

Algorithm

수학의 집합체!

Algorithm

수학의 집합체!

Ex) Machine Learning - Linear Regression

선형 회귀는 주어진 데이터 집합 $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^n$ 에 대해, 종속 변수 y 와 p 개의 설명 변수 x_j 사이의 **선형** 관계를 모델링한다. 모델은 다음과 같은 형태를 갖는다.

$$y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

주어진 식에서 β_i 는 각 독립변수의 계수이며, p 는 선형 회귀로 추정되는 모수의 개수이다. T 는 전치를 의미하고, $\mathbf{x}_i^T \boldsymbol{\beta}$ 는 \mathbf{x}_i 와 $\boldsymbol{\beta}$ 의 내적을 의미한다. ε_i 는 *오차항*, *오차 변수*로, 관찰되지 않은 확률 변수이며, 종속 변수와 독립 변수 사이에 오차를 의미한다.

이것이 선형 회귀라 불리는 것은, 종속변수가 독립변수에 대해 선형 함수(1차 함수)의 관계에 있을 것이라 가정하기 때문이다. 그러나 $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ 의 그래프가 직선이고 y_i 가 x_i 의 선형 함수일 것이라고 생각하는 것은 잘못이다. 예를 들어 다음과 같은 "선형 회귀"도 있기 때문이다. $\mathbf{y} = \beta_1 + \beta_2 x + \beta_3 x^2 + \varepsilon$ 는 x 와 x^2 에 관해 선형이기 때문에, x 축과 y 축을 가진 그래프가 직선상에 있지 않더라도 선형회귀라고 할 수 있다.

이 식은 벡터 형식으로 표현하면 다음과 같이 표현할 수 있다.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

이 식에서 각 항의 의미는 다음과 같다.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

몇 가지 중요한 용어를 확인하고 넘어가자.

- y_i 는 **응답 변수**, 종속 변수라 불린다 (**독립 변수와 종속 변수** 읽어보기.) 어떤 변수가 종속 변수가 되고, 어떤 변수가 독립 변수가 되는지는, 어떤 변수가 무엇에 직간접적으로 영향을 주느냐에 대한 가정을 따른다. 한편, 목적에 따라서는 의존 관계에 대한 뚜렷한 이유없이 한 변수가 다른 변수에 종속하는 것으로 가정하고 선형 회귀 분석을 하기도 한다.
- $x_{i1}, x_{i2}, \dots, x_{ip}$ 는 **입력 변수**, **예측 변수**, 독립 변수라 불린다 (**독립 변수와 종속 변수** 읽어보기. 독립 변수는 독립 확률 변수와는 다른 것이다.) 행렬 \mathbf{X} 는 설계 행렬이라 불리기도 한다.
 - 일반적으로 입력 변수에 상수가 포함된다. 예를 들어, x_{i1} 를 상수로 택한다 ($= 1 \ i = 1, \dots, n$) x_{i1} 앞에 붙는 상수 β 를 절편이라 부른다. 많은 선형 통계 모델에서 절편이 필요하며, 실질적으로 절편이 0인 경우에도 이를 포함해 모델링한다.
 - 때로 독립 변수는 다른 독립 변수 또는 데이터에 대해 비선형 함수이기도 하다. 이러한 경우에도 이 독립 변수가 파라미터 벡터 $\boldsymbol{\beta}$ 에 대해서만 선형이지만 하면 여전히 선형 모델이라 부른다.
 - 독립 변수 x_{ij} 는 확률 변수로 생각할 수도 있고, 또는 고정된 값으로 생각할 수도 있다. 경우에 따라 두 가지 중에 적합한 것을 선택해야 하지만, 두 가지 모두 같은 추정 과정을 거친다. 하지만 각각의 경우에서 해석은 다르다.
- $\boldsymbol{\beta}$ 는 p 차원 **파라미터 벡터**이다. 이것의 각 원소는 **회귀 계수**라고 불리기도 한다. 파라미터 벡터의 원소는 종속 변수에 대한 편미분으로 해석할 수도 있다.
- ε_i 는 **오차항**, **노이즈**이다. 이 변수는 종속 변수 y 에 대한 모든 오차 요인을 포함한다.