알아낼 수 있을까 ?



(a)



(b)



(c)

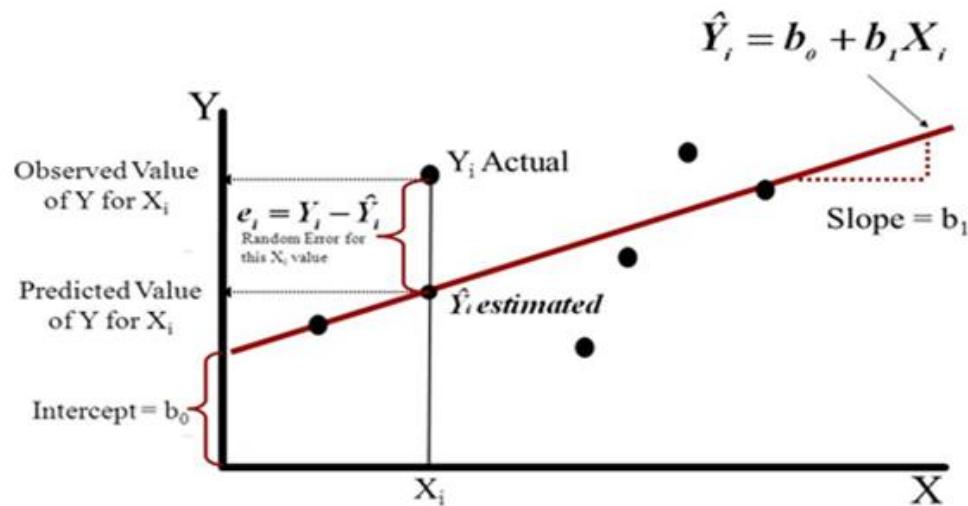
(b)가 가장 optimal한 직선임을 볼 수 있다.
→ Loss (실제값과 예측값의 차이) 최소화

Unit 01 | 선형 회귀분석

Least Square Method(LSE) - 단순선형회귀

- Loss를 최소화시키기 위한 방법

Simple Linear Regression Model

잔차($e = y - \hat{y}$)의 제곱합을 최소화

최적의 회귀계수(모수) 추정

$$L = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

Loss Func.

목적함수 값이 작을수록 좋은 모델

Unit 01 | 선형 회귀분석

Least Square Method(LSE) -단순선형회귀

$$L = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

↓ Partial differential for minimization

Normal Equation(정규방정식)

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))x_i = 0$$

→
Result

Least Squares Estimator(최소제곱 추정치)

$$\widehat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Unit 01 | 선형 회귀분석

Least Square Method(LSE) -다중선형회귀

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_k x_{1k} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_k x_{2k} + \varepsilon_2$$

⋮

⋮

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_k x_{nk} + \varepsilon_n$$



$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$



$$y = X\beta + \varepsilon \quad \varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$$

$$\sum_{i=1}^n \varepsilon_i^2 = \varepsilon' \varepsilon = (y - X\beta)'(y - X\beta)$$

$$= y'y + \beta'X'X\beta - 2\beta'X'y$$

↓ Partial differential for beta

$$\frac{\partial L}{\partial \beta} = 2X'X\beta - 2X'y = 0$$

$$\Rightarrow X'X\beta = X'y$$

$$\Rightarrow \beta = (X'X)^{-1}X'y$$

정규방정식

최소제곱
추정치

Unit 01 | 선형 회귀분석

Partition of Sum of Squares(제곱합 분해)

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

<SST> <SSR> <SSE>

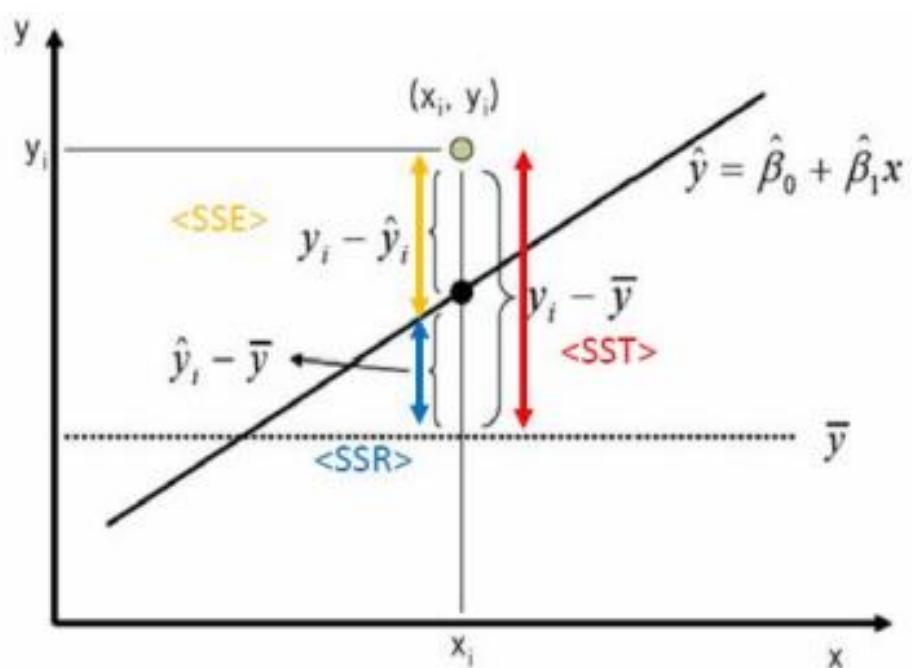
SST : 총 제곱합

SSR : 회귀 제곱합 (전체 제곱합 중 회귀식으로 설명가능)

SSE : 잔차 제곱합 (전체 제곱합 중 회귀식으로 설명불가)

→ 회귀식이 데이터를 잘 나타낼수록 SSR↑ SSE↓

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$



Unit 01 | 선형 회귀분석

Partition of Sum of Squares(제곱합 분해)

$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$
 $F > F(\alpha, p, n - p - 1)$ 이면 H_0 기각

	자유도	제곱합(SS)	제곱평균(MS)	F
회귀 SSR	p	SSR	$MSR = SSR / p$	MSR/MSE
잔차 SSE	$n-(p+1)$	SSE	$MSE = SSE / n-p-1$	
총 SST	n-1	$SST = SSR + SSE$		

회귀식이 설명하지 못하는 부분
 → 작을수록 Good !

단순선형회귀분석(p=1에서 잔차의 제약조건은 2개 !)
 → 자유도는 n-2

Contests

Unit 01 | 선형 회귀분석

Unit 02 | 회귀 진단

Unit 03 | 로지스틱 회귀분석

Unit 04 | 최대우도추정 & 평가지표

Unit 02 | 회귀 진단

회귀진단 (Regression Diagnostics)

- 1) 회귀모형의 가정이 타당한가?
- 2) 각각의 관측값이 모형 및 가정에 어떠한 영향을 미치는가?

[회귀모형 기본 가정]

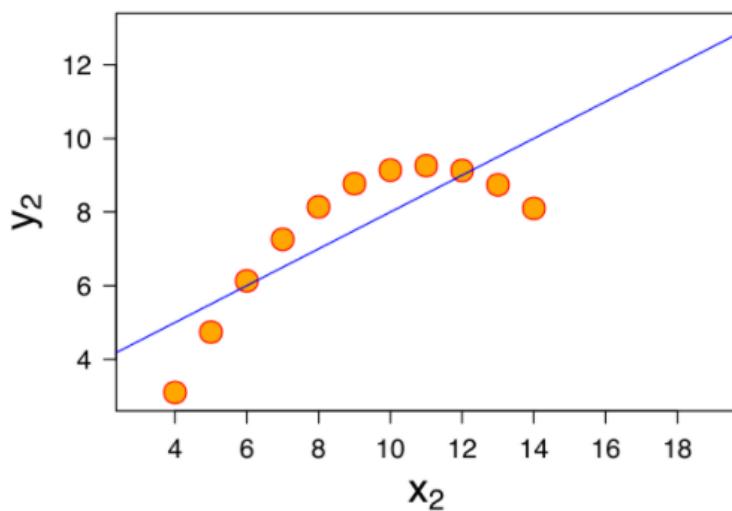
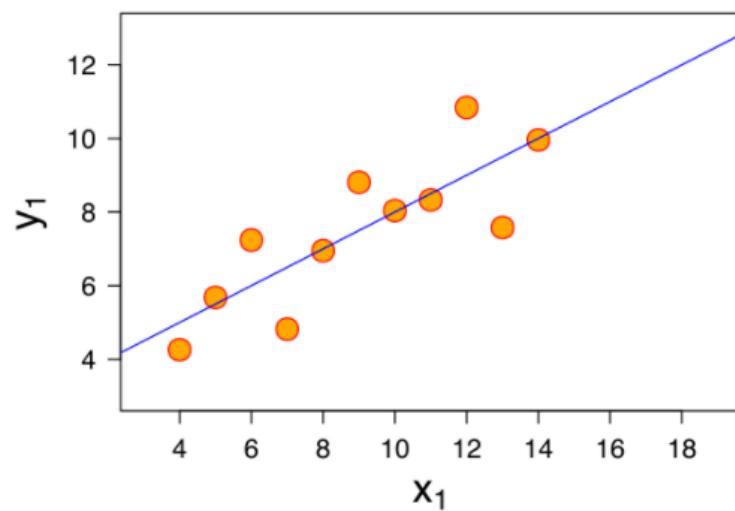
1. 선형성(Linearity) : 설명변수(X)와 반응변수(Y) 간 선형관계
2. 정규성(Normality) : 오차(Error)의 정규성
3. 등분산성(Homoscedasticity) : 오차의 등분산성
4. 독립성(Independence) : 오차의 독립성



