Regression

김도윤

Linear regression

- 다음과 같은 형식의 n개의 샘플이 주어 졌을 때,

  $$(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots (\mathbf{x_n}, y_n)$$

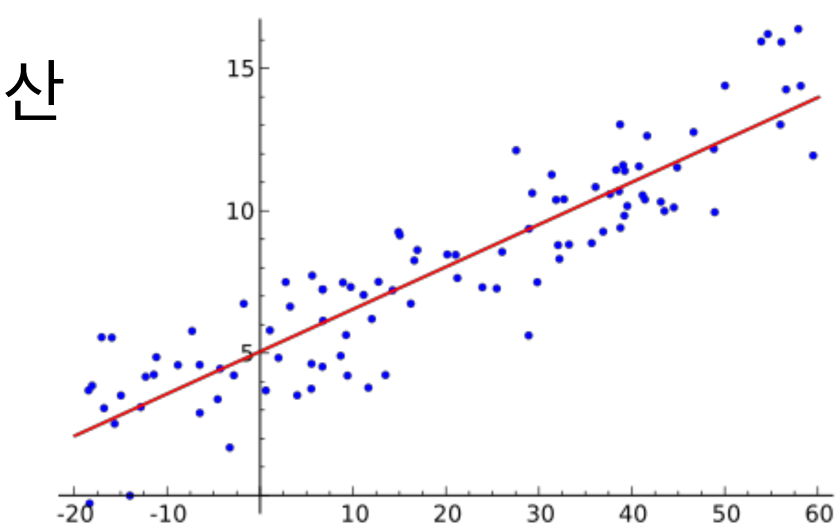
  x(m차원)와 y간의 관계를 선형 결합을 통해 상관관계를 설명하고자 하는 방법

  $$y = A\mathbf{x} + b$$

- 위와 같은 식으로 **x**와 y의 관계를 설명하며, 선형 회귀를 푼다는 것은 A와 b를 구한다는 뜻

- **선형 결합** : $A_1 x_1 + A_2 x_2 + \ldots + A_m x_m$
  위와 같이 **x**(m차원)의 각 변수에 대하여 다음과 같이 연산

- 2차원 선형 회귀 :
  x의 차원이 1차원인 경우
  (우리가 평소에 보는 오른쪽 그림을 상상하면 된다)



Using matrix inversion

- A와 b를 역행렬을 이용하여 구해보자!

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \dots & \dots \\ x_n & 1 \end{pmatrix} \cdot \begin{pmatrix} A \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$$

$$\begin{aligned} \mathbf{XA} &= \mathbf{Y} \\ \mathbf{A} &= \mathbf{X}^{-1}\mathbf{Y} \\ \mathbf{A} &= \mathbf{X}^{-1}\left(\mathbf{X}^{\mathbf{T}}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} \\ \mathbf{A} &= \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} \end{aligned}$$

- 단점
  - 데이터가 많아지면… 답 없다…
  - 진짜로 답이 없을 때도 있다**(X:singular matrix)**

행렬 분해 기법

- 조금은 더 나을 수 있는…(?) 기법이다.
- $$\mathbf{XA} = \mathbf{Y}$$
  $$\mathbf{A} = \mathbf{X}^{-1}\mathbf{Y}$$
  $$\mathbf{A} = \mathbf{X}^{-1}\left(\mathbf{X}^{\mathbf{T}}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$
  $$\mathbf{A} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

$\mathbf{X} = \mathbf{L}\mathbf{L}'$ **: Cholesky 분해**

$$\mathbf{L}\boxed{\mathbf{L}'\mathbf{A}} = \mathbf{Y} \longrightarrow \mathbf{LB} = \mathbf{Y} \longrightarrow \mathbf{B} = \mathbf{L}^{-1}\mathbf{Y}$$

$$\mathbf{L}'\mathbf{A} = \mathbf{B} \longrightarrow \mathbf{A} = \mathbf{L}'^{-1}\mathbf{B}$$

