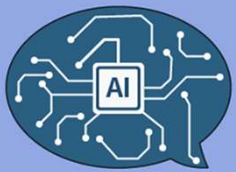
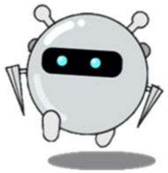


선형 회귀



OUTTA

AI, Machine Learning, and Deep Learning

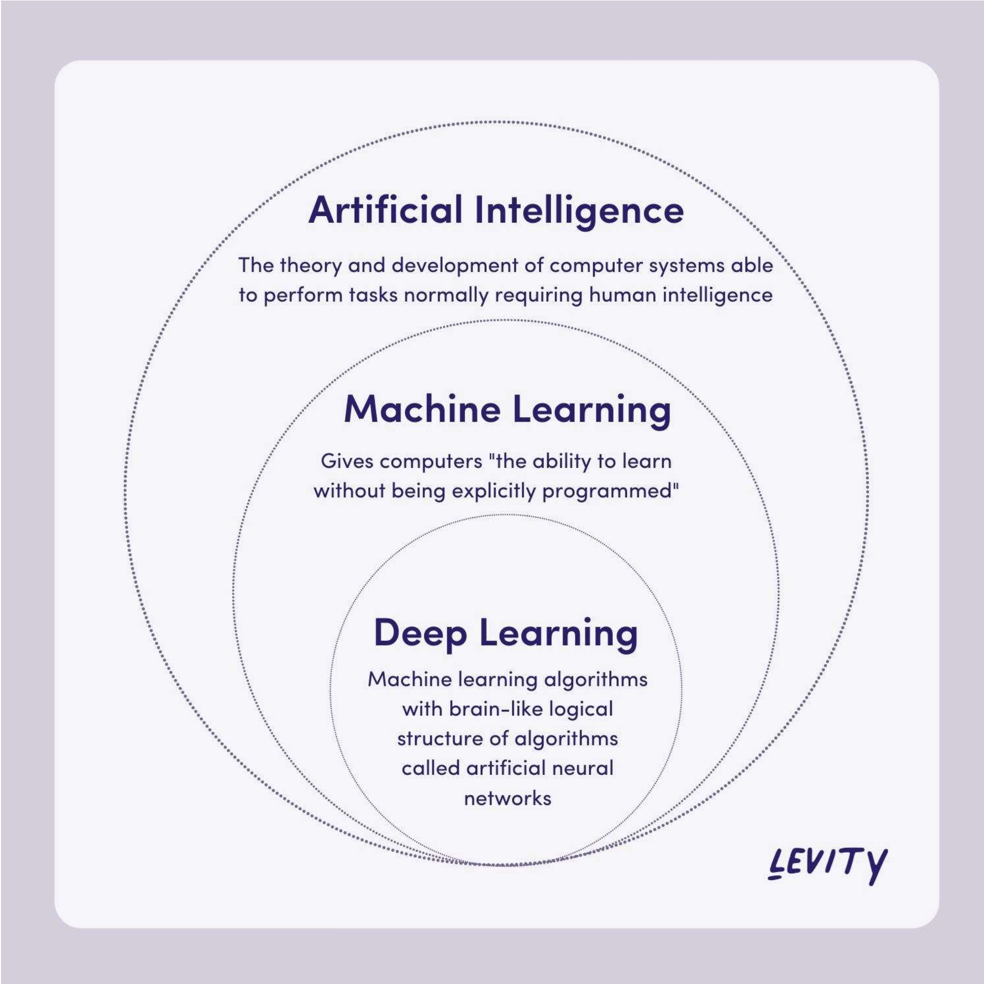


그림 출처: https://levity.ai/blog/difference-machine-learning-deep-learning

COUTA

What is a Machine Learning?

머신러닝 시스템 (ML system)은 한 번도 본 적 없는 데이터로부터 의미 있는 예측을 생산하기 위해 입력값들을 어떻게 조합할지 배웁니다.

