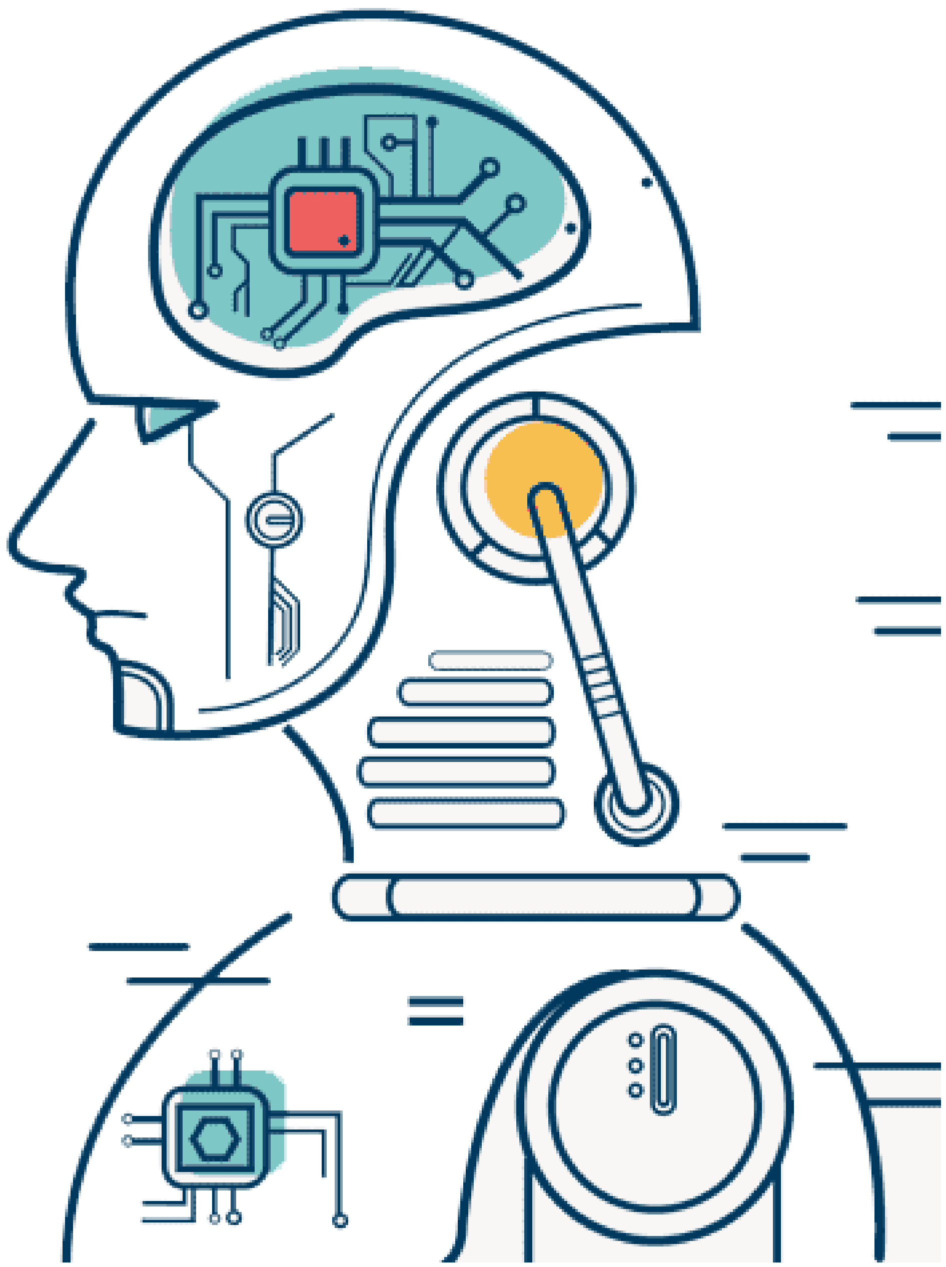
스마트인재개발원
Smart Human Resources Development

Machine Learning

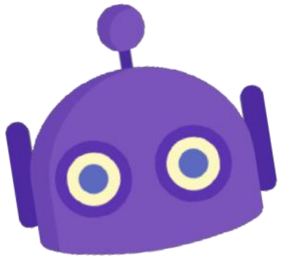
Chapter 5 지도 학습
(Linear Regression, Ridge, Lasso,
회귀평가지표, Scaling)



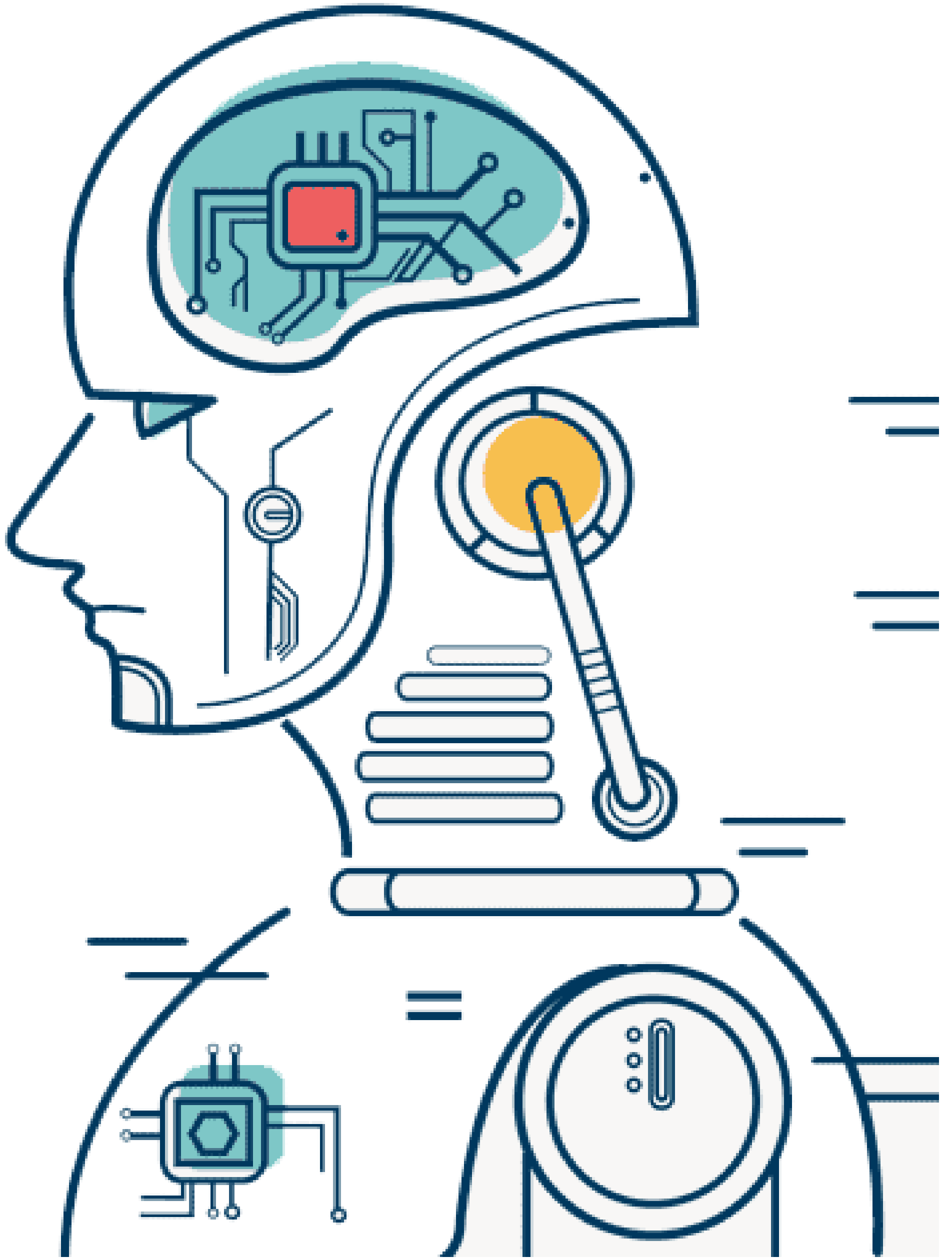
학습목표

- 선형회귀 모델을 이해하고 사용 할 수 있다.
- 회귀 모델의 평가방법을 알 수 있다.
- 데이터 스케일링의 필요성을 이해 할 수 있다.
- 다양한 스케일링 방법을 알 수 있다.

스마트인재개발원
Smart Human Resources Development



Linear Model
(Regression)



Linear Model (선형 모델)

- 입력 특성에 대한 선형 함수를 만들어 예측을 수행
- 다양한 선형 모델이 존재한다
- 분류와 회귀에 모두 사용 가능

Linear Model (선형 모델)

