



🔍 파이썬 머신러닝 기초

머신러닝 - 회귀

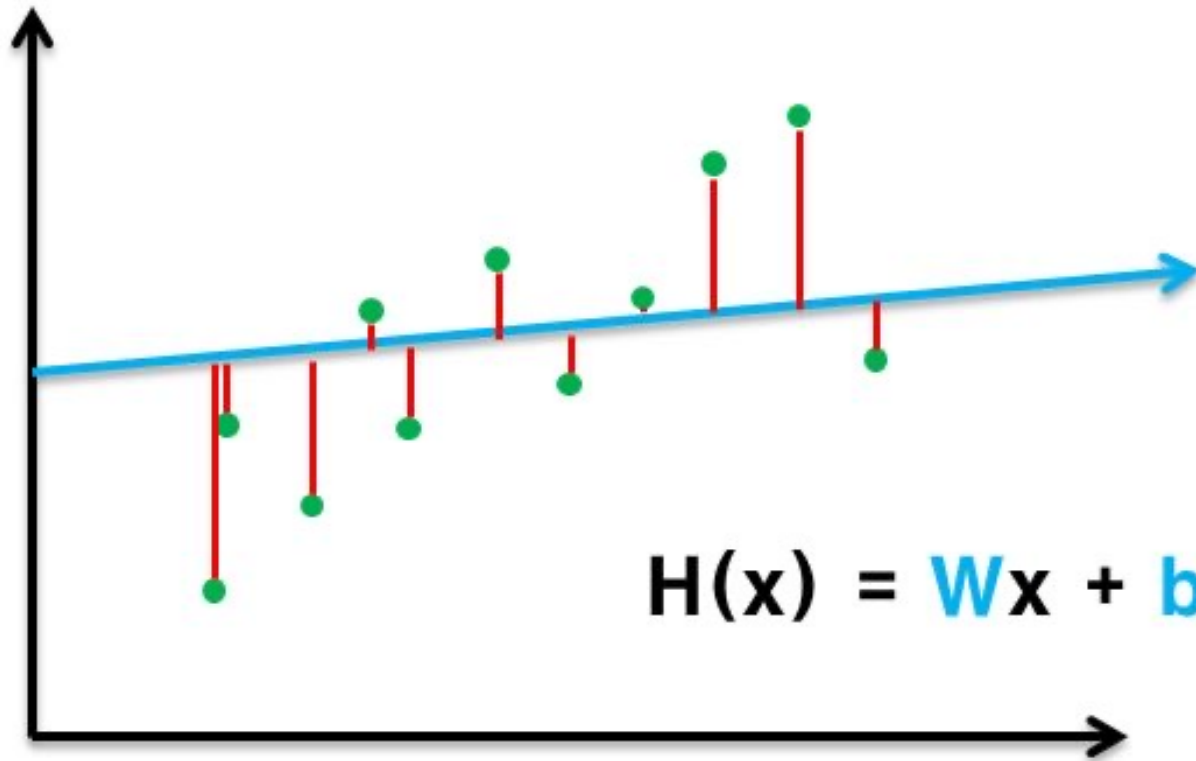
회귀 (regression)