Unit 02 | 회귀 진단

그래프적 방법

1. 선형성(설명변수와 반응변수 간 선형관계) 판단

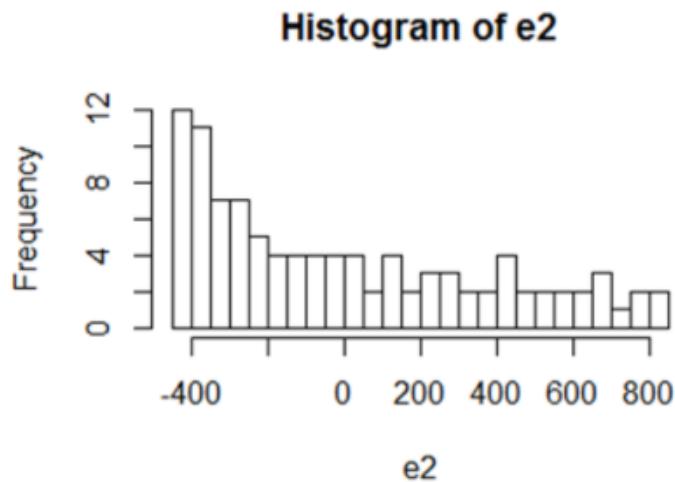
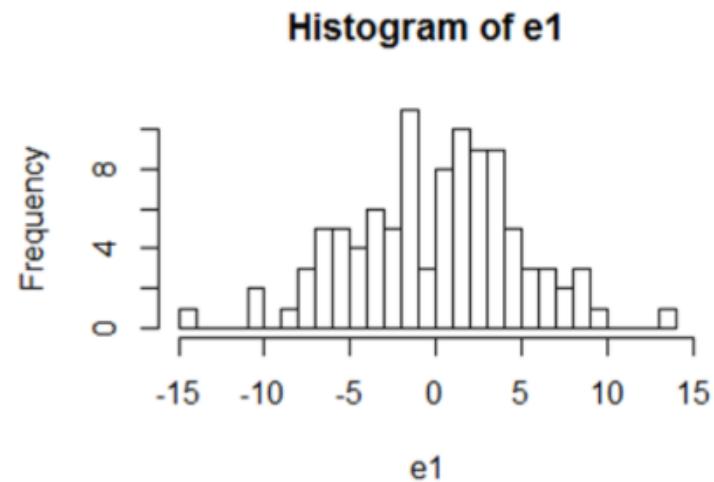


- **산점도**(Scatter plot)을 통해 선형성 판단 가능
- x_1 과 y_1 간에는 선형관계가 존재하지만, x_2 와 y_2 사이엔 선형관계가 있다고 보기 어려움

Unit 02 | 회귀 진단

그래프적 방법

2. 정규성(오차가 정규분포를 따르는지) 판단



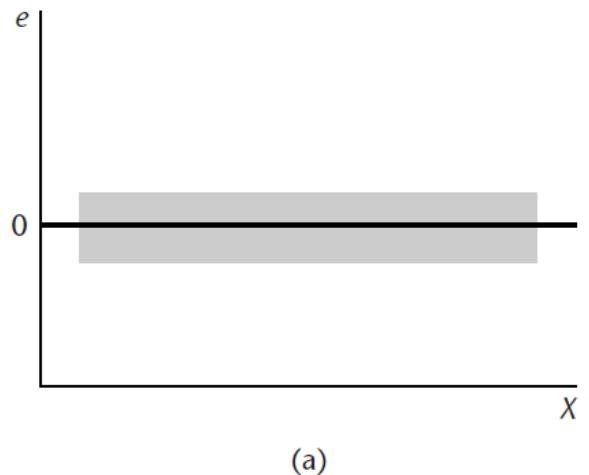
✓ 용어정리
모수 → 오차
표본 → 잔차

- 잔차의 히스토그램 → 오차의 정규성 파악 가능
- e1은 정규성 가정을 만족하고, e2은 정규성 가정을 위배한다고 볼 수 있음
- [R] Shapiro-Wilk Normality Test

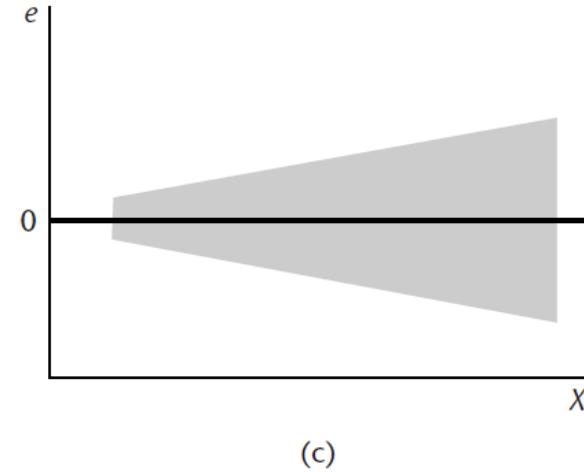
Unit 02 | 회귀 진단

그래프적 방법

3. 등분산성(오차의 분산이 일정한지) 판단



(a)



(c)

설명변수에 대한 그림으로 오차의 등분산성 판단 가능

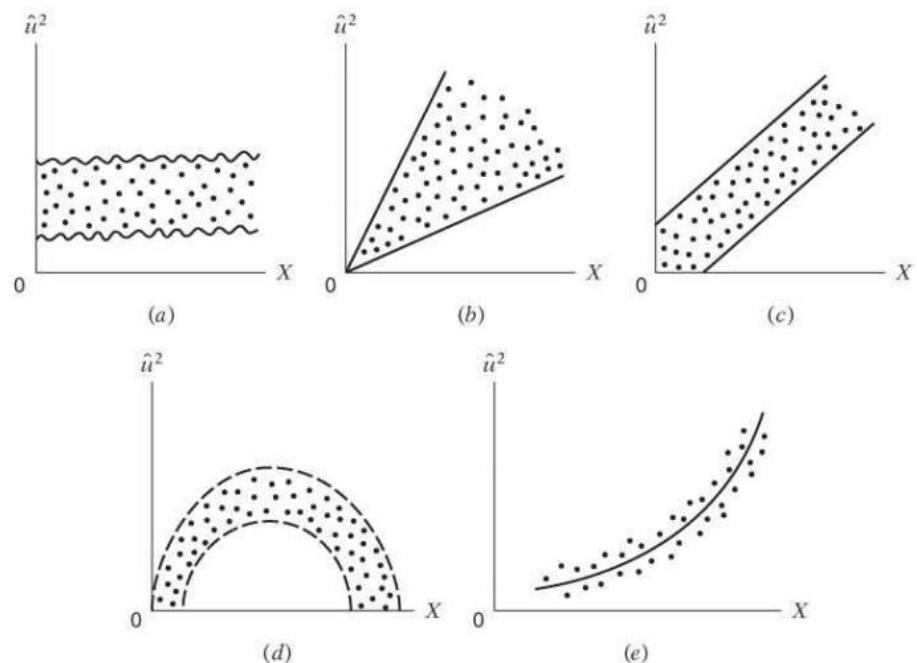
band width가 일정해서
등분산성 가정 만족

band width가 넓어져서
등분산성 가정을 만족 X

Unit 02 | 회귀 진단

그래프적 방법

4. 독립성(오차가 서로 독립인지) 판단



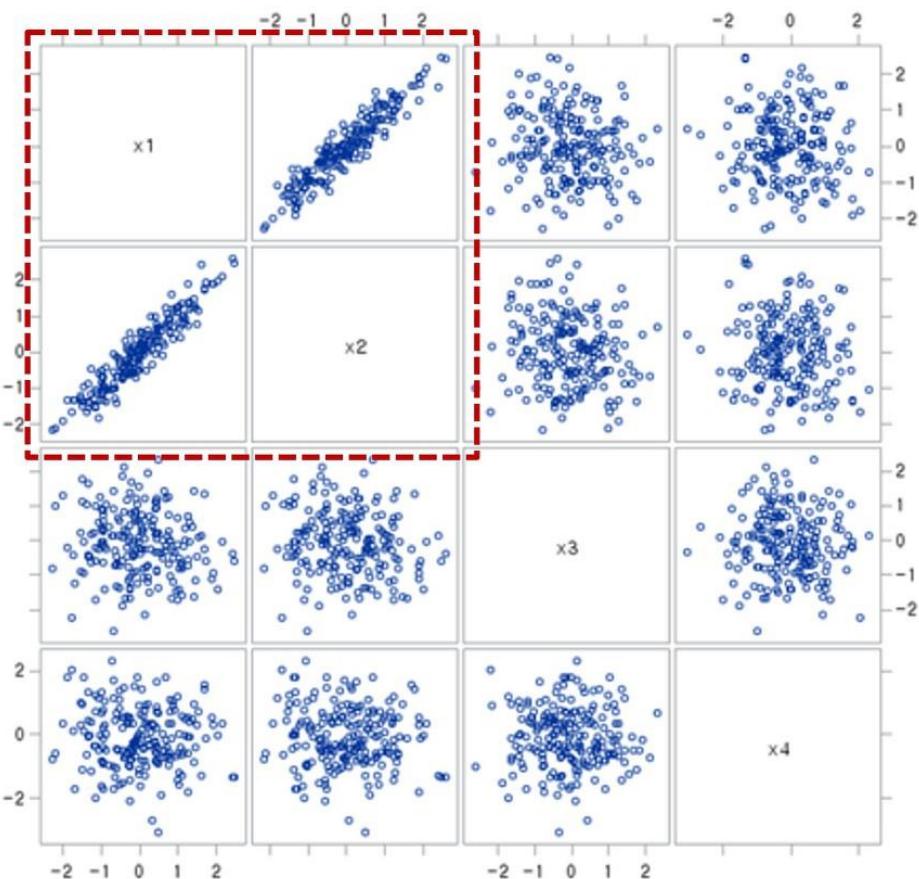
- 설명변수와의 상관성과 자기상관성 확인 → 독립성 판단
- 잔차에 패턴이 존재한다면 독립적이지 않음
- a : 잔차에 어떤 패턴도 X → 독립적이라 판단 가능
- Durbin-Watson 검정, ACF

Unit 02 | 회귀 진단

다중공선성(Multicollinearity)

- 독립변수들 간에 강한 상관관계가 나타나는 문제
- 회귀분석의 전제가정인 **독립성**을 위배
- 상관관계 분석, 설명변수 산점도, heatmap으로 확인
- 다중공선성 진단통계량 : **VIF** (Variance Inflation Factor)
 - VIF > 10** 이면 다중공선성이라고 판단
 - 다른 설명변수들과 상관관계가 강할수록 VIF ↑

$$VIF_i = \frac{1}{1 - R_i^2}$$



Unit 02 | 회귀 진단

다중공선성(Multicollinearity) 제거

제거 이유

- 설명변수 간 독립적이지 않으면 회귀계수의 추정이 불안정함.
- 추정값이 존재하지 않거나, 추정값의 분산이 커지는 문제점을 가져올 수 있음.

