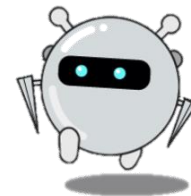


# 머신러닝 기초와 선형회귀



진행자 : 현시은

# 머신러닝의 용어

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 인공지능? 머신러닝? 딥러닝?

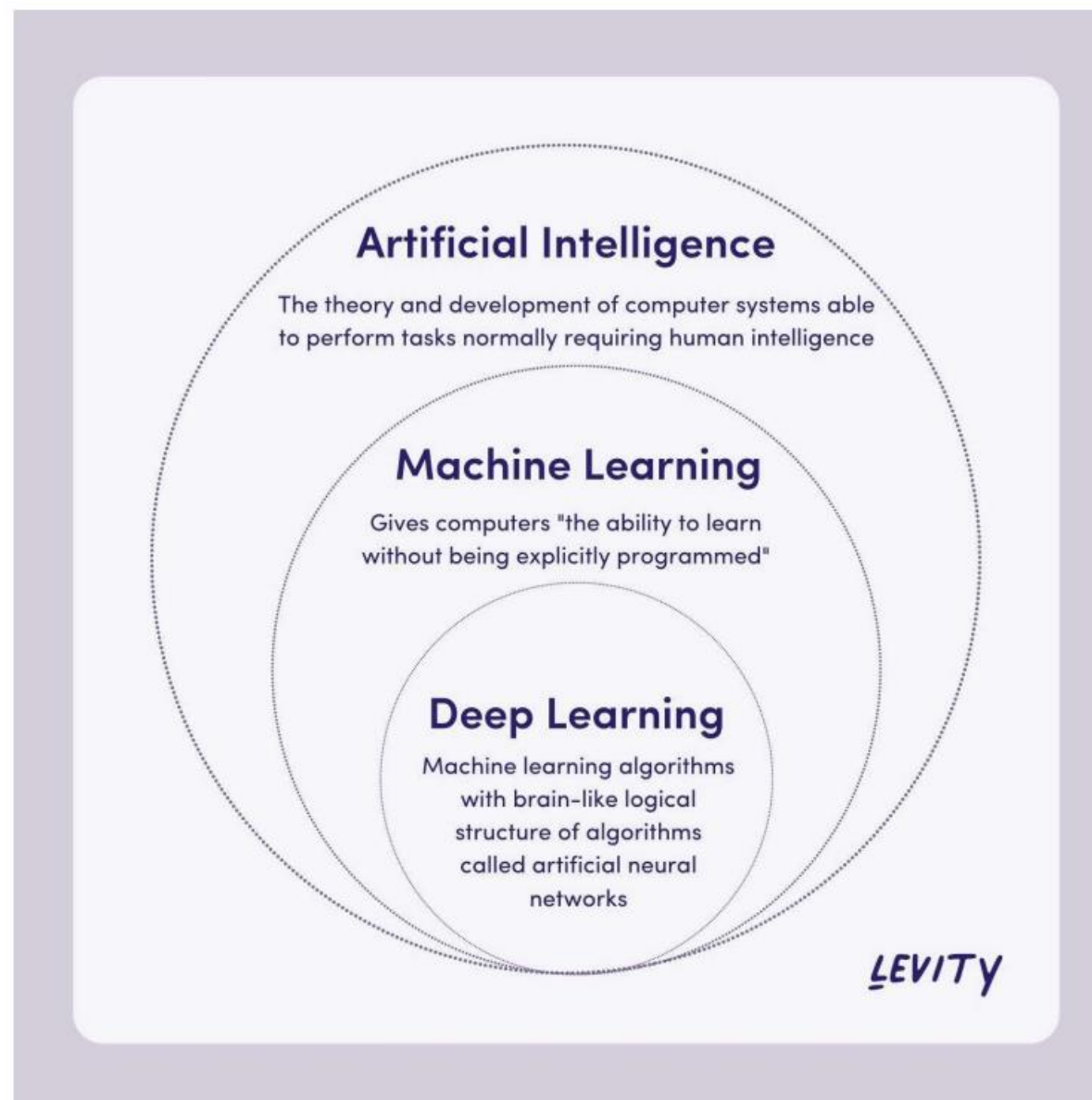


그림 출처 :  
<https://levity.ai/blog/difference-machine-learning-deep-learning>

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 머신 러닝이란?