x(hour)	y(score)
9	90
8	80
4	40
2	20

시험성적 데이터

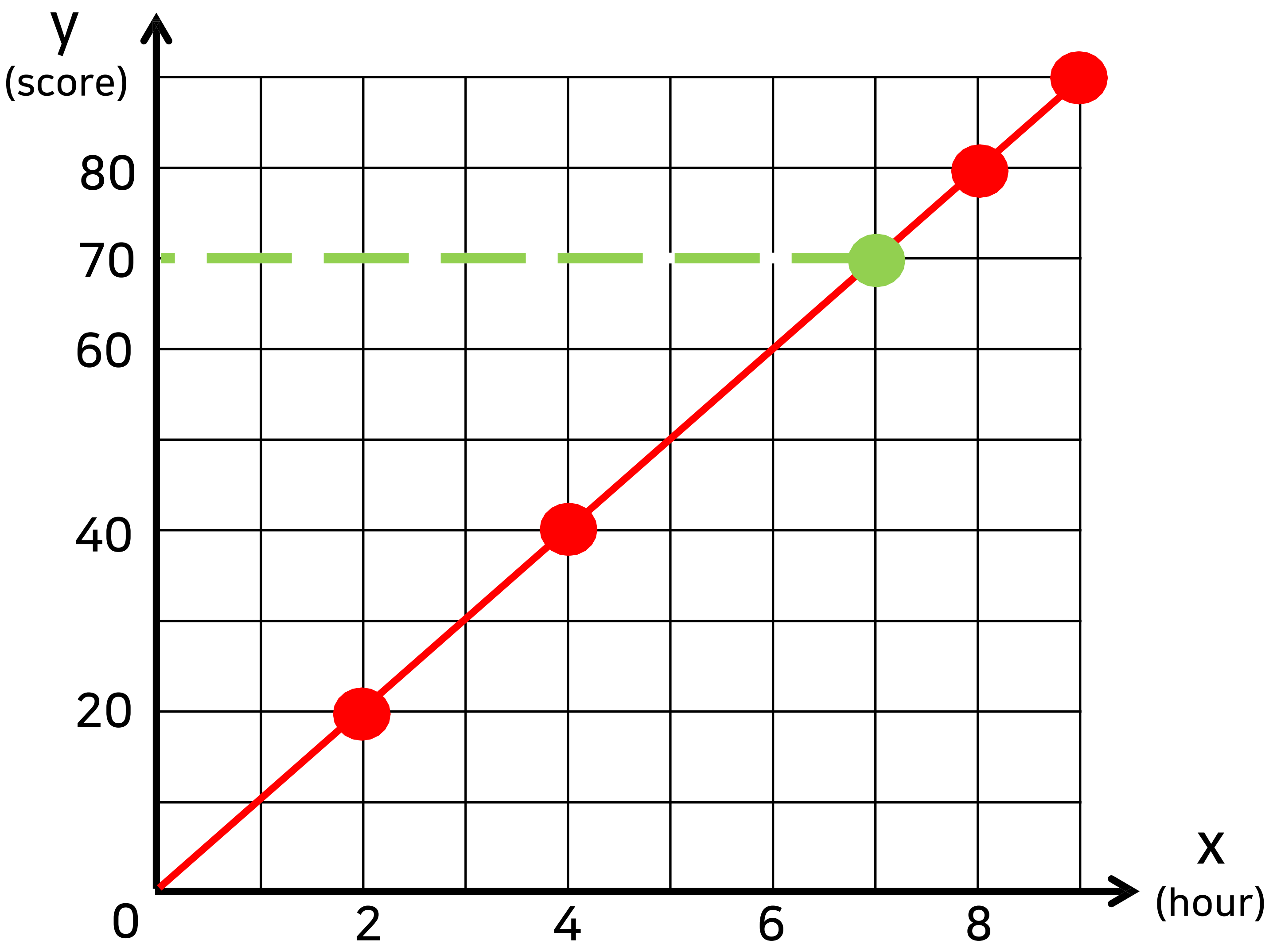
7시간 공부 할 경우
성적은 몇 점 일까?

Linear Model - Regression

x(hour)	y(score)
9	90
8	80
4	40
2	20

$$y = ax + b$$

$$y = 10x + 0$$



Linear Model - Regression

기울기 (가중치)

절편 (편향)

$$y = ax + b$$

종속(응답) 변수

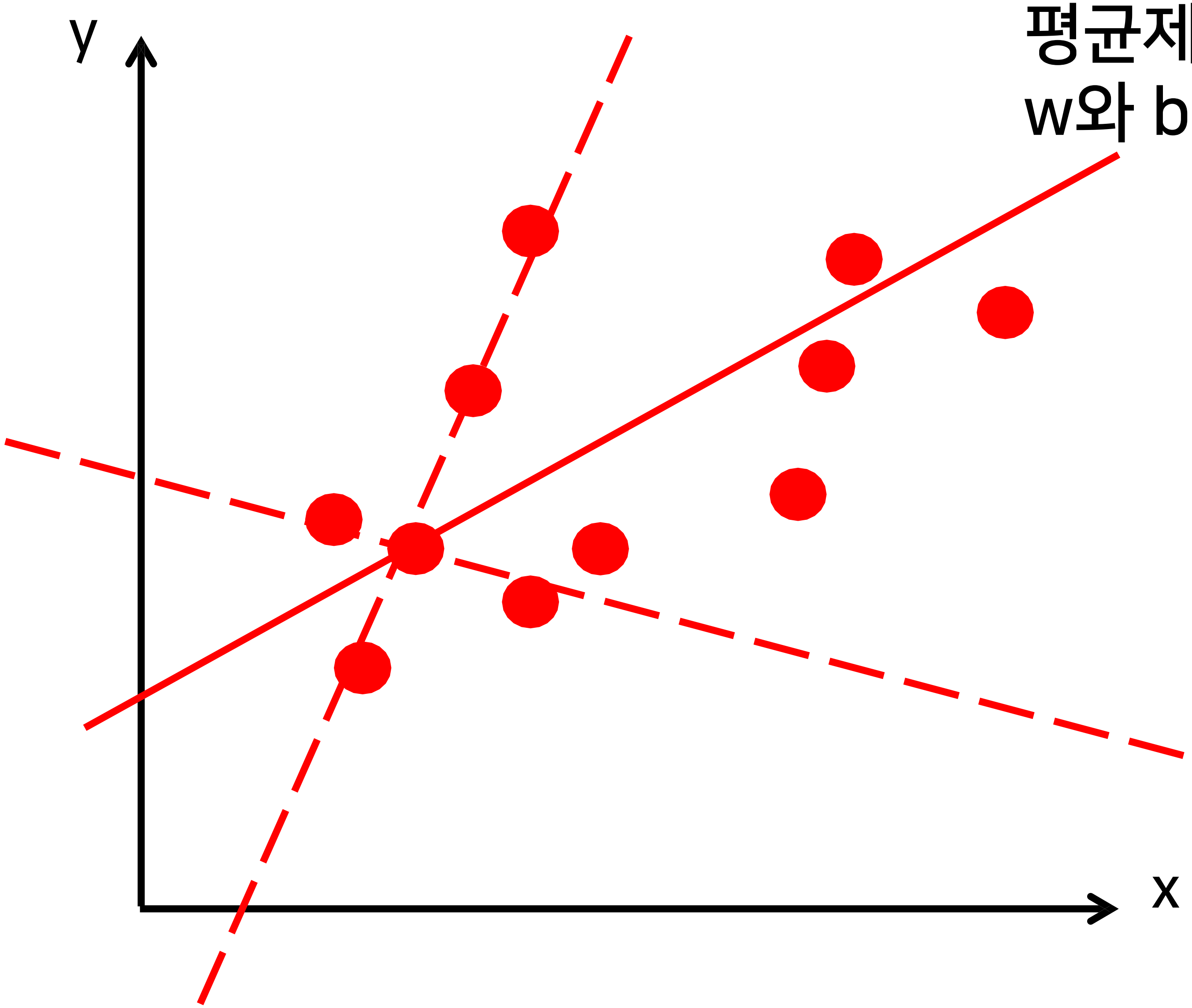
독립(입력) 변수

선형 회귀 함수

$$y = w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_p x_p + b$$

- w : 가중치(weight), 계수(coefficient)
- b : 편향(bias), 절편(intercept)
-
- 모델 w 파라미터 : model.coef_
- 모델 b 파라미터 : model.intercept_

Linear Model – Regression (MSE)



평균제곱오차가 최소인
w와 b를 찾는다

Linear Model – Regression (MSE)

스마트인재개발원
Smart Human Resources Development

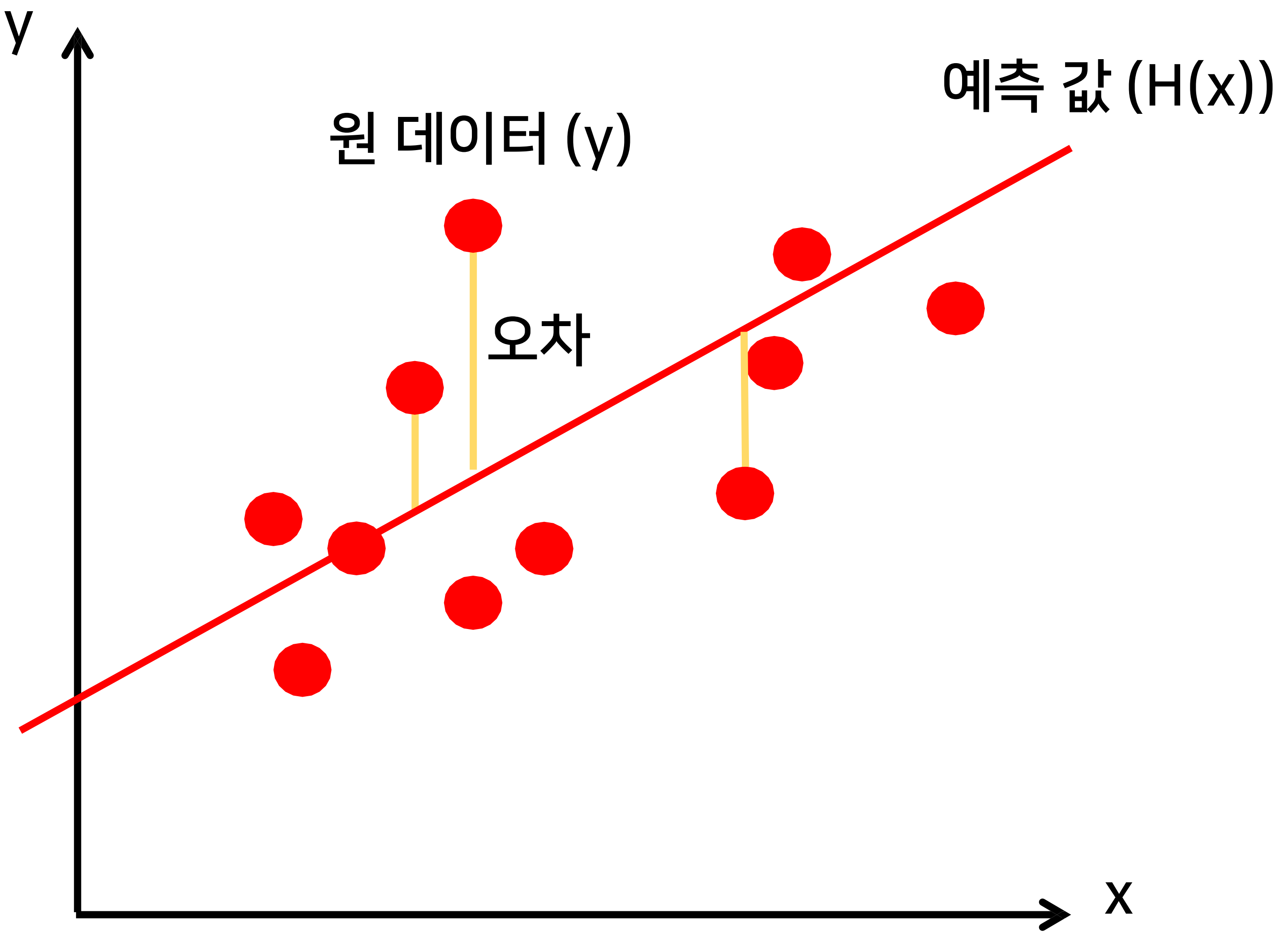
Cost function

비용함수 : 수식을 검증

H(x) = w * x + b

오차 = H(x) - y

절대값 또는
제곱 값 사용



Linear Model – Regression (MSE)