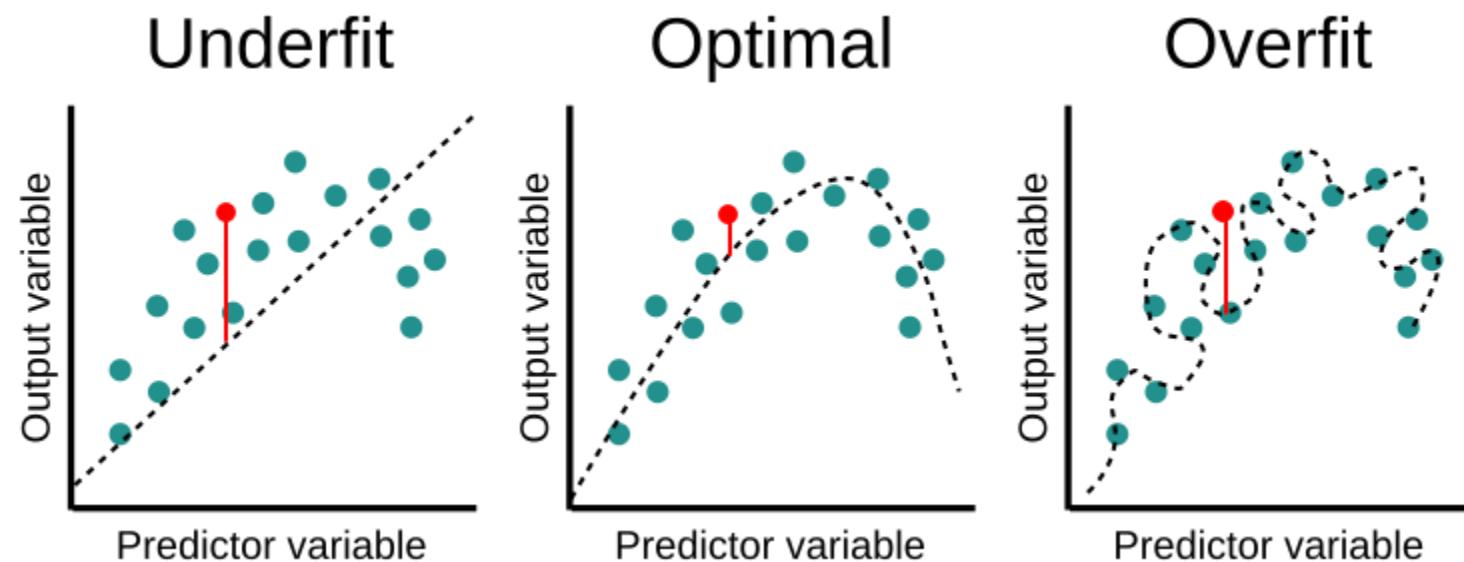
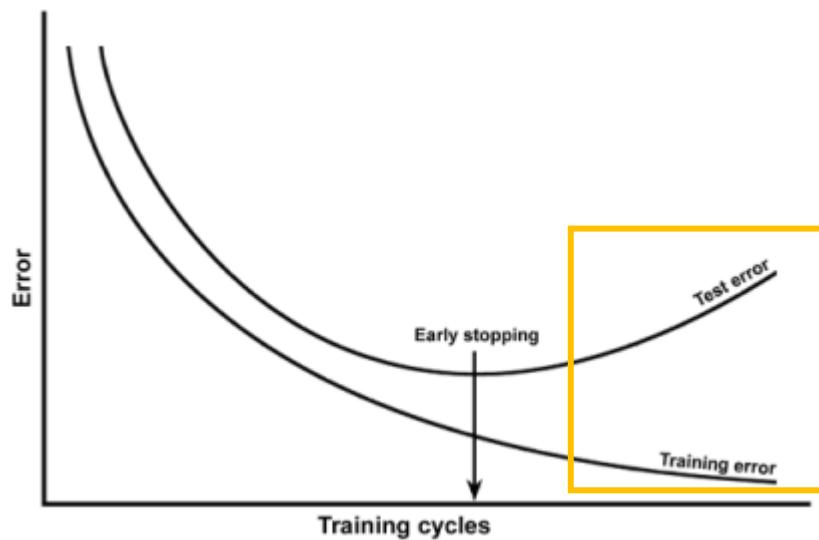
제거 방법

- 더 많은 데이터 수집
- 다중공선성을 유발하는 주요 변수 2개를 찾아, 각 변수 제거시 R-squared의 변동 확인하여 제거해도 결정계수가 유지되는 변수 제거
- PCA(Principal Component Analysis, 주성분 분석) → 차원 축소
- Ridge / Lasso Regression

Unit 02 | 회귀 진단

과적합(Overfitting)

- 학습 데이터에 과하게 학습하여 실제 데이터에 대한 오차가 증가하는 현상



Unit 02 | 회귀 진단

정규화(Regularization)

- 모델이 복잡해질수록 **penalty**를 크게 하고자 목적함수에 항을 하나 더 추가
- 과적합된 모델을 일반성을 갖추도록 하기 위하여 사용

Ridge
Regression

Lasso
Regression

ElasticNet
Regression

Unit 02 | 회귀 진단

Ridge Regression(L2 Regression)

$$\beta_1, \beta_2, \dots, \beta_p$$

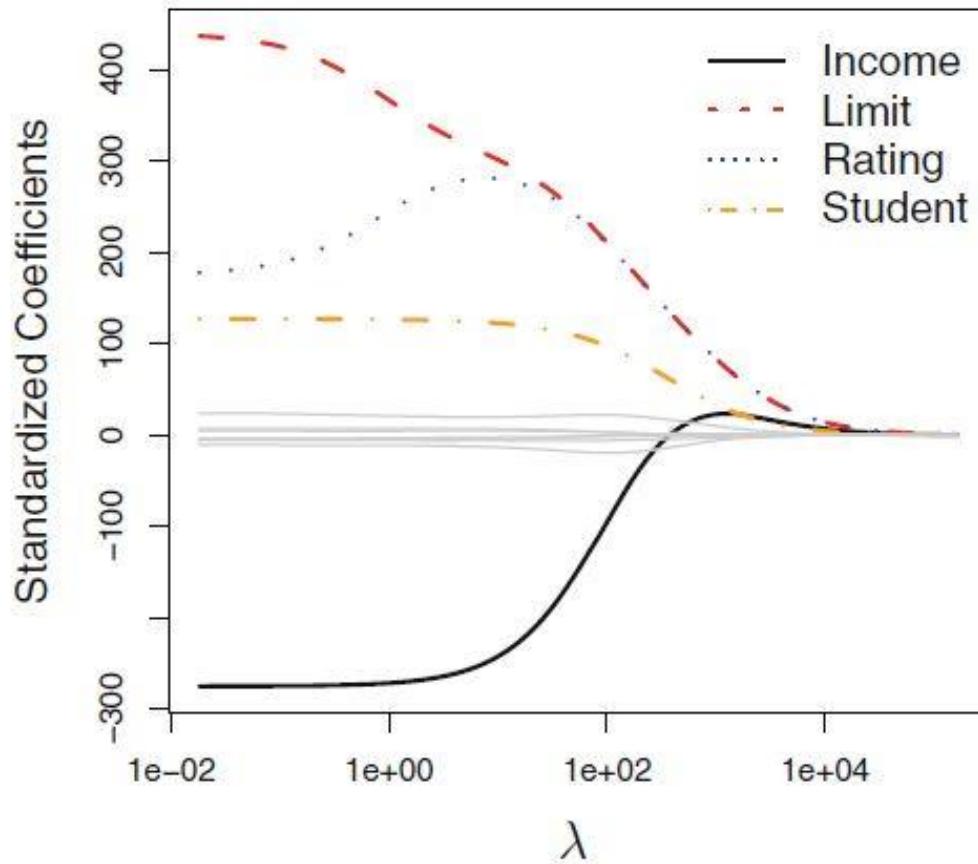
$$L(\beta) = \min_{\beta} \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{(1) \text{ Training accuracy}} + \lambda \underbrace{\sum_{j=1}^p \beta_j^2}_{(2) \text{ Generalization accuracy}}$$

- Training Accuracy에 Generalization Accuracy를 추가
➤ 회귀계수(β)에 제약을 줄 수 있게 됨

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^p \beta_j^2$$

Unit 02 | 회귀 진단

Ridge Regression(L2 Regression)



$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^p \beta_j^2$$

- $\lambda \uparrow \rightarrow$ 계수를 많이 줄이는데 집중
 - $\lambda \downarrow \rightarrow$ 기존 최소 제곱법 문제
 - β^2 을 사용하기 때문에 완전히 0으로 수렴하지 X
- ✓ 변수의 크기가 결과에 큰 영향을 미치기 때문에,
변수를 **스케일링**을 해주는 작업이 필요할 수 있다.

Unit 02 | 회귀 진단

Lasso Regression(L1 Regression)

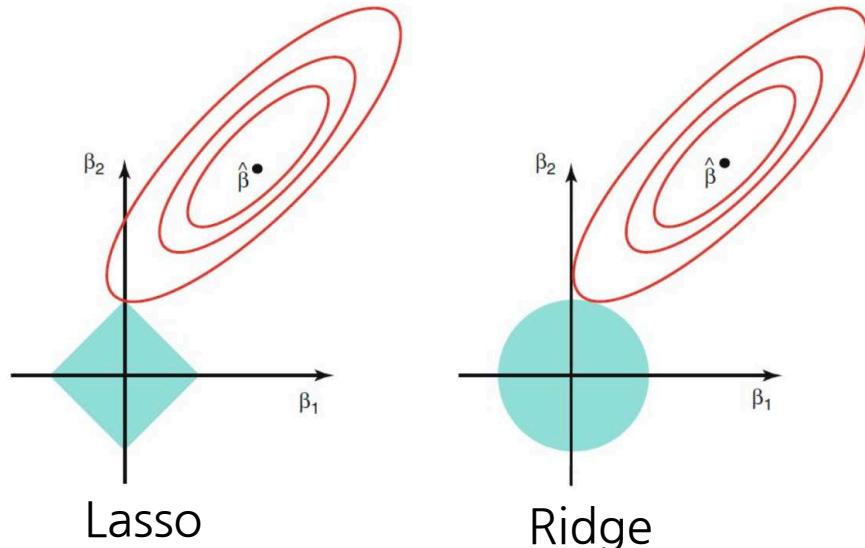
- Ridge Regression과 다른점은 패널티 항에 절대값의 합을 주었다는 것!