머신러닝은 인공지능의 하위 집합으로, 많은 양의 데이터를 제공하여 명시적으로 프로그래밍하지 않고 신경망과 딥 러닝을 사용하여 시스템이 자율적으로 학습하고 개선할 수 있게 해줍니다. (Google Cloud 사이트에서 발췌)

### <머신러닝의 과정>

데이터를 활용하여 머신 러닝 모델을 학습시킨다. 머신 러닝 모델의 학습은 손실함수를 최적화하는 것으로 이루어진다. 손실함수의 최적화를 통해 문제 상황에 가장 적합한 모델 파라미터를 찾는 것이 모델의 학습 및 최적화이다.

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 중요 키워드

데이터

모델

학습

모델 파라미터

손실함수

최적화

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 데이터

데이터는 Label(또는 Target)과 Feature로 구성되어 있다.

Label(또는 Target) : 우리가 예측할 대상을 의미한다. 함수에서  $y$  값, 즉 출력되는 값을 의미한다.

Feature : 우리가 예측을 위해 사용할 변수를 의미한다. 함수에서  $x$  값, 즉 입력되는 값을 의미한다. 주로 다변수인 경우가 대부분으로  $x_1, x_2, \dots, x_n$ 으로 나타낸다.

머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

데이터

<예시>

키	몸무게	시력	재산	성적	성별
180	60	1.0	A	100	남
170	50	2.0	C	90	여
160	80	1.5	B	80	남
150	70	2.0	A	50	남
121	59	0.5	B	70	여
153	85	0.1	A	50	여
169	53	0.6	A	60	남
158	49	0.3	B	80	남
201	100	0.9	C	90	여
147	50	1.0	A	80	남

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

데이터

<예시>

$$y = f(x)$$

$$y = g(x_1, x_2, \dots, x_n)$$



- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 데이터

데이터는 3가지 종류로 구성되어 있다.

- Training 데이터
- Validation(또는 evaluation) 데이터
- Test(또는 inference) 데이터

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 데이터

- Training 데이터 : 실제로 모델을 최적화할 때 쓰이는 훈련 데이터

- Validation(또는 evaluation) 데이터 : 모델이 잘 최적화되어가고 있는지 확인하기 위한 데이터

- Test(또는 inference) 데이터 : 훈련이 다 완료되고, 모델의 성능을 최종적으로 검증하기 위한 데이터

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 데이터

- Training 데이터 : 연습 문제

- Validation(또는 evaluation) 데이터 : 모의 고사

- Test(또는 inference) 데이터 : 실전 시험

머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

데이터

<예시>

키	몸무게	시력	재산	성적	성별
180	60	1.0	A	100	남
170	50	2.0	C	90	여
160	80	1.5	B	80	남
150	70	2.0	A	50	남
121	59	0.5	B	70	여
153	85	0.1	A	50	여
169	53	0.6	A	60	남
158	49	0.3	B	80	남
201	100	0.9	C	90	여
147	50	1.0	A	80	남

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 모델

인공지능에서 '모델'은 데이터를 기반으로 복잡한 현실 세계의 패턴을 학습하고, 이를 바탕으로 예측이나 결정을 내리는 알고리즘의 구조

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 모델

<예시>

$$y = f(x)$$

$$y = g(x_1, x_2, \dots, x_n)$$

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 학습(Training)

- 데이터로부터 패턴, 특성, 규칙을 배우고, 이를 이용하여 예측이나 결정을 수행하기 위한 알고리즘을 훈련시키는 과정

- 즉, 학습이란 최적의 모델 파라미터를 찾는 과정이고, 우리의 목표는 학습을 **최대한 잘** 시키는 것!

- 그럼 모델 파라미터가 뭔데?

- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 모델 파라미터(Model Parameter)

- 모델이 학습 과정을 통해 데이터로부터 학습하고 조정하는 내부 변수를 의미한다. 이러한 파라미터들은 모델이 어떻게 입력 데이터를 해석하고 예측을 수행하는지를 결정한다.



- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 모델 파라미터

<예시>

$$y = f(x)$$

$$y = g(x_1, x_2, \dots, x_n)$$

- 머신러닝의 용어

회귀 모델

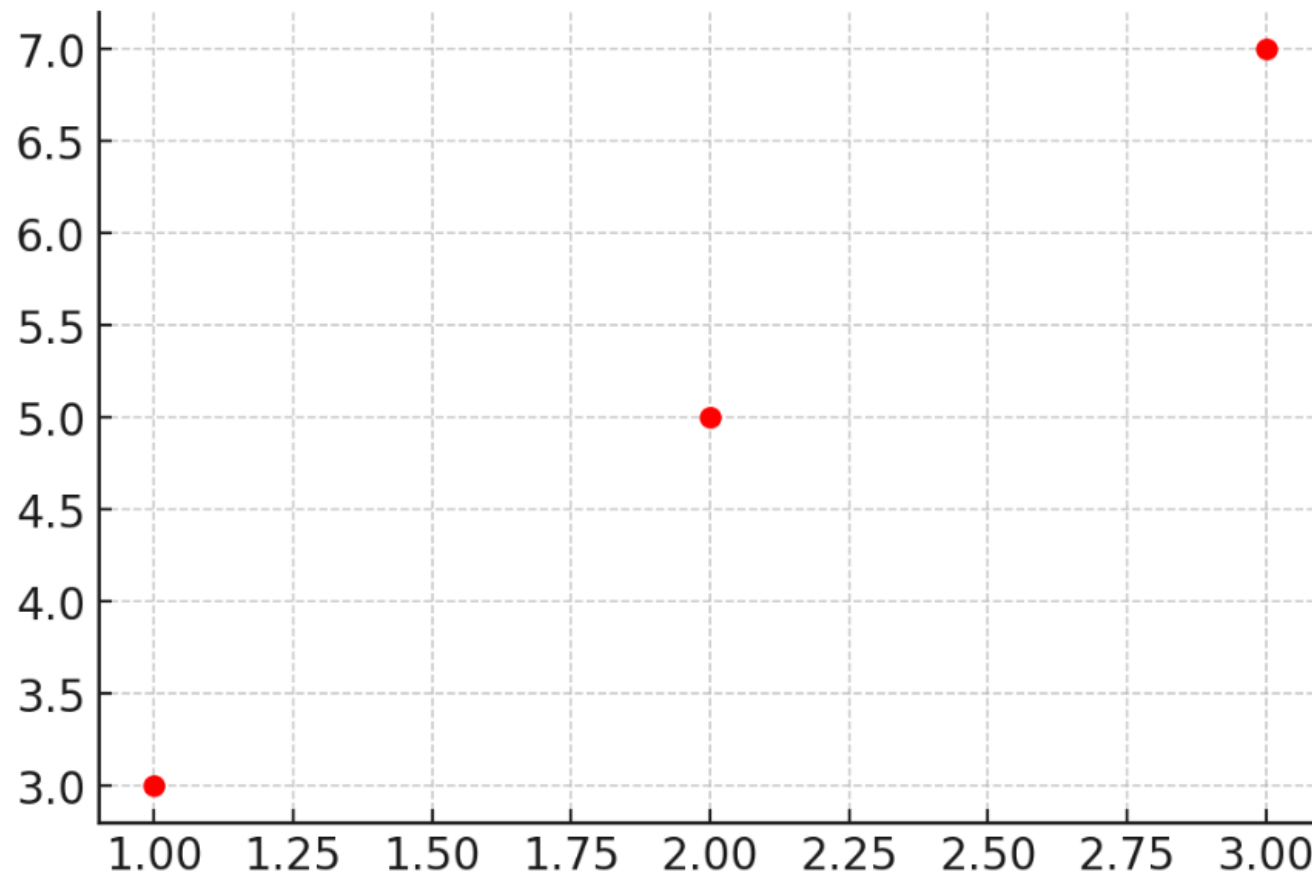
선형 모델

최소 제곱법

선형 회귀

## 모델 파라미터(Model Parameter)를 학습한다?

(1, 3), (2, 5), (3, 7)이라는 데이터 3개가 있다고 가정하자.