| 데이터를 대표하는 선(직선, 곡선 등)을 만들어보자

| 여러 개의 독립변수와 한 개의 종속변수 간의 상관관계를 모델링하는 기법을 통칭

| $Y = W_1 * X_1 + W_2 * X_2 + ... W_n * X_n$ 와 같이 수식으로 표현 가능

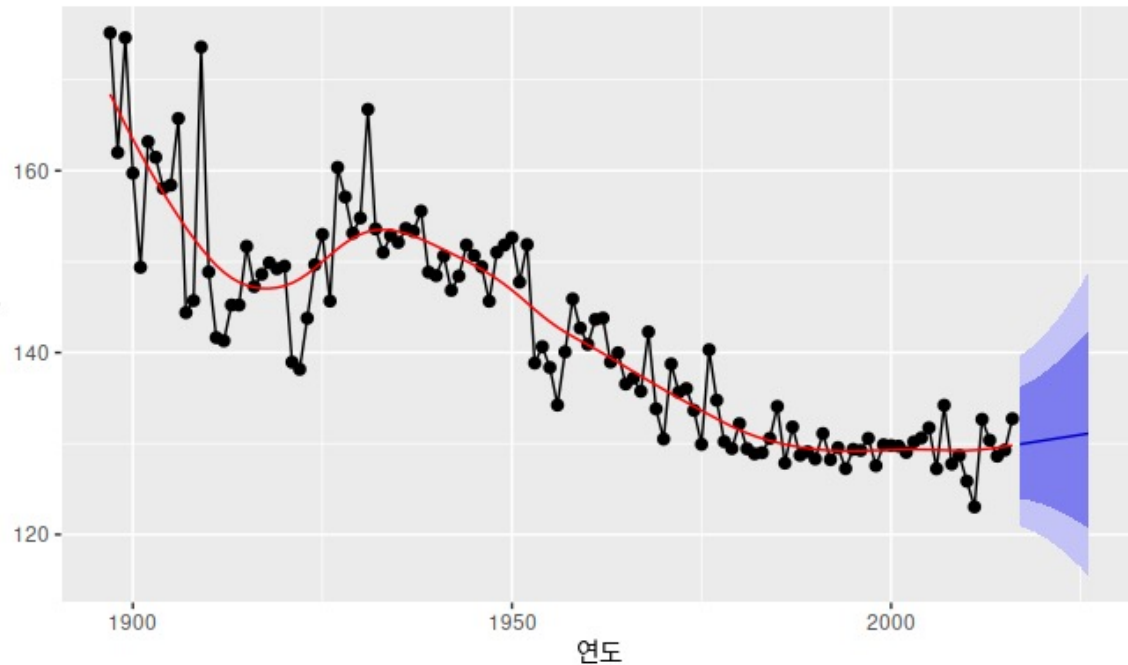
회귀 (regression)



| 회귀 계수 - coefficient

| 절편 : b = bias, intercept

회귀 (regression)



| 회귀 계수의 선형/비선형 여부

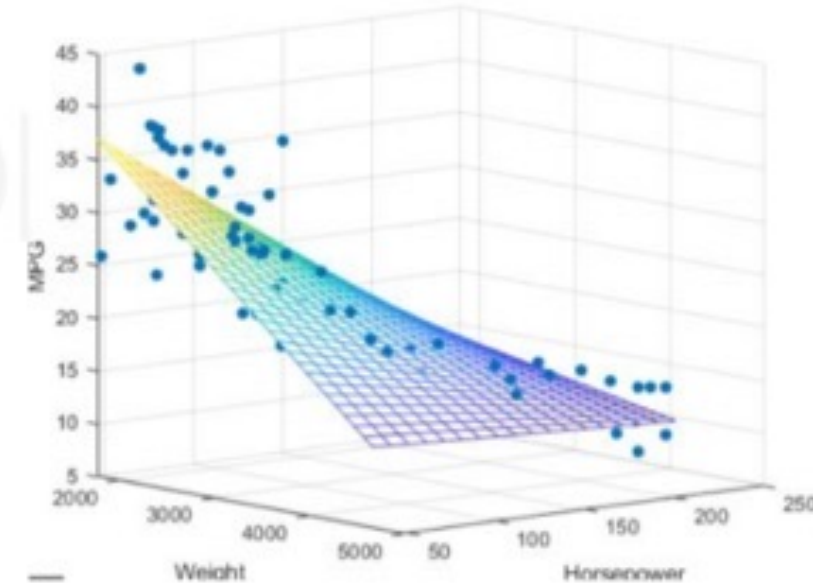
- 선형회귀
- 비선형회귀

| 독립변수의 개수

- 1개 : 단일 회귀
- 여러 개 : 다중회귀

| 종속변수의 개수

회귀 (regression)