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^p |\beta_j|.$$

라그랑지 승수를 없애고 제약식이 있는 최적화 문제로 변경

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq s$$

- 최적값은 모서리 부분에서 나타날 확률이 릿지에 비해 높아 몇몇 유의미하지 않은 변수들에 대해 계수를 0에 가깝게 추정
- 작은 값의 파라미터를 0으로 만들어 해당 변수를 삭제한다는 점이 차이점



Unit 02 | 회귀 진단

선형 회귀분석 마무리

1. 회귀모형 설정 : 반응변수 및 주요 설명변수 파악
2. 선형성 검토 : 산점도를 통해 상관관계 파악
3. 설명변수 검토 : 각 설명변수 분포 확인 + 다중공선성 점검
4. 모델 적합 : 모델 회귀계수 추정 및 모형 적절성 검토
5. 변수 선택 : 주요 설명변수 선택
6. 적합된 모형 검토 : 오차에 대한 기본 가정 확인
7. 최종 모형 선택

CONTENTS

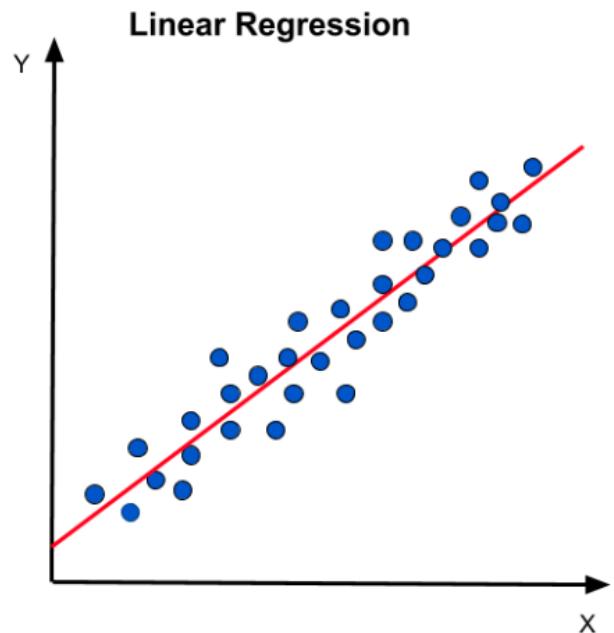
Unit 01 | 선형 회귀분석

Unit 02 | 회귀 진단

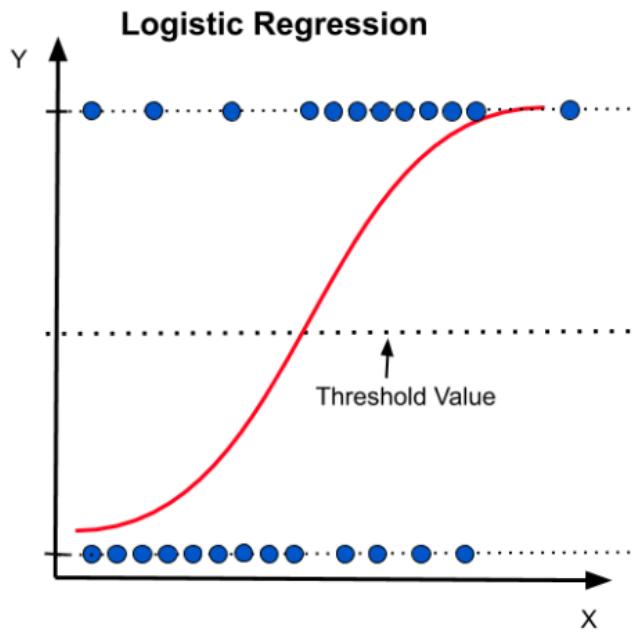
Unit 03 | 로지스틱 회귀분석

Unit 04 | 최대우도추정 & 평가지표

Unit 03 | 로지스틱 회귀분석



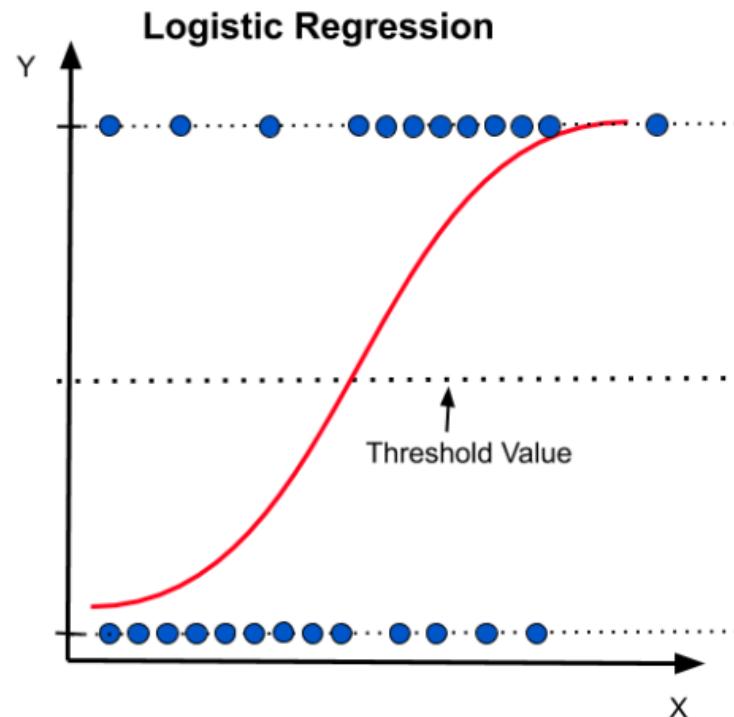
Y가 **연속형** 변수일 경우 적합
Regression



Y가 **범주형** 변수일 경우 적합
Classification

Unit 03 | 로지스틱 회귀분석

Logistic Regression(로지스틱 회귀분석)



- 범주형 데이터를 대상으로 하는 회귀분석
- 일종의 Classification(분류) 기법
- 새로운 관측치가 들어올 때, 기존 범주 중 하나로 예측

Ex)

1. 제품이 불량인지, 정상인지
2. 고객이 이탈고객인지 잔류고객인지
3. 카드거래가 정상인지 사기인지

Unit 03 | 로지스틱 회귀분석

Odds

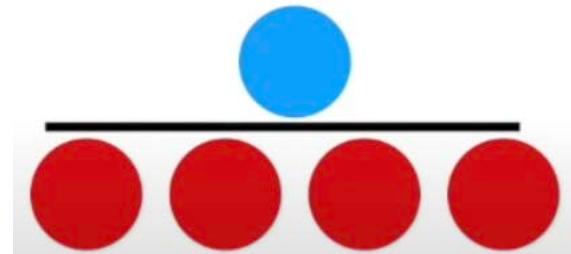
어떤 일이 일어나지 않을 확률(빈도) 대비 어떤 일이 일어날 확률(빈도)

Odds 값의 범위 : [0 , ∞)

Odds 는 확률이 아니다.



남자 1 여자 4



남자일 Odds = 1/4

Unit 03 | 로지스틱 회귀분석

Odds

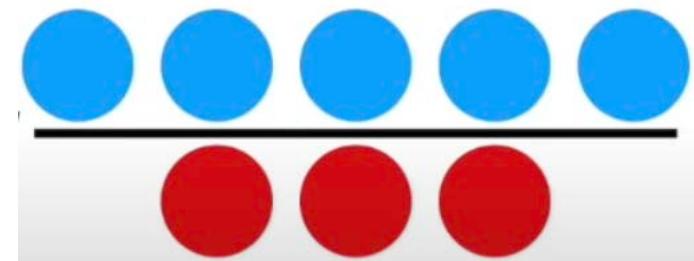
어떤 일이 일어나지 않을 확률(빈도) 대비 어떤 일이 일어날 확률(빈도)

Odds 값의 범위 : [0 , ∞)

Odds 는 확률이 아니다.



남자 5 여자 3



남자일 Odds = 5/3

Unit 03 | 로지스틱 회귀분석

Logistic Regression(로지스틱 회귀분석)

- Y 는 질적변수이므로 일반적인 회귀분석과 달리 회귀계수들과 변수들의 단순 선형결합으로 표현될 수 없다
 $\rightarrow Y=1$ 일 확률을 구하는 방향으로 접근한다.

주어진 $X=x$ 에 대해 $Y=1$ 일 조건부 확률을 π 라고 하면

$$\pi := P(Y = 1|X = x) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}$$

- (i) 지수함수는 항상 0보다 크고, π 의 분모가 분자보다 크므로 $0 < \pi < 1$ 이다.
- (ii) 자연스럽게 $Y = 0$ 일 확률은 $1 - \pi$ 이다.
- (iii) $Y = 0$ 일 확률 대비 $Y = 1$ 일 Odds는

$$\frac{\pi}{1 - \pi} = \frac{\frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}}{\frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}$$

Unit 03 | 로지스틱 회귀분석

Logistic Regression(로지스틱 회귀분석)

- Y 는 질적변수이므로 일반적인 회귀분석과 달리 회귀계수들과 변수들의 선형결합으로 표현될 수 없다
 $\rightarrow Y = 1$ 일 확률을 구하는 방향으로 접근한다.
- $0 < \pi < 1$ 이므로 $\pi / (1 - \pi)$ 의 범위는 $[0, \infty)$ 이다.

