14기 정규세션

ToBig's 13기 이재빈

Regression Analysis

회귀분석

Contents

머신러닝 ?

기계학습 !

사람이 하나부터 열까지 직접 가르치는 기계를 의미하는 것이 아니라,
학습할 거리를 일단 던져 놓으면 이걸 가지고 스스로 학습하는 기계

def 머신러닝

"만약 어떤 작업 T에서 경험 E를 통해 성능 측정 방법인 P로 측정했을 때 성능이 향상된다면,
이런 컴퓨터 프로그램은 학습을 한다고 말한다" – Tom Mitchell

□와 △에 들어갈 정수는?

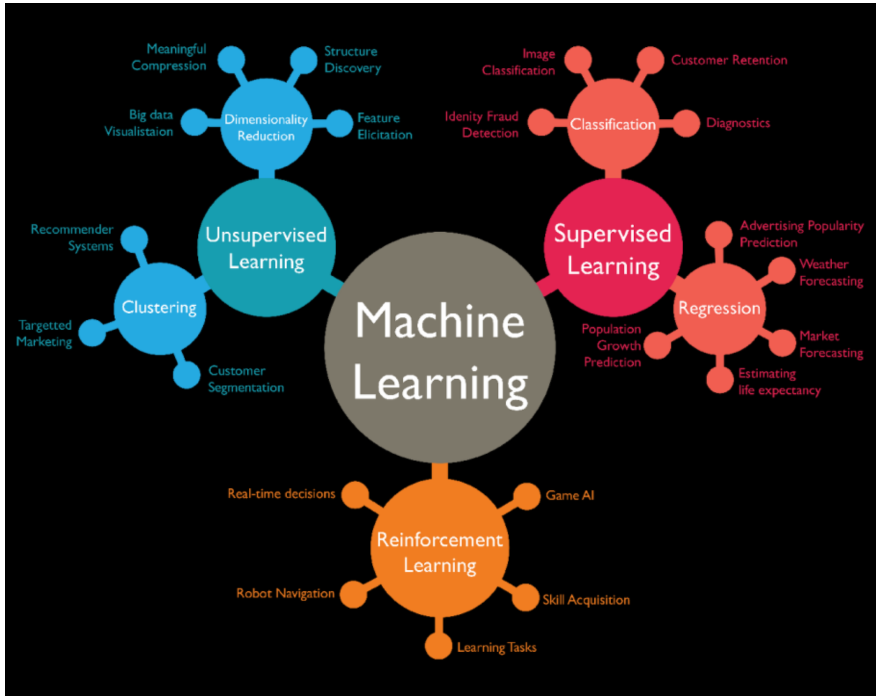
- 3 x □ + 2 x △ = 1
- 1 x □ + 4 x △ = -3
- 5 x □ + 5 x △ = 0
- 8 x □ + 3 x △ = 5

[작업 T] □와 △ 구하기
[성능 P] 수식이 정확할 확률
[경험 E] 입력값 (3, 2) (1, 4) (5, 5) (8, 3)를 입력,
출력값 1, -3, 0, 5를 도출하도록 학습
* **학습** : 경험 E를 통해 가중치(□=1, △=-1)를 찾는 것

머신러닝 알고리즘 종류

1. **지도학습 (Supervised Learning)**
   - 입력과 결과값 (label : 정답) 이용한 학습
   - **회귀** (Regression), **분류** (Classification)
   - ex. **선형 / 로지스틱 회귀**, KNN, SVM, Decision Tree

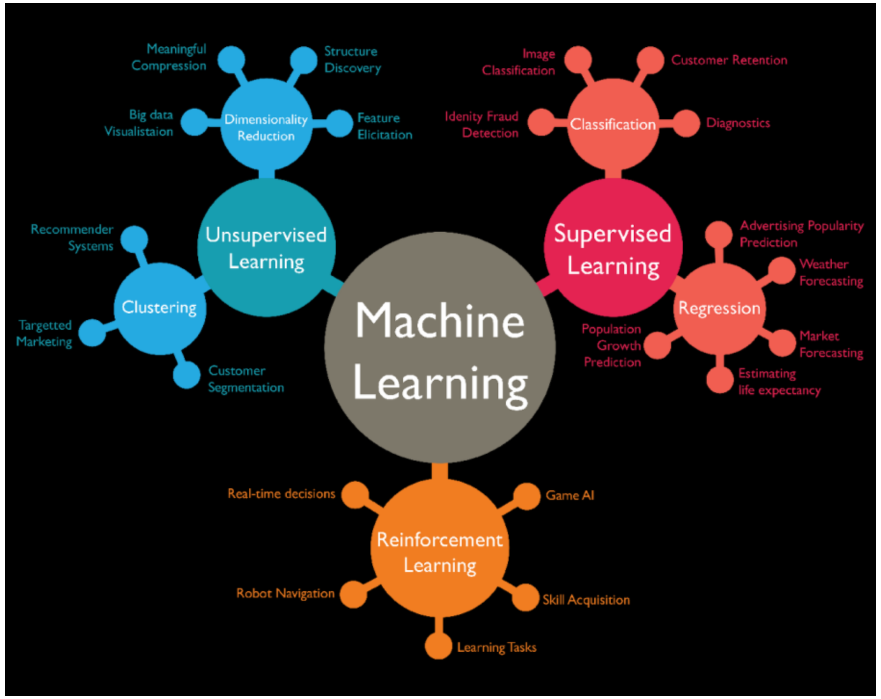
2. **비지도학습 (Unsupervised Learning)**
   - 입력만을 이용한 학습
   - 군집화 (Clustering)
   - ex. K-means Clustering



3. **강화학습 (Reinforcement Learning)**
   - Agent가 주어진 State에서 Action을 취했을 때, 이로부터 얻는 Reward를 최대화하는 방향으로 학습

머신러닝 알고리즘 종류

1. **지도학습 (Supervised Learning)**
   - 입력과 결과값 (label : 정답) 이용한 학습
   - **회귀** (Regression), **분류** (Classification)
   - ex. **선형 / 로지스틱 회귀**, KNN, SVM, Decision Tree



2. **비지도학습 (Unsupervised Learning)**
   - 입력만을 이용한 학습
   - 군집화 (Clustering)
   - ex. K-means Clustering

3. **강화학습 (Reinforcement Learning)**
   - Agent가 주어진 State에서 Action을 취했을 때, 이로부터 얻는 Reward를 최대화하는 방향으로 학습

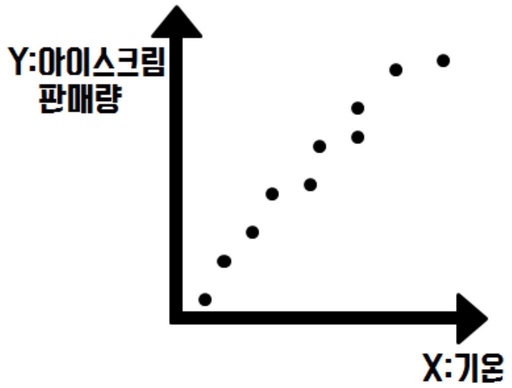
머신러닝과 통계 ?

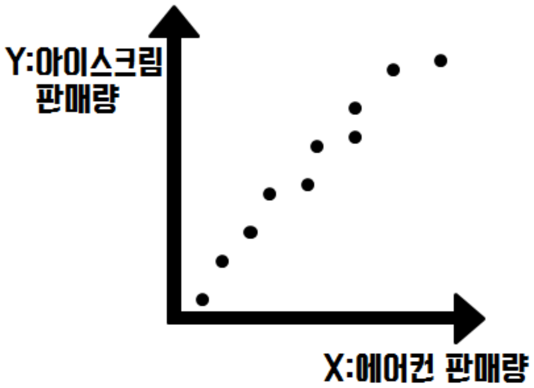
	머신러닝	통계학
특징	CS : 머신러닝 정확한 예측(Prediction) 내년에 병원에 갈 사람들의 숫자 예측	Stat : 데이터 마이닝 해석 가능성 (Interpretability) 사람들이 왜 병원에 가는지 이유 분석
모델	지도학습 (예측모델)	선형회귀분석
x , y	x : feature, y : label	x : 독립변수, y : 종속변수
Parameter (θ) 구하는 과정	학습	회귀식의 추정

어떤 데이터를 쓰고, 변수를 어떻게 (Feature Engineering) 집어넣어야
유사한 다른 상황에서도 비슷한 결과를 기대할 수 있는 모델이 나올 수 있을까 ?
-> 통계를 알면 머신러닝을 제대로 활용할 수 있습니다 ^_^

Contents

들어가기 앞서 1. 인과관계 vs 상관관계





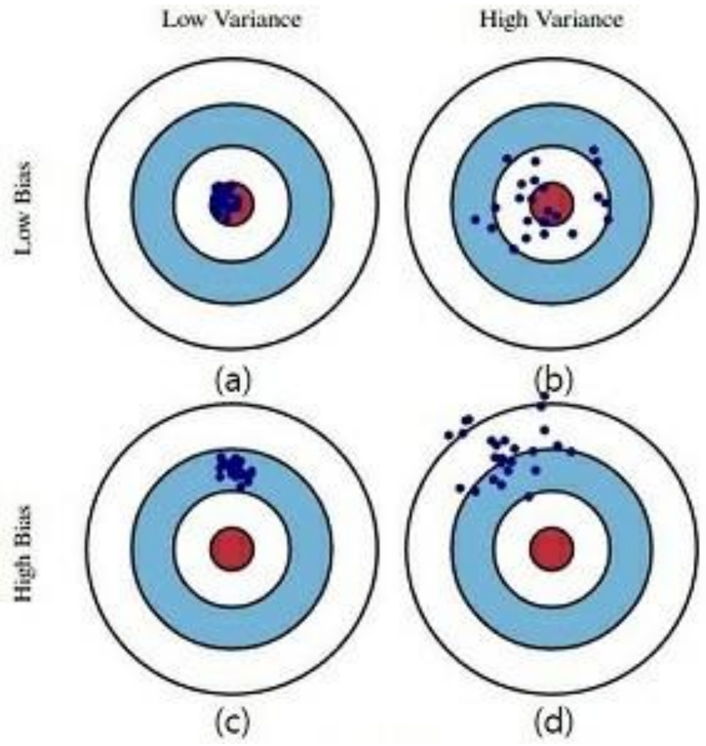
<u>인과관계</u>
어떠한 것이 원인이 되어서 결과가 나타나는 것

<u>상관관계</u>
한 쪽이 변화함에 따라서 다른 한 쪽도 증가하거나 감소하는 어떠한 관계의 추세

- 기온이 올라갈수록 아이스크림 판매량이 많아진다
- 에어컨 판매량이 많아질수록 아이스크림 판매량이 많아진다 ?