| 회귀 계수의 선형/비선형 여부

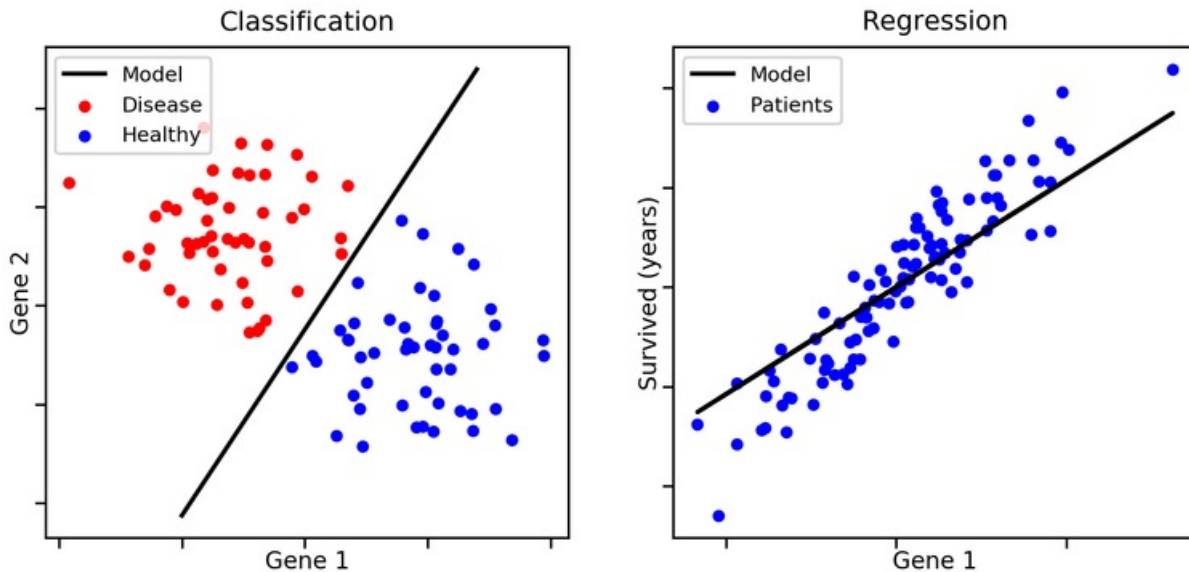
- 선형회귀
- 비선형회귀

| 독립변수의 개수

- 1개 : 단일 회귀
- 여러 개 : 다중회귀

| 종속변수의 개수

회귀 (regression)



| 지도학습은 분류, 회귀 두 가지로 나뉜다

| 분류는 예측값이 카테고리나 같은 이산형 클래스 값

| 회귀는 연속형 숫자값

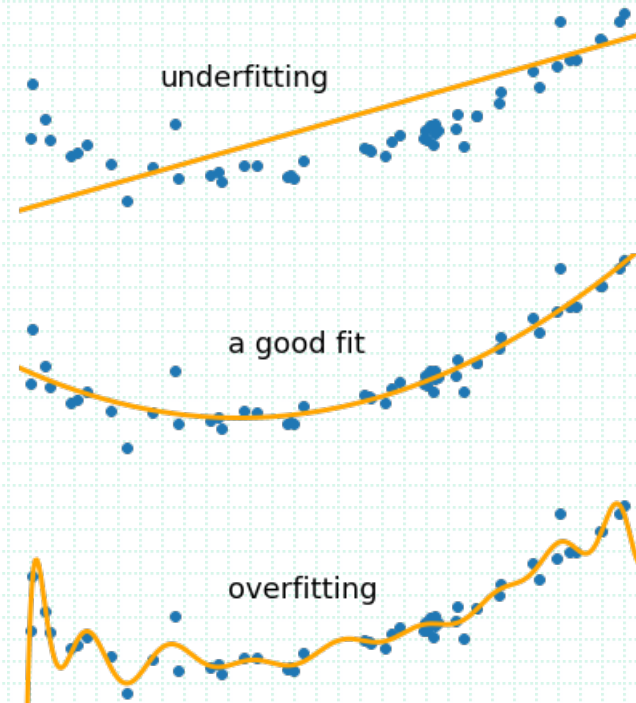
회귀 (regression)

