3주차 : 회귀분석 기초 (선형 회귀 및 릿지, 라쏘)

9기 김경덕 & 최지희

Kuggle

Contents

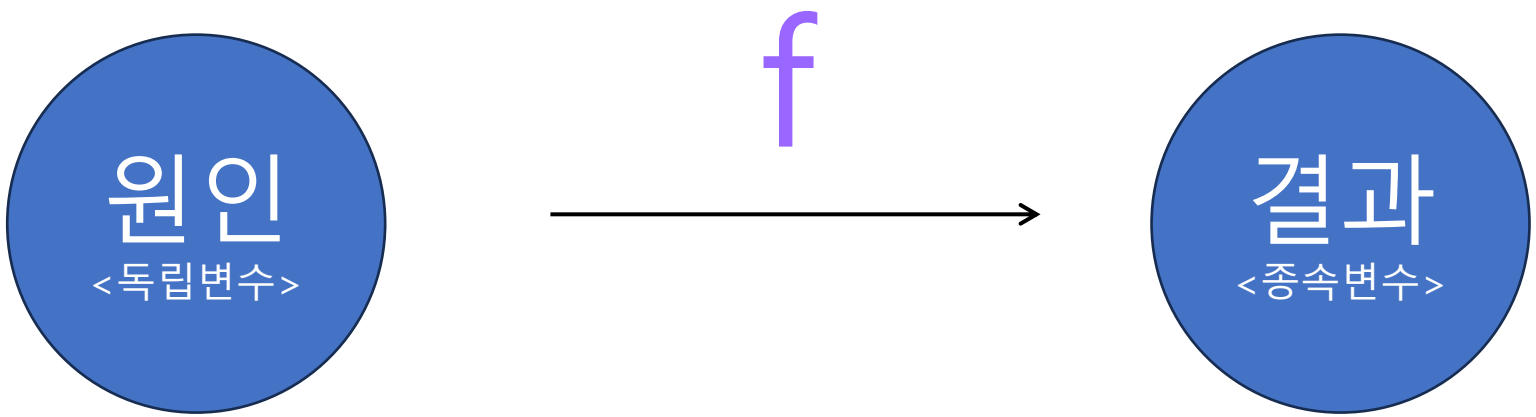
1. 회귀분석 정의, 분류

2. 과(대)/과소 적합

3. 회귀 평가 지표

4. 규제 선형 모델 - 릿지, 라쏘

1. 회귀분석이란?



X1 : 아파트 방의 개수

X2 : 아파트 위치

X3 : 아파트 주변 학군

Y : 아파트 가격

1. 회귀분석이란?

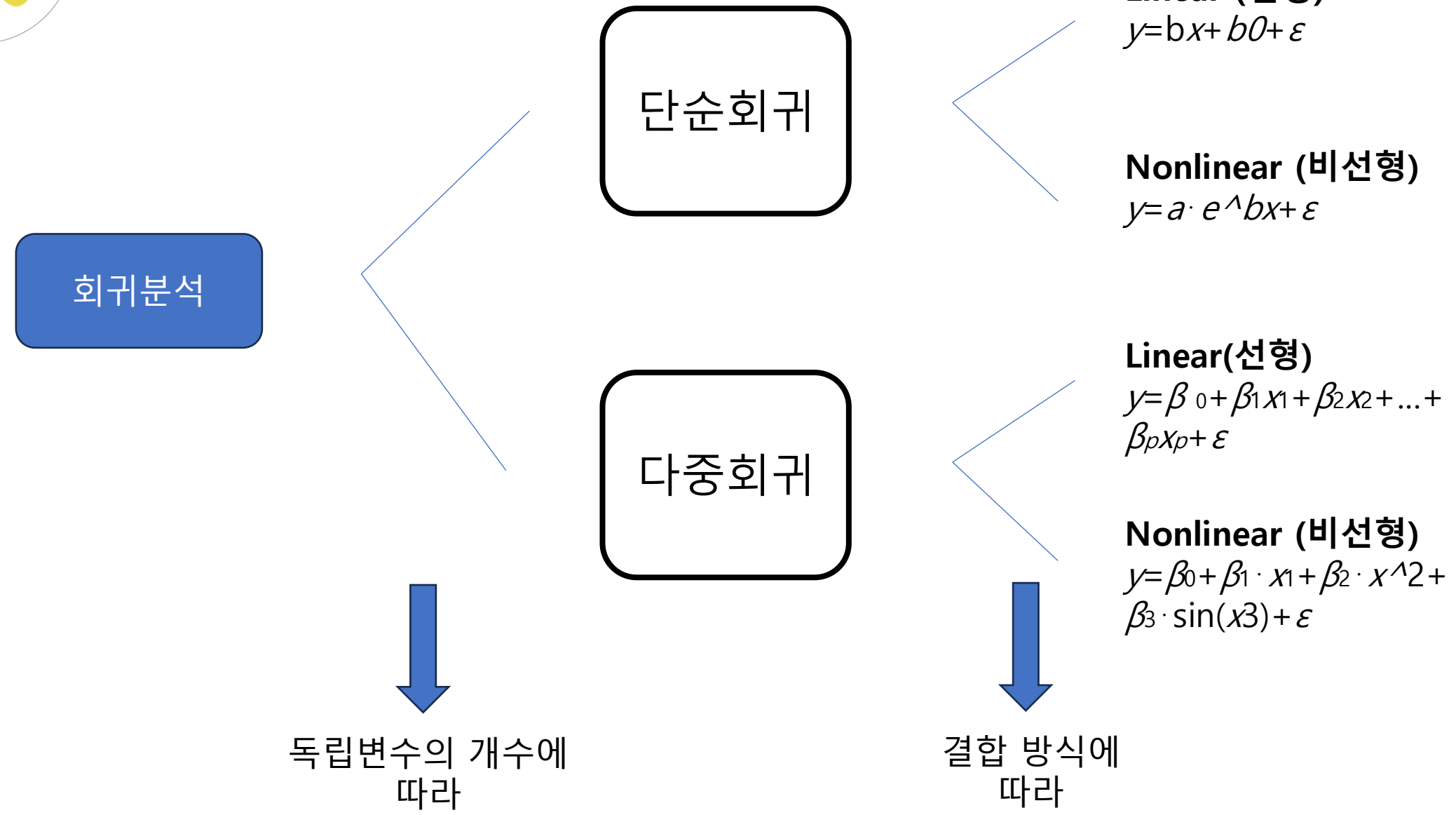
회귀분석 (Regression Model)