-> 상관관계가 인과관계를 의미하는 것이 아니다

들어가기 앞서 2. Bias vs Variance



붉은 영역: 참값 / 파란 점 : 추정값

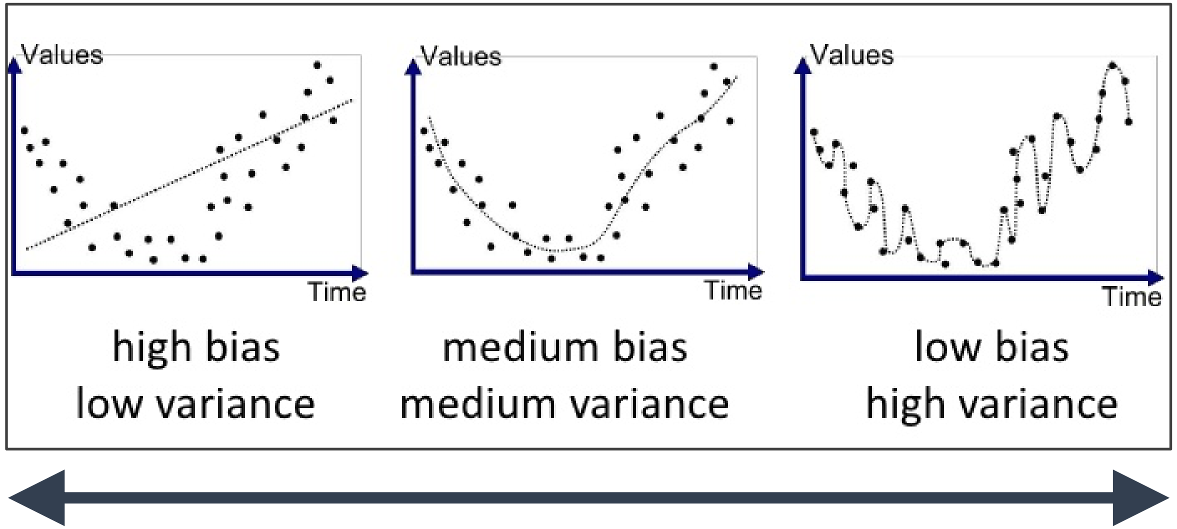
Bias

- $Bias(\hat{y})^2 = E[(E(\hat{y}) - y)^2]$
- 모델에 데이터를 넣어 나오는 예측값이 '전반적으로' 실제값을 얼마나 정확하게 예측하는지

Variance

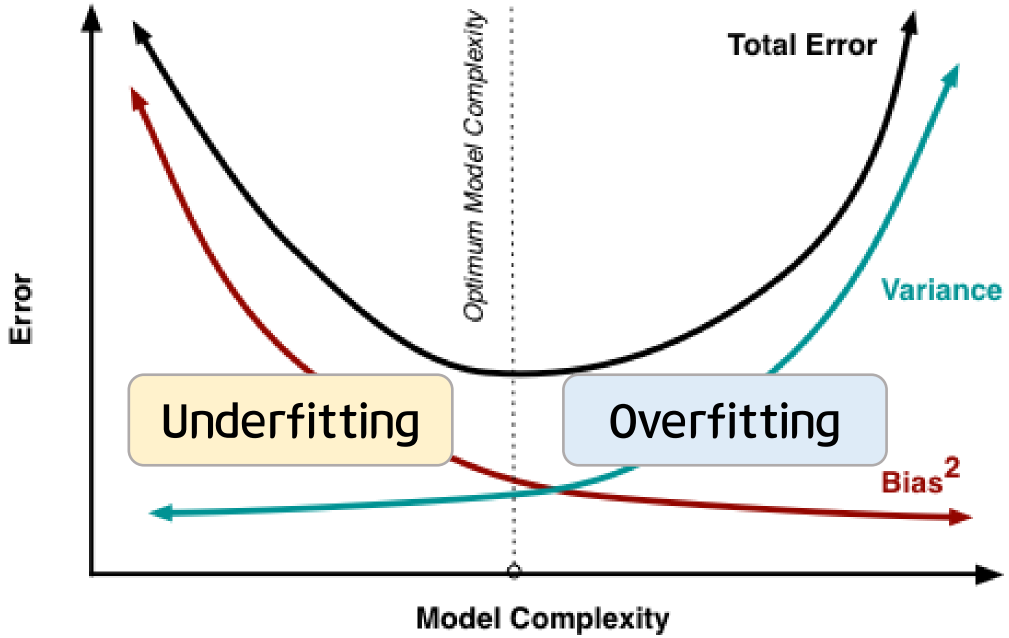
- $Var(\hat{y}) = E[(\hat{y} - E(\hat{y})^2]$
- 모델에 데이터를 넣어 나오는 예측값이 얼마나 큰 변동성을 가지는지

들어가기 앞서 3. Underfitting vs Overfitting



Underfitting
단순한 모델

Overfitting
복잡한 모델



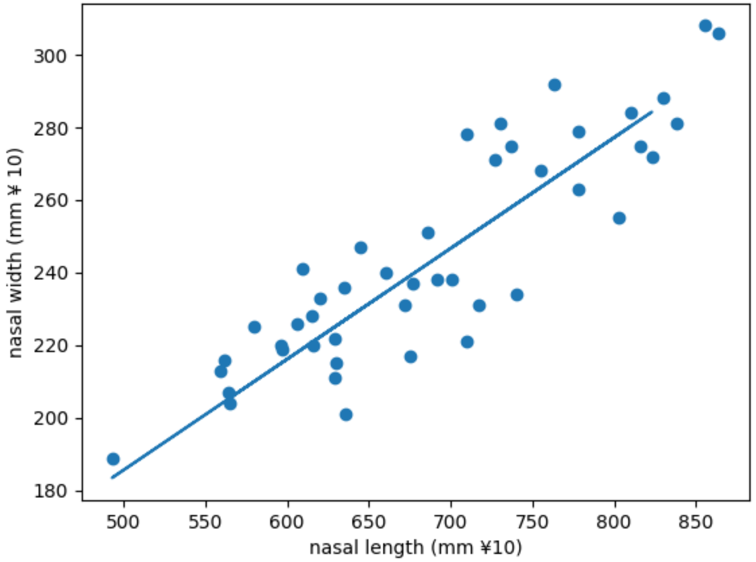
선형회귀분석

- **변수와 변수 사이의 선형 상관 관계를 도출하고자 하는 기법**
- ex) 해결한 과제 수(x)에 따른 행복감(y)

- 한 개 이상의 독립변수(x)들이 종속변수(y)에 미치는 영향을 추정하는 기법
- 변수들 사이의 관계를 밝히고 모형을 적합하여 관심있는 변수를 예측하거나 추론할 수 있음

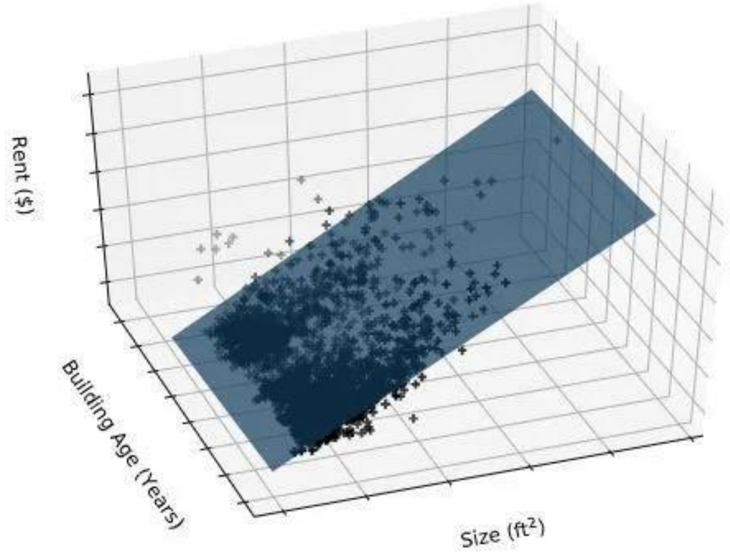
- x : **영향을 주는 변수 – 독립변수, 설명변수**
- y : **영향을 받는 변수 – 종속변수, 반응변수**

Unit 02 | 선형 회귀분석



단순선형회귀

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



다중선형회귀

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

오차 vs 잔차

ε_i

오차 (error)

모집단에서 회귀식을 얻었다면, 그 회귀식을 통해 얻은 예측값과 실제 관측값의 차이

* 모집단 (population) : 연구자가 조사하고 싶은 집단 전체

e_i

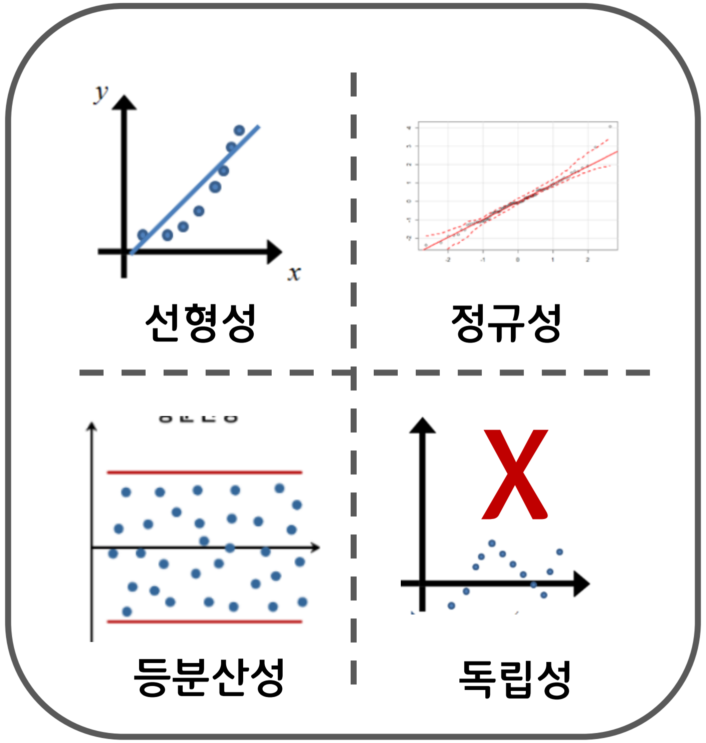
잔차 (residual)

표본에서 회귀식을 얻었다면, 그 회귀식을 통해 얻은 예측값과 실제 관측값의 차이

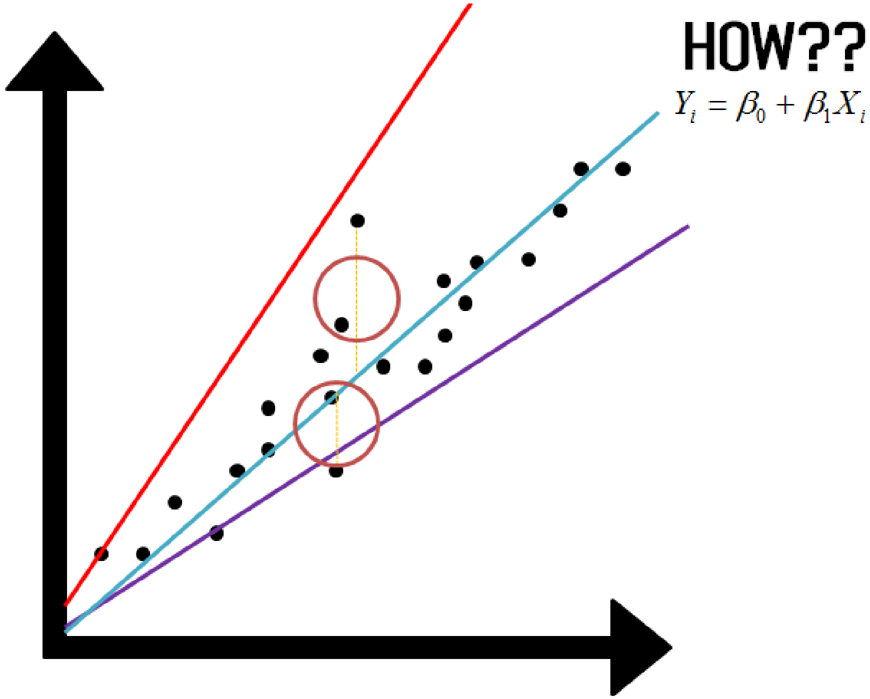
* 표본 (sample) : 연구자가 측정 또는 관찰한 결과들의 집합

회귀분석 기본 가정

1. 선형성 : 설명변수(X)와 반응변수(Y)가 선형 관계에 있다
2. 정규성 : $\varepsilon_i \sim N(0, \sigma^2)$, 오차 ε_i 는 정규분포를 따른다
3. 등분산성 : 오차 ε_i 의 분산은 σ^2로 항상 동일하다
4. 독립성 : 오차 ε_i 는 서로 독립이다 (iid)