| 선형 회귀 모델의 유형들

→ 규제란 일반적 선형 회귀의 과적합 문제를 해결하기 위해 회귀 계수에 패널티를 적용하는 것

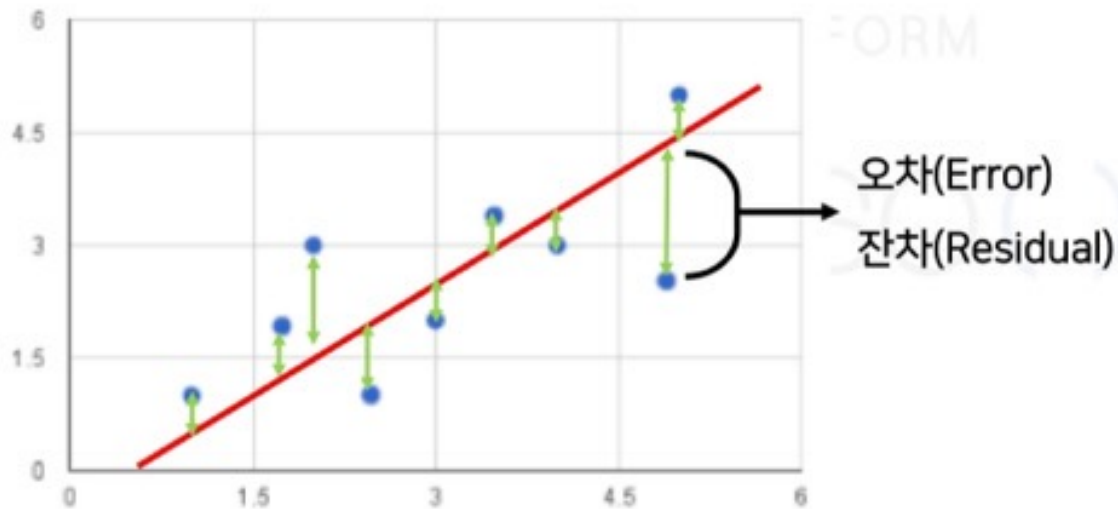
- 일반 선형회귀
- 릿지
- 라소
- 엘라스틱넷

cf) 로지스틱회귀



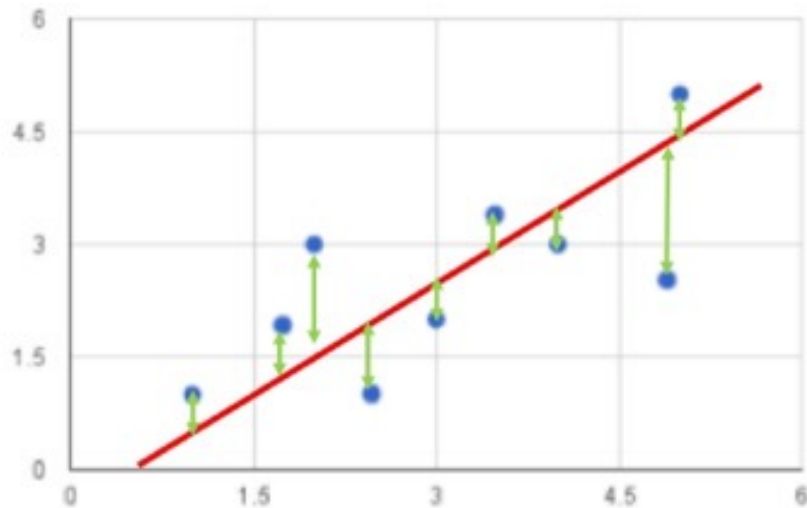
회귀 (regression)

- | 실제 값과 회귀 모델의 차이에 따른 오류 값을 남은 오류, 잔차라고 부른다
- | 우리는 이러한 잔차의 합이 최소가 되는 모델을 만들고 싶다