: 독립변수들과 종속변수 간의 관계를 모델링하고 예측하는 것

$Y=f(x1, x2, x3, x4, \ldots\ldots xn)$

X3 : 아파트 주변 학군

1. 회귀분석의 분류


회귀분석
단순회귀
다중회귀
Linear (선형)
y=bx+b0+ε
Nonlinear (비선형)
y=a·e^bx+ε
Linear(선형)
y=β0+β1x1+β2x2+…+βpxp+ε
Nonlinear (비선형)
y=β0+β1·x1+β2·x^2+β3·sin(x3)+ε
독립변수의 개수에 따라
결합 방식에 따라

1. 회귀분석의 분류

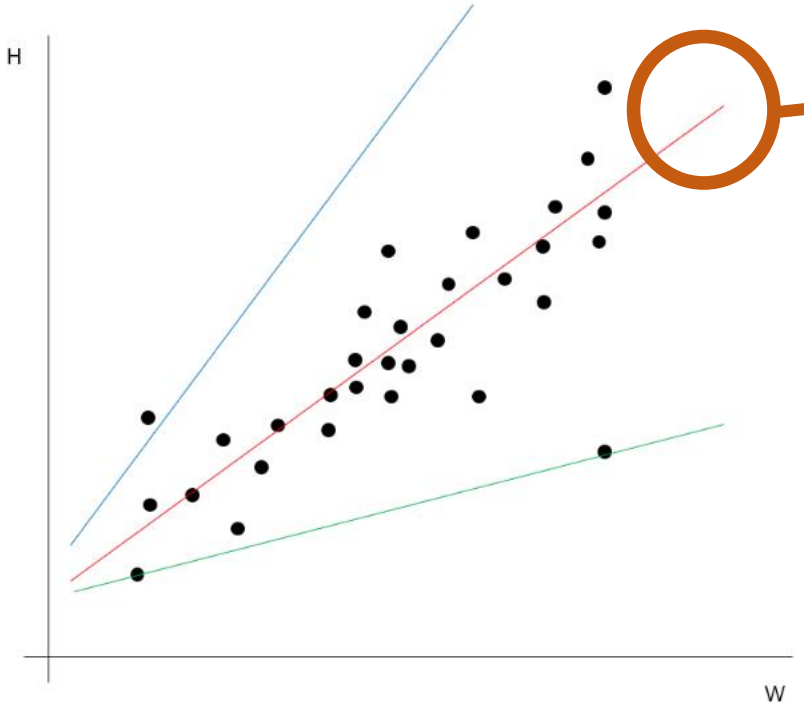
단순선형회귀(Simple linear regression)

단순선형회귀 모델 $Y = \beta_0 + \beta_1 X + \varepsilon$

- β_0: constant, intercept
- β_1: slope, coefficient
- ε: error, 오차, x로 설명되지 않는 어떤 것

1. 단순회귀

단순선형회귀(Simple linear regression)



최적의 회귀직선

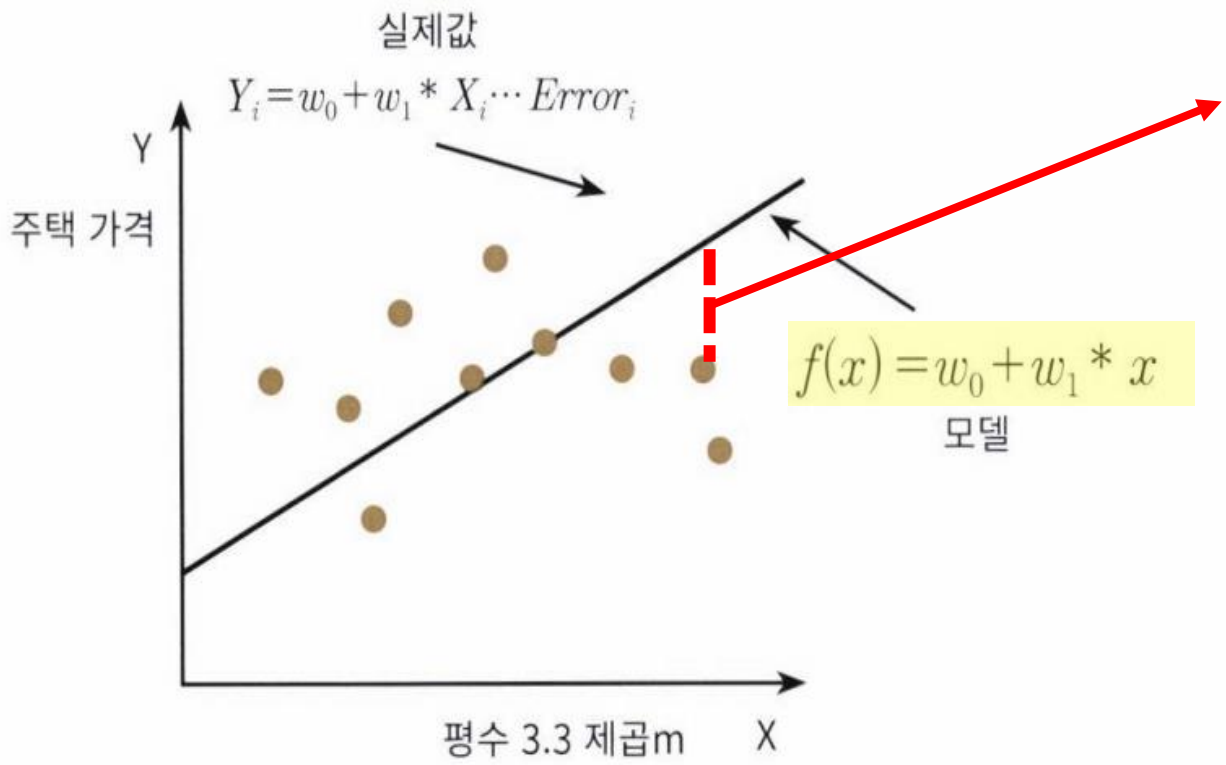
실제 관측된 값(점)과
우리가 고려하고 있는 여러가지 직선들
사이의 거리를 측정하여
그 거리(=오차)를 가장 작게 해주는 선