- 머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

선형 회귀

## 손실함수와 최적화

- 손실함수 : 모델의 예측값과 실제 값 사이의 차이를 측정하는 함수이다.  
이 함수는 모델이 얼마나 잘 혹은 못하고 있는지를 수치적으로 나타낸다.

- 최적화 : 손실함수를 어떻게 가장 작게 (혹은 가장 크게) 할 것인지에 대한  
다양한 수학적 방법들을 의미한다.

# 회귀 모델

## 머신러닝의 용어

- 회귀 모델

### 선형 모델

### 최소 제곱법

### 선형 회귀

## 인공지능에서 Task란?

- AI 시스템이 수행해야 하는 특정한 작업이나 문제를 의미
- 대표적인 Task 2개를 소개하면 다음과 같다.

Regression(회귀): 연속적인 값을 예측하는 것이 목표이다.

· ex) 2023년 1월 1일부터 2023년 12월 31일 까지의 평균 기온 데이터를 이용하여 2024년 1월 1일의 평균 기온을 예측하는 Task

Classification(분류): 특정 카테고리 혹은 이산적인 값을 예측하는 것이 목표이다.

· ex) 주어진 동물 사진을 보고 그 동물이 개인지 고양이인지 분류하는 Task

## 머신러닝의 용어

- 회귀 모델

### 선형 모델

### 최소 제곱법

### 선형 회귀

## 인공지능에서 Task란?

- AI 시스템이 수행해야 하는 특정한 작업이나 문제를 의미
- 대표적인 Task 2개를 소개하면 다음과 같다.

Regression(회귀): 연속적인 값을 예측하는 것이 목표이다.

· ex) 2023년 1월 1일부터 2023년 12월 31일 까지의 평균 기온 데이터를 이용하여 2024년 1월 1일의 평균 기온을 예측하는 Task

Classification(분류): 특정 카테고리 혹은 이산적인 값을 예측하는 것이 목표이다.

· ex) 주어진 동물 사진을 보고 그 동물이 개인지 고양이인지 분류하는 Task

## 머신러닝의 용어

- 회귀 모델

### 선형 모델

### 최소 제곱법

### 선형 회귀

## Regression

- Input 변수를 기반으로 output 변수를 예측하거나 추정하는 방법
- 출력 값으로 산술적인 예측값(실수값)이 나온다.

(참고로, Classification(분류) 모델은 이산적인 값이 나온다.)

### Regression(회귀) 모델의 종류

- 선형 회귀 (Linear Regression) : 두 변수의 관계를 설명하는 선형 함수를 찾는 것
- 로지스틱 회귀 (Logistic Regression) : 시스템이 일반적으로 클래스 예측에 매핑하는 0.0에서 1.0 사이의 확률을 생성하는 것

# 선형 모델



머신러닝의 용어

회귀 모델

• **선형 모델**

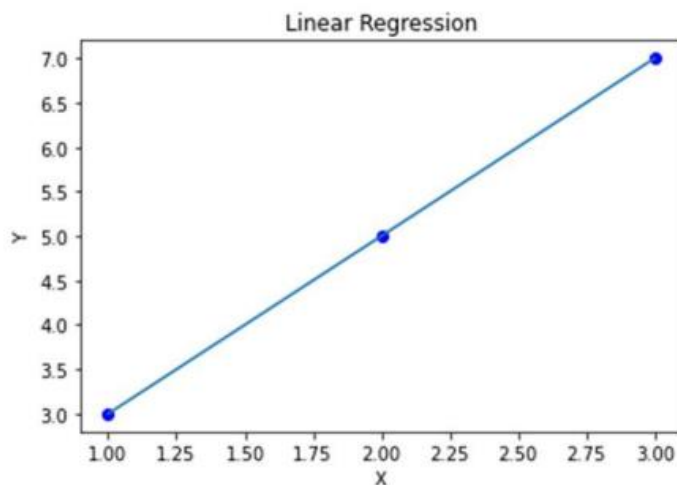
최소 제곱법

선형 회귀

## 선형 회귀와 선형 모델이란?

- 선형 회귀 : 두 변수의 관계를 설명하는 선형 함수를 찾아내는 것

ex)  $X = [1, 2, 3]$ ,  $Y = [3, 5, 7]$ ,  $X = 4$  일 때  $Y$  값은?