스마트인재개발원
Smart Human Resources Development

평균제곱오차 (MSE : Mean Squared Error)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

평균제곱근오차 (RMSE : Root Mean Squared Error)

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Linear Model – Regression (MSE)

평균제곱오차(MSE)가 최소가 되는 w와 b를 찾는 방법

1. 수학 공식을 이용한 해석적 방법 (Ordinary Least Squares)
2. 경사하강법 (Gradient Descent Algorithm)

수학 공식을 이용한 해석적 방법 (Ordinary Least Squares)

$$a\sum x^2 + b\sum x = \sum xy$$
$$a\sum x + bn = \sum y$$

$$a = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - \Sigma X \Sigma X}$$

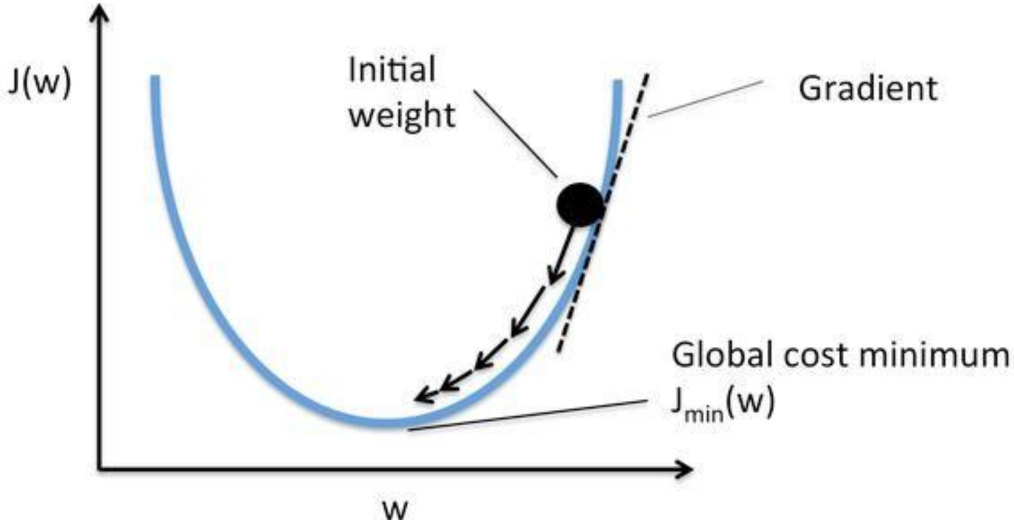
$$b = \frac{\Sigma X^2 \Sigma Y - \Sigma X \Sigma XY}{n\Sigma X^2 - \Sigma X \Sigma X}$$

x(hour)	y(score)
1	1
2	2
3	3

LinearRegression 클래스로 구현되어 있다.

경사하강법 (Gradient Descent Algorithm)

경사하강법 (Gradient Descent Algorithm)



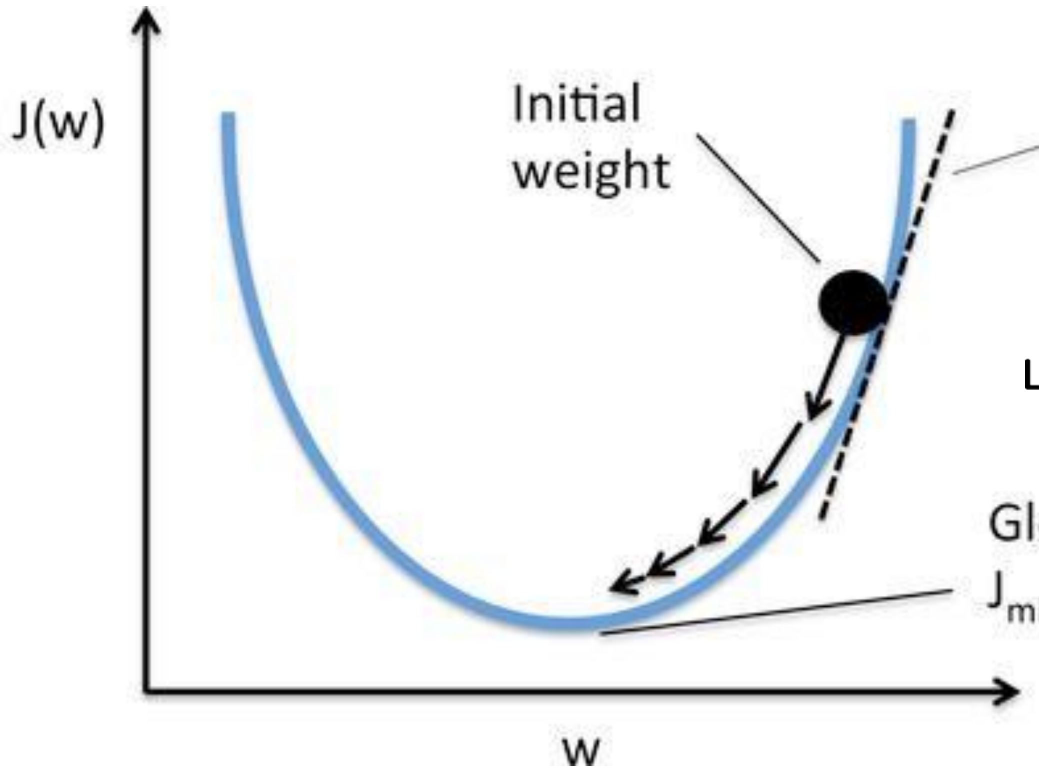
비용함수의 기울기(경사)를 구하여 기울기가 낮은 쪽으로
계속 이동하여 값을 최적화 시키는 방법

경사하강법 (Gradient Descent Algorithm)

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

Gradient

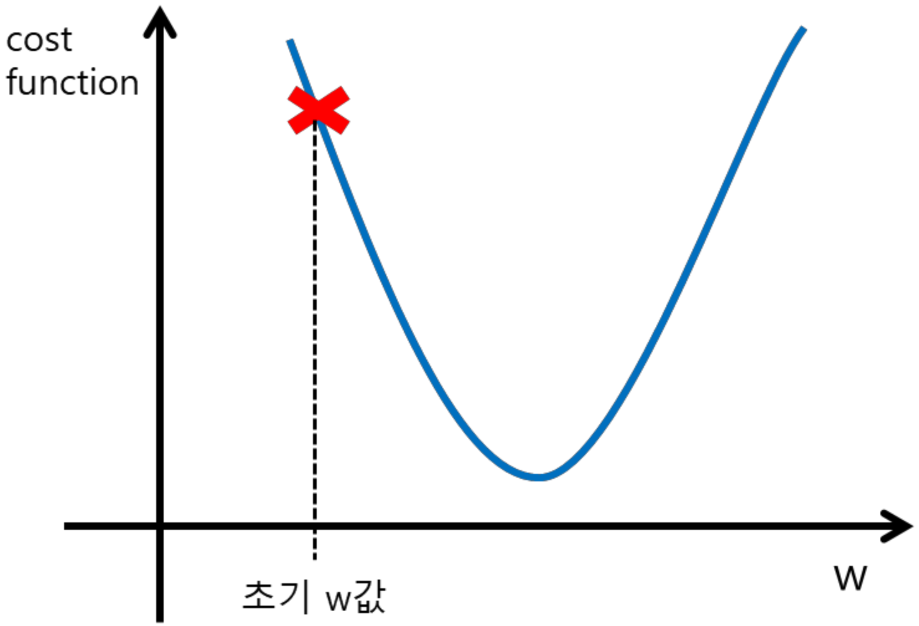
Learning rate



경사하강법 (Gradient Descent Algorithm)

(1) 우선 임의로 w값을 하나 선정

- 운이 아주 좋으면 최적의 값이겠지만 그렇지 않을 확률이 훨씬 더 큼
- 대부분 최적의 w값과는 거리가 먼 것이 설정



경사하강법 (Gradient Descent Algorithm)

(1) 최적의 w값을 찾아가기 위해서 시작점에서 손실 곡선의 기울기를 계산
비용함수를 w에 대해서 편미분
(2) 파라미터를 곱한 것을 초기 설정된 w값에서 빼 줌

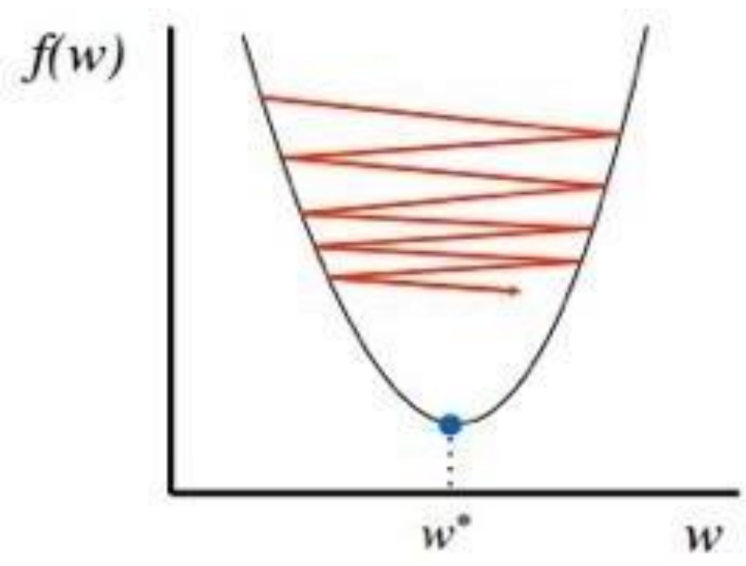
- 학습률(learning rate) : 기울기의 보폭
- 학습률이 너무 작으면 최적의 w를 찾는데 오래 걸리고 크면 건너뛰어 버리고 발산할 수 있음

$$w' = w - a\frac{\partial e}{\partial w} \qquad b' = b - a\frac{\partial e}{\partial b}$$