-> 최소제곱의 아이디어

사진 출처 : https://www.blog.ecsimsw.com/entry/%EC%84%A0%ED%98%95-%ED%9A%8C%EA%B7%80-Linear-Regression

1. 단순회귀

RSS



실제 값과 모델 사이의 오류 값 : 잔차

-> 절댓값이여서 모두 양수

편차의 합=0 -> 제곱합 , 절댓값 활용

최소제곱합 : 미분 이용

$$RSS(w_0, w_1) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + w_1 * x_i))^2$$

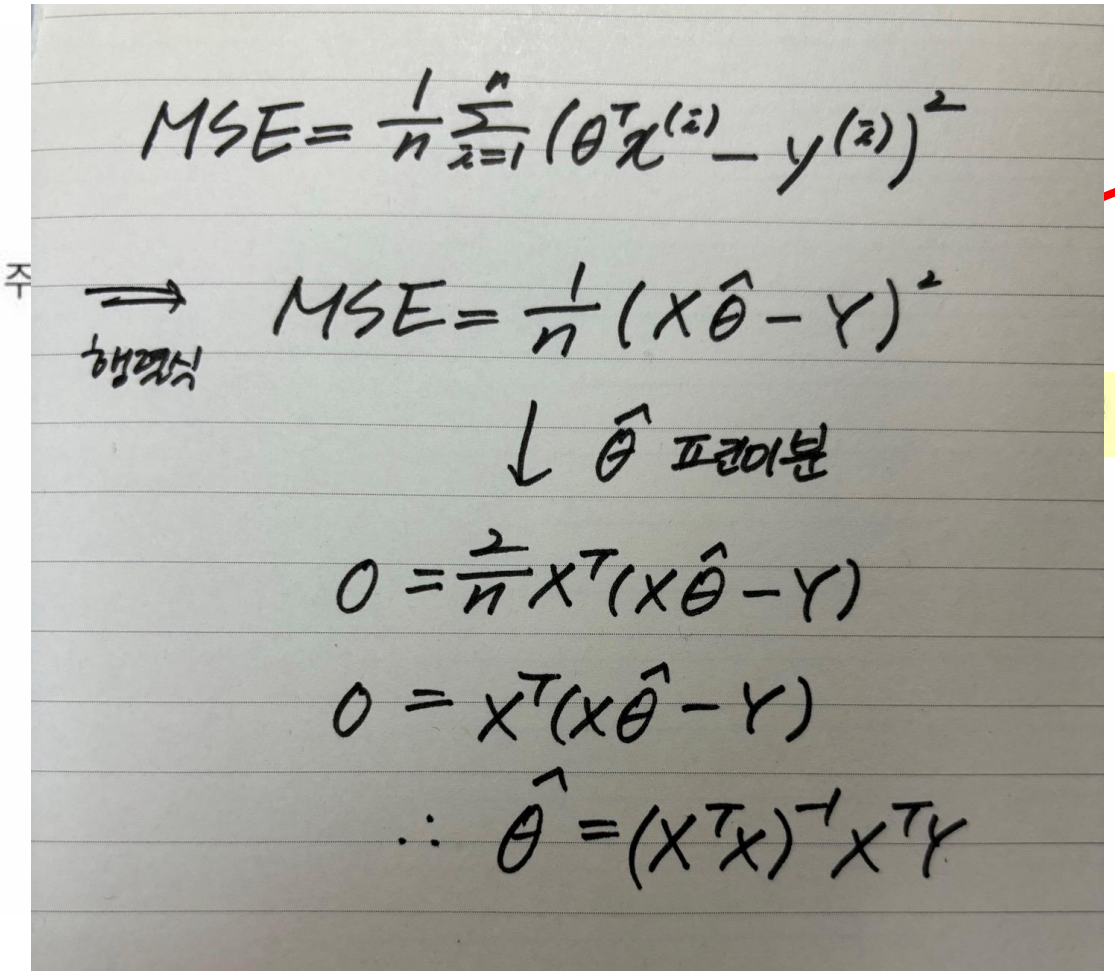
(i는 1부터 학습 데이터의 총 건수 N까지)

Loss function 는 최소화 하려는 식을 지칭

-> Loss function을 최소화 하는 w0, w1 찾기 문제

1. 단순회귀

RSS – 정규방정식



실제 값과 모델 사이의 오류 값 : 잔차

-> 절댓값이여서 모두 양수

편차의 합=0 -> 제곱합 , 절댓값 활용

최소제곱합 : 미분 이용

$$RSS(w_0, w_1) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + w_1 * x_i))^2$$