➔  $H(W, b) = Wx + b$  (목표 :  $W = 2$ ,  $b = 1$ )

머신러닝의 용어

회귀 모델

• **선형 모델**

최소 제곱법

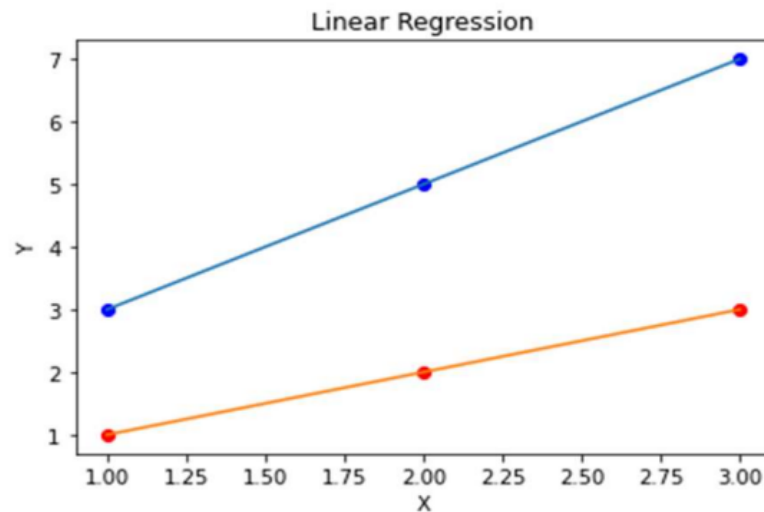
선형 회귀

## 선형 회귀와 선형 모델이란?

- 선형 회귀 : 두 변수의 관계를 설명하는 선형 함수를 찾아내는 것

ex) [가설 초기화]

- $W = 1, b = 0$
- 얼마나 잘못되었는가? → 손실 함수(Loss Function)
- 손실 함수를 **최소화** 하는 것이 목표!