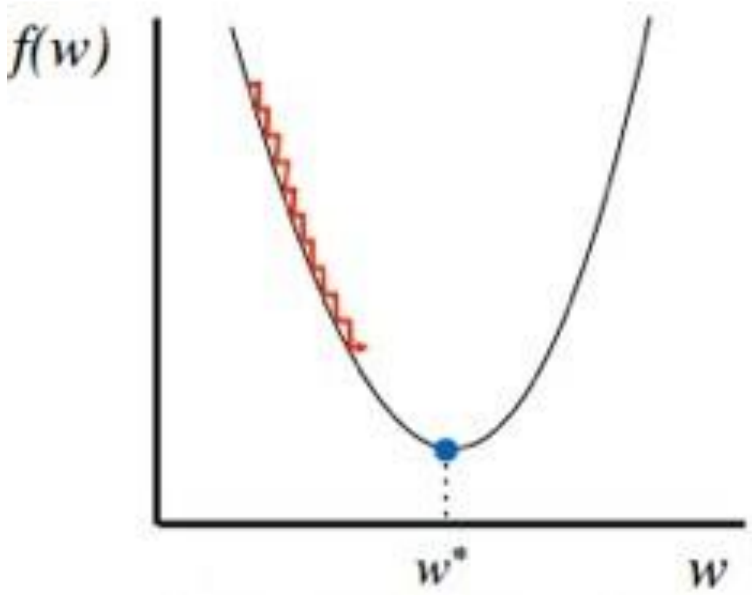
Linear Model – Regression(Gradient descent algorithm)

경사하강법 (Gradient Descent Algorithm)

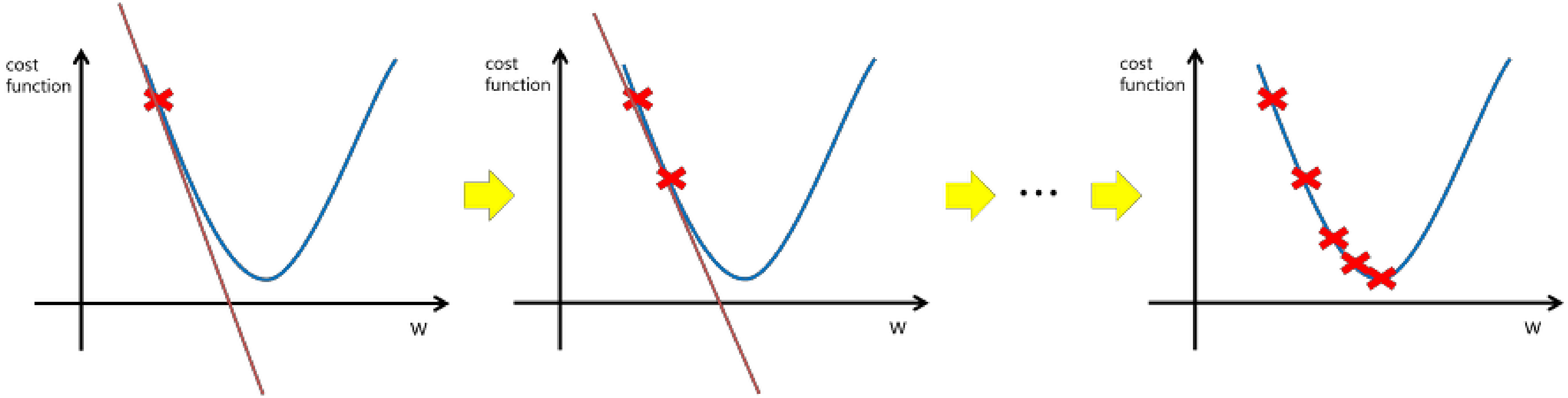
Learning rate가 큰 경우



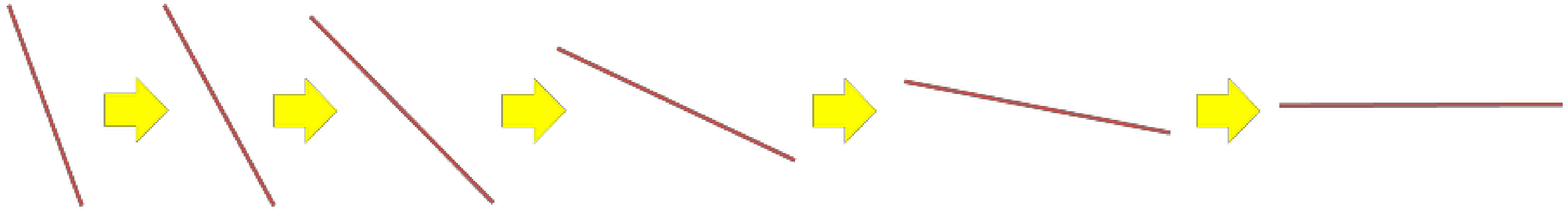
Learning rate가 작은 경우



경사하강법 (Gradient Descent Algorithm)



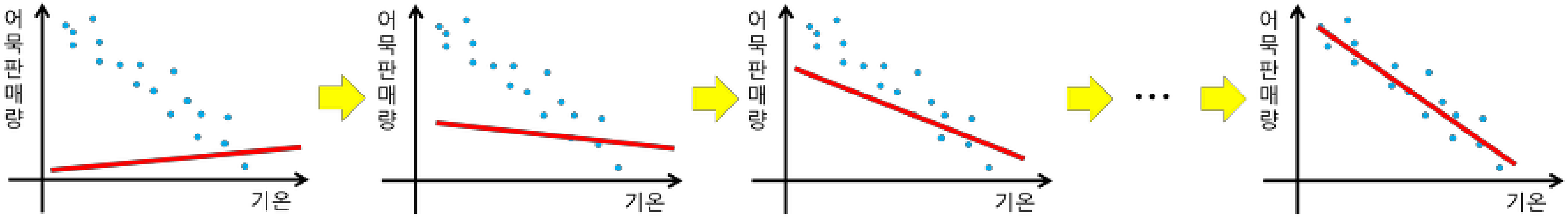
W값 갱신 과정



경사가 점차 감소되는 현상을 이용하므로 경사감소법!

Linear Model – Regression(Gradient descent algorithm)

경사하강법 (Gradient Descent Algorithm)



초기 w, b값은 데이터를 잘 반영하는 일차함수식을 만들지 못했지만,

점차적으로 데이터를 잘 반영해내는 값들로 갱신

LinearRegression 사용하기

주요 매개변수(Hyperparameter)

scikit-learn의 경우

```
SGDRegressor(max_iter, eta0)
```

- 가중치 업데이트 횟수 : max_iter
- 학습률 : eta0

경사하강법으로 학습하는
SGDRegressor 사용하기

Linear Model – Regression

Linear Model 장점

- 결과예측(추론) 속도가 빠르다.
- 대용량 데이터에도 충분히 활용 가능하다.
- 특성이 많은 데이터 세트라면 훌륭한 성능을 낼 수 있다.

Linear Model 단점

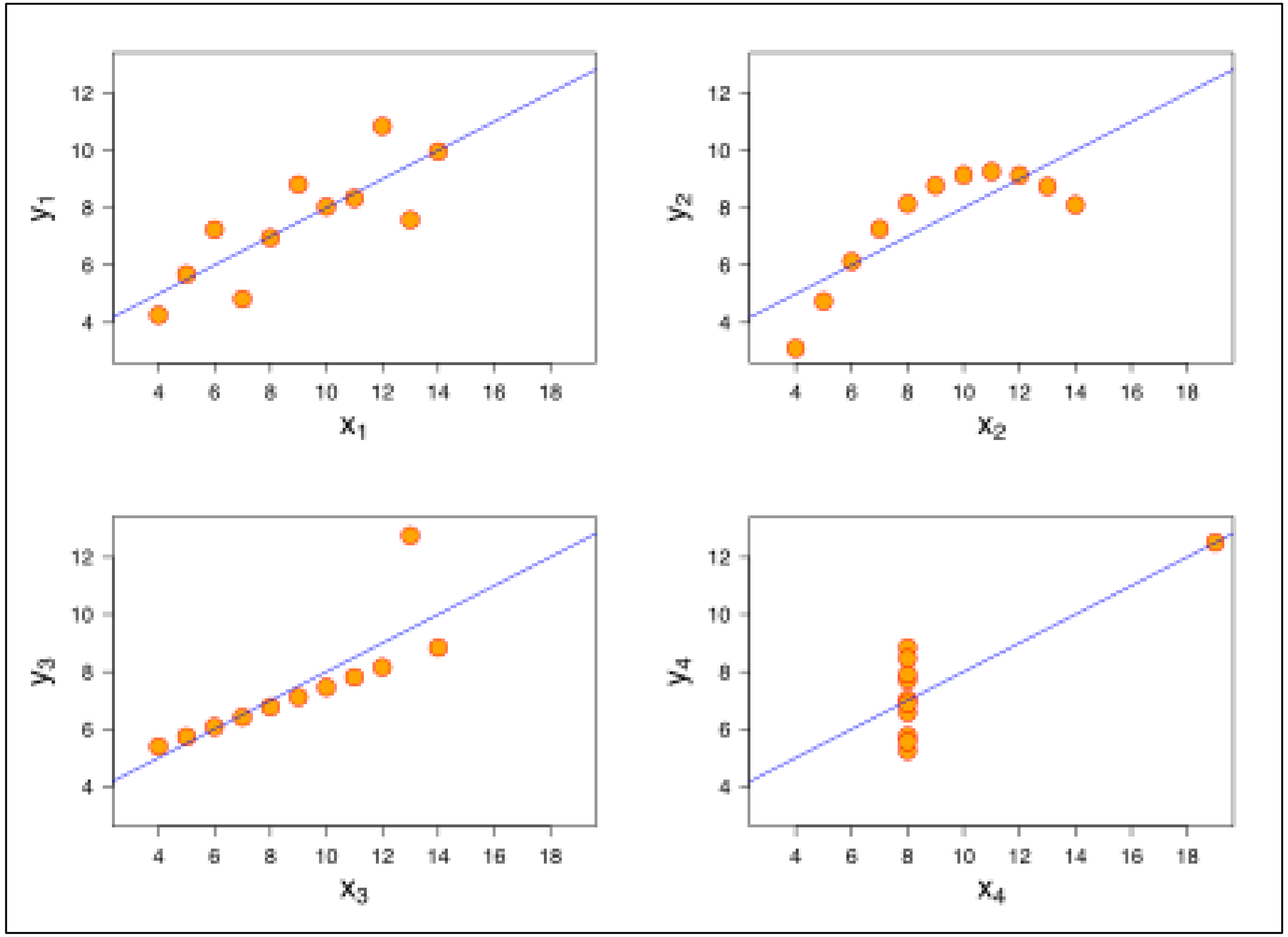
- 특성이 적은 저차원 데이터에서는 다른 모델의 일반화 성능이 더 좋을 수 있다. ➡ 특성확장을 하기도 한다.
- LinearRegression Model은 복잡도를 제어할 방법이 없어 과대적합 되기 쉽다.