{i는 1부터 학습 데이터의 총 건수 N까지}

Lc $\hat{\theta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$ 식을 지칭

-> w0, w1 찾기 문제

1. 단순회귀

정규방정식

$$RSS(w_0, w_1) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + w_1 * x_i))^2$$

(i는 1부터 학습 데이터의 총 건수 N까지)

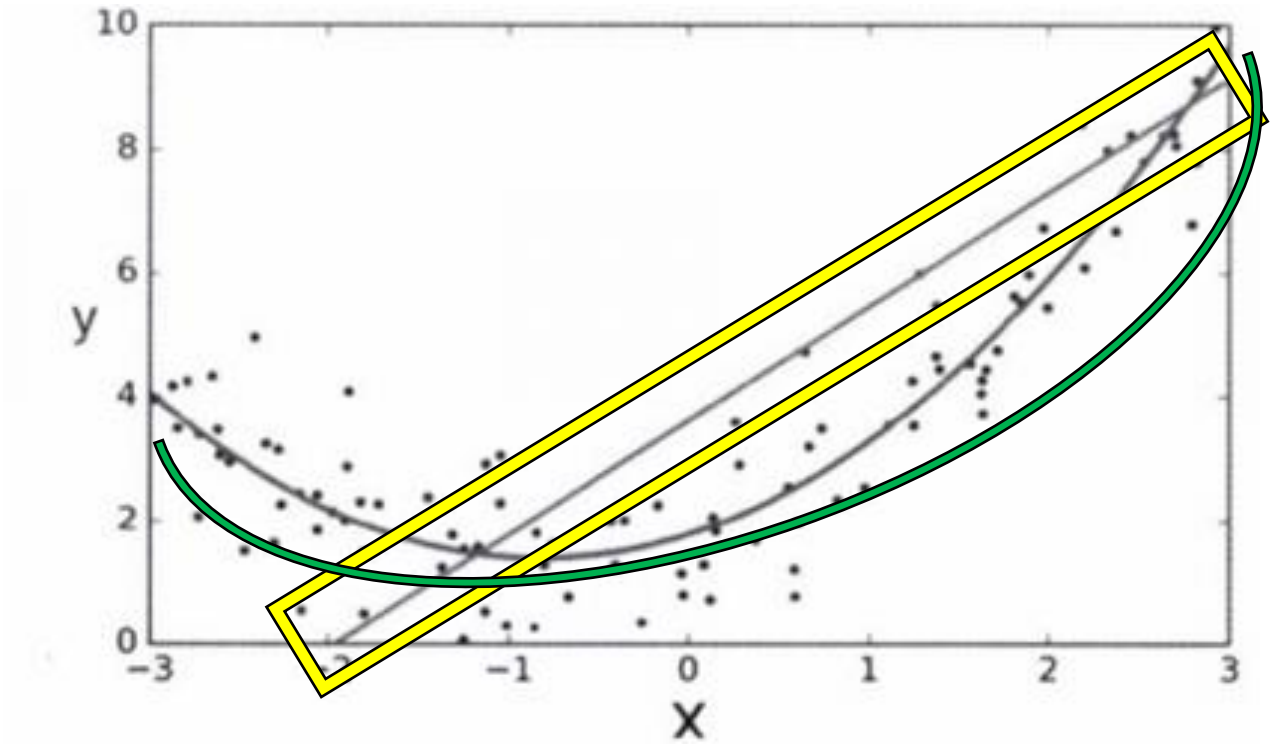
Loss function 는 최소화 하려는 식을 지칭

-> Loss function을 최소화 하는 w0, w1 찾기 문제

회귀에서 이 RSS는 비용이라 하며
이 비용을 최소로 하게 하는 W_0 ,W_1을
학습을 통해 찾는 것이 머신러닝 기반 회귀의
핵심 사항이다.