머신러닝의 용어

회귀 모델

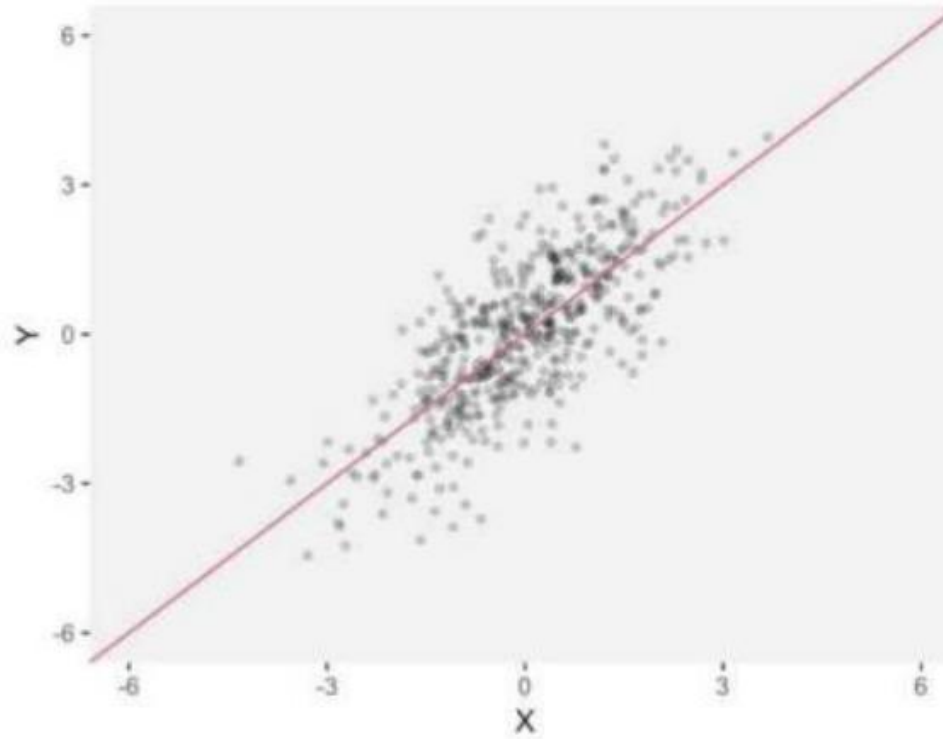
• **선형 모델**

최소 제곱법

선형 회귀

## 왜 선형 회귀 인가?

- 실제 데이터의 측정값에는 노이즈가 포함될 수밖에 없고, 이런 노이즈들로부터 다시 원래의 선형 연속함수로 돌아가는 과정이기 때문.



머신러닝의 용어

회귀 모델

- **선형 모델**

최소 제곱법

선형 회귀

## 선형 모델(Linear model) 세우기

- 모델 :  $F(m, b ; x) = mx + b$
- 모델 파라미터 :  $m$  (slope),  $b$  (intercept)
- 모델 파라미터를 최적화하는 것이 목표
- 손실함수를 최적화함으로써 모델 파라미터를 최적화한다.
- 선형 모델의 최적화에는 '손실함수' 중 하나인 잔차제곱합(RSS)이 사용됨
- '최소제곱법'은 손실함수를 최적화하는 방법 중 하나

## 최소 제공법

머신러닝의 용어

회귀 모델

선형 모델

- **최소 제곱법**

선형 회귀

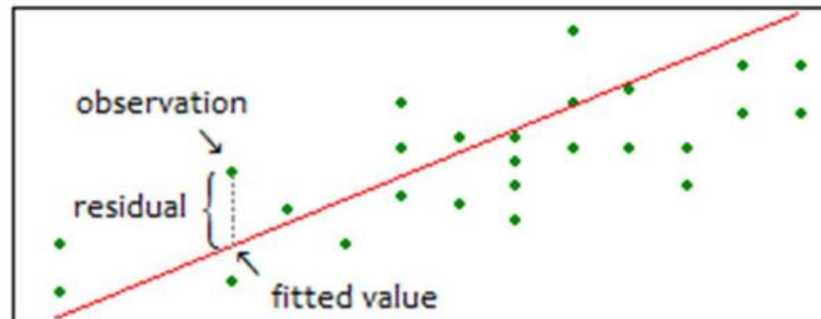
## 최소 제곱법 : 잔차제곱합을 최소로 하는 것

- Least Square Method (최소제곱법) 이란?

; 최소 제곱법은 선 또는 곡선에서 Residual (잔차) 의 제곱의 합을 줄여 데이터 점 집합에 가장 적합한 곡선 또는 가장 적합한 선을 찾는 프로세스

- Residual (잔차) 란?

; 예측 값과 실제 값의 차이



머신러닝의 용어

회귀 모델

선형 모델

- **최소 제곱법**

선형 회귀

## 잔차제곱합이란?

- N개의 데이터셋
- **True data points** :  $(x_i, y_i^{(true)}), 0 \leq i \leq N - 1$
- **Expected data points** :  $(x_i, y_i^{(pred)}), y_i^{(pred)} = mx_i + b, 0 \leq i \leq N - 1$
- 잔차(Residual) :  $d_i = (y_i^{(true)} - y_i^{(pred)})$
- 잔차 제곱합 (RSS) :

$$\sum_{n=0}^{N-1} d_i^2$$

선형 회귀 모델의 학습을 통해  
모델 파라미터 m, b값을 조절하여  
RSS를 최소화하고자 함

머신러닝의 용어

회귀 모델

선형 모델

- 최소 제곱법

선형 회귀

## 잔차제곱합이란?

- 잔차 제곱합 (RSS) :

$$\sum_{n=0}^{N-1} d_i^2$$

- **Q) 잔차의 합 대신 제곱을 사용하는 이유?**
  - 잔차의 합은 선형 회귀 모델의 오차를 대표할 수 없다. 잔차 부호의 통일이 필요하다.
- **Q) 잔차의 절댓값 합 대신 제곱합을 사용하는 이유?**
  - 절댓값을 사용한 경우 잔차 부호는 통일되나 최적화 과정이 제곱에 비해 복잡하다.
  - 일반적으로 절댓값 함수는 미분 불가능한 반면 제곱합은 미분이 수월하다.