Odds
$$\frac{\pi}{1 - \pi} = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}$$

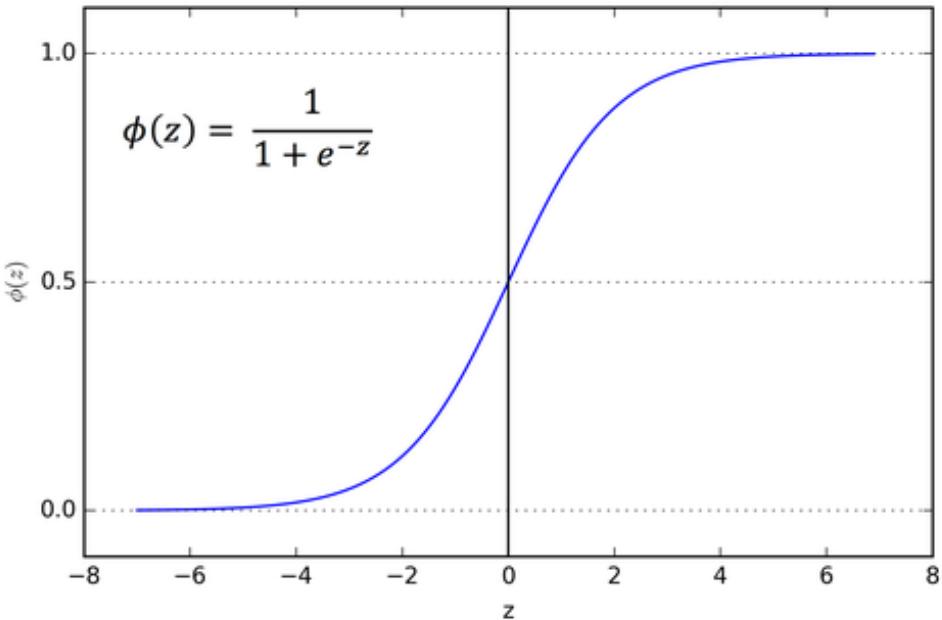
- 양변에 로그를 취하면 (로짓 변환), 범위가 $(-\infty, \infty)$ 로 변경된다.

$$\ln\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = z$$

- 설명변수(x_1, \dots, x_p)의 정보가 담겨있는 z 값을 이용하여 y 값(0 or 1)을 예측하자

Unit 03 | 로지스틱 회귀분석

Logistic Function(로지스틱 함수)



- 정의역 : 실수 전체
- Output : (0,1)

- 음의 무한대에서 양의 무한대까지의 실수값을 0~1 사이의 실수값으로 일대일대응 시키는 시그모이드 함수

$$\pi := P(Y = 1|X = x)$$

로짓변환 → 출력값의 범위 $(-\infty, \infty)$

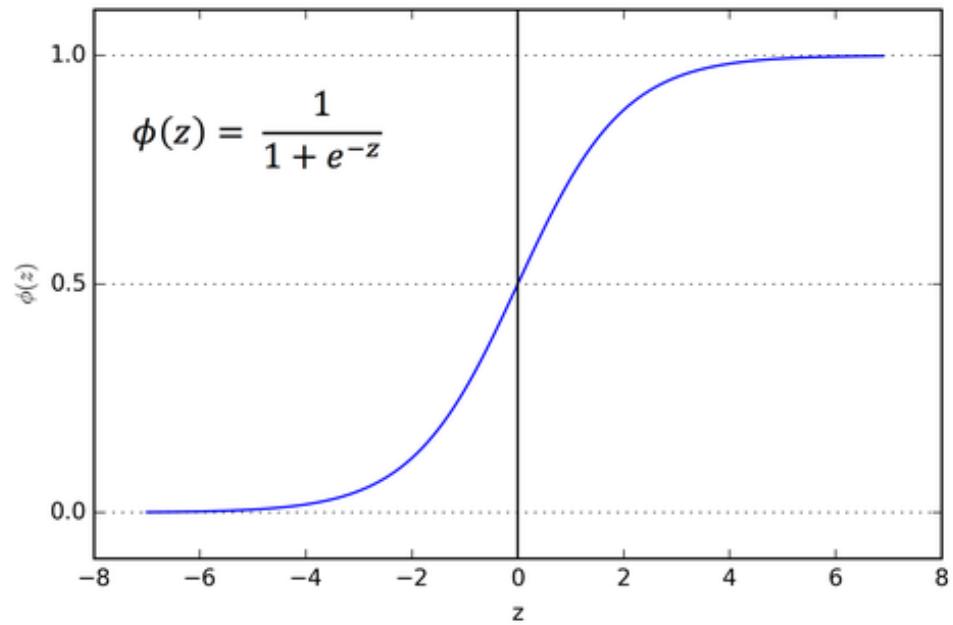
$$z = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

Odds : 실패 확률 대비
성공 확률 비율 $[0, \infty]$

$$\text{logitstic}(z) = \mu(z) = \frac{1}{1 + \exp(-z)}$$

Unit 03 | 로지스틱 회귀분석

선형 판별함수



$$\text{logistic}(z) = \mu(z) = \frac{1}{1 + \exp(-z)}$$

- 로지스틱 함수를 사용하는 경우,
 - If $z = 0$, $\mu = 0.5$
 - If $z > 0$, $\mu > 0.5 \rightarrow \hat{y} = 1$
 - If $z < 0$, $\mu < 0.5 \rightarrow \hat{y} = 0$

→ Z가 분류 모형의 판별함수(decision function)의 역할을 한다고 볼 수 있음

Unit 03 | 로지스틱 회귀분석

회귀 계수의 해석

- Linear Regression : 설명변수가 1만큼 증가할 때의 반응변수의 변화량

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- Logistic Regression : 설명변수가 1만큼 증가할 때의 log(Odds)의 변화량

$$\log(Odds) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Contests

Unit 01 | 선형 회귀분석

Unit 02 | 회귀 진단

Unit 03 | 로지스틱 회귀분석

Unit 04 | 최대우도추정 & 평가지표

Unit 04 | 최대우도추정 & 평가지표

회귀 계수의 추정

- 선형 회귀분석 → 최소제곱합(LSE) 이용
- 로지스틱 회귀분석 → 최대 우도 추정법 (**MLE**) 이용



MLE (Maximum Likelihood Estimation, 최대 우도법)

- 비선형의 회귀식은 최소제곱합을 사용하여 추정할 수 없음
- Likelihood를 Maximize하는 Parameter를 추정하는 방법

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta | \mathcal{D}) = \operatorname{argmax}_{\theta} \prod_{i=1}^n p(\mathcal{D}_i | \theta)$$

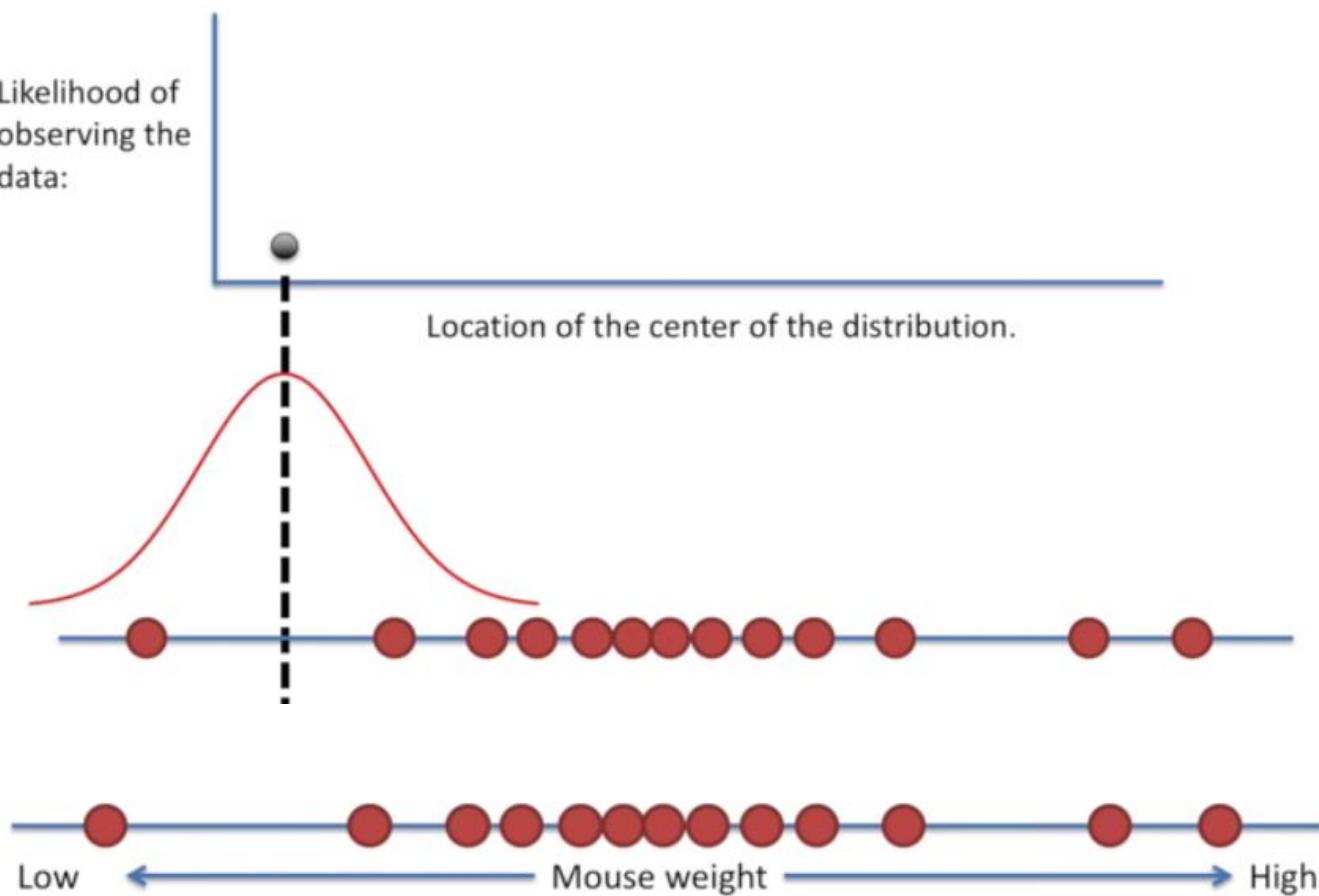
Unit 04 | 최대우도추정 & 평가지표

MLE(최대 우도 추정)

def

각 관측값에 대한 총 가능성(모든 가능성의 곱)이
최대가 되게하는 분포를 찾는 것

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta | \mathcal{D}) = \operatorname{argmax}_{\theta} \prod_{i=1}^n p(\mathcal{D}_i | \theta)$$