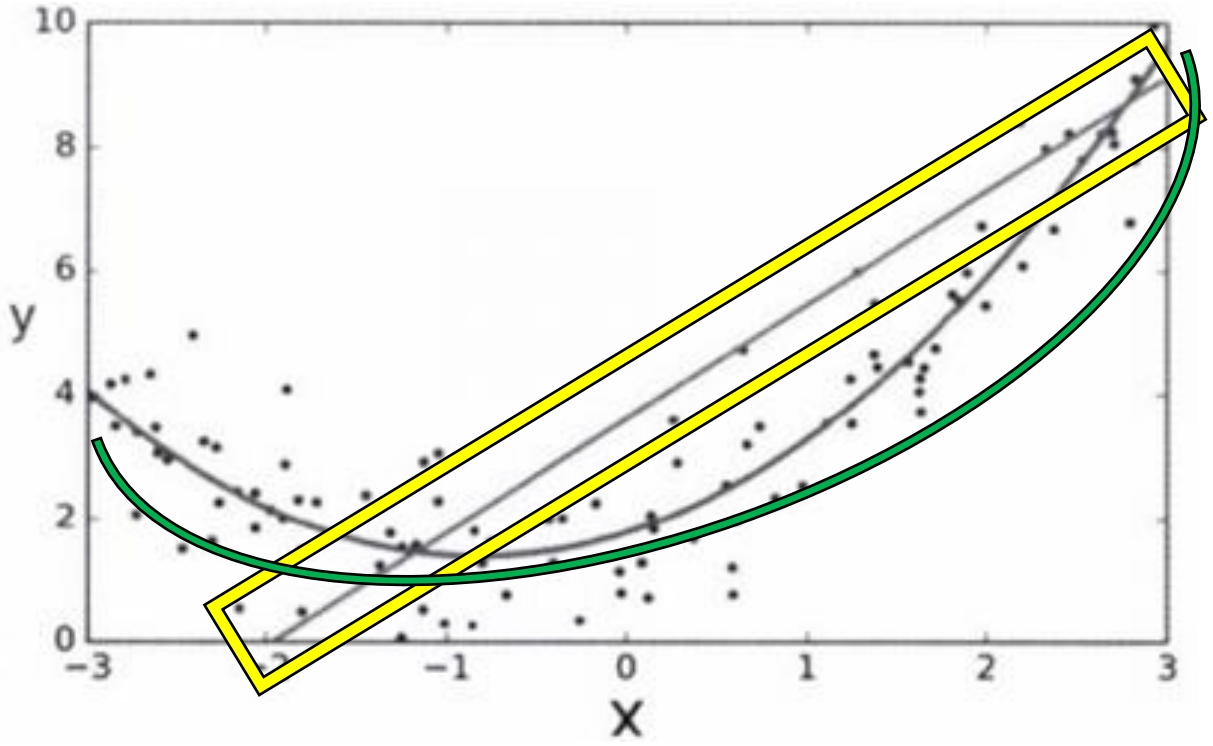
2. 다중회귀

다항회귀



2. 다중회귀

다항회귀



선형회귀

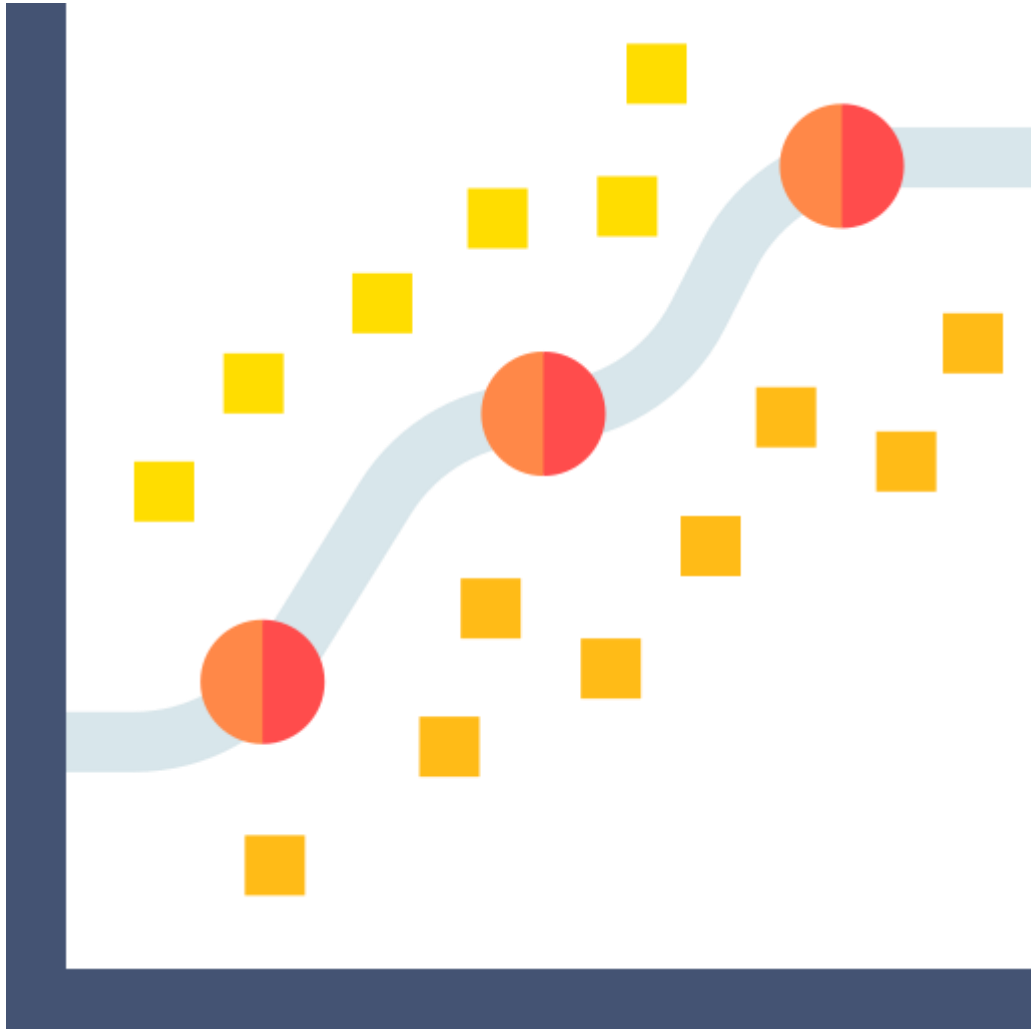
$$y=w0+w1*z1+w2*z2+w3*z3+w4*z4+w5*z5$$

다항회귀

$$y=w0+w1*x1+w2*x2+w3*x1*x2+w4*x1^2+w5*x2^2$$

2. 과(대)/과소 적합

다항회귀



다항회귀

- 독립변수가 다항식으로 표현
- 복잡한 비선형 관계 모델링
- 적절한 **차수** 선택이 중요