머신러닝의 용어

회귀 모델

선형 모델

- 최소 제곱법

선형 회귀

## 손실 함수

- 손실함수

$$\mathcal{L}(\mathbf{m}, \mathbf{b}; (x_n, y_n^{(true)})_{n=0}^{N-1}) = \sum_{n=0}^{N-1} \left( y_n^{(true)} - F(\mathbf{m}, \mathbf{b}; x_n) \right)^2$$

를 최소화하는 변수  $\mathbf{m}$ ,  $\mathbf{b}$ 를 찾아야 하며

이러한 접근 방식을 **최소제곱법**이라 한다.

- $\mathbf{m}^*, \mathbf{b}^* = \operatorname{argmin}_{\mathbf{m}, \mathbf{b} \in R} \mathcal{L}$

머신러닝의 용어

회귀 모델

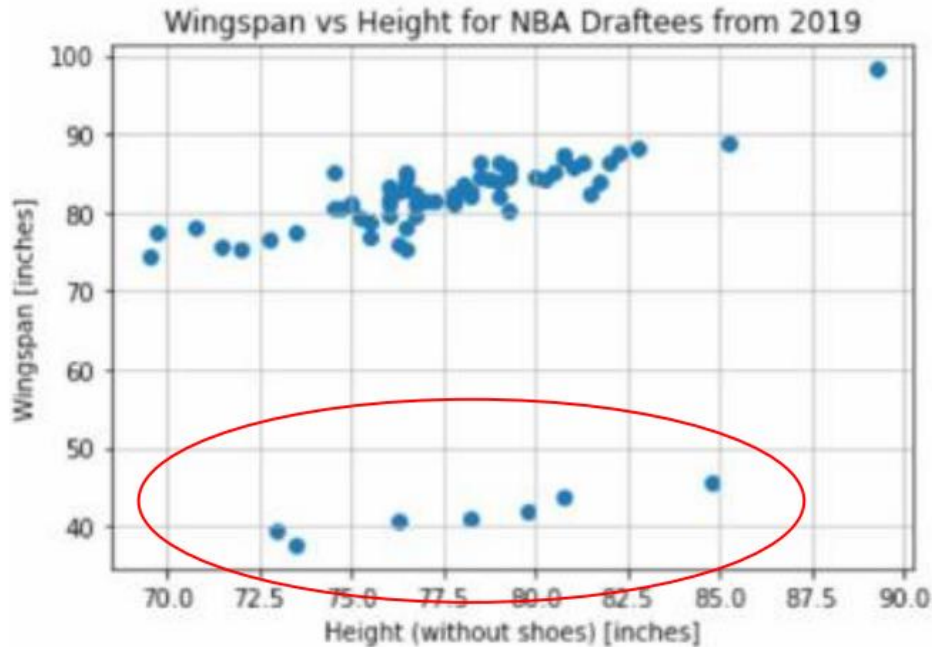
선형 모델

- 최소 제곱법

선형 회귀

## 최소 제곱법의 한계

- Outlier가 많이 존재하는 데이터에서는 최소제곱법을 적용할 수 없다.