Unit 04 | 최대우도추정 & 평가지표

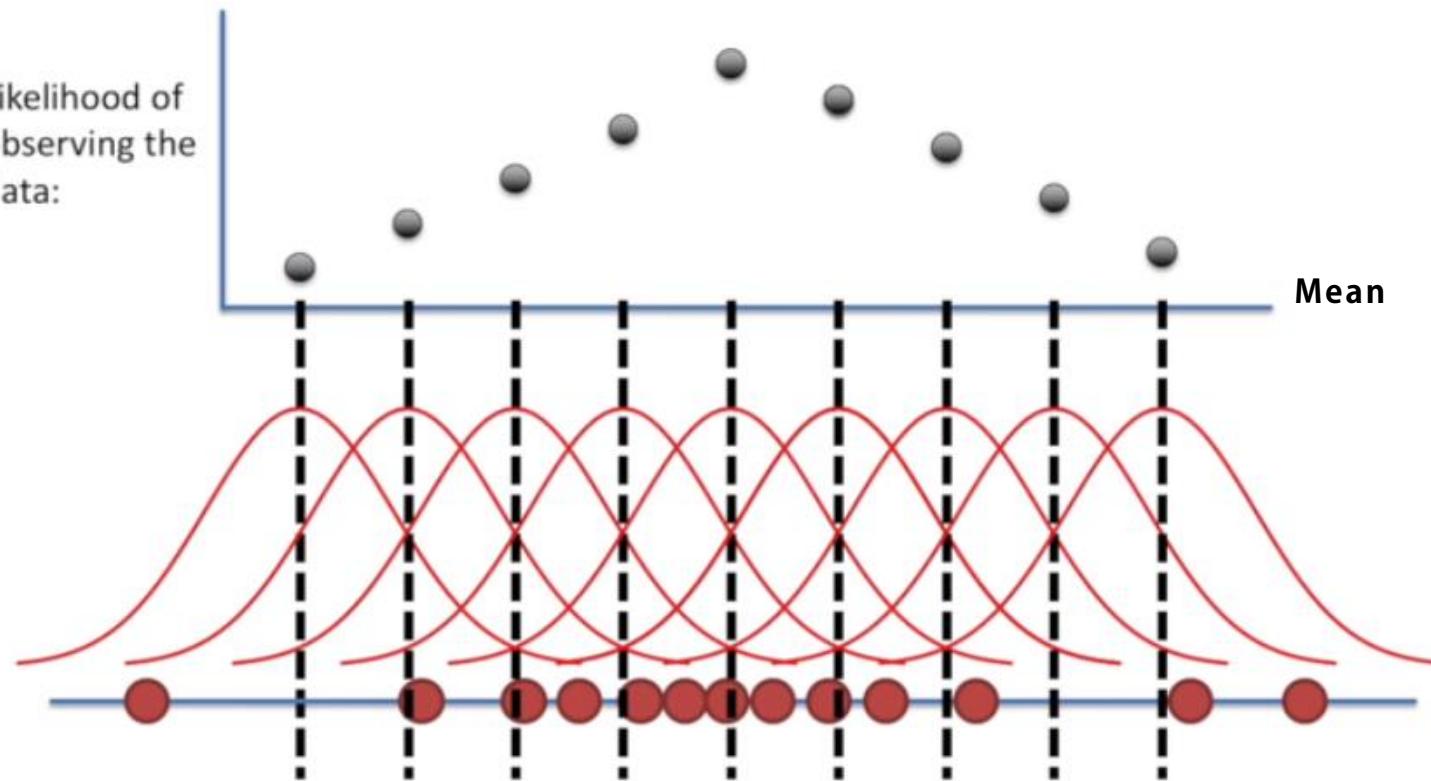
MLE(최대 우도 추정)

def

각 관측값에 대한 총 가능성(모든 가능성의 곱)이
최대가 되게하는 분포를 찾는 것

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta | \mathcal{D}) = \operatorname{argmax}_{\theta} \prod_{i=1}^n p(\mathcal{D}_i | \theta)$$

Likelihood of
observing the
data:



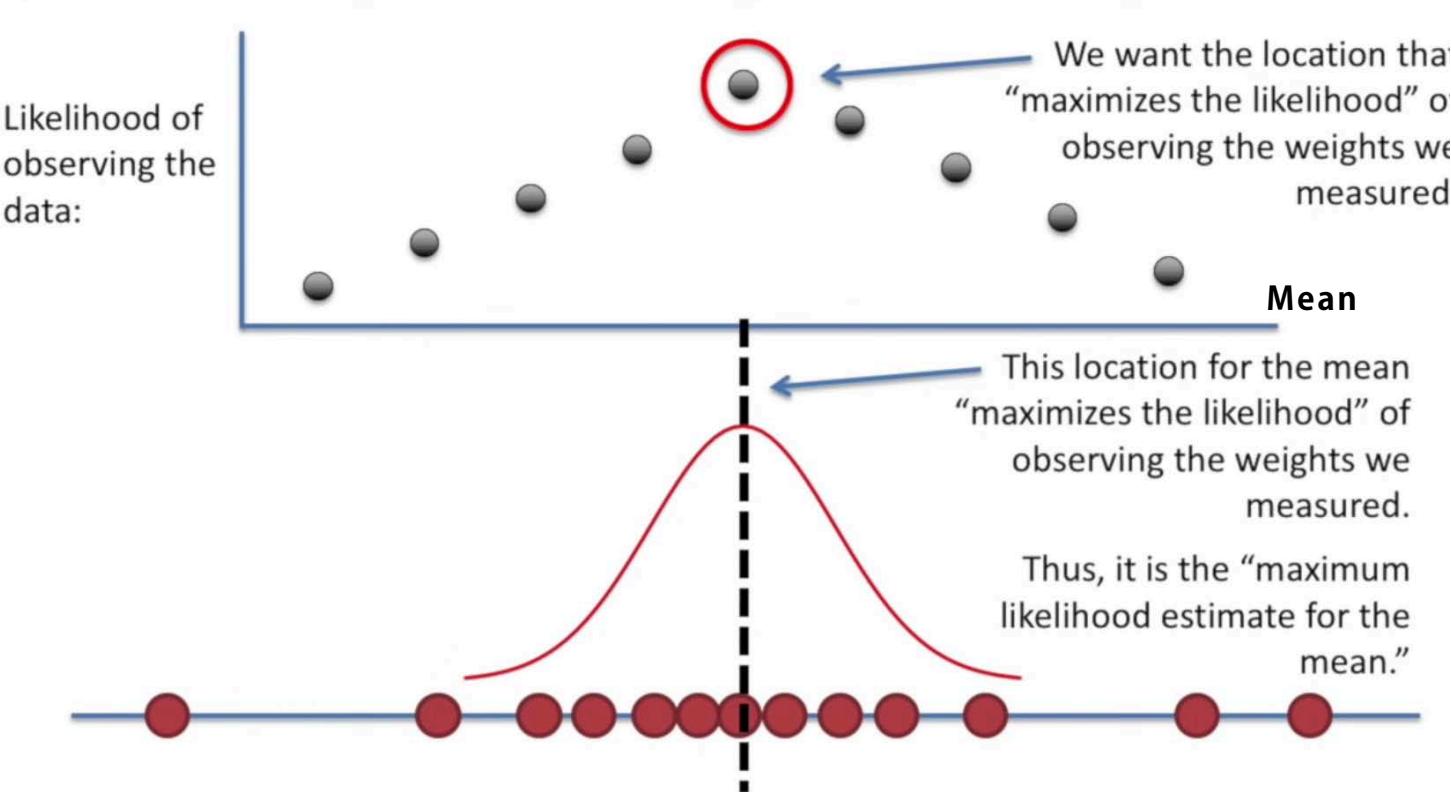
Unit 04 | 최대우도추정 & 평가지표

MLE(최대 우도 추정)

def

각 관측값에 대한 총 가능성(모든 가능성의 곱)이
최대가 되게하는 분포를 찾는 것

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta | \mathcal{D}) = \operatorname{argmax}_{\theta} \prod_{i=1}^n p(\mathcal{D}_i | \theta)$$



Unit 04 | 최대우도추정 & 평가지표

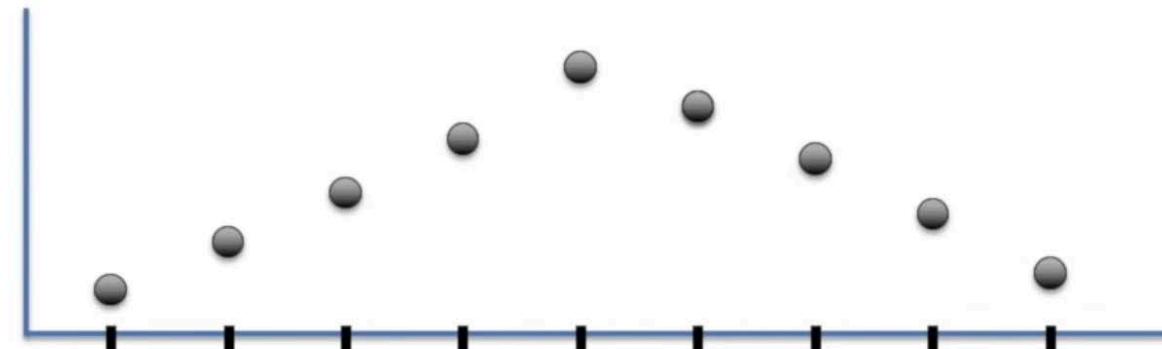
MLE(최대 우도 추정)

def

각 관측값에 대한 총 가능성(모든 가능성의 곱)이
최대가 되게하는 분포를 찾는 것

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathcal{L}(\theta | \mathcal{D}) = \operatorname{argmax}_{\theta} \prod_{i=1}^n p(\mathcal{D}_i | \theta)$$

Likelihood of
observing the
data:



Unit 04 | 최대우도추정 & 평가지표

MLE(최대 우도 추정)

$$L(\theta) = p(X|\theta) = \prod_{n=1}^N p(x_n|\theta)$$

수리적 편의(미분 용이)를 위해 양변에 log를 취한 후, -를
붙인 Negative log likelihood (for minimize)

$$E(\theta) = -\ln L(\theta) = -\sum_{n=1}^N \ln p(x_n|\theta)$$

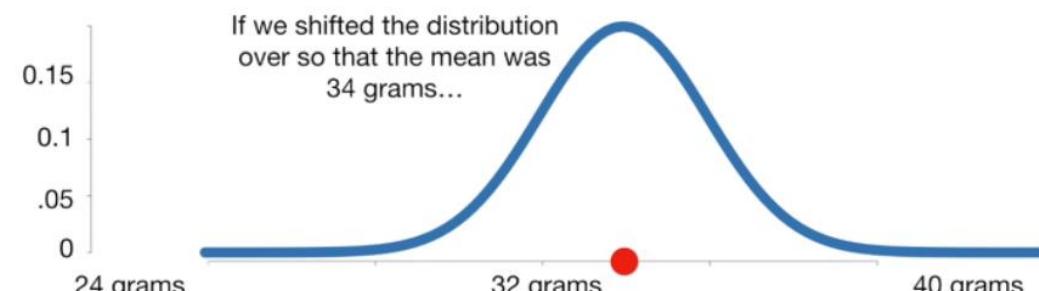
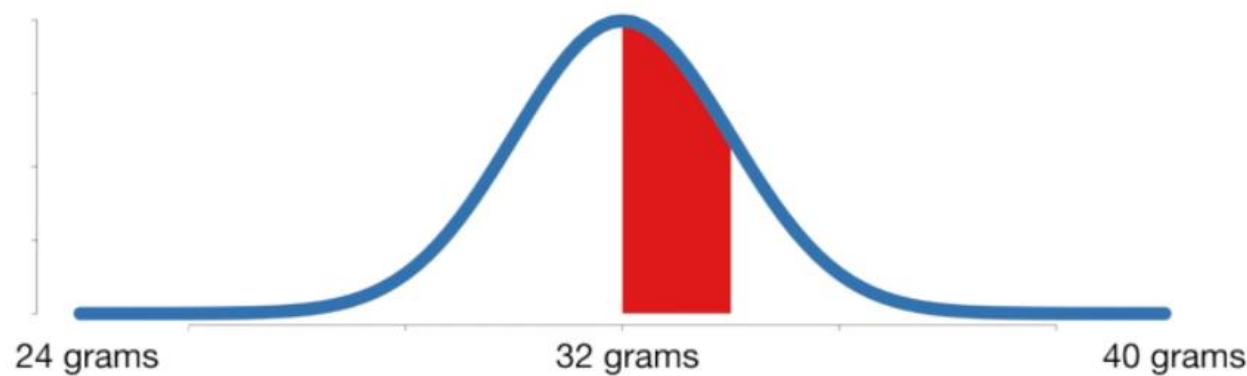
편미분 값이 0이 되는 모수 찾기

$$\frac{\partial}{\partial \theta} E(\theta) = -\frac{\partial}{\partial \theta} \sum_{n=1}^N \ln p(x_n|\theta) = -\sum_{n=1}^N \frac{\frac{\partial}{\partial \theta} p(x_n|\theta)}{p(x_n|\theta)} \stackrel{!}{=} 0$$

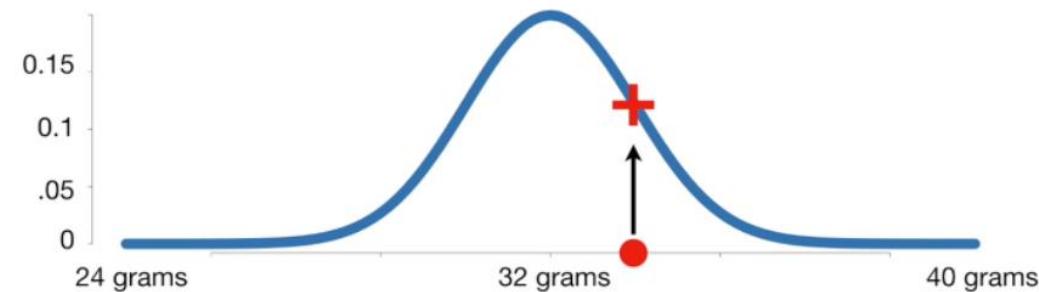
Unit 04 | 최대우도추정 & 평가지표

Probability(확률) VS Likelihood(우도, 가능성)

$pr(\text{weight between 32 and 34 grams} \mid \text{mean} = 32 \text{ and standard deviation} = 2.5)$



$L(\text{mean} = 32 \text{ and standard deviation} = 2.5 \mid \text{mouse weighs 34 grams})$



Probability

- 주어진 확률분포에서, 관측값 또는 관측구간이 분포 안에서 얼마의 확률로 존재하는가를 나타내는 값

Likelihood

- 주어진 관측값이, 어떤 확률 분포에서 왔을지에 대한 확률
- 데이터가 특정 분포로부터 생성될 확률

Unit 04 | 최대우도추정 & 평가지표

Probability(확률) VS Likelihood(우도, 가능성)

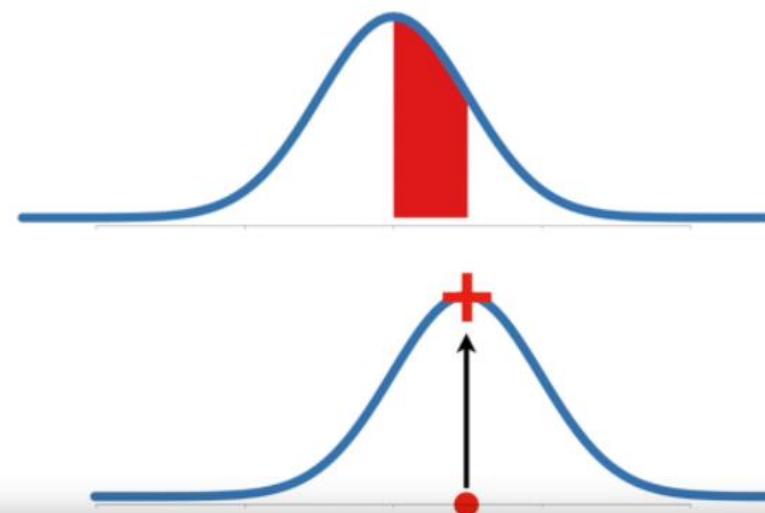
Probabilities are the areas under a fixed distribution...

$pr(\text{data} | \text{distribution})$

Likelihoods are the y-axis values for fixed data points with distributions that can be moved...

$L(\text{distribution} | \text{data})$

In summary...

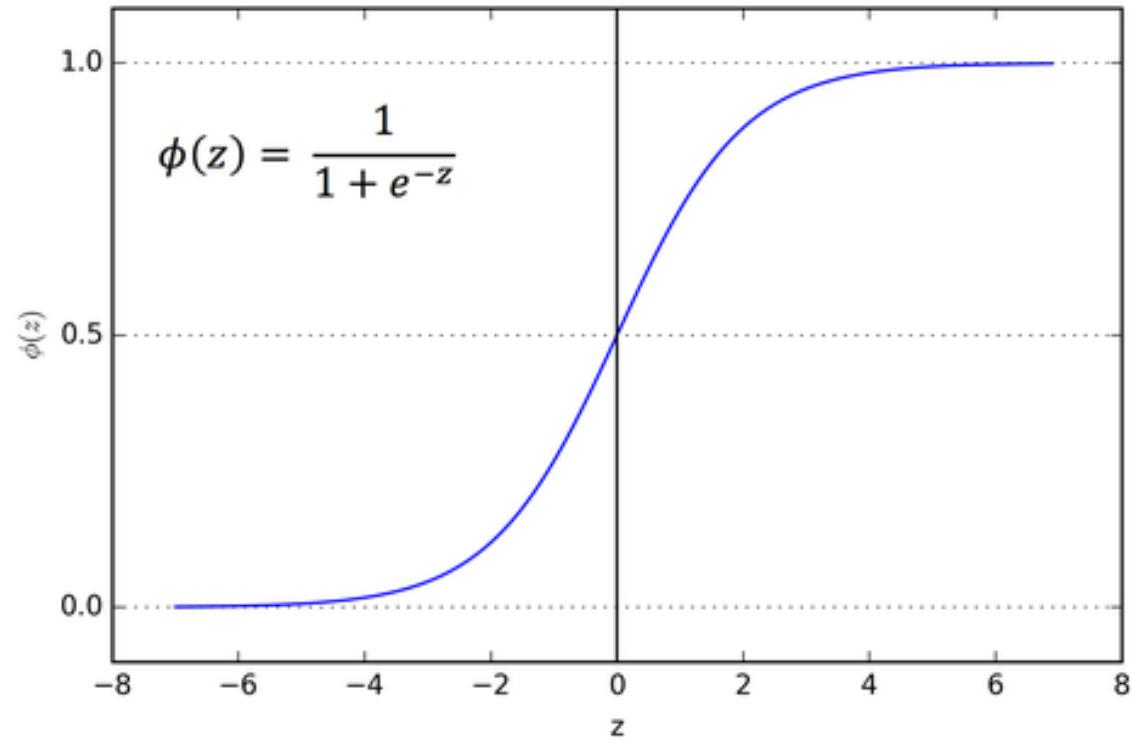


확률 : 주어진 확률분포에서 해당 관측값이 나올 확률

가능도 : 주어진 관측값이 특정 확률분포에서 나왔을 확률

Unit 04 | 최대우도추정 & 평가지표

최종 로지스틱 회귀모델 with 최적의 파라미터

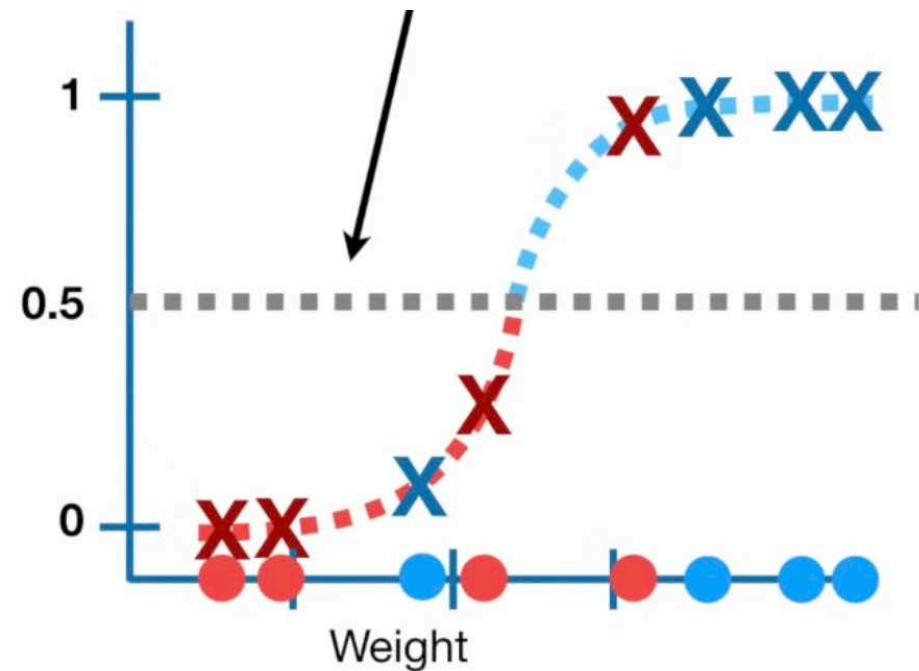


$$\begin{aligned}\pi(X) &= f(X) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k)}} \\ &= \frac{1}{1 + e^{-(\hat{\beta} X)}}\end{aligned}$$