회귀분석 목적



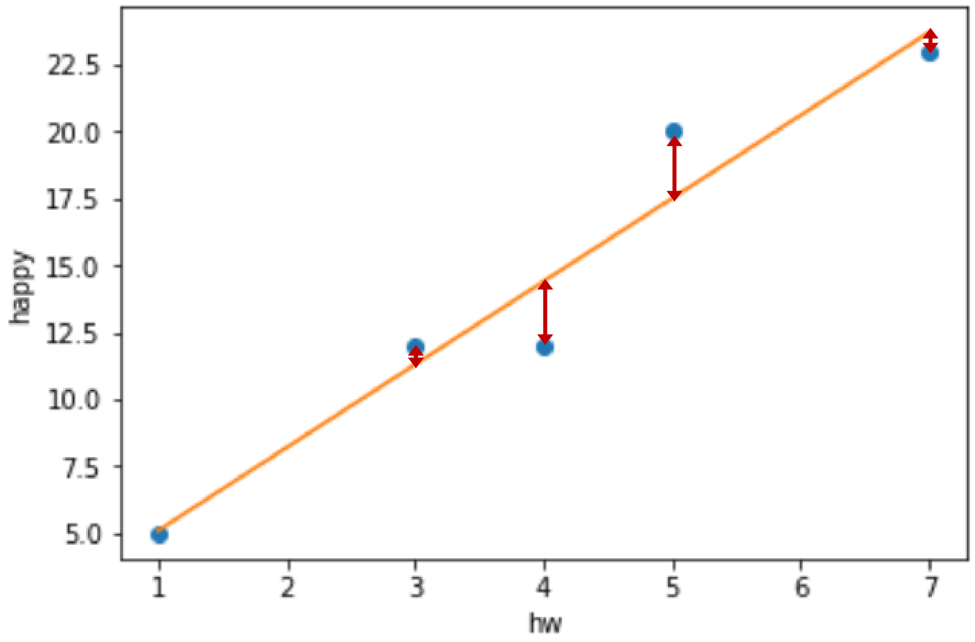
회귀식을 도출해 내서,
예측의 정확성을 높이고자 함 !

이를 위해서는,
모수를 정확하게 **추정**하는 것이 중요 !

Q. 빨강 파랑 보라 .. 가장 좋은 회귀식이 뭘까요 ?

A. 회귀식과 실제 관측 값의 차이(**잔차**) 가
가장 작은 것이 좋은 모델

최소제곱법 (LSE)



회귀식이 예측한 값과 실제 값의 차이 최소화

회귀식이 예측한 $\hat{y}$ 값과 실제 y값의 차이 (=잔차)의 제곱합을 최소화하는 알고리즘

-> 최적 모수 추정

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

Loss Function

최소제곱법 (LSE)

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

Loss
Function
목적함수

↓ 최소화하므로 편미분 = 0

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

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

<최소제곱 추정치>

$$\widehat{\beta_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

최소제곱법 (LSE)

<적합된 회귀식>

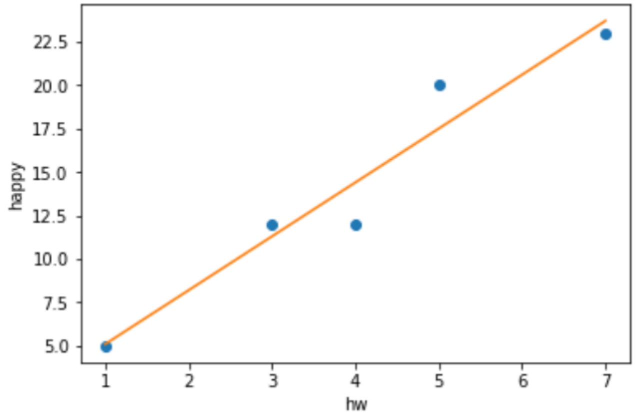
$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

- 예측 : x_i에 값을 대입해 종속변수 예측
- 해석 : $\widehat{\beta_0}$ = intercept (절편)

  $\widehat{\beta_1}$ = x_i가 한 단위 증가할 때 종속변수의 증가(감소)량
- $\hat{\beta}$: 추정된 회귀계수

hw	happy
1	5
3	12
4	12
5	20
7	23

yhat
5.1
11.3
14.4
17.5
23.7



y = 2.0 + 3.1 x

과제를 하나 더 해결할수록, 행복함이 3.1만큼 증가

최소제곱법 with 행렬

$$\underset{n \times 1}{Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}, \quad \underset{n \times (p+1)}{X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}}, \quad \underset{(p+1) \times 1}{\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}, \quad \underset{n \times 1}{\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}$$

절편 변수1 변수p

road_name	road_bunji1	road_bunji2	Close_date	Close_result	point.y	point.x
해운대해변로	30.0	NaN	2018-06-14 00:00:00	배당	35.162717	129.137048
마린시티2로	33.0	NaN	2017-03-30 00:00:00	배당	35.156633	129.145068

$$Y = X\beta + \varepsilon$$

n : data 개수
p : feature 개수 (column / variable)

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 & \beta_1 x_{11} & \ldots\beta_p x_{1p} \\ \beta_0 & \beta_1 x_{21} & \ldots\beta_p x_{2p} \\ \vdots & \vdots & \vdots \\ \beta_0 & \beta_1 x_{n1} & \ldots\beta_p x_{np} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

최소제곱법 with 행렬

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

목적함수

$$= (\mathrm{y} - \mathrm{X}\beta)'(\mathrm{y} - \mathrm{X}\beta)$$

↓ 최소화하므로 미분 = 0

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

< Normal Equation > 정규방정식

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

Unit 02 | 선형 회귀분석

예시

hw	happy
1	5
3	12
4	12
5	20
7	23

?

$$Y = \begin{bmatrix} 5 \\ 12 \\ 12 \\ 20 \\ 23 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 7 \end{bmatrix}$$

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

예시

$$(X'X) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 5 & 7 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 7 \end{bmatrix} = \begin{bmatrix} 5 & 20 \\ 20 & 100 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{500-400} \begin{bmatrix} 100 & -20 \\ -20 & 5 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'y = \frac{1}{500-400} \begin{bmatrix} 100 & -20 \\ -20 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 5 & 7 \end{bmatrix} \begin{bmatrix} 5 \\ 12 \\ 12 \\ 20 \\ 23 \end{bmatrix} = \begin{bmatrix} 2 \\ 3.1 \end{bmatrix}$$

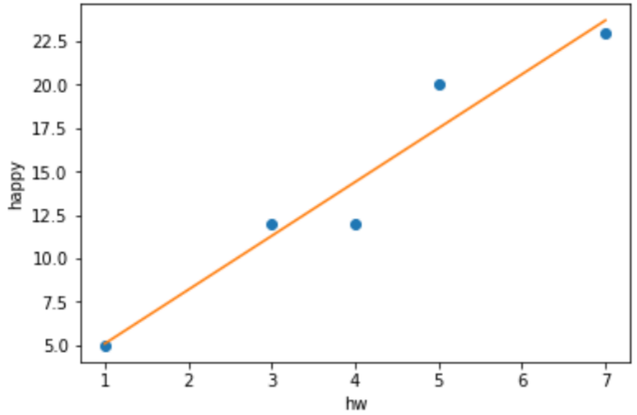
hw	happy
1	5
3	12
4	12
5	20
7	23

예시

1. $$\hat{y} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 7 \end{bmatrix} \begin{bmatrix} 2 \\ 3.1 \end{bmatrix} = \begin{bmatrix} 5.1 \\ 11.3 \\ 14.4 \\ 17.5 \\ 23.7 \end{bmatrix}$$

hw	happy
1	5
3	12
4	12
5	20
7	23

yhat
5.1
11.3
14.4
17.5
23.7

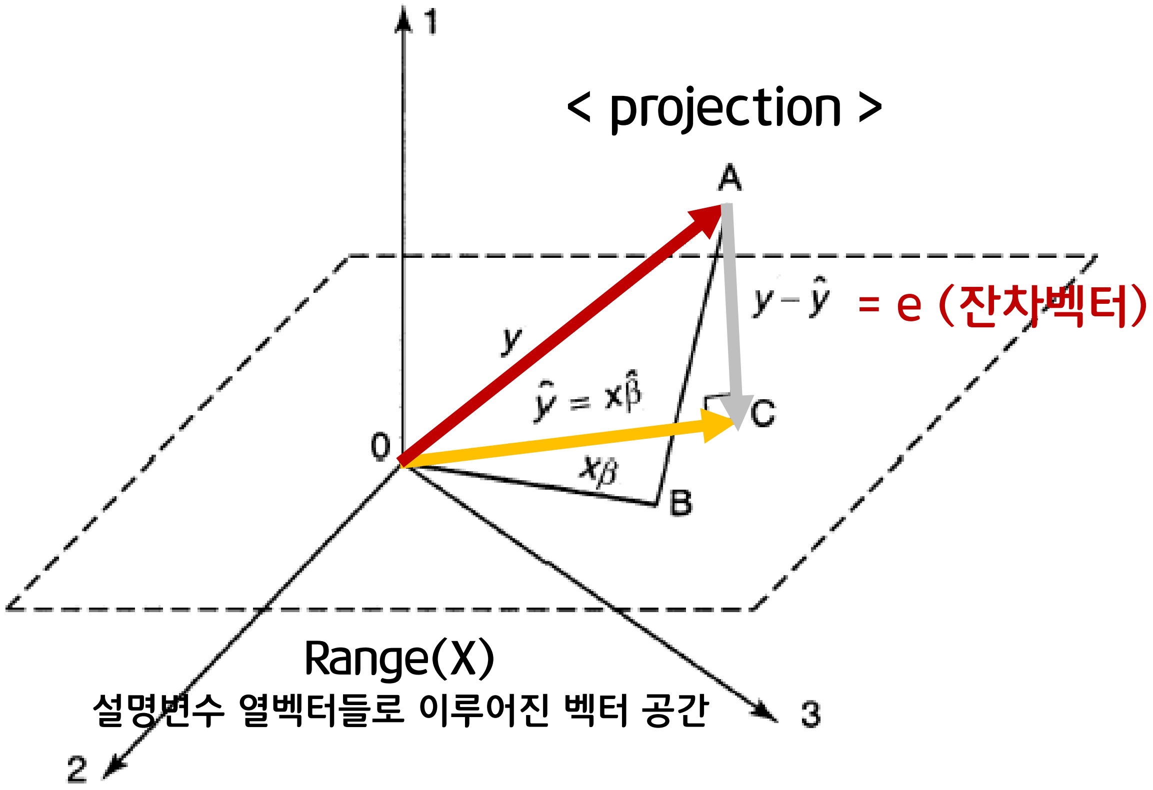


y = 2.0 + 3.1 x

실제 – 예측 차이 최소화하는 식 추정

2. 새로운 관측값 (x=10) : $$\widehat{y_0} = [1 \quad 10] \begin{bmatrix} 2 \\ 3.1 \end{bmatrix} = 33$$

정규방정식 기하학적 의미로 유도하기



잔차벡터 $e = y - X\hat{\beta}$ 와 Range(X)의 거리가 **최소**가 되도록 하려면?

