4 [Regression]

changwon Data analysis Study group

Intro. Regression

회귀분석 : (통계학적 의미) 여러 개의 독립변수와 한 개의 종속변수 간의 상관관계를 모델링하는 기법
(Machine learning) 주어진 피처와 결정 값 데이터 기반에서 학습을 통해 최적의 회귀 계수를 찾아내는 것

단순 선형 회귀 : $Y = wx + b + eij$

다중 선형 회귀 : $Y = (w1 * x1) + (w2 * x2) + \cdots + (wn * xn) + b + eij$

(독립변수) (회귀계수) (절편) (오차항)

회귀계수 연속형 값(= 선형) vs. 회귀계수 이산형 값(=비선형) => 선형 회귀 / 비선형 회귀
독립변수 1개 vs. 독립변수 n개 => 단순(단일) 회귀 / 다중 회귀
회귀계수에 페널티 값 적용 X vs. O => 회귀식 / 규제항 회귀식

Intro. Regression

회귀분석 : (통계학적 의미) 여러 개의 독립변수와 한 개의 종속변수 간의 상관관계를 모델링하는 기법
(Machine learning) 주어진 피처와 결정 값 데이터 기반에서 학습을 통해 최적의 회귀 계수를 찾아내는 것

단순 선형 회귀 : $Y = wx + b + eij$

다중 선형 회귀 : $Y = (w1 * x1) + (w2 * x2) + \cdots + (wn * xn) + b + eij$

(독립변수) (회귀계수) (절편) (오차항)

회귀계수 연속형 값(= 선형) vs. 회귀계수 이산형 값(=비선형) => 선형 회귀 / 비선형 회귀

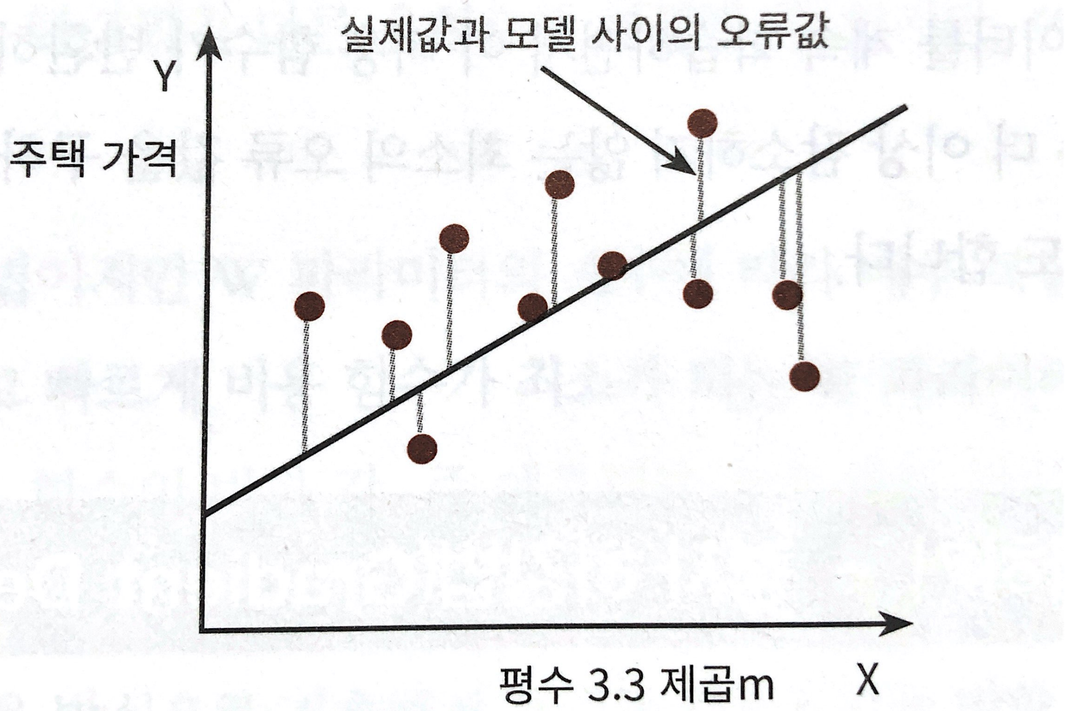
독립변수 1개 vs. 독립변수 n개 => 단순(단일) 회귀 / 다중 회귀

회귀계수에 페널티 값 적용 X vs. O => 회귀식 / 규제항 회귀식

1. 회귀 이해

단순 선형 회귀 : $Y = wx + b + e_{ij}$ (실제 값과 회귀 모델의 차이 == 잔차)

즉, 최적의 회귀 모델을 만드는 것은 전체 데이터의 잔차 합이 최소가 되는 모델을 만든다는 것을 의미



단순 합을 계산할 경우 부호(+-)에 따른 값의 상쇄 발생

1) 절대값을 취한후 오차합 계산 (MAE)

2) 오류값을 제곱하여 합산 (RSS)

1.2 Residual Sum of Square (a.k.a RSS, cost function)

단순 선형 회귀 : $Y = wx + b + e_{ij}$ (실제 값과 회귀 모델의 차이 == 잔차)

즉, 최적의 회귀 모델을 만드는 것은 전체 데이터의 잔차 합이 최소가 되는 모델을 만든다는 것을 의미

$\therefore Y = wx + b$

$x = [1.2, 1.5, 2.3]$

$y = [1.5, 2.0, 2.5]$

$w = 2, \quad b = 0$

w_1

$Y = x^T w + b + \epsilon_{ij}$

$[1.2, 1.5, 2.3] \begin{bmatrix} 2 \end{bmatrix}$ (행렬곱)