회귀 (regression)

| 오류 값의 합은 보통 RSS(Residual Sum of Square)을 통해 구해진다



$$\sum (\text{실제 값} - \text{예측 값})^2$$

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

$$= \text{Error}^2$$

$$= \text{RSS (Residual Sum of Squares)}$$

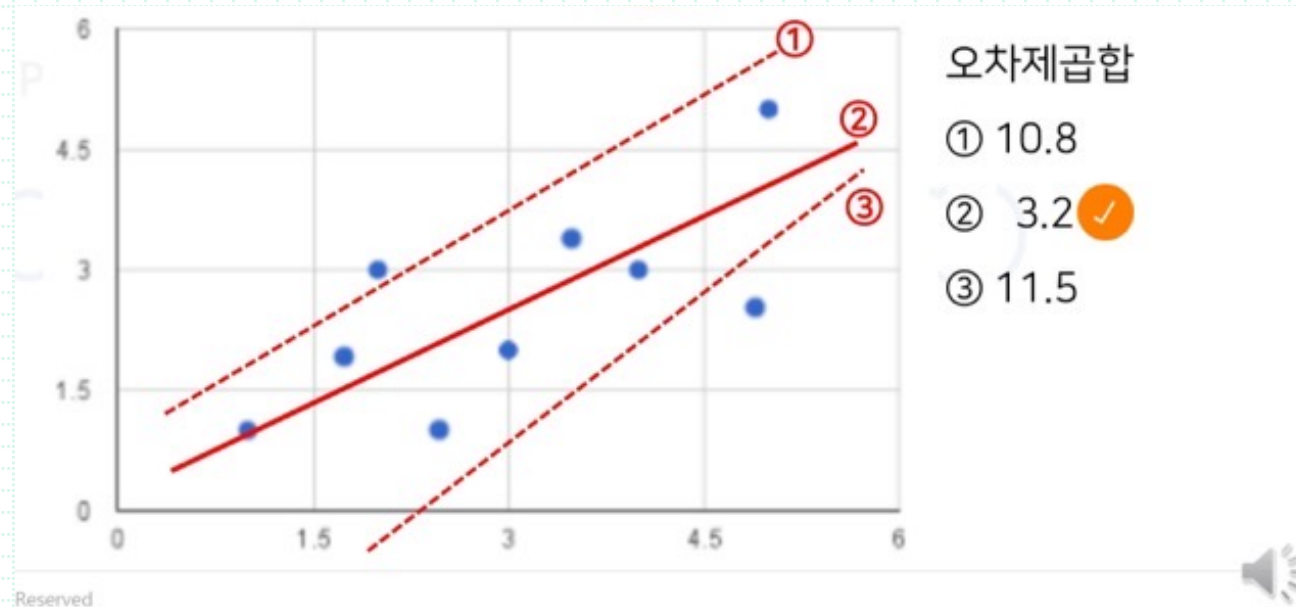
$$= \text{오차제곱합}$$

\hat{Y}
predicted
values of Y

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

회귀 (regression)

| 결국 우리는 RSS가 최소가 되는 직선을 구현하는 모델을 만들고 싶다!

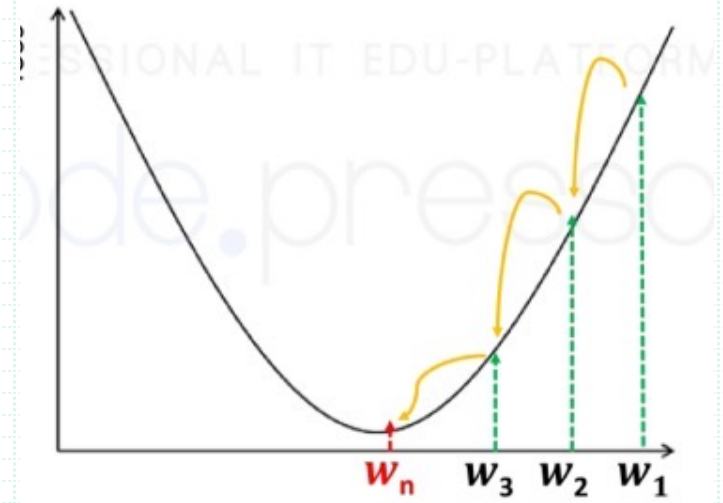


회귀 (regression)

| 어떻게 비용함수가 최소가 되는 W 파라미터를 구할 수 있을까?