# 선형 회귀

머신러닝의 용어

회귀 모델

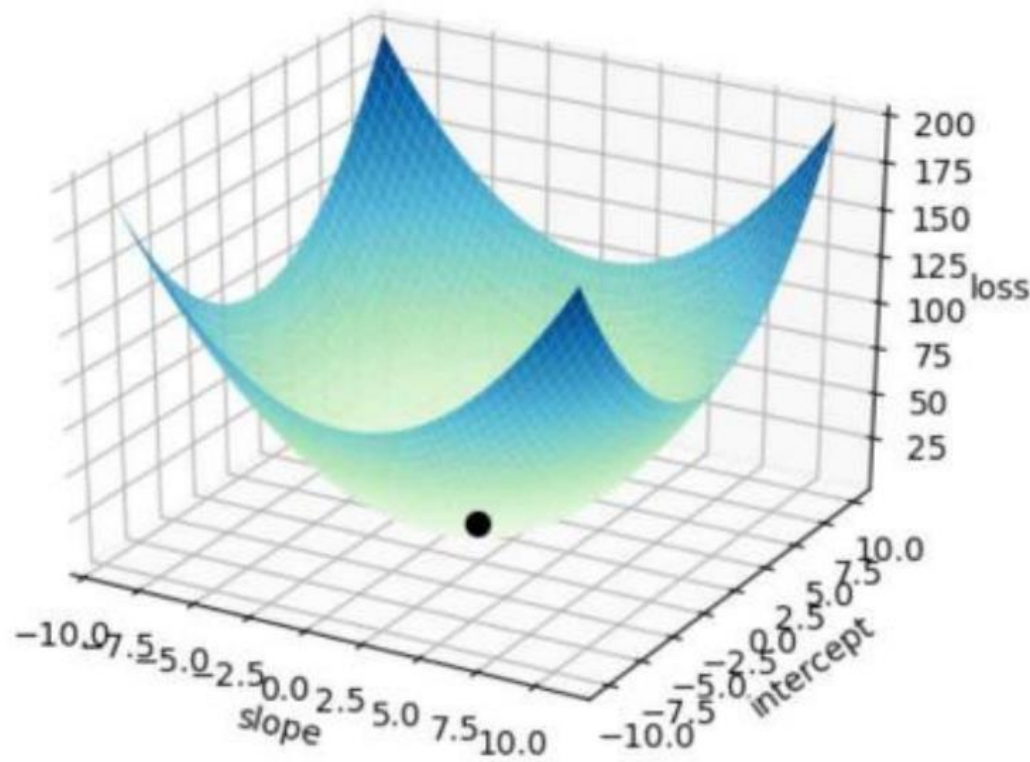
선형 모델

최소 제곱법

- 선형 회귀

$m^*, b^*$  구하기

$$\mathcal{L}(m, b; (x_n, y_n^{(true)})_{n=0}^{N-1}) = \sum_{n=0}^{N-1} (y_n^{(true)} - F(m, b; x_n))^2$$



$\nabla \mathcal{L} = 0$  를 만족하는  $m^*, b^*$  를 찾는다.

머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

- 선형 회귀

## $m^*, b^*$ 구하기

$$\mathcal{L} = \sum_{n=0}^{N-1} (y_n - (mx_n + b))^2$$

$$= \sum_{n=0}^{N-1} (m^2 x_n^2 + b^2 + y_n^2 + 2bmx_n - 2mx_n y_n - 2by_n)$$

손실함수를  $m, b$ 에 대해 편미분한다.

$$\frac{\partial \mathcal{L}(m, b)}{\partial m} = \sum_{n=0}^{N-1} (2mx_n^2 + 2bx_n - 2x_n y_n) = 0 \quad - \quad (1)$$

$$\frac{\partial \mathcal{L}(m, b)}{\partial b} = \sum_{n=0}^{N-1} (2b + 2mx_n - 2y_n) = 0 \quad - \quad (2)$$

머신러닝의 용어

회귀 모델

선형 모델

최소 제곱법

- 선형 회귀

$m^*, b^*$  구하기

$$\frac{\partial \mathcal{L}(m, b)}{\partial m} = \sum_{n=0}^{N-1} (2mx_n^2 + 2bx_n - 2x_n y_n) = 0 \quad - \quad (1)$$

$$\frac{\partial \mathcal{L}(m, b)}{\partial b} = \sum_{n=0}^{N-1} (2b + 2mx_n - 2y_n) = 0 \quad - \quad (2)$$

1과 2를 연립하면..

$$m^* = \frac{\sum_{n=0}^{N-1} x_n y_n - \frac{1}{N} \sum_{n=0}^{N-1} x_n \sum_{n=0}^{N-1} y_n}{\sum_{n=0}^{N-1} x_n^2 - \frac{1}{N} (\sum_{n=0}^{N-1} x_n)^2} \quad b^* = \bar{y} - m^* \bar{x}$$

이렇게 구한  $m^*, b^*$  를 '최소제곱추정량' 이라고 한다.



감사합니다