$= [2.4, 3, 4.6]$

$= \frac{(0.81 + 1, 4.31)}{3} = \frac{6+2}{3}$

RSS(1) $= 2.04$

1.3 RSS 최소화 (feat. Gradient Descent)

단순 선형 회귀 : $Y = wx + b + eij$ (실제 값과 회귀 모델의 차이 == 잔차)

즉, 최적의 회귀 모델을 만드는 것은 전체 데이터의 잔차 합이 최소가 되는 모델을 만든다는 것을 의미

$$RSS(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{(w_1 \cdot x_i + w_0)}_{(A)} - y_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(A^2 - 2Ay_i + y_i^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(w_1^2 x_i^2 + 2w_1 x_i w_0 + w_0^2 - 2w_1 x_i y_i - 2w_0 y_i + y_i^2)$$

= 즉, 편미분 결과값을 반복적으로 보정 하면서
w1, w0 값 업데이트
(이전 w1, w0 - 새로운 w1, w0)

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

$$= \frac{2}{n}\sum_{i=1}^{n} x_i \underbrace{(y_i - (w_1 x_i + w_0))}_{-1}$$

$$\frac{\partial RSS(w_0, w_1)}{\partial w_0} = \frac{1}{n}\sum_{i=1}^{n}(2w_i w_1 + 2w_0 - 2y_i)$$

$$= -\frac{2}{n}\sum_{i=1}^{n}\underbrace{(y_i - (w_i w_1 + w_0))}_{-1}$$

1.4 회귀평가 지표

회귀평가 지표는 실제 값과 회귀 예측값의 차이를 기반으로한 지표

1) MAE (Mean Absolute Error) : | 실제값과 예측값 차이 | / N

2) MSE (Mean Squared Error) : (실제값과 예측값 차이)^2 / N

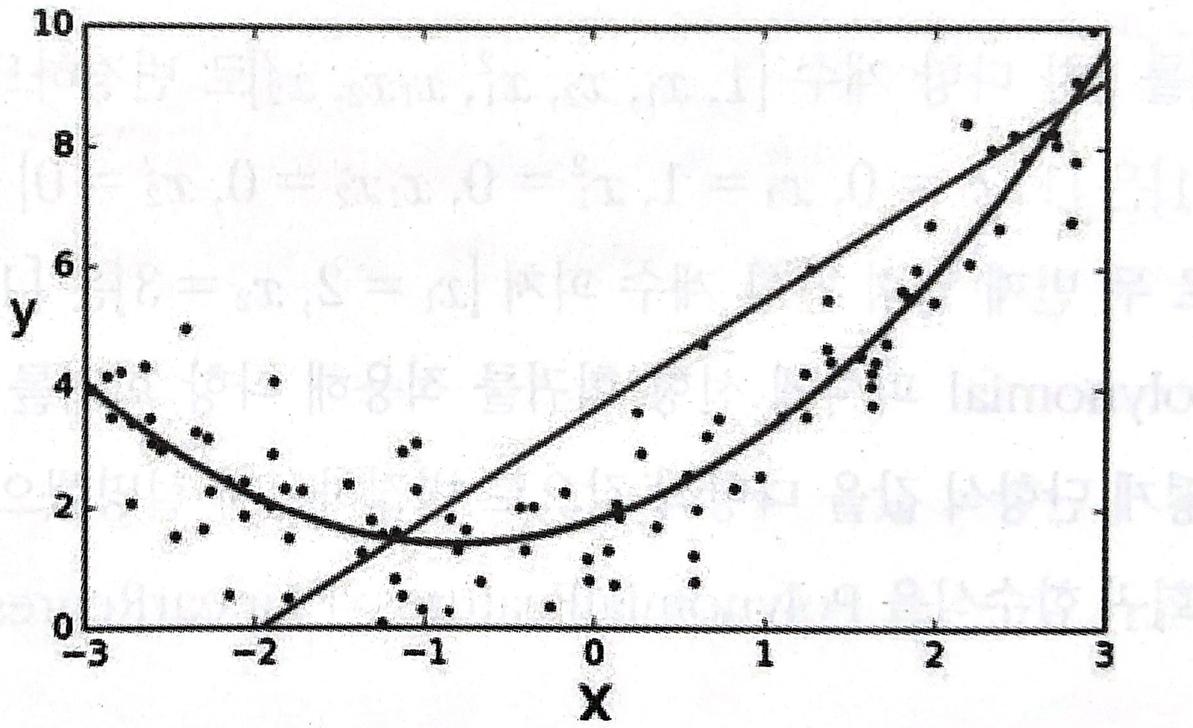
3) RMSE (Mean Squared Error) : ((실제값과 예측값 차이)^2 / N)^(1/2)

4) R^2 (R square) : 예측값 분산 / 실제값 분산

2. 다항회귀 (Polynomial Regression)

독립변수가 단항식이 아닌 2차 3차 방정식과 같은 다항식으로 표현되는 회귀 모형

비선형 회귀로 혼동하기 쉬우나, 차수가 다를 뿐 선형 회귀



〈 주어진 데이터 세트에서 다항 회귀가 더 효과적임 〉