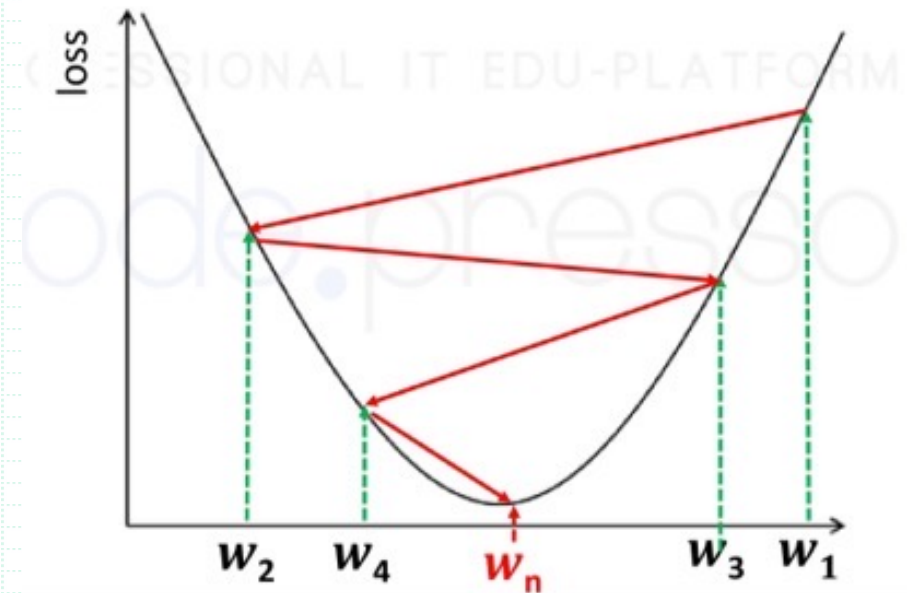
⇒ 경사하강법(Gradient Descent)

| 결과적으로 기울기가 0인 지점을 찾아야 하는 것



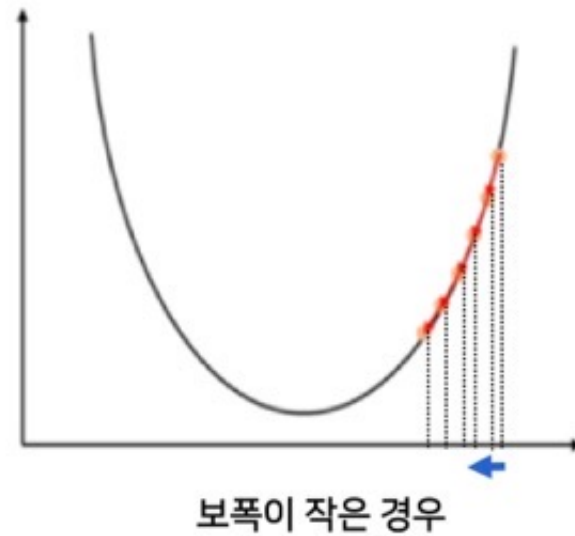
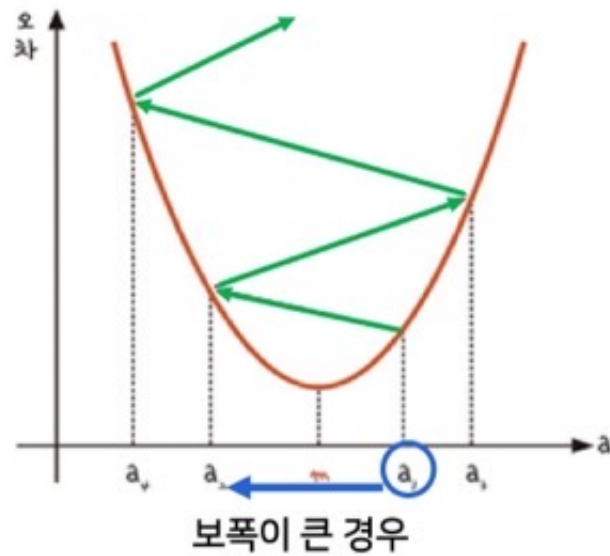
회귀 (regression)