⬇

모델 정규화(Regularization)을 통해 과대적합을 제어한다.

Linear Model – Regression

Linear Model 단점



Linear 모델을 이용해 보스턴 집값
데이터를 이용하여
주택 가격을 예측 해보자.

Linear Model – Regression (MSE)

RMSE/MSE의 단점

- 예측 대상의 크기에 영향을 받음

삼성전자와 NAVER의 주가예측 RMSE가 5000이 나왔다면
두 모델의 성능은 동일한가요 ?

R2 Score

- MSE,RMSE의 문제점
  - 예를 들어 키를 예측하였는데 MSE 값이 5.7cm이 나왔다고 하면 이것의 성능이 얼마나 우수한지 다른 경우와 비교하기 어렵다.
  - 몸무게를 예측하였는데 MSE값이 3.8kg이라면 얼마나 우수한 것인가?

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

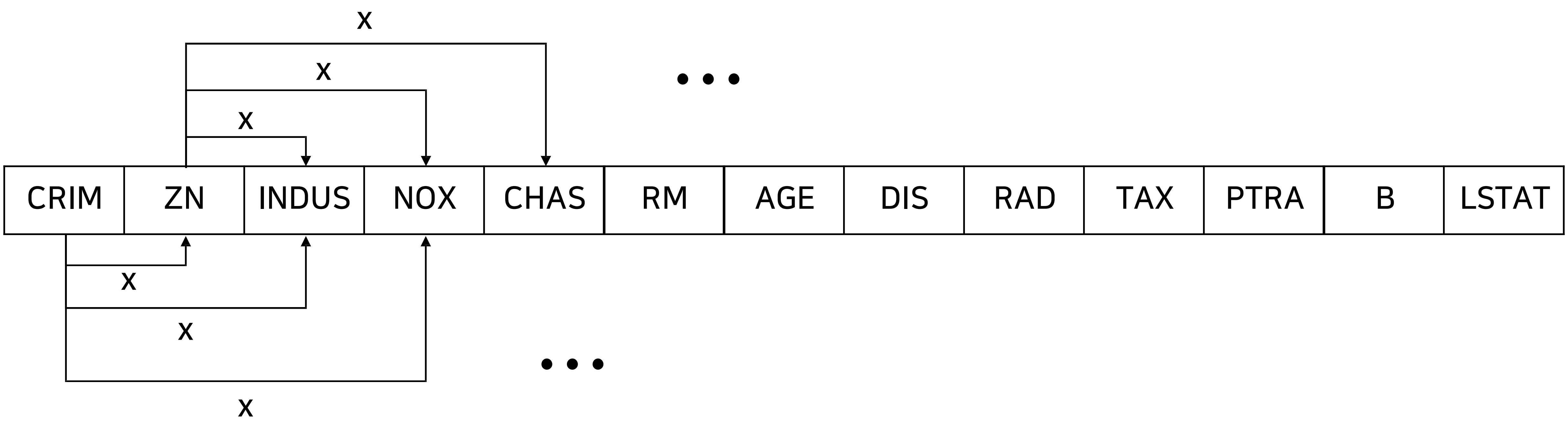
$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$ 표준편차의 합

예측값 f_i가 평균값과 같아지면 스코어는 0이 되어 좋지 않은 모델
예측값 f_i가 원래값과 같아지면 스코어는 1이 되어 우수한 모델

Linear Model – 보스턴 주택 값 예측 실습

스마트인재개발원
Smart Human Resources Development

확장 보스톤 집값 데이터셋



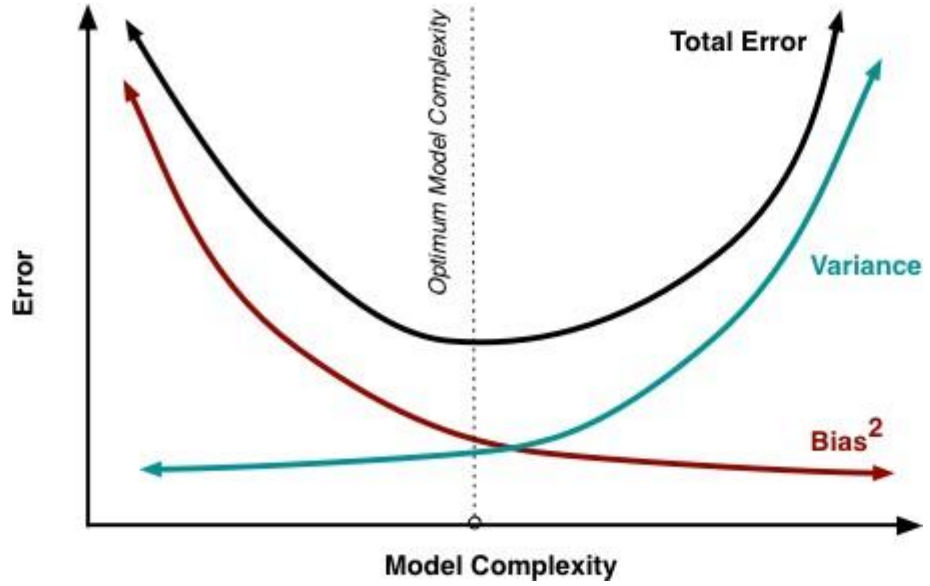
확장 보스턴집값을 적용하여
학습해보자

Linear Model – Regularization

스마트인재개발원
Smart Human Resources Development

모델 정규화

- Overfitting 문제 해결
  - 데이터의 복잡도 줄이기
  - 정규화를 통한 분산 감소



Linear Model – Regularization

모델 정규화

- w가 크다면 입력 x가 조금만 달아져도 y가 크게 변함
- **w에 규제를 주어 영향을 줄이도록 하는 것**.

- L1 규제 : Lasso

  w의 모든 원소에 똑같은 힘으로 규제를 적용하는 방법. 특정 계수들은 0이 됨.
  특성선택(Feature Selection)이 자동으로 이루어진다.

- L2 규제 : Ridge

  w의 모든 원소에 골고루 규제를 적용하여 0에 가깝게 만든다.

Linear Model – Regularization

스마트인재개발원
Smart Human Resources Development

정규화 : cost 함수

규제의 강도

L1 규제 : Lasso

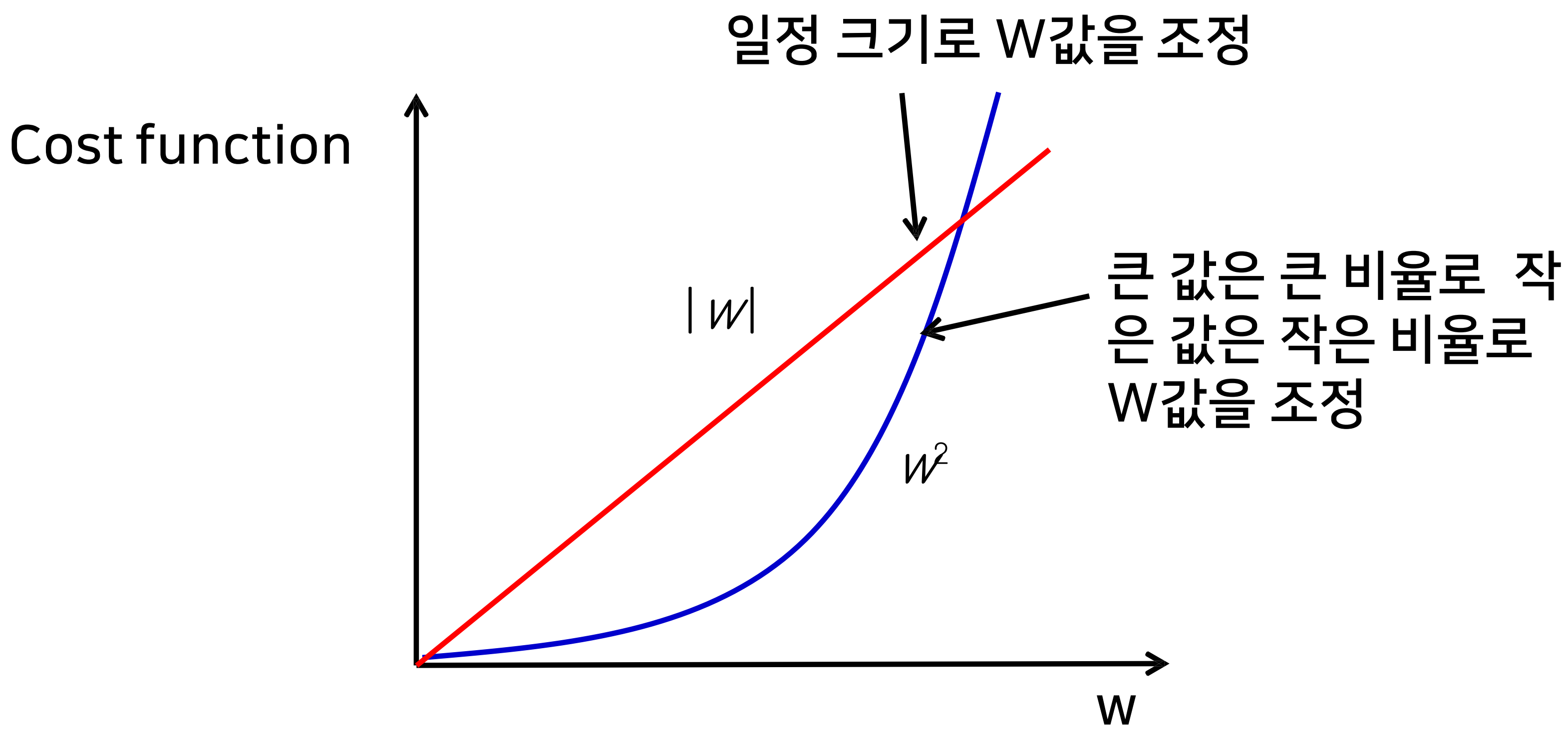
$$J(w)_{LASSO} = \sum_{i=1}^{n}(y^{(i)} - y^{\wedge(i)})^2 + \lambda \underbrace{\sum_{j=1}^{m}|w_j|}_{\text{L1 규제}}$$