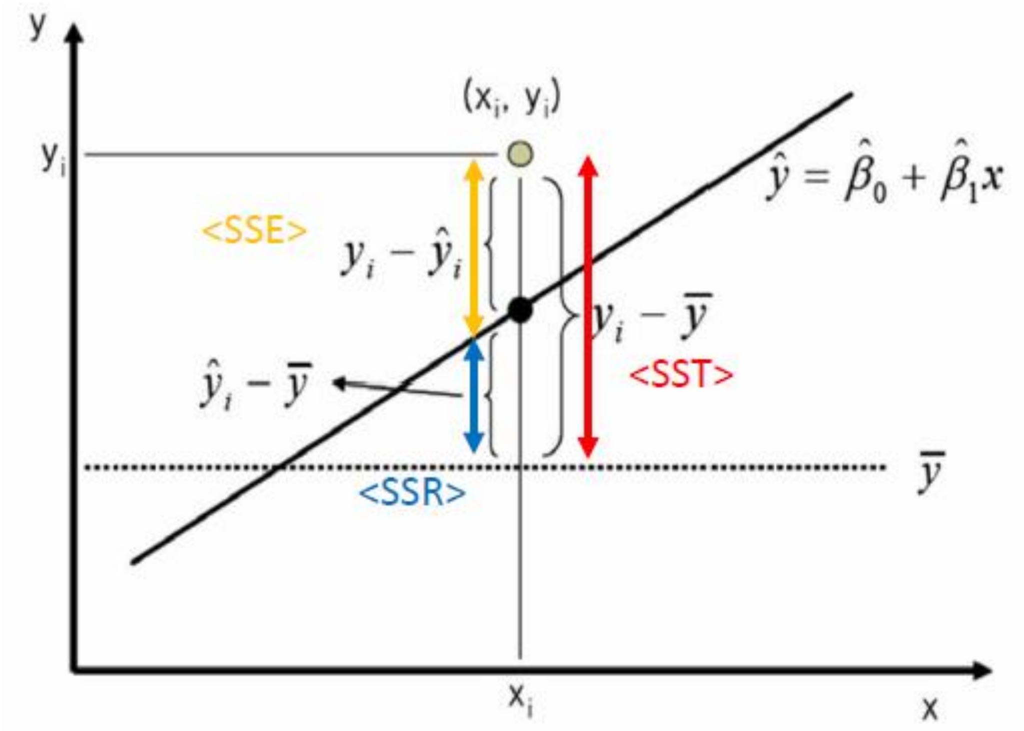
-> 두 벡터가 **수직**이어야 함 !

$$X'(y - X\beta) = 0$$
$$X'y - X'X\beta = 0$$
$$\therefore \ \hat{\beta} = (X'X)^{-1}X'y$$

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 : 총제곱합
SSR : 회귀제곱합 (전체 제곱합 중 회귀식으로 설명할 수 있는 부분)
SSE : 잔차제곱합 (전체 제곱합 중 회귀식으로 설명하지 못하는 부분)

-> 회귀식이 데이터를 잘 설명할수록 SSR 증가 (SSE감소)

회귀분석 표 해석

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

$H0: \beta_1 = \beta_2 = \dots = \beta_p = 0$
F > F(α, p, n-p-1) 이면 H0 기각

MSE = 회귀식이 설명하지 못하는 부분
-> MSE 값이 작을수록 좋음

(cf) 단순회귀 제약조건 2개 = 1+1개
1. $\sum$ 잔차 = 0   2. $\sum$ 잔차 * x_i = 0

Contents

Unit 03 | 모형 진단

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    OPS   R-squared:                       0.915
Model:                            OLS   Adj. R-squared:                  0.914
Method:                 Least Squares   F-statistic:                     1931.
Date:                Tue, 28 Jul 2020   Prob (F-statistic):               0.00
Time:                        02:03:49   Log-Likelihood:                 254.44
No. Observations:                1633   AIC:                            -490.9
Df Residuals:                    1624   BIC:                            -442.3
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
year           0.3380      0.018     18.524      0.000       0.302       0.374
BB             0.3019      0.059      5.151      0.000       0.187       0.417
HBP            0.1914      0.043      4.411      0.000       0.106       0.277
SO             0.0439      0.051      0.854      0.393      -0.057       0.145
height         0.2135      0.032      6.701      0.000       0.151       0.276
age_year       0.2850      0.024     11.762      0.000       0.237       0.333
HR             0.0194      0.009      2.064      0.039       0.001       0.038
SB             0.0052      0.007      0.749      0.454      -0.008       0.019
H              0.0293      0.013      2.217      0.027       0.003       0.055
==============================================================================
Omnibus:                      580.341   Durbin-Watson:                   1.987
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8336.581
Skew:                           1.255   Prob(JB):                         0.00
Kurtosis:                      13.780   Cond. No.                         17.3
==============================================================================
```

OLS : ordinary least square

- R-squared / Adj. R-squared
- F-statistics
- AIC / BIC

- Coef p값

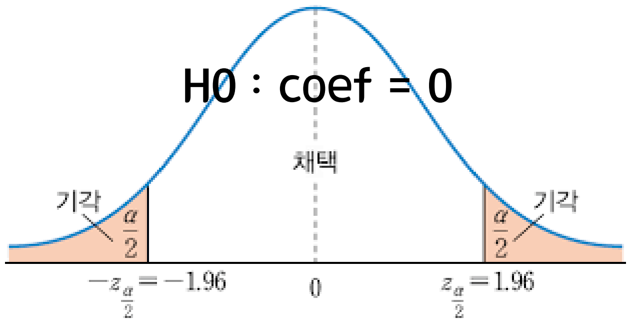
- Durbin-Watson (오차의 자기상관)
- Condition Number (다중공선성)

변수 선택 : Test on Coefficient

	coef	std err	t	P>|t|	[0.025	0.975]
year	0.3380	0.018	18.524	0.000	0.302	0.374
BB	0.3019	0.059	5.151	0.000	0.187	0.417
HBP	0.1914	0.043	4.411	0.000	0.106	0.277
SO	0.0439	0.051	0.854	0.393	-0.057	0.145
height	0.2135	0.032	6.701	0.000	0.151	0.276
age_year	0.2850	0.024	11.762	0.000	0.237	0.333
HR	0.0194	0.009	2.064	0.039	0.001	0.038
SB	0.0052	0.007	0.749	0.454	-0.008	0.019
H	0.0293	0.013	2.217	0.027	0.003	0.055

p값 > 유의수준 (보통 0.05) = 통계적으로 의미 없는 추정값

Coef : 추정한 회귀계수 값
P > |t| : 추정된 회귀계수의 p값
(유의확률, 양측검정)



https://m.blog.naver.com/vnf3751/220830413960

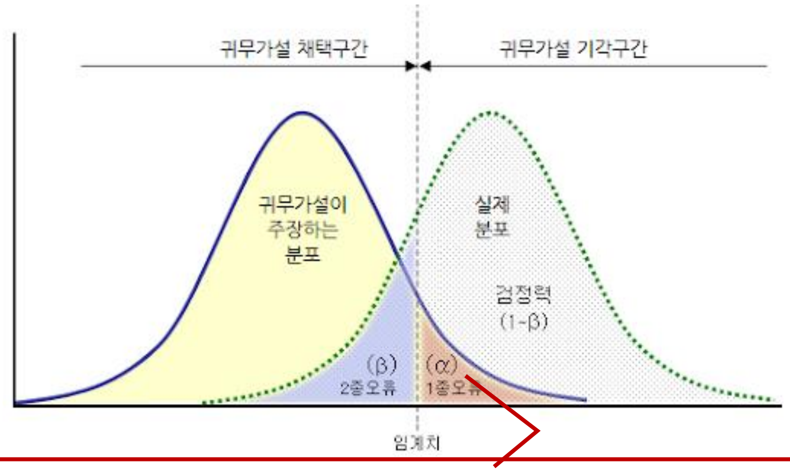
(cf) p값? 귀무가설?? 유의확률..???

p값이 0.05보다 작으므로
95% 유의수준 하에서 귀무가설을 기각한다

가설검정
모집단의 특징에 대한 통계적 가설을
추출된 표본을 통하여 검토하는 추론 방법

귀무가설 (H0) : 기각하고자 하는 사실
대립가설 (H1) : 일반적으로 주장하고자 하는 사실

-> H0를 기각함으로써, H1을 입증한다 !

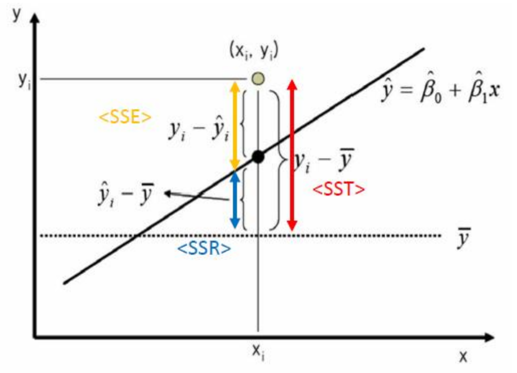


- p값 : 귀무가설이 옳다는 가정 하에, 검정통계량이 계산될 확률
- 귀무가설을 기각시키려면 -> 귀무가설이 거짓이어야 하는데
  -> 그렇지 않을 경우도 대비해야 함 !
- 귀무가설이 옳은데 실수로 기각될 확률,
  즉, 1종 오류를 범하게 될 확률 최소화
- 1종 오류의 상한선 (=유의수준) 미리 설정

모형 선택 기준 : R-squared

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

전체 제곱합 중 회귀식으로 설명 가능한 부분
-> 결정계수가 크면 클수록 좋음 !



SST : 총제곱합
SSR : 회귀제곱합
SSE : 잔차제곱합

$$adj\ R^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

Adjusted R square(조정된 결정계수)

설명변수를 추가하면 SSR 이 항상 커져 결정계수가 항상 증가
따라서 설명변수의 개수가 다른 모델과 단순 비교 불가

-> 설명변수의 개수를 고려하여,
설명변수가 증가하면 값이 감소하도록 패널티를 줌

-> 설명변수를 추가했는데 adjusted R square가
감소하지 않는다면 패널티를 감수할만큼 설명을 잘하는 변수

(cf) 설명변수가 증가하면 R-squared 값이 증가하는 이유

	Mean	LSE
Model1	$Ey = \beta_0 + \beta_1 x_1$	${b_0}^*, {b_1}^*$
Model2	$Ey = \beta_0 + \beta_1 x_1 + \beta_2 x_2$	b_0, b_1, b_2

by definition of LSE,

$$SSE(M2) = \sum(y_i - b_0 - b_1 x_1 - b_2 x_2)^2 \le \sum(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2)^2 = S(\beta_0, \beta_1, \beta_2)$$
$$if\ \beta_0 = {b_0}^*, \beta_1 = {b_1}^*, \beta_2 = 0\ \ then\ \ S({b_0}^*,\ {b_1}^*, 0) = SSE(M1)$$
$$\therefore SSE(M2) \le SSE(M1)$$

모형 선택 기준 : AIC / BIC

- AIC = $-2\log(\text{likelihood}) + 2p = \; n\ln\frac{SSE}{n} + 2p$
- BIC = $-2\log(\text{likelihood}) + \text{plog(n)} = n\ln\frac{SSE}{n} + p\ln n$

에러 포함!

- 선형 회귀에서는 AIC = $n\ln\frac{\boxed{RSS}}{n} + \boxed{2p}$, BIC = $n\ln\frac{\boxed{RSS}}{n} + \boxed{p}\ln n$

penalty

- likelihood는 가장 크게 하면서 + 변수의 개수가 가장 적은 최적의 모델
- 설명변수를 추가했는데 AIC, BIC 값이 증가하지 않으면 좋은 변수 (변수의 설명력 > penalty)

- 값이 작을수록 좋음 !

< 후보 모형 선택 기준 >

1. adj R^2 = $1-\frac{SSE/(n-p)}{SST/(n-1)}$: 클수록 좋음
2. MSE = $\frac{SSE}{n-p}$: 작을수록 좋음
3. R^2 = $1-\frac{SSE}{SST}$: p 같다면 클수록 좋음
4. AIC = $n\ln\frac{SSE}{n}+2p$, BIC = $n\ln\frac{SSE}{n}+p\ln n$
   : 작을수록 좋음

```
sion Results
========================================
③  R-squared:                      0.915
①  Adj. R-squared:                 0.914
②  F-statistic:                    1931.
   Prob (F-statistic):              0.00
   Log-Likelihood:                254.44
④  AIC:                           -490.9
   BIC:                           -442.3