- | 미분을 이용한 기울기 계산으로 방향성을 가질 수 있다
- | 기울기가 + 이면 음의 방향, -면 양의 방향



회귀 (regression)

| 그리고 여기서 보폭을 얼마나 이동할 것인가에 대한 Learning Rate도 중요하다

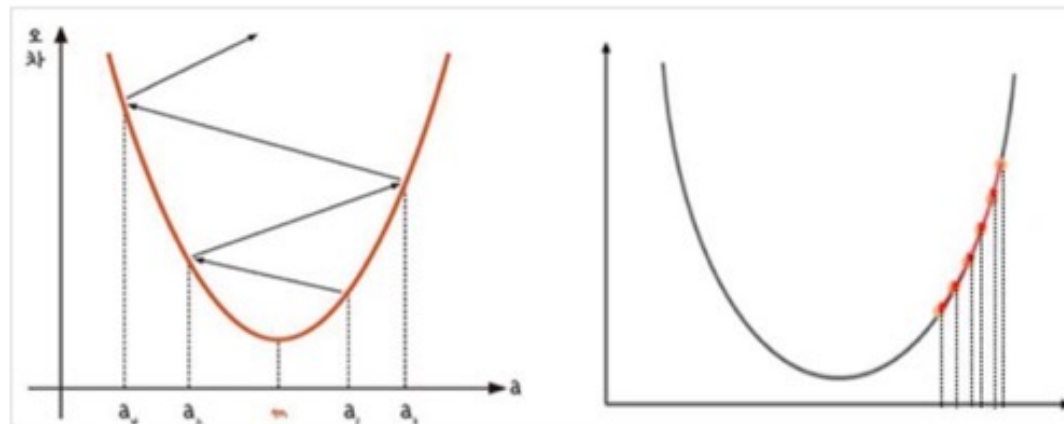


회귀 (regression)

| 그리고 여기서 보폭을 얼마나 이동할 것인가에 대한 Learning Rate도 중요하다

학습률(Learning Rate) : 하이퍼 파라미터

$$x_n = x_{n-1} - \alpha \nabla f(x_{n-1})$$



회귀 (regression)

| Sclearn의 LinearRegression 을 사용해보자

```
class sklearn.linear_model.LinearRegression(fit_intercept=True, normalize=False  
                                              copy_X = True, n_jobs=1)
```


| fit_intercept : Boolean, 절편을 계산할 것인가?

| normalize : Boolean, 입력 데이터 세트를 정규화할 것인가?


회귀 (regression)

| 회귀 평가지표

- **MSE(Mean Squared Error)**
- 정답과 예측 값 차이의 제곱의 평균

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



- **MAE(Mean Absolute Error)**
- 정답과 예측 값 차이의 절대값의 평균

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$


- **RMSE(Root Mean Squared Error)**

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})}$$

- **MAPE(Mean Absolute Percentage Error)**

$$MAPE = \frac{100}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$


회귀 (regression)

| 회귀 평가지표

• R^2 (R-Squared, 결정계수)

- 회귀 모델의 설명력을 표현하는 지표
- 추정한 회귀선이 주어진 자료에 대해 얼마나 적합한가를 의미함
- 독립변수들 간의 영향력의 정도를 정량화한 수치



$$R^2 = \frac{\text{예측 값의 분산 (SSR)}}{\text{실제 값의 분산 (SST)}}$$

$$= 1 - \frac{SSE}{SST}$$



$(0 < R^2 < 1)$

1에 가까울수록
예측 정확도 높음

- 총 변동 (SST, Total Sum of Squares)
- 오차에 의한 변동 (SSE, Error(Residual) Sum of Squares)
- 회귀선에 의한 변동 (SSR, Regression Sum of Squares)

그림출처: <https://www.dxydyt.com/that-venerable-f-test-2/>

회귀 (regression)

- | 사실 우리 사회에서 단항회귀 문제로만 해결되지 않은 문제들이 있다
- | 따라서 문제를 다항회귀적으로 표현하고 그 모델을 만드는 것이 중요
- | 다항회귀란 독립변수의 단항식이 아닌 2차, 3차 방정식과 같은 다항식으로 표현되는 것
- | 그리고 비선형회귀와 다항회귀 둘을 혼동하는 경우가 많다 이 점을 유의!