➔ 어떻게 머신러닝을 학습시킬 수 있을지, 최적의 학습은 어떻게 이뤄지는 지 등을 다루는 것

➔ 손실함수를 최적화하여 문제 상황에 가장 적합한 모델 파라미터를 찾아 모델을 완성시키는 것

OUTTA

기본 Machine Learning 용어들

- **Label 이란?**

; Label은 우리가 예측할 대상입니다. 단순 선형 회귀에서는 y 변수입니다.

ex) 사진에서 보여진 동물의 종, 오디오 클립의 의미 등

- **Feature 이란?**

; Feature은 입력 변수 입니다. 단순 선형 회귀에서는 x 변수입니다. 더 섬세한 ML 프로젝트일 수록 대단히 많은 feature을 가지며 이를 $x_1, x_2, \cdots, x_n$ 으로 나타냅니다.

ex) 이메일 내용의 글자, 수신자의 주소 등 (spam/ham mail 분류에서)

- **Example 이란?**

; Example은 Data의 특정한 예시입니다. 이는 Unlabeled, Labeled 두 종류로 나눠질 수 있으며, Model을 Train하기 위해 사용됩니다. ; labeled examples: {features, label}: (x, y)

OUTTA

기본 Machine Learning 용어들

- ## Model 이란?

; Model은 Feature와 Label간의 관계를 정의합니다.

- Training 이란?

; Training은 모델을 만들거나 학습하는 것을 의미합니다. 즉, 모델에게 labeled된 example 들을 제공하고 모델이 feature와 label 사이 관계를 배우도록 하는 것을 의미합니다.

- Inference (추론) 이란?

; Inference는 unlabeled 된 example 들에 trained model을 적용하는 것을 의미합니다. 즉, trained된 모델이 의미 있는 예측값 (y') 을 만들도록 하는 것을 의미합니다.

OUTTA

기본 Machine Learning 용어들

- Model 이란?

; Model은 Feature와 Label간의 관계를 정의합니다.

- Model Parameter 란?

; 올바른 예측과 결정을 얻기 위해 조정하는 변수들

- Loss Function (손실 함수) 란?

; 모델의 질을 평가하는 함수

➔손실함수를 최적화하여 문제 상황에 가장 적합한 모델 파라미터를 찾아 모델을 완성시키는 것

OUTTA

기본 Machine Learning 용어들

- **Regression (회귀) 란?**

; Regression model은 연속적인 값을 예측합니다.

ex) 서울시 집 **값** 예측, 광고를 클릭할 **확률**, 2달 동안 다이어트에 성공할 **인원 수** 등

- **Classification (분류) 란?**

; Classification model은 이산적인 값을 예측합니다.

ex) 주어진 메일이 스팸 메일인지 아닌지, 주어진 이미지의 동물이 개인지 고양이 인지 등

OUTTA

회귀 모델

Regression Model

AI

OUTTA

- **회귀 모델**
- 선형 모델
- 최소 제곱법
- 선형 회귀

Regression Model (회귀 모델 이란)?

- **Regression Model (회귀 모델) 이란?**

- Input 변수를 기반으로 Output 변수를 예측하거나 추정하는 방법 (Regression Analysis)
- 산술적 예측 (Numerical Prediction)을 생성하는 모델