========================================
```

모형 선택 과정 : 변수선택법

1. **전진선택법 (forward selection)**
   p-value가 작은 순서대로 변수를 축차적으로 추가
2. **후진제거법 (backward elimination)**
   p-value가 큰 순서대로 변수를 축차적으로 제거

   -> 두 방법은 추가 설명력이 변화하더라도, 한번 선택한(빠진) 변수를 다시 제거(선택)할 수 없음

3. **단계적 선택법 (stepwise) : 전진선택 + 후진제거**
   변수의 추가/제거 모두 가능

Dummy Variable

- 범주형 변수를 회귀분석에 사용할 수 있도록 변환한 것
- 범주 : 학년 (1,2,3) , 혈액형 (A, B, O, AB)

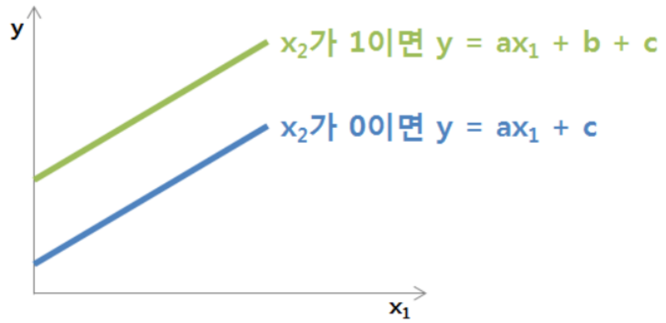
- 연속형 변수처럼 만들어서, 회귀분석에 사용

- 더미변수 개수 : 범주의 개수 – 1
  기준이 되는 변수를 정해, 이 값을 제외하고 더미변수 생성

직업
가수
배우
개그맨
배우
가수

→

직업_가수	직업_개그맨
1	0
0	0
0	1
0	0
1	0



Intercept만 변화하게 됨 !

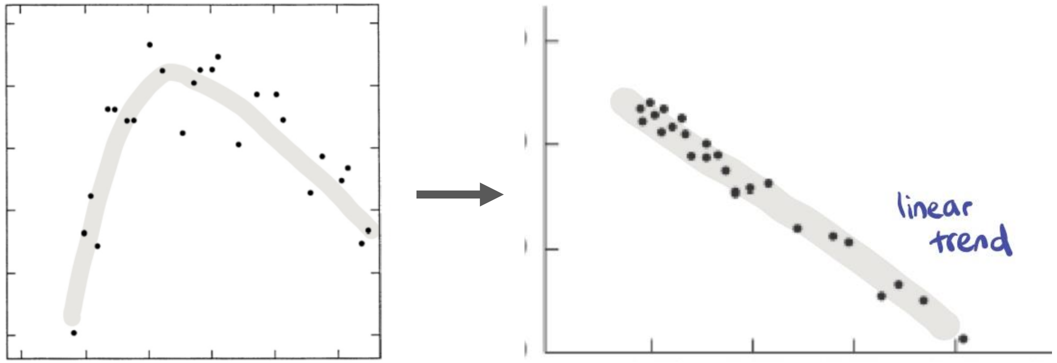
교호작용 (interaction)

- **한 변수의 효과가 다른 변수의 수준에 의존하는 경우**
- **단일 변수만으로는 알 수 없는 변수들간의 상호작용 고려**

- **보통 범주형*범주형 , 범주형*연속형 변수의 관계만 고려**
  -> 연속형*연속형의 경우 해석의 모호함이 생길 수 있기 때문에 !

- ex) 흡연을 하면 건강 -3, 음주를 하면 건강 -2 ⇔ 흡연과 음주를 동시에 하는 사람은 ?
  흡연과 음주를 동시에 한 결과 더 많은 악영향 (-10)이 끼친다면, 교호작용 변수를 고려해야 함 !
- 건강(Y) ~ 흡연 + 음주 ⟶ 건강(Y) ~ 흡연 + 음주 + 흡연*음주

변수 변환



비선형적인 함수 관계를 선형으로 바꿔 다룰 수 있다

ex) $\log(x)$, $\sqrt{x}, x^2$, ⋯

다만, 계수 해석에 매우 유의해야 함 !

$$\log(write) = \beta_0 + \beta_1 * female + \beta_2 * read + \beta_3 * math$$

lgwrite	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.114718	.0195341	5.87	0.000	.076194	.153242
read	.0066305	.0012689	5.23	0.000	.0041281	.0091329
math	.0076792	.0013873	5.54	0.000	.0049432	.0104152
intercept	3.135243	.0598109	52.42	0.000	3.017287	3.253198

Female : Dummy Variable

$e^{0.1147}$=1.12 : 약 12% 정도 더 높은 점수를 받는다

변수 변환 : normalization

min-max normalization

0 ~ 1 사이의 값 가짐

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Z-score normalization

Mean, variance를 고려해 scaling

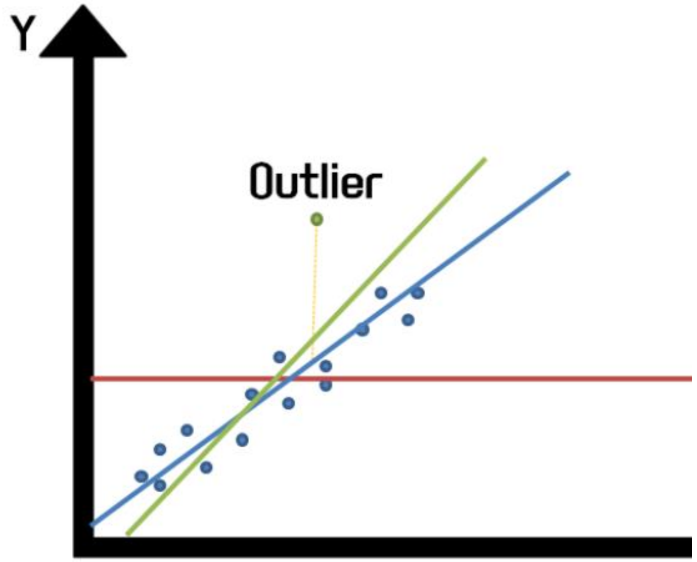
$$x_i = \frac{x_i - \mu}{\sigma}$$

이상점 (outlier)

- 자료 중 전체 형태에서 동떨어져서, 큰 잔차를 갖는 관측값
- 주어진 회귀 모델에 의해 잘 설명되지 않는 데이터

- R-student 잔차의 값이 크면, 이상점으로 판단

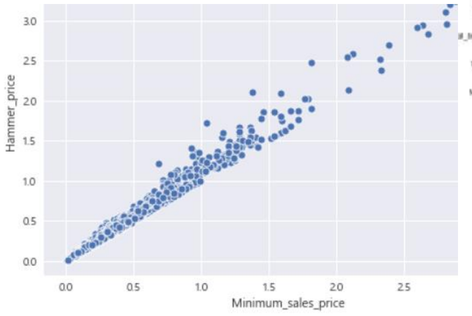
$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2(1-h_{ii})}}$$

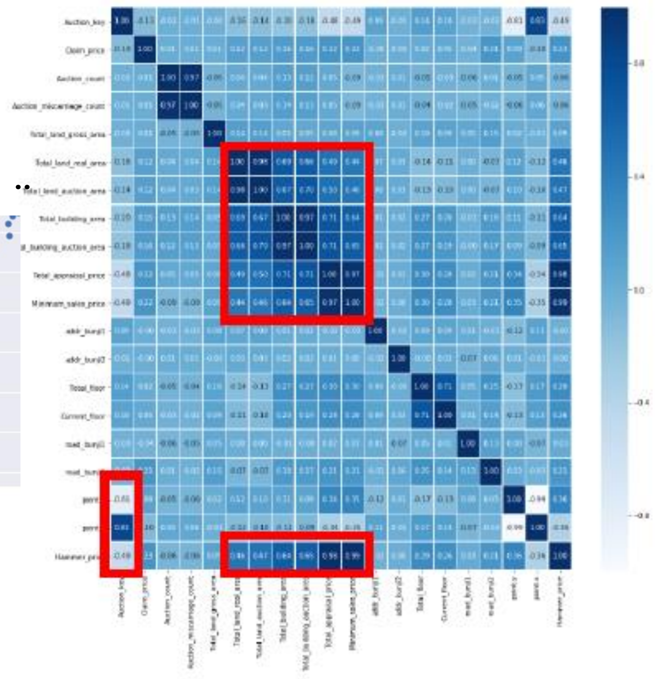


다중공선성 (Multicollinearity)

- 설명변수 간 **상관관계**가 강해 설명변수의 일부를 다른 설명변수의 **선형결합**으로 표현가능한 것

ex) X2=2X1, X3=X1+X2, ..





- 상관계수, scatter plot, heatmap 등 확인

- 다중공선성 진단통계량 : VIF
  - i 번째 설명변수를 다른 설명변수들로 회귀한 성능
  - 다른 설명변수들과 상관관계가 강할수록 VIF 값이 큼
  - **VIF > 10 이면 다중공선성 존재한다고 판단 !**

$$\mathrm{VIF}_i = \frac{\sigma^2}{(n-1)\mathrm{Var}[X_i]} \cdot \frac{1}{1-R_i^2}$$