L2 규제 : Ridge

$$J(w)_{Ridge} = \sum_{i=1}^{n}(y^{(i)} - y^{\wedge(i)})^2 + \lambda \underbrace{\sum_{j=1}^{m}{w_j}^2}_{\text{L2 규제}}$$

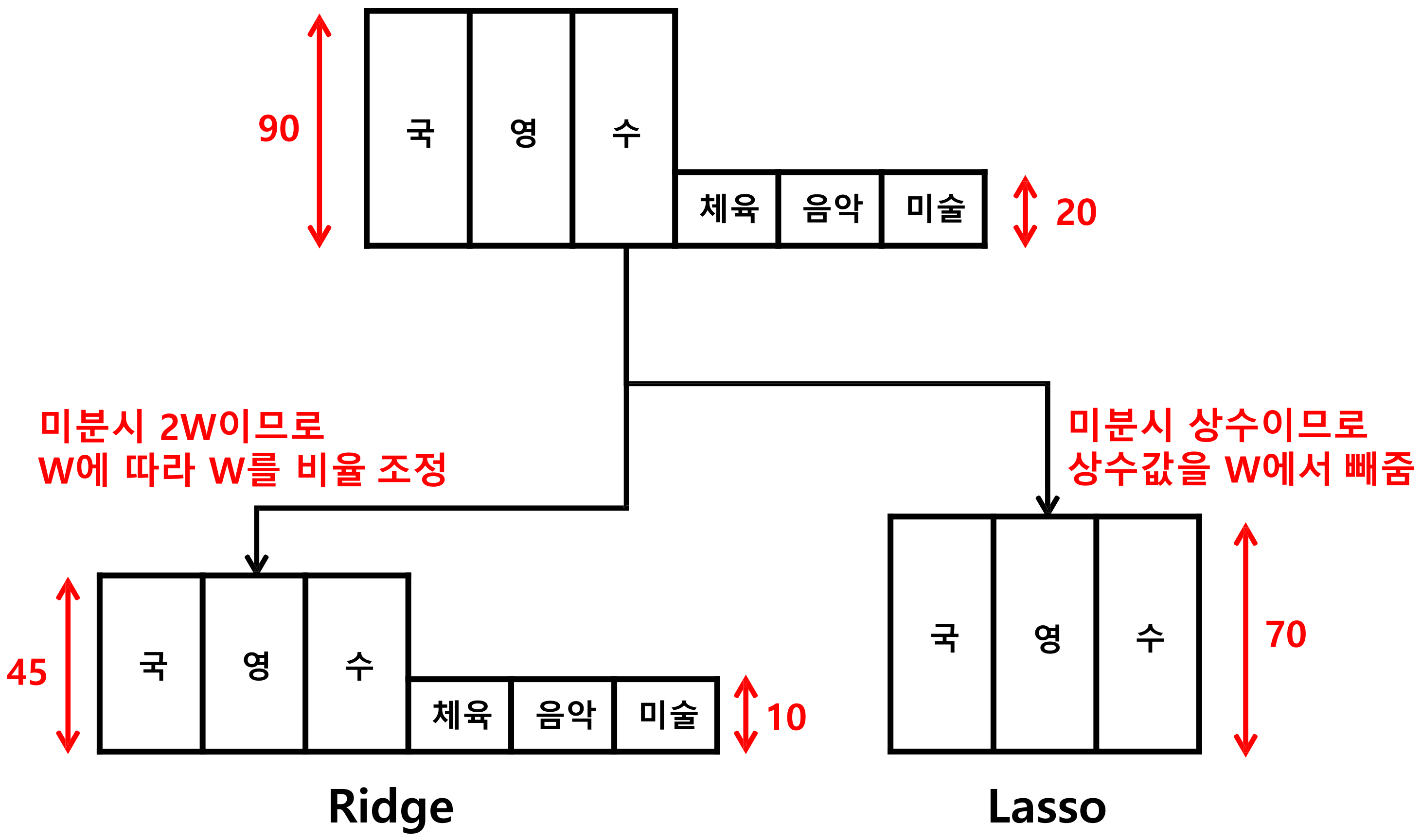
Linear Model – Regularization

정규화 : cost 함수



Linear Model – Regularization

정규화 : cost 함수



Linear Model – Regularization

스마트인재개발원
Smart Human Resources Development

구분	릿지회귀	라쏘회귀
제약식	L_2 norm	L_1 norm
변수선택	불가능	가능
solution	closed form	명시해 없음
장점	변수간 상관관계가 높아도 좋은 성능	변수간 상관관계가 높으면 성능↓
특징	크기가 큰 변수를 우선적으로 줄임	비중요 변수를 우선적으로 줄임

주요 매개변수(Hyperparameter)

scikit-learn의 경우

Ridge(alpha)

Lasso(alpha)

- 규제의 강도 : alpha

Linear Model – Regularization

스마트인재개발원
Smart Human Resources Development

총 104개의 특성을 라쏘 회귀 모델을 만들기 위해 사용

↓

α=1로 설정했더니 104개의 가중치 중에서 50개가 0이 되면서 특성은 단 54개만 사용

↓

훈련셋에서의 점수와 테스트셋에서의 점수를 보니 과소적합

↓

복잡도를 높이기 위해서 α=0.0001로 설정했더니 가중치 중에서 0개가 0이 되면서 104개의 특성이 사용

↓

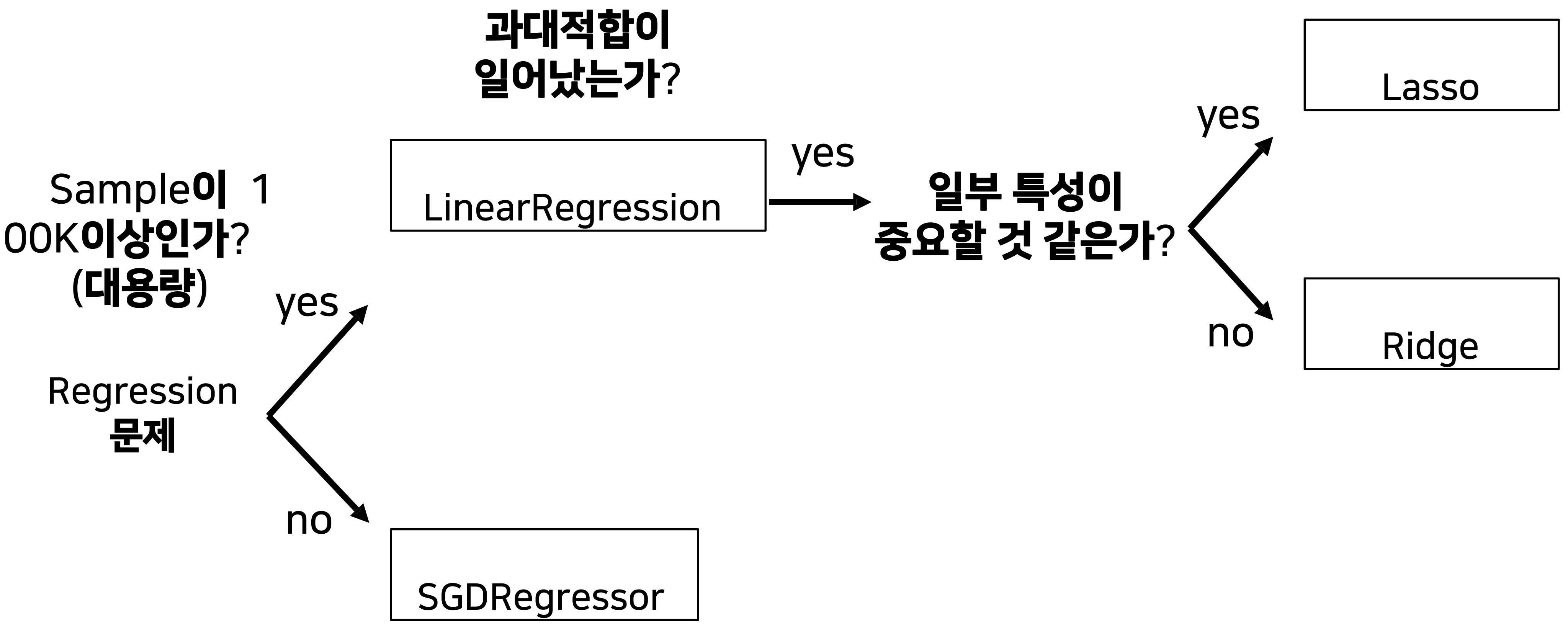
훈련셋과 테스트셋에서의 점수를 보니 훈련셋과 테스트셋이 많이 좋아짐

↓

가장 좋은 모델인지 확인하기 위해 다시 복잡도를 α=0.1로 설정했더니 105개의 가중치 중에서 25개가 0이 되면서 79개의 특성이 사용 → 훈련점수와 테스트점수가 조금씩 떨어짐

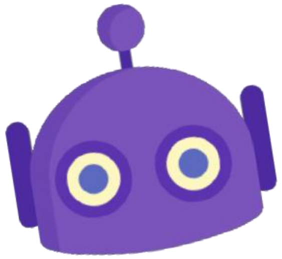
Ridge와 Lasso 모델을 이용하여
확장 보스턴 집값의 과대적합 문제를
해결해보자.