텐서플로의 선형 회귀 방식

- 위의 방식들은 일반적이지 않다
- 다음과 같은 최적화 문제를 이용하여 구하는 편임

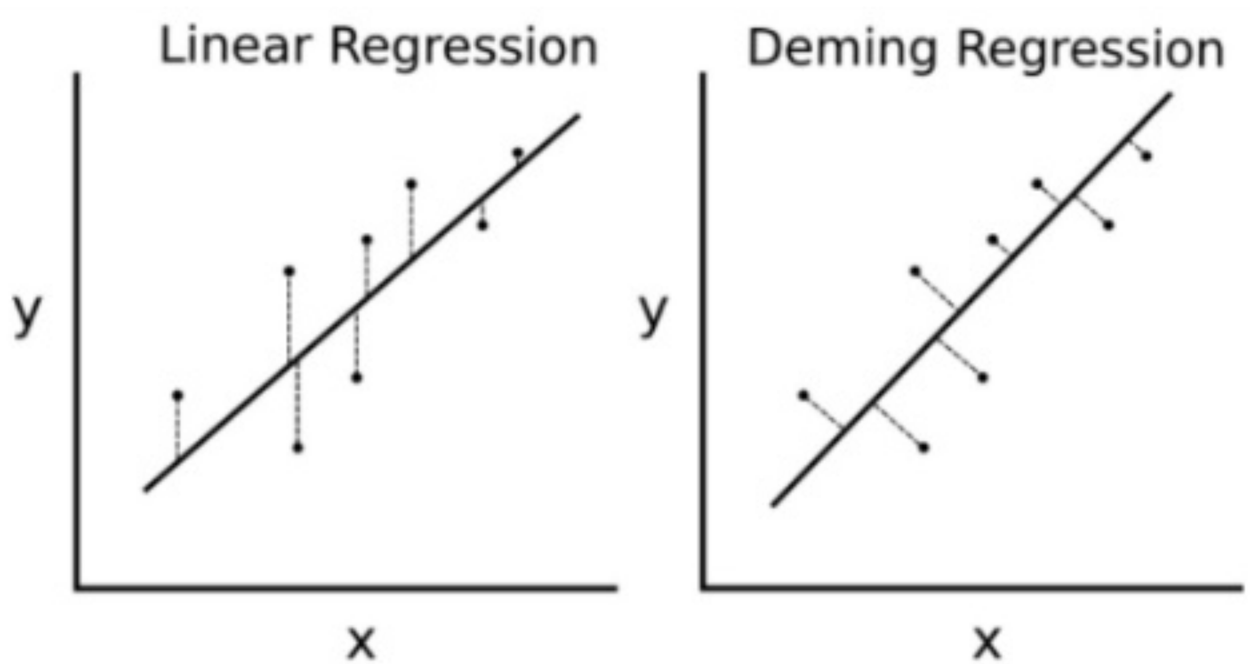
$$let\ L(\mathbf{p}, \mathbf{p}') = \sum_{i=1}^{n} (p_i - p_i')^2,$$
$$minimize\ L(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{y})$$

- 위의 L을 비용 함수 혹은 손실 함수 등이라 부름. 위의 예에선 제곱항을 이용하여 계산 했으므로 L2 비용 함수라 부르며, 절대값을 이용하여 계산하면 L1 비용 함수라 함
- 최적화 문제의 해결은
  Analytic 한 방식으로는 Linear programming, quadratic programming,
  Iterative 한 방식으로는 gradient descent 등이 있음
- Gradient descent 만 알면 충분함

Deming regression

- 기존 Linear regression은 y에 대하여 target, prediction 간의 차이를 비용함수로 이용하였음

- Deming regression은 회귀식과 target간의 거리를 비용함수로 이용함



Lasso, ridge, elastic net regression

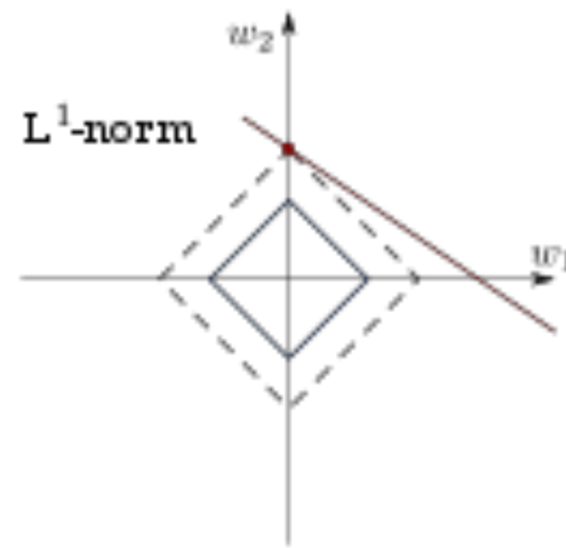
- Regression의 방법 보다는 비용함수에 regularization 항을 추가하여 robustness를 높이자는 방법임

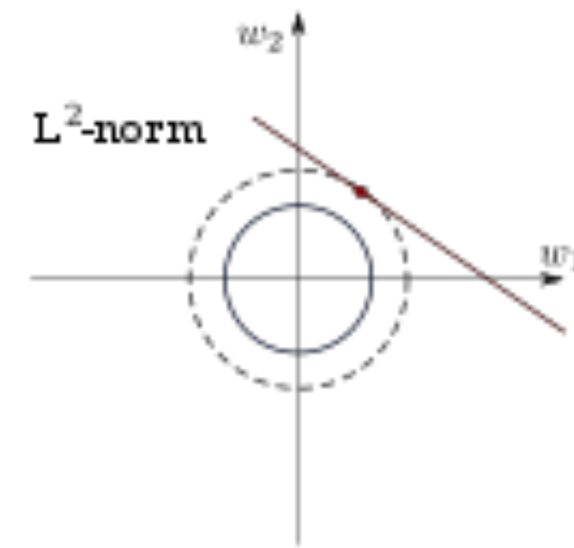
- 비용 함수

Lasso $$L(\mathbf{p}, \mathbf{p}') = \sum_{i=1}^{n} (p_i - p_i')^2 + \lambda ||\mathbf{A}||_1$$

Ridge $$L(\mathbf{p}, \mathbf{p}') = \sum_{i=1}^{n} (p_i - p_i')^2 + \lambda ||\mathbf{A}||_2$$

Elastic net $$L(\mathbf{p}, \mathbf{p}') = \sum_{i=1}^{n} (p_i - p_i')^2 + \lambda_1 ||\mathbf{A}||_1 + \lambda_2 ||\mathbf{A}||_2$$





Logistic regression

- 선형 회귀는 y(real value, 실수)의 값을 예측하는 데 사용
- Logisitic regression은 선형회귀와 유사하나 True/false를 예측하는데 사용함. 선형 회귀의 결과 값에 대하여 활성함수(예: 시그모이드)를 사용하여 0-1 사이의 값을 출력함
- 시그모이드 함수
  -