(cf) 다중공선성을 제거하는 이유 ?

- 설명변수 간 독립적이지 않으면 회귀계수의 추정이 불안정하게 됨 !
- 추정값이 존재하지 않거나, 추정값의 분산이 매우 매우 커지거나 …

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

설명변수끼리 완벽한 선형관계가 존재하면
이 부분이 Full rank가 아니어서 역행렬 존재하지 않음

완벽한 선형관계가 아니더라도, 강한 다중공선성이 존재하면
이 부분이 작아서 역행렬을 취하면 값이 매우 커짐

-> 회귀 계수의 분산이 매우 커지게 되어 불안정한 추정이 됨

다중공선성 제거 방법

1. 더 많은 데이터 수집
2. 상관계수 가장 높은 변수 제거
3. 데이터 centering : $(x_i - \bar{x})$
4. PCA : 차원 축소 (dimension reduction) -> 향후 띵강 있을 예정 ..
5. Ridge / Lasso Regression

Regularization

- 학습 알고리즘이 현재 train data에 둔감하게 만들어, overfitting을 피하는 방식
- LSE : unbiased estimator ⇔ Regularization : biased 하지만 smaller variance를 갖는 estimator
- 모델이 복잡해질수록 penalty를 크게 주도록, 목적 함수에 항을 하나 더 추가

1. Ridge Regularization

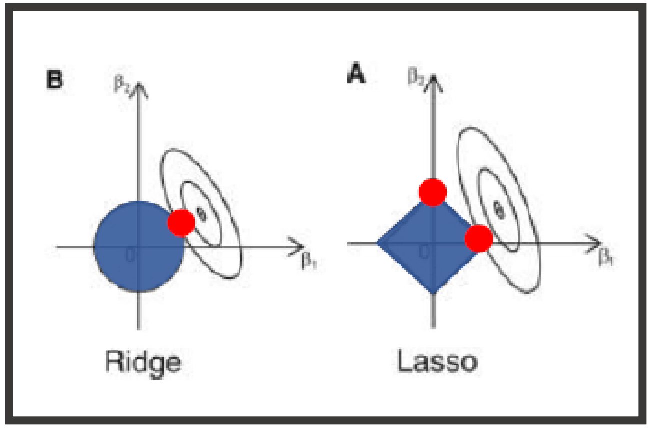
$$E(w) = \frac{1}{2}\sum_{n=0}^{N}(train_n - g(x_n, w))^2 + \boxed{\frac{\lambda}{2}\|w\|^2}$$

2. Lasso Regularization

$$E(w) = \frac{1}{2}\sum_{n=0}^{N}(train_n - g(x_n, w))^2 \boxed{+ \lambda|w|}$$

$$\frac{d}{dw}E(w) = 0$$

기존 Loss 식 뒤에 회귀계수 크기에 대한 제약조건 term이 붙은 새로운 Loss
새로운 Loss를 최소화시킴(최소제곱법)
W는 $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$, (회귀계수 열벡터)



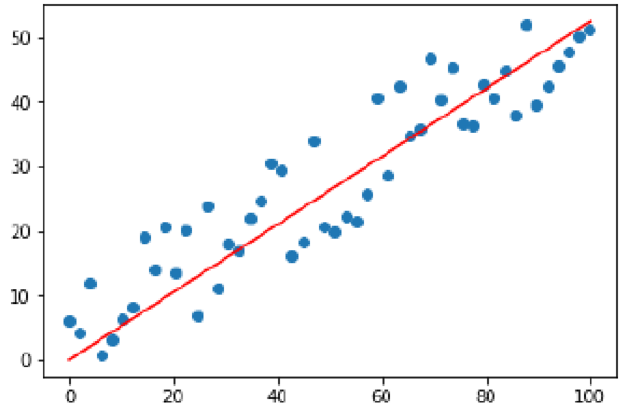
< 마무리 > 선형회귀분석

1. 회귀 모형 설정 : 반응변수 및 주요 설명변수 파악
2. 선형성 검토 : 산점도를 통해 상관관계 파악
3. 설명변수 검토 : 각 변수들의 분포 확인 + 다중공선성 파악
4. 모델 적합 : 모형의 회귀계수 추정 및 모형의 적절성 검토
5. 변수 선택 : 중요 설명변수 선택
6. 적합된 모형 검토 : 오차 가정 체크
7. 최종 모형 선택

Contents

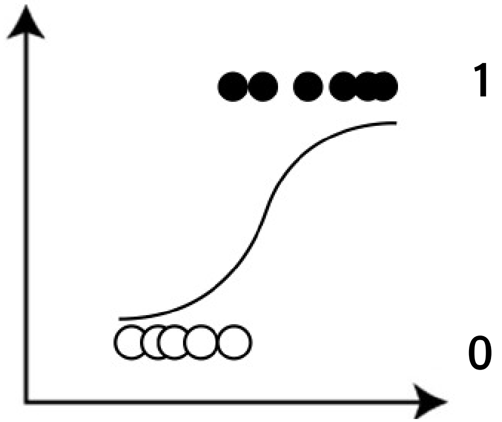
Unit 04 | 로지스틱 회귀분석

Linear Regression



연속형 Y 예측
Regression

Logistic Regression



범주형 Y 분류
Classification

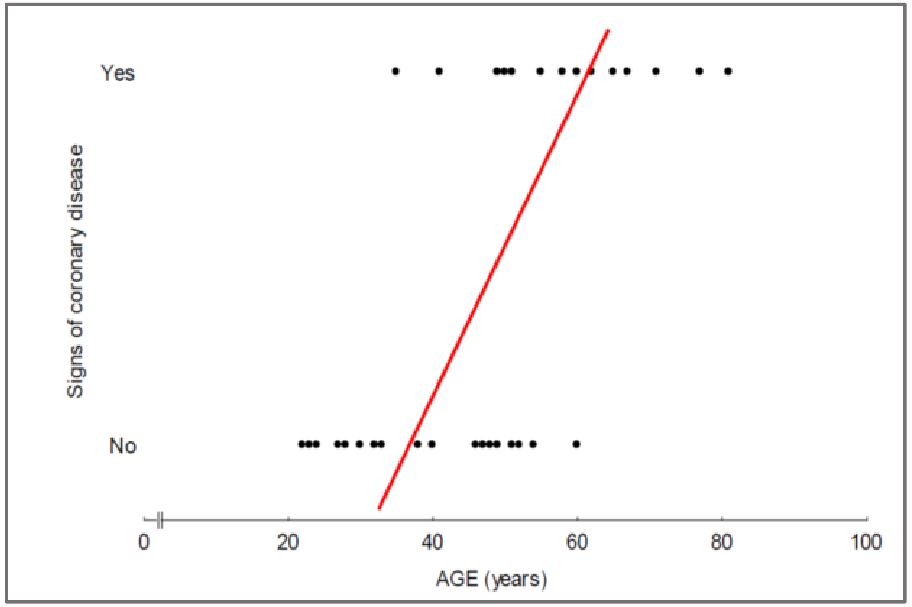
로지스틱 회귀분석

새로운 관측치가 있을 때, 이를 기존 범주 중 하나로 예측 (범주 예측)

로지스틱 회귀모델의 예시 : "분류"

- 제품이 불량인지 정상인지
- 고객이 이탈고객인지 잔류고객인지
- 카드 거래가 정상인지 사기인지
- 내원 고객이 질병이 있는지 없는지

범주형 변수를 선형회귀로 예측한다면 …?



범위가 일치하지 않음 !

1. 선형회귀 (-inf, +inf)
2. 로지스틱 0 / 1

중간 범주가 없고, 숫자가 아무런 의미를 지니지 않게 됨

-> Y가 범주형(categorical) 변수일 때는
다중선형회귀 모델을 그대로 적용할 수 없다 !

로지스틱 함수 = 확률값 예측 !

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \ Y_i$ = 0 or 1

Assume $E(\varepsilon_i) = 0$, $E(Y_i) = \beta_0 + \beta_1 X_i$

$$P(Y_i = 1) = \pi_i$$
$$P(Y_i = 0) = 1 - \pi_i$$

$$E(Y_i) = 1 * \pi_i + 0 * (1 - \pi_i) = \pi_i$$

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i$$

X값이 주어졌을 때,
출력변수 Y가 1의 값을 가질 확률

로지스틱 함수 = 확률값 예측 !

hw	happy
1	0
3	0
4	0
5	1
7	1

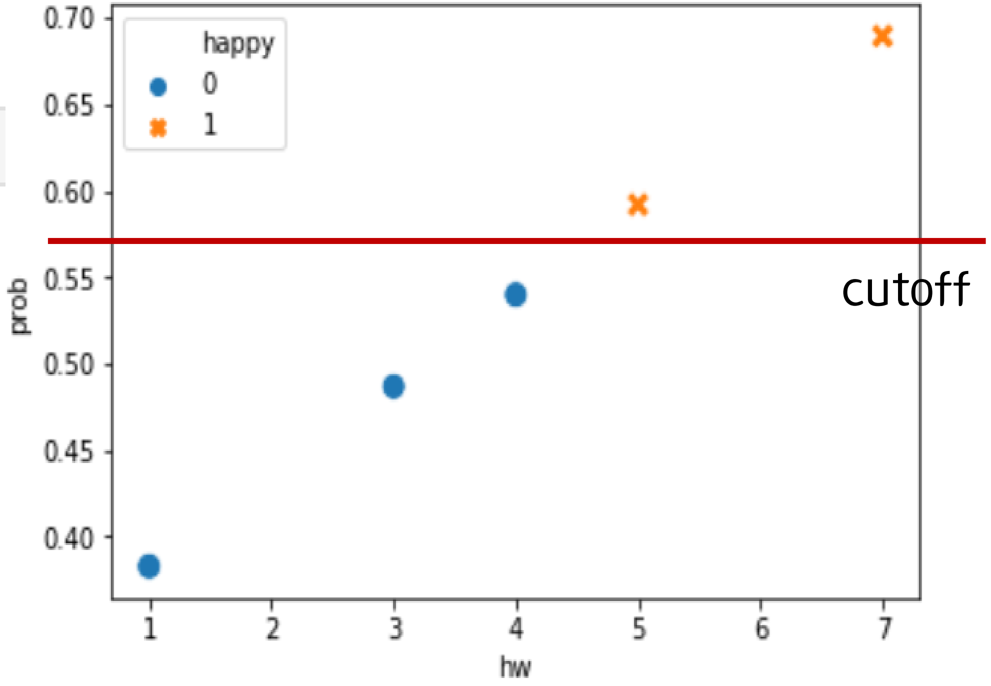
해결한 과제 수(x)에 따른 행복함 여부(y)