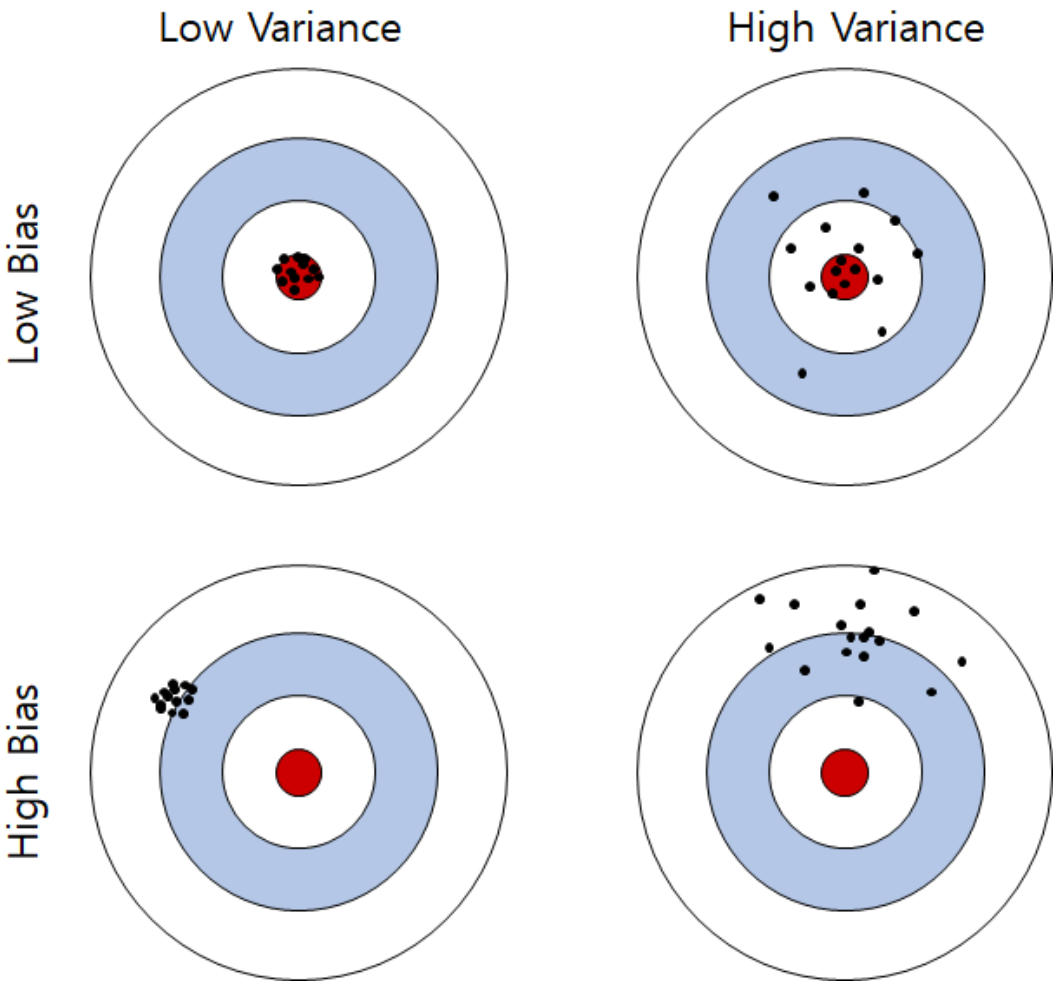
차수(degree)가 커질수록

장점 : 복잡한 변수(피처) 모델링 가능

단점 : 예측 정확도가 떨어질 수 있음

2. 과(대)/과소 적합

편향과 분산

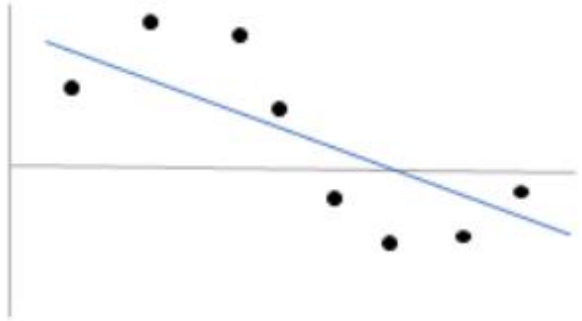


편향(Bias) :
모델의 예측값과 실제 값의 차이

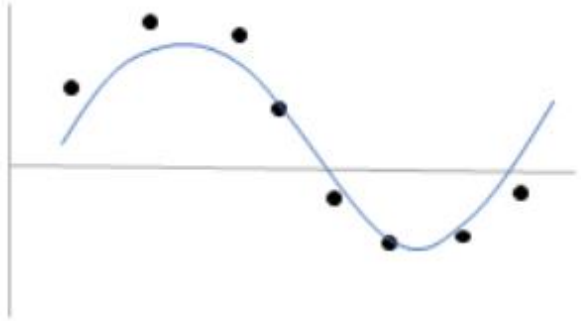
분산(Variance) :
동일한 모델에 다른 데이터셋을 사용할 때, 예측값 간의 변동성