Unit 04 | 최대우도추정 & 평가지표

Cutoff (Threshold)

- 분류(Classification)를 위한 기준 (logistic regression에서 cutoff를 0.5로 설정하였음)
- 로지스틱 함수로 구한 확률이 cutoff 이상이면 1, cutoff 이하이면 0으로 분류
- Cutoff을 조정하여 성능 조절 가능



		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	3	1
	Is Not Obese	1	3

Unit 04 | 최대우도추정 & 평가지표

Model Evaluation(1)

Accuracy

		예측결과	
		Positive	False
실제값	Positive	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

※ True : 옳은 예측(정답) / False : 틀린 예측(오답)

- 예측결과가 True일 때, 실제값도 True인 것
- 실제 분포가 **편향** 되어 있는 경우엔 적합하지 않음

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}}$$

- 웹사이트 판매량 데이터
- 학습 데이터 : 99% 물건 사지 않음 ($Y=0$), 1% 물건 구매($Y=1$)
 - 실제 데이터와 무관하게 $Y=0$ 이라고 예측할 확률이 높아짐
 - 99%의 정확도를 가지기에 좋은 결과처럼 보임

Unit 04 | 최대우도추정 & 평가지표

Model Evaluation(2)

Precision(정밀도)

- 모델이 True로 분류한 것 중에서 실제값이 True인 비율

$$\text{Precision} = \frac{TP}{TP + FP}$$

		예측결과	
		Positive	False
실제값	Positive	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

※ True : 옳은 예측(정답) / False : 틀린 예측(오답)

Recall(재현율)

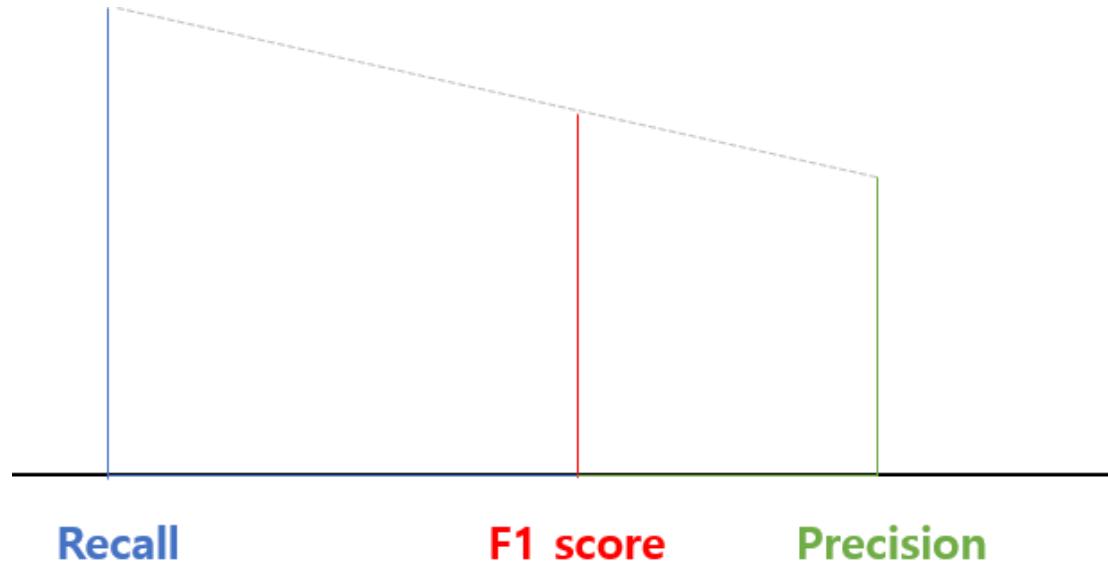
- Sensitivity
- 실제 True인 것 중에서 모델이 True라고 분류한 것의 비율

$$\text{Recall} = \frac{TP}{TP + FN}$$

(cf) Precision과 Recall은 Trade-Off 관계

Unit 04 | 최대우도추정 & 평가지표

Model Evaluation(3)



F1 Score

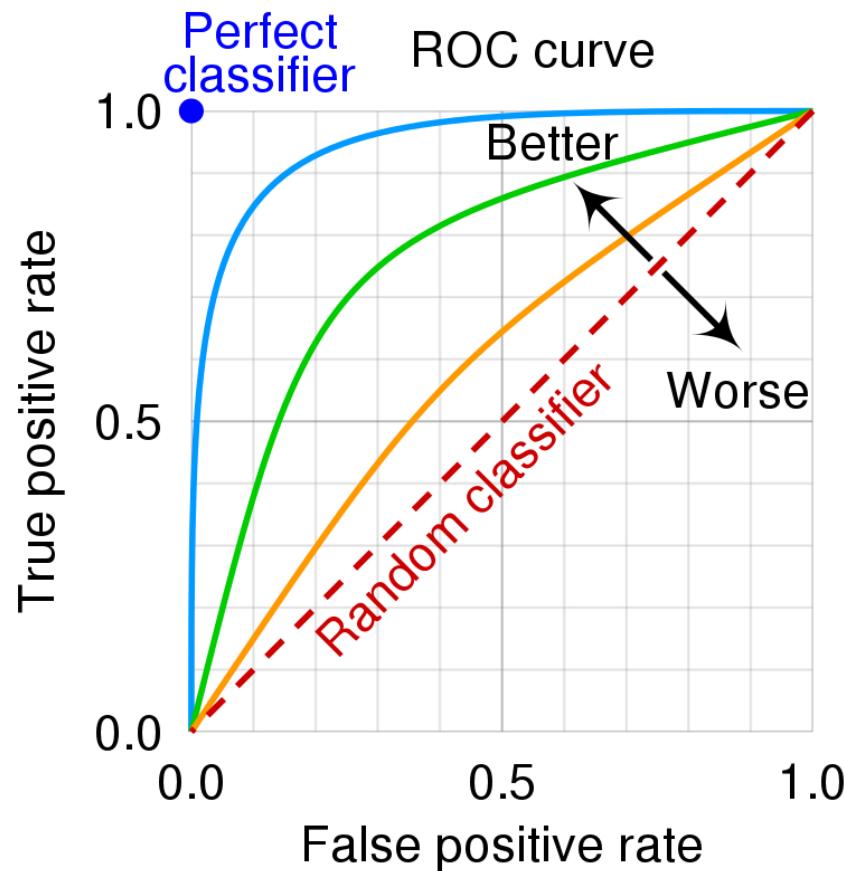
- Precision과 Recall의 조화 평균

$$F1 \text{ Score} = 2 \times \frac{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- 불균형한 데이터에서 잘 동작

Unit 04 | 최대우도추정 & 평가지표

Model Evaluation(4)



ROC Curve

- 각각의 Confusion Matrix에서 FPR, Recall(Sensitivity) 값 계산
 - 그래프가 좌 상단에 위치할수록 좋은 모델
 - FPR (False Positive) = $\frac{FP}{FP+TN}$
 - Recall (True Positive) = $\frac{TP}{TP+FN}$
 - AUC(=Area Under Curve) : ROC Curve 아래 면적
- ✓ 1에 가까울 수록 좋은 모델

		예측결과	
		Positive	False
실제값	Positive	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

Unit 04 | 최대우도추정 & 평가 지표

로지스틱 회귀분석 마무리

1. 범주형 반응변수 Y 분류를 위한 기법
2. 로지스틱 함수의 출력값은 0과 1 사이
3. Logit 함수 : $\log(\text{Odds}) = \log(p/(1-p))$
4. Beta1 : $\log(\text{Odds})$ 의 변화량
5. 최대 우도 추정법(MLE)으로 최적의 parameter 찾기
6. Recall, Precision, F1-score등의 기준으로 Classification 성능 개선

과제

[과제 1]

- LSE normal equation, MSE 구현

[과제 2] 회귀분석 – Used Car Priced Prediction

- Ch 1, Ch 2를 토대로 자유롭게 회귀분석 & 회귀진단 진행
- 주석으로 설명 및 근거 자세하게 달아주세요 ☺

[과제 3] 로지스틱 회귀분석 – Credit Card Fraud Detection

- 파이썬 sklearn 패키지를 활용해 로지스틱 회귀분석 진행
- 성능지표 계산 및 해석
 - Sklearn → mean accuracy, f1 score 등
 - confusion matrix → tp, fp, fn, tn 값
- 성능 개선 시도 (어떤 성능지표를 기준으로 했는지, 해당 지표 선택 이유 등)
- 주석으로 설명 및 근거 자세하게 달아주세요 ☺

Reference

[강의안]

- 투빅스 14기 강재영님 강의안
- 투빅스 15기 장아연님 강의안
- 투빅스 16기 이예림님 강의안
- 투빅스 17기 이지수님 강의안
- 투빅스 18기 김희경님 강의안
- 연세대학교 응용통계학과 김현태 교수님 <회귀분석> 강의안

[교재]

- Michael H. Kutner, Christopher J. Nachtsheim, John Neter, <Applied Linear Regression Models>

[참고 자료]

- [선형, 로지스틱] 데이터 사이언스 스쿨 4장, 6장 (<https://datascienceschool.net/intro.html>)
- [로지스틱] <https://ratsgo.github.io/machine%20learning/2017/04/02/logistic/>
- [Ridge/Lasso Regression] [Ridge regression\(능형 회귀\) 간단한 설명과 장점 \(tistory.com\)](#)
- [회귀진단] [Regression\(03\) – 회귀진단 | DataLatte's IT Blog \(heung-bae-lee.github.io\)](#)
- [MLE] <https://www.youtube.com/watch?v=XepXtl9YKwc>

Q & A

들어주셔서 감사합니다.