```
model.predict_proba(X.values.reshape(-1,1))
```

```
array([[0.61728793, 0.38271207],
       [0.5132455 , 0.4867545 ],
       [0.4602007 , 0.5397993 ],
       [0.40804243, 0.59195757],
       [0.31064148, 0.68935852]])
```

0 1

확률값 예측

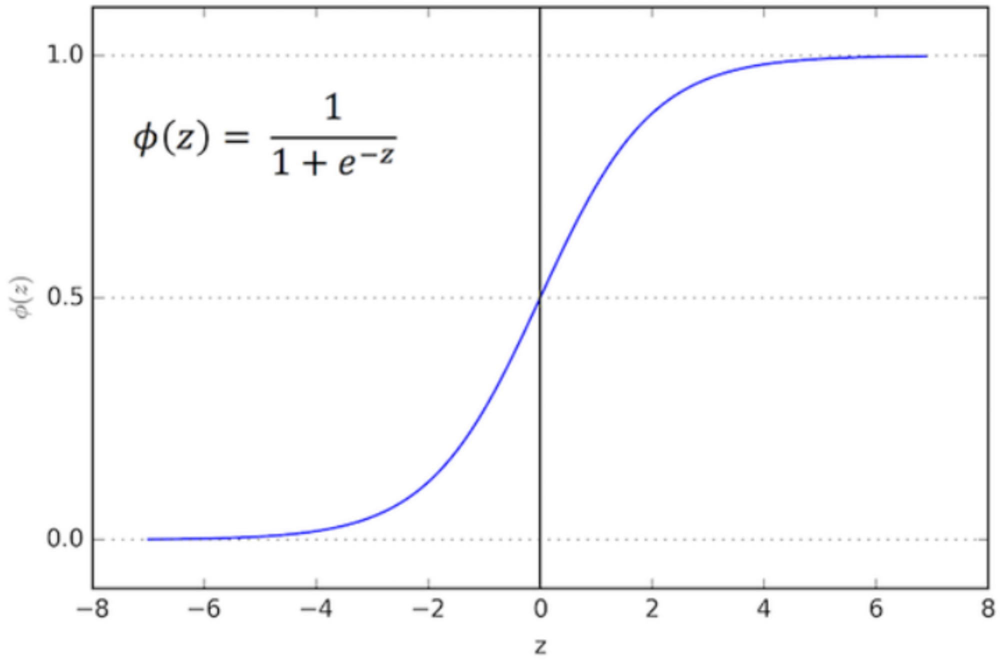


S-curve

$$f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

로지스틱 (Logistic) 함수
시그모이드 (Sigmoid) 함수

Output 범위 : (0, 1)
Input 값에 대해 단조증가 (or 단조감소)



관측치 x가 범주 1에 속할 확률

Odds : beta1의 의미

$$E(y) = \pi(X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

β_1의 해석 직관적이지 못함 !

$$Odds = \frac{p}{1-p} = \frac{\text{성공확률}}{\text{실패확률}}$$

Odds (승산)

성공 확률을 p로 정의할 때,
실패 대비 성공 확률의 비율

로짓 변환 (Logit Transformation)

$$\log(Odds) = \log\left(\frac{\pi(X=x)}{1-\pi(X=x)}\right) = \log\left(\frac{\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}{1-\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}\right) = \beta_0 + \beta_1 x$$

β_1 의 의미 : x가 한 단위 증가했을 때, log(Odds)의 증가량

beta1 해석

- log-odds at x : $\hat{\eta}(x) = \log\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = \beta_0 + \beta_1 x$

- log-odds at x+1 : $\hat{\eta}(x+1) = \log\left(\frac{\hat{\pi}(x+1)}{1-\hat{\pi}(x+1)}\right) = \beta_0 + \beta_1(x+1)$

- Difference in the log-odds : $\hat{\eta}(x+1) - \hat{\eta}(x) = \widehat{\beta_1}$

(cf) Odds Ratio

$$\widehat{O_R} = \frac{Odds_{x+1}}{Odds_x} = e^{\widehat{\beta_1}}$$

OR > 1 : 입력변수가 목표변수에 양의 방향으로 영향을 미침
OR < 1 : 입력변수가 목표변수에 음의 방향으로 영향을 미침

ex) OR = 1.2 : 설명변수 한 단위 증가함에 따라 반응변수 발생확률이 20% 증가한다

회귀 계수의 해석

- Linear Regression : 설명변수가 1만큼 증가함에 따른 반응변수의 변화량

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \cdots + \widehat{\beta_p}x_p$$

- Logistic Regression : 설명변수가 1만큼 증가함에 따른 로그 오즈의 변화량

$$\log(Odds) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \cdots + \widehat{\beta_p}x_p$$

로지스틱 확률값 구하기 !

log(Odds) 값을 반응변수로 사용한 회귀식을 추정한 뒤,
구한 odds를 기준으로 역산을 통해 **확률값** 예측 !

$$\log(Odds) = \log\left(\frac{p}{1-p}\right) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \cdots + \widehat{\beta_k}x_k$$

$$\Rightarrow \quad \frac{p}{1-p} = e^{\widehat{\beta_0}+\widehat{\beta_1}x_1+\cdots+\widehat{\beta_k}x_k} \quad \Rightarrow \quad p = \frac{1}{1+e^{-(\widehat{\beta_0}+\widehat{\beta_1}x_1+\cdots+\widehat{\beta_k}x_k)}} = \sigma(x,\beta)$$

MLE (Maximum Likelihood Estimtion) : 최대 우도 추정법

선형회귀분석(최소제곱법)과 달리, MLE로 계수를 추정한다 !

$$\pi(X) = \frac{1}{1 + e^{-(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_k} x_k)}}$$

왜? 비선형 함수라서 !

$$\hat{\theta} = \underset{\theta}{argmax}\, L(\theta)$$

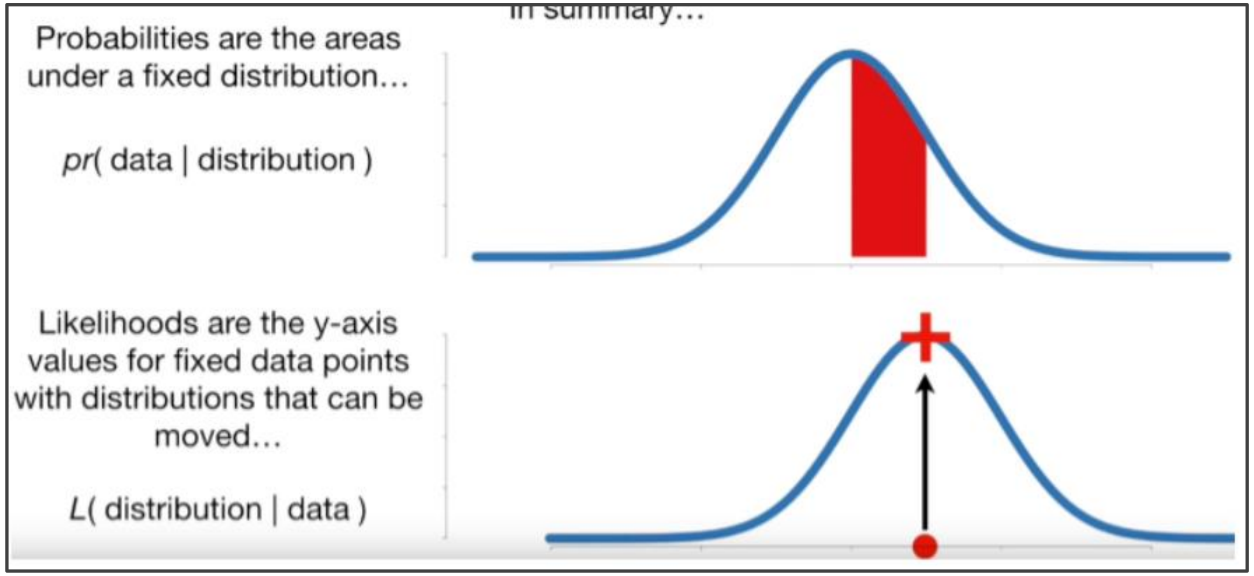
Likelihood를 최대화하는 parameter 추정 !

Likelihood

- Probability
  주어진 확률분포에서 해당 관측값이 나올 확률

- Likelihood
  어떤 값이 관측되었을 때,
  이것이 어떤 확률분포에서 왔을지에 대한 확률

  데이터가 있을 때, 어떤 결과가 일어날 가능성 !



- $L(\theta) = f_\theta(x_1, x_2, \dots, x_n) \xrightarrow[\text{i.i.d.}]{\text{Joint distribution}} L(\theta) = \prod_i f_\theta(x_i)$

MLE in 로지스틱 회귀

- 관측값 y_i에서의 확률 분포 : $f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, i = 1, 2, \ldots, n$

- Likelihood Function : $L(\mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

- log-likelihood : $\ln L(\mathbf{y}, \boldsymbol{\beta}) = \ln \prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\pi_i)$ →

MLE in 로지스틱 회귀

$$\ln L = \sum y_i(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) - \sum \ln(1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k})$$

(cf) $\log(Odds) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ & $\pi_i = \frac{1}{1+e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$

log-likelihood function이 **최대**가 되는 **파라미터 β** 결정 !

- log-likelihood 함수는 비선형 함수이므로, 선형회귀 모델처럼 명시적인 해가 존재하지 않음
- 따라서 Gradient Descent 등의 수치 최적화 알고리즘을 이용해 해를 구합니다 !

Cross Entropy

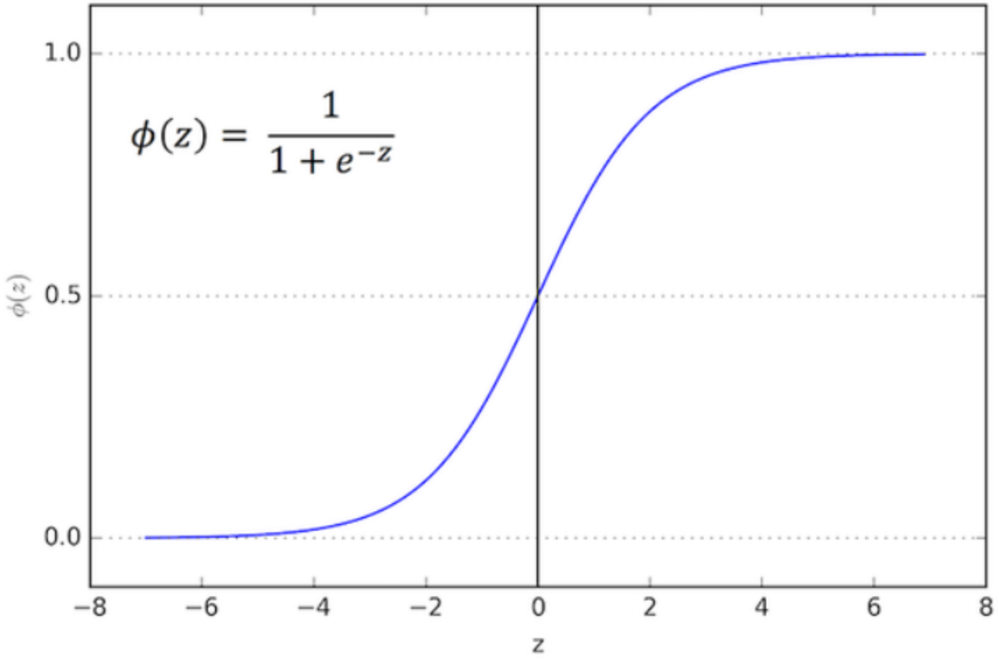
- 분류에서의 학습을 위한 손실함수 = 입력값과 출력분포의 차이를 최소화

- Likelihood : $L(\mathbf{y},\boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$
- log-likelihood : $\log\big(L(p)\big) = \sum_{i=1}^{n}[y_i \log(p) + (1-y_i)\log(1-p)]$

- Cross-Entropy Loss : $\boxed{-\log(L) = -\sum_{i=1}^{n}[y_i \log(p) + (1-y_i)\log(1-p)]}$

- Cross-Entropy Loss Minimize = log-likelihood Function Maximize

최종 로지스틱 모델



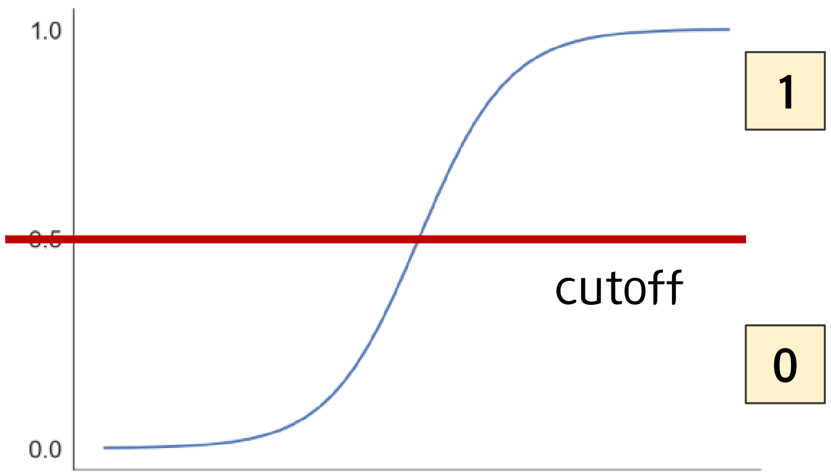