Linear Model – Regression

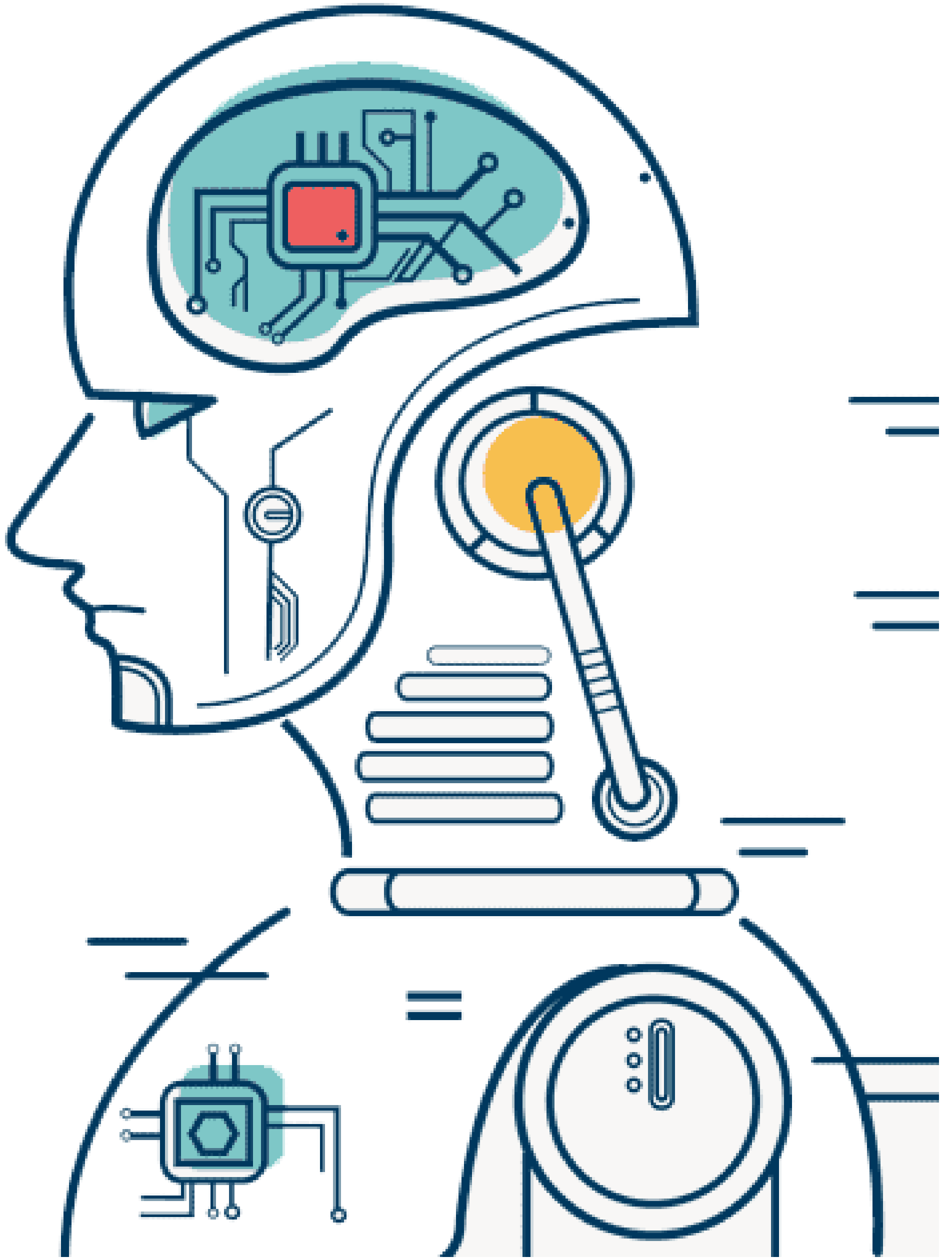


이전 실습의 Lasso **모델의 최적의** alpha **값을 구해보자. 파라미터 튜닝 과정을 그래프로 그려보자**

스마트인재개발원
Smart Human Resources Development



데이터 스케일링
(Data scaling)



데이터 스케일링(Data scaling) 정의

데이터 스케일링 (Data scaling)

- 특성(Feature)들의 범위(range)를 정규화 해주는 작업
- 특성마다 다른 범위를 가지는 경우 머신러닝 모델들이 제대로 학습되지 않을 가능성이 있다.
  (KNN, SVM, Neural network 모델, Clustering 모델 등)

시력	키
0.2	178
1.0	156
0.5	168
0.3	188
0.6	149

시력과 키를 함께 학습시킬 경우
키의 범위가 크기때문에 거리값을
기반으로 학습할 때 영향을 많이 준다.

데이터 스케일링(Data scaling) 종류

장점

- 특성들을 비교 분석하기 쉽게 만들어 준다.
- Linear Model, Neural network Model 등에서 학습의 안정성과 속도를 개선시킨다.

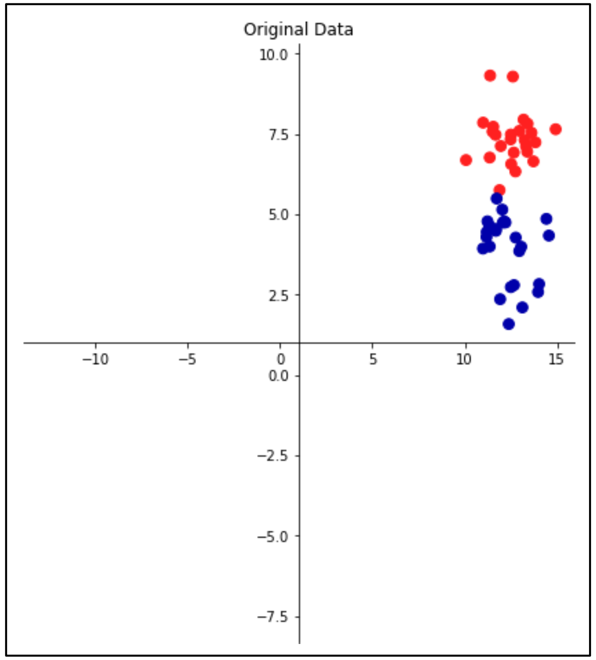
단점

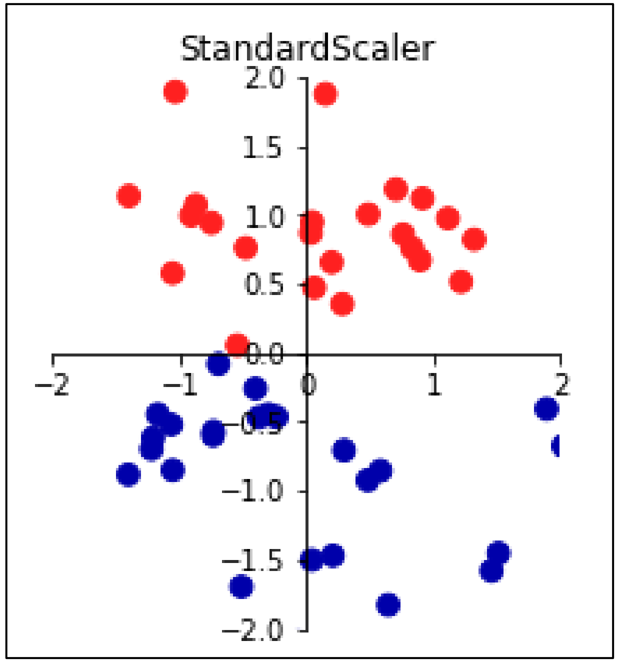
- 하지만 특성에 따라 원래 범위를 유지하는게 좋을 경우는 scaling을 하지 않아도 된다.

데이터 스케일링(Data scaling) 종류

StandardScaler

- 변수의 평균, 분산을 이용해 정규분포 형태로 변환 (평균 0, 분산 1)
- 데이터가 정규분포인 경우에 사용

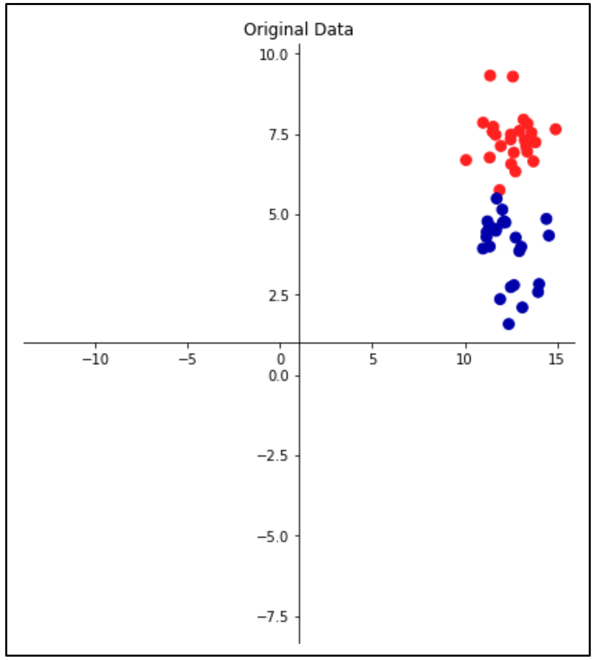


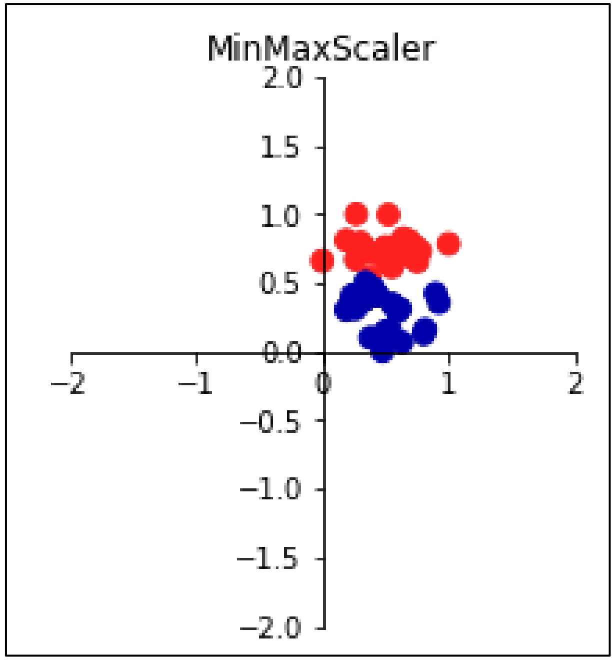


데이터 스케일링(Data scaling) 종류

MinMaxScaler

- 변수의 최대값 1로, 최소값을 0으로 하여 변환 (0 ~ 1 사이 값으로 변환)
- 데이터가 비정규분포인 경우에 사용
- 이상치(Outlier)에 크게 영향을 받는다 > 이상치가 있는 경우 사용 못함.