2. 과(대)/과소 적합

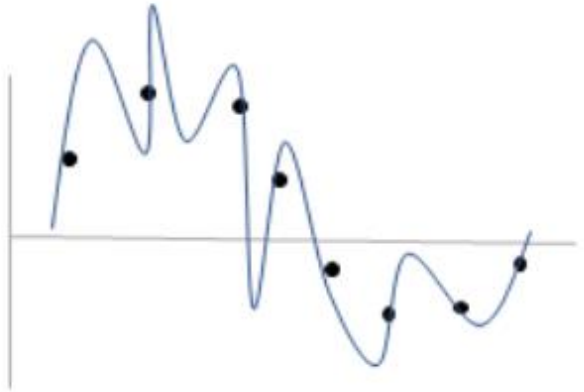
편향과 분산



High Bias – Low Variance



Medium Bias – Variance



Low Bias – High Variance

과(대) 적합
고편향

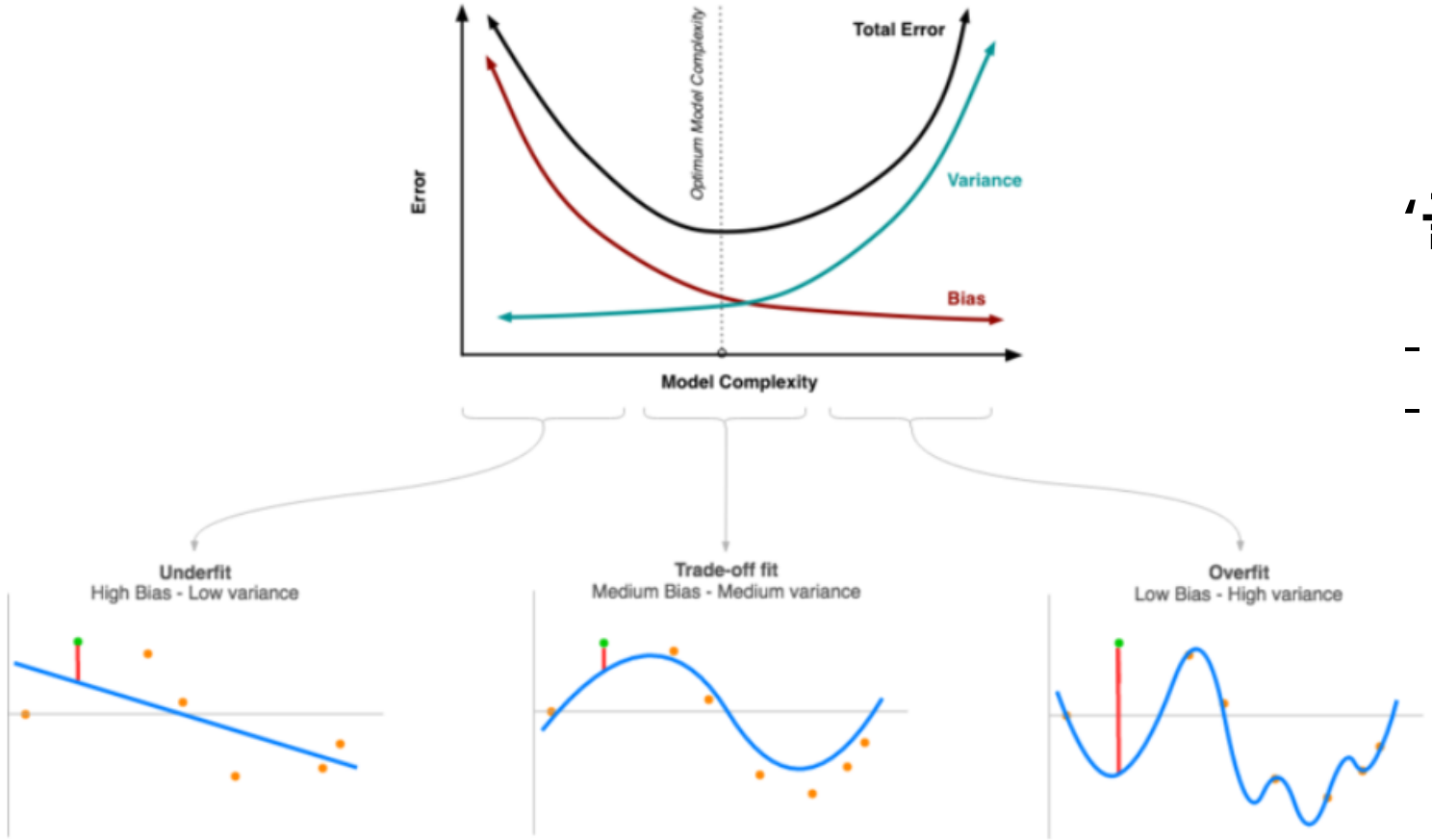
과소 적합
고분산

사진 출처 : https://m.blog.naver.com/ckdgus1433/221594203319

2. 과(대)/과소 적합

편향-분산 트레이드오프

편향-분산 트레이드오프



'골디락스' 지점

- 최적화 지점
- 편향은 낮추고 분산은 높여 전체 오류가 가장 낮아지는 점

3. 회귀 평가 지표

회귀평가지표

회귀 평가 지표 – 회귀의 성능을 평가하는 지표

평가 지표	설명	수식
MAE	Mean Absolute Error(MAE)이며 실제 값과 예측값의 차이를 절댓값으로 변환해 평균한 것입니다.	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert Yi - \hat{Y}i\rvert$
MSE	Mean Squared Error(MSE)이며 실제 값과 예측값의 차이를 제곱해 평균한 것입니다.	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)$
RMSE	MSE 값은 오류의 제곱을 구하므로 실제 오류 평균보다 더 커지는 특성이 있으므로 MSE에 루트를 씌운 것이 RMSE(Root Mean Squared Error)입니다.	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2}$
R^2	분산 기반으로 예측 성능을 평가합니다. 실제 값의 분산 대비 예측값의 분산 비율을 지표로 하며, 1에 가까울수록 예측 정확도가 높습니다.	$R^2 = \frac{\text{예측값 } Variance}{\text{실제값 } Variance}$

3. 회귀 평가 지표

회귀평가지표

RMSE를 구하는 이유?

MSE의 단점이 뚜렷하기 때문이다.