최적 파라미터를 적합시킨 모델

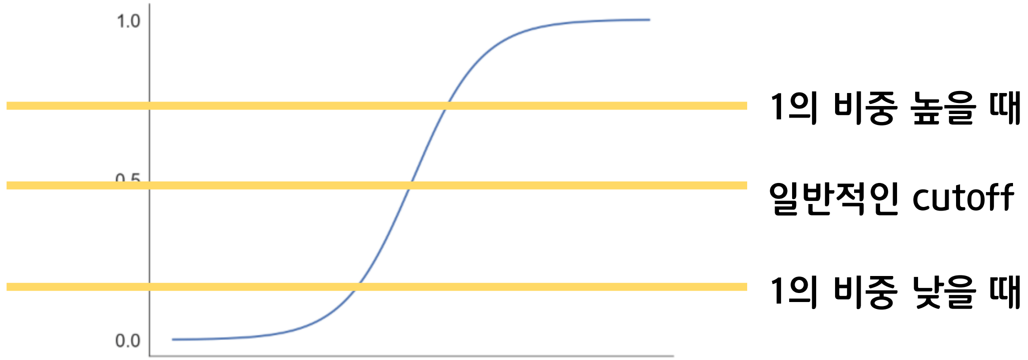
$$\pi(X) = f(X) = \frac{1}{1 + e^{-(\widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_k} x_k)}}$$

$$= \frac{1}{1 + e^{-\hat{\beta} X}}$$

Cutoff (=Threshold)

- Classification 을 위한 기준값
- 로지스틱 함수로부터 구한 성공확률이 cutoff 이상이면 1 / cutoff 이하이면 0 으로 분류





✓ 사전 확률을 고려한 cutoff
✓ 검증 데이터의 성능을 최대화하는 cutoff

https://ratsgo.github.io/machine%20learning/2017/04/02/logistic/

(cf) Multiclass 범주에서의 로지스틱 회귀

- Binary Classification : log(Odds)
- Multiclass Classification : Baseline logit model

$$\log \frac{P(Y=1|X=\vec{x})}{P(Y=3|X=\vec{x})} = \beta_1^T \vec{x}$$

$$\log \frac{P(Y=2|X=\vec{x})}{P(Y=3|X=\vec{x})} = \beta_2^T \vec{x}$$

Y=3을 기준으로 하는 baseline logit model
P(Y=3) = 1 – P(Y=1) – P(Y=2)
-> 두 개의 계수 추정만 이루어지게 됨 !

앞에서 했던 것처럼 로그 확률비를 확률의 형태로 변환하고,
일반화된 형태를 취하면, 다음과 같은 형태를 보임

$$P(Y=c) = \frac{e^{\vec{\beta_c}^T \vec{x}}}{\sum_{k=1}^{K} e^{\vec{\beta_k}^T \vec{x}}}$$

c번째 범주에 속할 확률
⇔
Neural Network의 활성화 함수로 쓰이는
Softmax 함수와 동일한 형태 !

Model Evaluation

Confusion Matrix

	P' 1 (Predicted)	N' 0 (Predicted)
1 P (Actual)	True Positive	False Negative
0 N (Actual)	False Positive	True Negative

1. True / False: 예측이 정확한가(T) 아닌가(F)?
2. Positive / Negative :
   1로 예측하면 Positive, 0으로 예측하면 Negative

Accuracy : 정확도

- 예측 결과가 실제와 얼마나 동일한지 측정
- 실제 분포가 skewed 되어 있는 경우 적합하지 않음

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

ex. Y = 질병 유무

질병이 없는 경우(Y=0)가 질병이 있는 경우(Y=1)보다 훨씬 많을 것!
이 때 분류 모형을 학습시키게 되면 Y=0일 때를 더 많이 학습하게 됨
-> 실제 데이터와 무관하게 Y=0이라고 예측할 확률이 커짐

즉, Accuracy는 TN, TP를 한번에 고려하므로,
TN은 높지만 TP가 낮은 경우는 고려하지 못하게 됨 !

Precision and Recall

		실제 정답 Actual	
		P	N
분류 결과 Predicted	P	True Positive	False Positive
	N	False Negative	True Negative

Confusion Matrix

Precision : 정밀도

- True라고 분류한 것 중에서 실제 True인 것의 비율

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

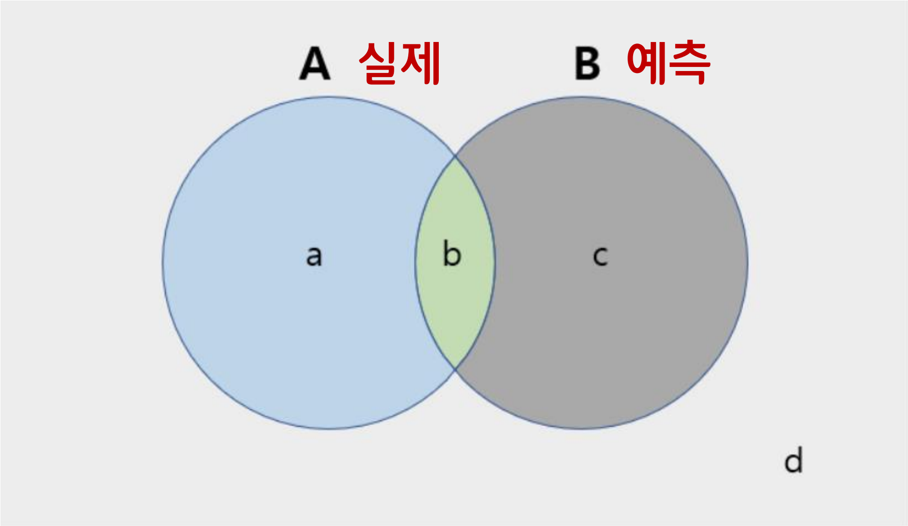
Recall (= sensitivity) : 재현율

- 실제 True인 것 중에서 True라고 예측한 것의 비율

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

(cf) Precision과 Recall은 Trade-off 관계

-> 두 개의 값을 동시에 높일 수 없다!



Precision = $\frac{b}{b+c}$, Recall = $\frac{b}{a+b}$

a 부분이 c로 다 흡수된다면..?

		실제 정답	
		True	False
분류 결과	True	TP(20)	FP(40)
	False	FN(30)	TN(10)

Precision = $\frac{20}{60}$ = 33.3%

Recall = $\frac{20}{50}$ = 40%

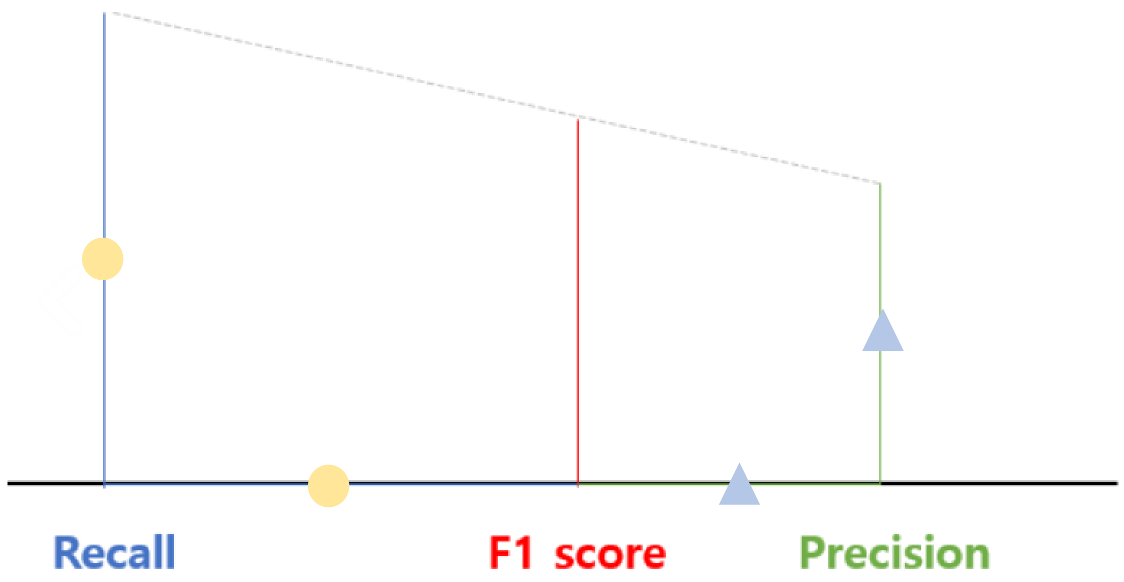
		실제 정답	
		True	False
분류 결과	True	TP(20)	FP(80)

Precision = $\frac{20}{100}$ = 20%

Recall = $\frac{20}{20}$ = 100%

F1 score

Precision과 Recall의 조화평균

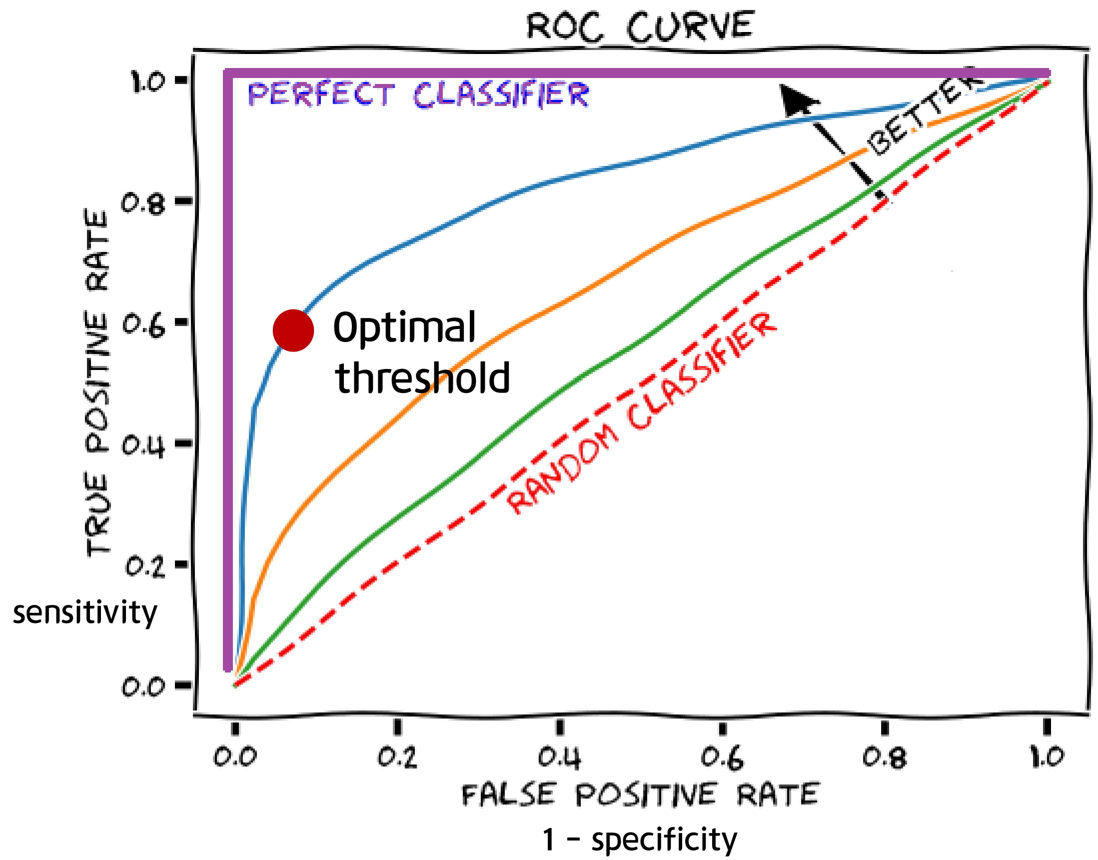


F1 score

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

$$= 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ROC Curve



여러 cutoff value 값을 기준으로
-> confusion matrix에서 sensitivity, specificity 계산
-> 값을 기준으로 그림을 그린 것

AUC (=Area Under Curve)
= ROC curve의 넓이 (0.5≤AUC≤1)
= 값이 클수록 모델의 성능이 좋다

$$\text{Sensitivity} = \frac{TP}{TP+FN}\ ,\ \ \text{Specificity} = \frac{TN}{FP+TN}$$

< 마무리 > 로지스틱 회귀분석

1. 범주형 변수 Y 분류
2. f(x) = 1 / (1+exp(-X*beta)) : 확률값 예측
3. Logit = log(Odds) = log(p/(1-p))
4. Beta1 = log(Odds)의 변화량
5. 최적 parameter (beta) = MLE 통해 !
6. 목적함수 Cross Entropy : 입력값과 출력값의 차이 최소화
7. Cutoff value 를 통해 Classification 성능을 바꿀 수 있다

Assignment

<과제1> 행렬 구현

- LSE normal equation, MSE 구현 (Assignment1 파일에서 함수 구현하기)

<과제2> 회귀분석 : Used Car Price Prediction

- 자유롭게 EDA, 전처리 및 파생변수 생성
- 회귀분석의 기본 가정 검토
- 변수 제거, 선택 시 이유 설명
- 다중공선성 확인, 처리
- 모델 평가

실습 때 제가 했던 것처럼 해 주세요 !! ㅎ_ㅎ

Assignment

<과제3> 로지스틱 회귀분석 : Credit Card Fraud Detection

- sklearn 패키지를 사용해 로지스틱 회귀모형 적합
- 성능지표 계산 (sklearn.metrics 및 confusion matrix 이용)
- 최적의 cutoff 값을 ROC 커브를 사용해 찾아보고, 다시 예측 진행해 성능 평가하기

해당 데이터셋은 사기 탐지 관련 데이터이기 때문에, 상당히 imbalance 합니다.
이에 기반하여 새롭게 cutoff 를 찾은 후, 해석을 상세하게 달아주세요!

Reference

<회귀분석>

투빅스 12기 이홍정님 강의자료, 투빅스 11기 심은선님 강의자료 / 투빅스 2기 김상진님 강의자료

이화여자대학교 통계학과 임용빈 교수님 강의

유의수준과 p값 : https://m.blog.naver.com/vnf3751/220830413960 , https://adnoctum.tistory.com/332

<로지스틱 회귀분석>

투빅스 12기 이유진님 강의자료, 투빅스 11기 이영전님 강의자료

rat's go blog : https://ratsgo.github.io/machine%20learning/2017/04/02/logistic/

Probability & Likelihood : https://jjangjjong.tistory.com/41 , https://www.youtube.com/watch?v=pYxNSUDSFH4

분류 성능 평가지표 : https://sumniya.tistory.com/26

Q & A

들어주셔서 감사합니다.