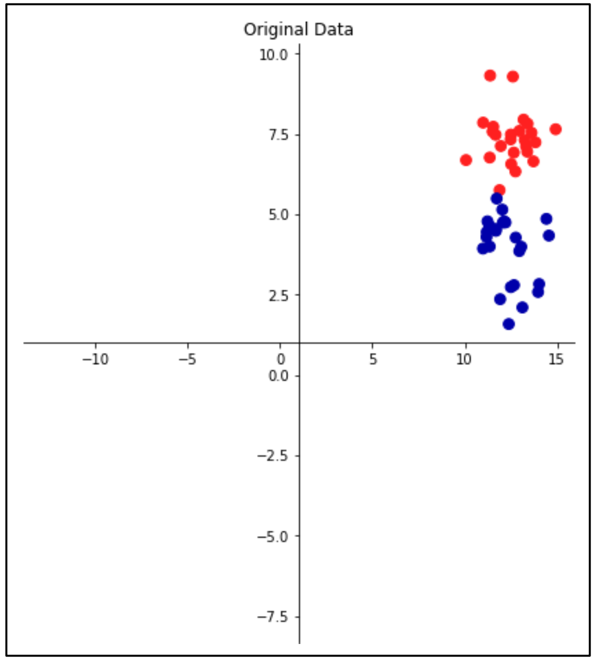


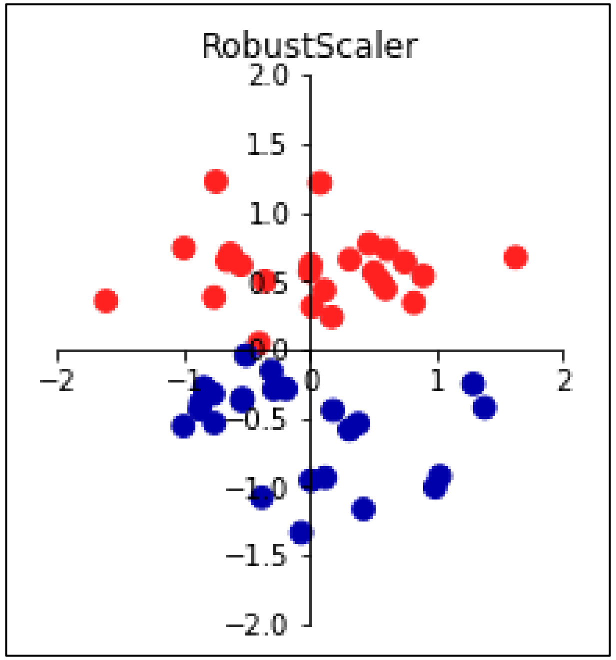


데이터 스케일링(Data scaling) 종류

RobustScaler

- 변수의 1/4 지점을 0으로 3/4 지점을 1로 하여 변환
- 이상치(Outlier)가 있는 경우 사용



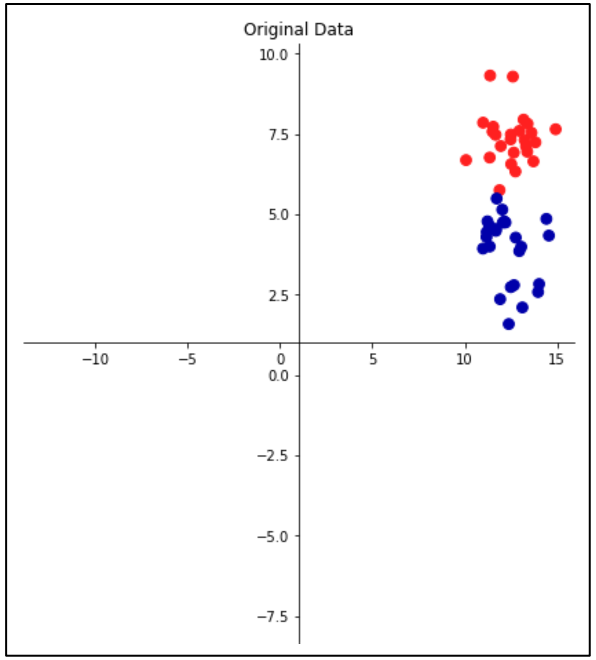


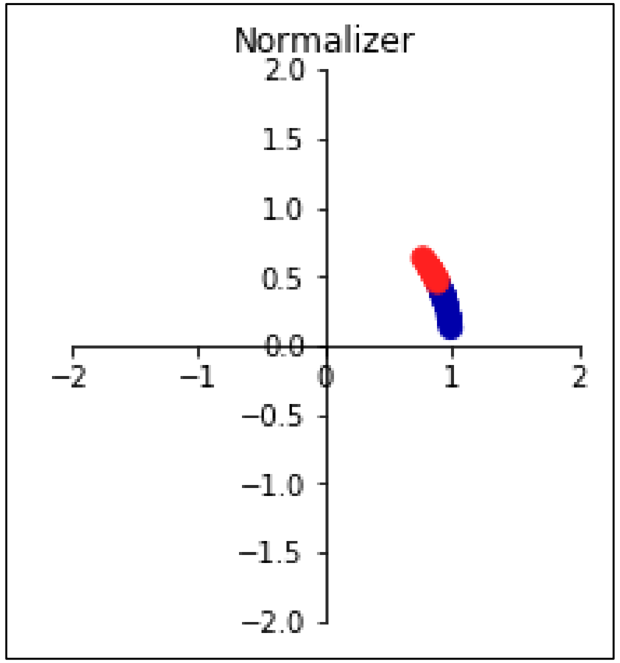
데이터 스케일링(Data scaling) 종류

스마트인재개발원
Smart Human Resources Development

Normalizer

- 특성 벡터의 유클리디안 길이가 1이 되도록 조정 (지름이 1인 원에 투영)
- 특성 벡터의 길이는 상관 없고 데이터의 방향(각도)만 중요할 때 사용.



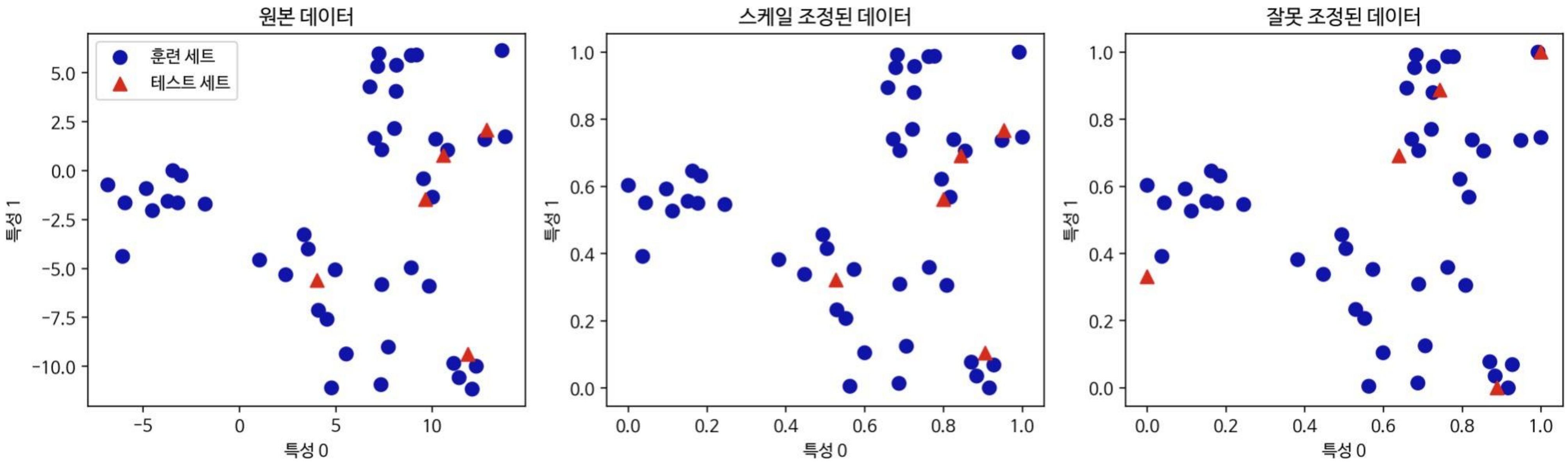


데이터 스케일링(Data scaling) 사용 시 주의점

스마트인재개발원
Smart Human Resources Development

주의점

- 훈련세트와 테스트세트에 같은 변환을 적용해야 한다.
- 예를 들어 훈련세트의 평균과 분산을 이용해 훈련세트를 변환하고, 테스트세트의 평균과 분산을 이용해 테스트세트를 각각 변환하면 잘못된 결과가 나올 수 있다.
  > 값의 범위가 다를 수 있으므로



유방암 데이터를 학습한 KNN 모델에
scaler를 적용하여 결과 확인