회귀 (regression)

| 그리고 w 값들의 연산이 아닌 독립 변수 x_i 들의 연산으로 이루어지는 것이다

$$y = w_0 + w_1x + w_2x^2 + \dots + w_dx^d$$

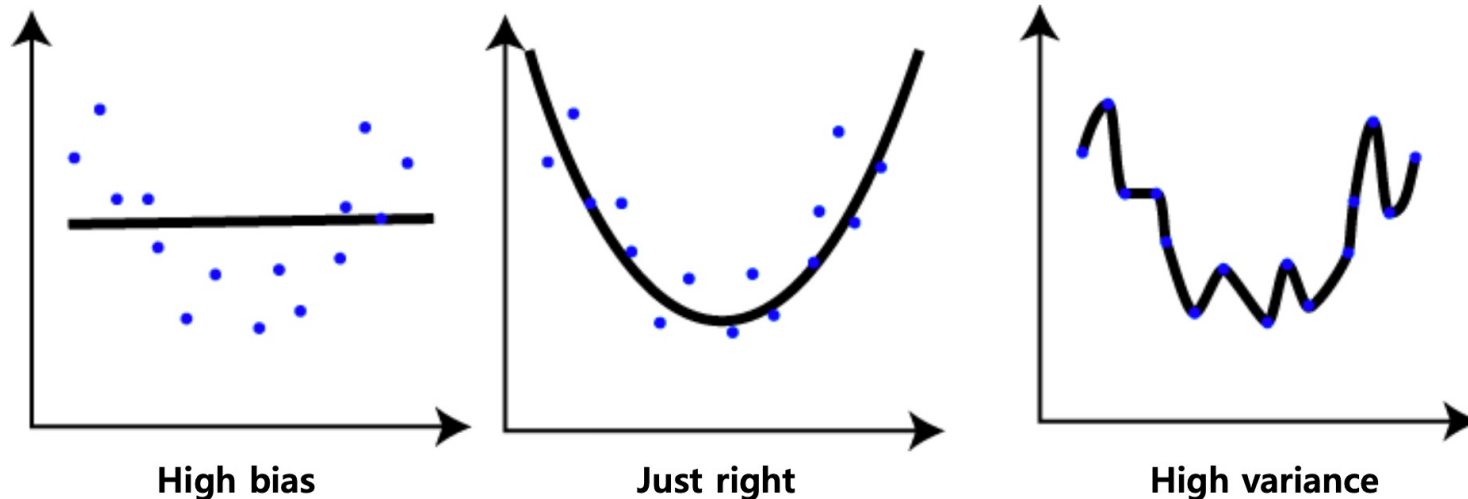
| 사이킷런에서는 다항회귀를 위한 클래스를 명시적으로 제공하지 않는다.

| 그러나 다항회귀 역시 선형회귀이기 때문에 비선형 함수를 선형 모델에 적용시키는 방법을 사용해 구현

회귀 (regression)

| 다항 회귀의 차수를 높일 수록 학습 데이터에만 너무 맞춘 학습이 이루어져서 정작 테스트 환경에서는 예측도가 떨어짐

⇒ 과적합 문제가 발생한다는 것



회귀 (regression)

- | 따라서 차수를 인위적으로 낮추어야 하는 경우가 있다
- | 이를 통해 균형잡힌 모델을 만드는 것이 중요

회귀 (regression)

- | 편향과 분산의 트레이드오프 (Bias – Variance Trade off)
- | 차수가 1인 다항회귀는 한 방향으로 치우침 → 고편향(High Bias)
- | 차수가 높은 다항회귀는 너무 다 다루려고 함 → 고분산(High Variance)

회귀 (regression)

| 편향과 분산의 트레이드오프 (Bias – Variance Trade off)