1. 오차의 합을 제곱한 것이기 때문에 에러의 차원과 MSE의 차원이 서로 다름

2. 제곱값이기 때문에 값이 매우 커질 수 있음

   -> 루트만 씌웠을 뿐인데 단점을 해결할 수 있음

3. 회귀 평가 지표

R^2

결정계수 R^2

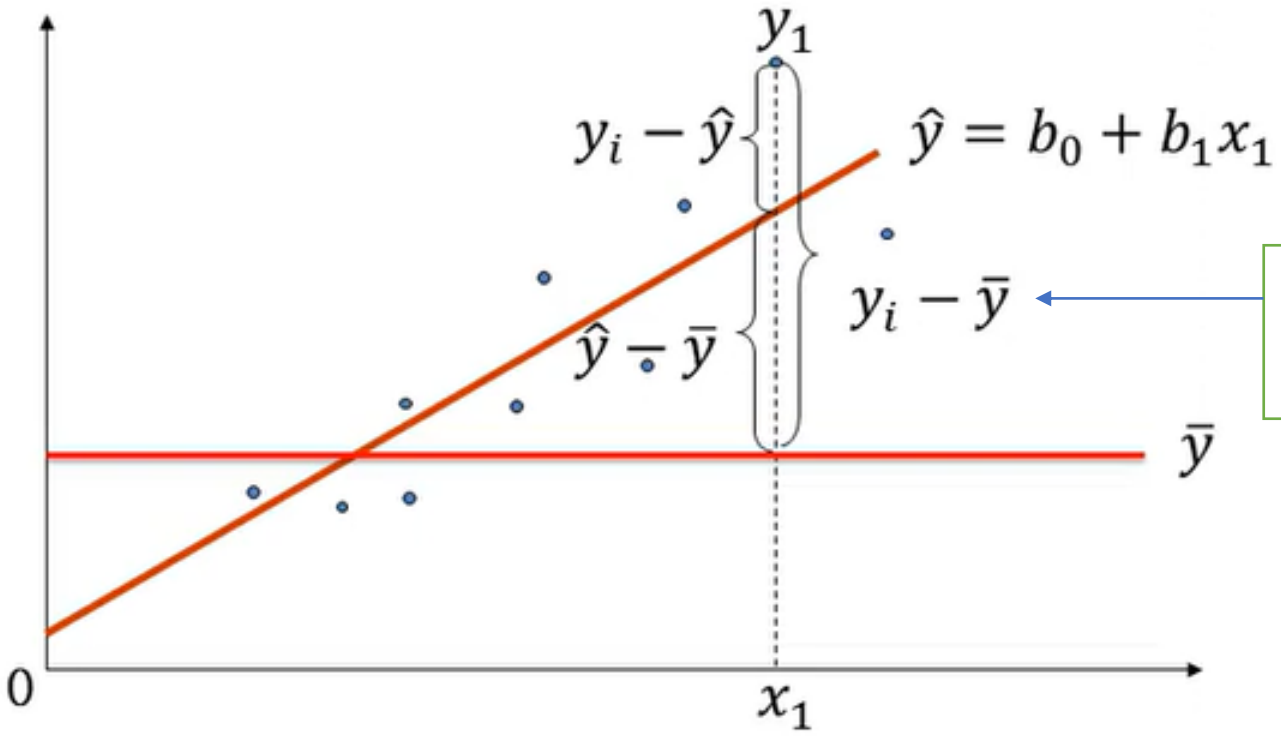
- 회귀선에 의해 종속변수가 설명되어지는 정도를 나타낸 것
- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- $0 \leq R^2 \leq 1$

(0에 가까우면 데이터를 잘 설명하지 못하는 회귀직선,

1에 가까우면 데이터를 잘 설명하는 회귀직선)

3. 회귀 평가 지표

합의 제곱 분해



평균에서 관측된 y값의 편차

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y_i - \hat{y})^2$$

SST = SSR + SSE

4. 규제 선형 모델

규제 필요

Dim이 커지면
(Feature가
많아지면)

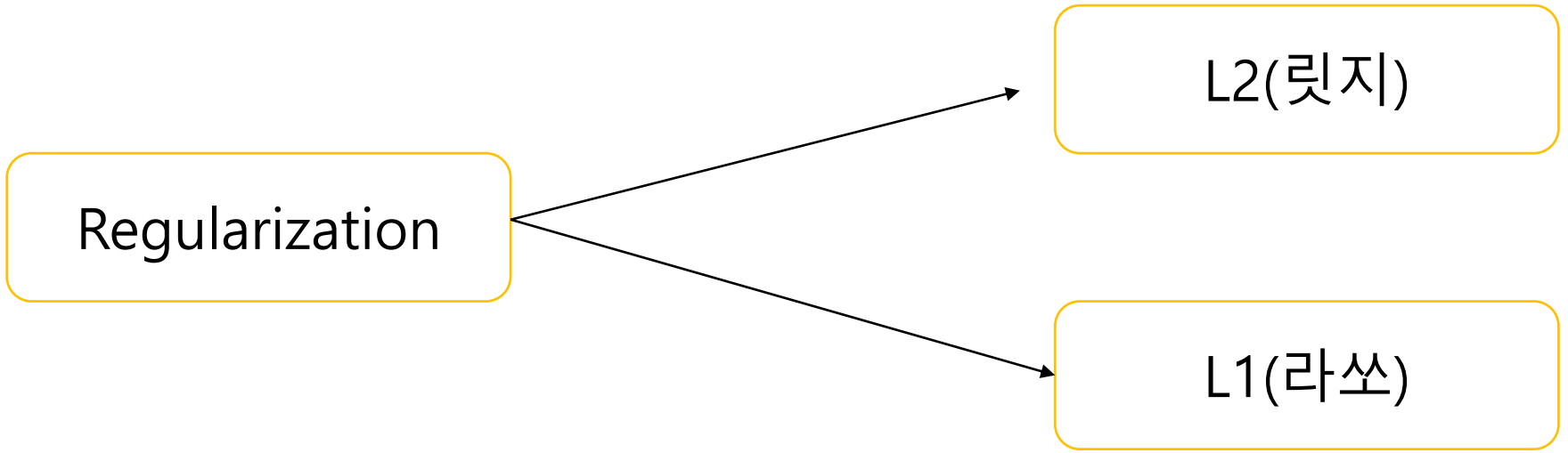
RSS는 작아지고
회귀계수에 영향

🚨 **과적합 문제, 테스트데이터에서 예측 성능 저하**

4. 규제 선형 모델

릿지 & 라쏘 회귀

** 규제는 선형 회귀의 과적합 문제를 해결하기 위해 회귀 계수에 제한을 주는 방식



1. 릿지(Ridge) -> 상대적으로 큰 회귀계수 값을 작게 만드는 규제 모델

2. 라쏘(LASSO) -> 예측 영향력이 작은 회귀계수를 0으로 만드는 규제 모델

감사합니다!