  cf) Classification Model 은 Class prediction을 생성

- **Regression Model 종류**

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

OUTTA

선형 회귀 모델

OUTTA

회귀 분석

• 선형 모델

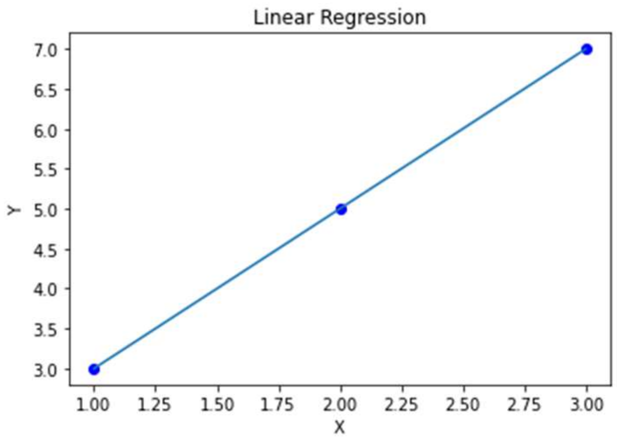
최소 제곱법

선형 회귀

Linear Regression

- Linear Regression (선형 회귀) 란?
  - 두 변수의 관계를 설명하는 선형 함수를 찾아내는 것

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



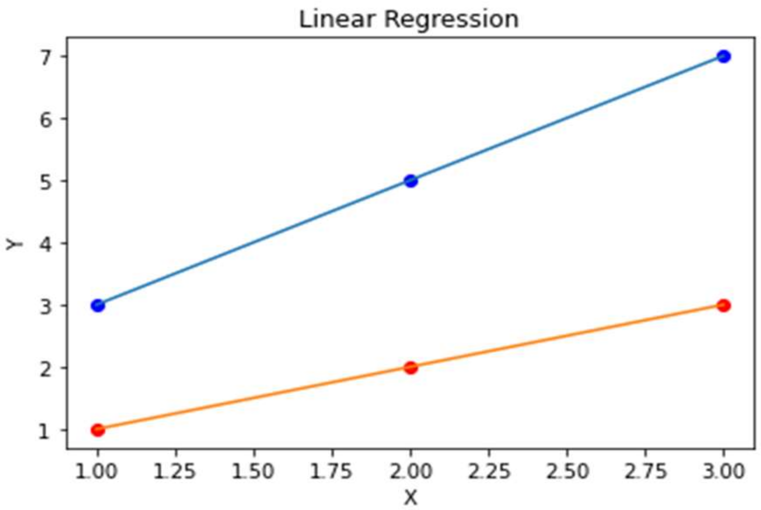
➔ H(W, b) = Wx + b (목표 : W = 2, b = 1)

OUTTA

Linear Regression

ex) [가설 초기화]

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



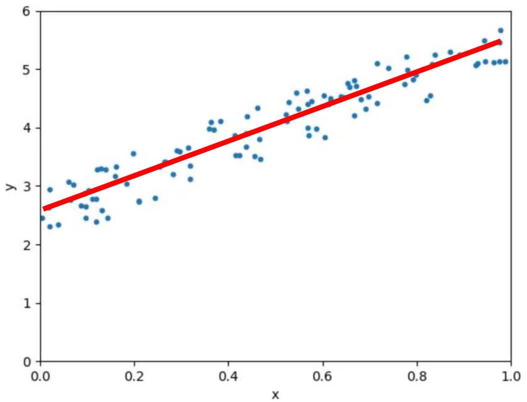
Linear Regression

- Linear Regression (선형 회귀) 란?

- 두 변수의 관계를 설명하는 선형 함수를 찾아내는 것

- 왜 "선형 회귀"인가? :

  ➔ 실제 데이터의 측정값에는 노이즈가 포함될 수 밖에 없고, 이런 노이즈들로부터 다시 원래의 선형 연속함수로 돌아가는 과정이기 때문.



Linear Model 세우기

- 모델) $\boldsymbol{F}(\boldsymbol{m}, \boldsymbol{b}; \boldsymbol{x}) = \boldsymbol{mx} + \boldsymbol{b}$
- 모델 파라미터) $\boldsymbol{m}$ $(\boldsymbol{slope})$, $\boldsymbol{b}$ $(\boldsymbol{intercept})$
- 모델 파라미터 결정) 최적화 ⇔ 최소제곱법

최소 제곱법

OUTTA

회귀 분석

선형 모델

• **최소 제곱법**

선형 회귀

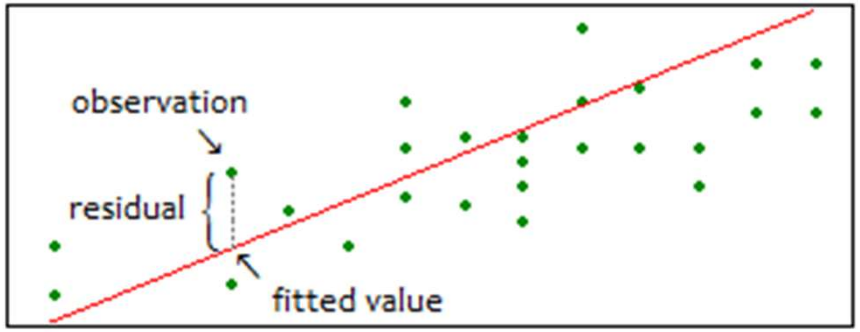
Least Square Method

- 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



OUTTA

회귀 분석

선형 모델

• **최소 제곱법**

선형 회귀

Residual sum of squares (잔차 제곱합)

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

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

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

OUTTA

Residual sum of squares (잔차 제곱합)

- 잔차 제곱합 **(RSS) :**

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

- **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.

Loss Function (손실 함수)
; Least Square method

- 손실함수

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

를 최소화하는 변수 **m, b**를 찾아야 하며

이러한 접근 방식을 최소제곱법이라 한다.

- $m^*, b^* = argmin_{m,b \in R}\, \mathcal{L}$

회귀 분석

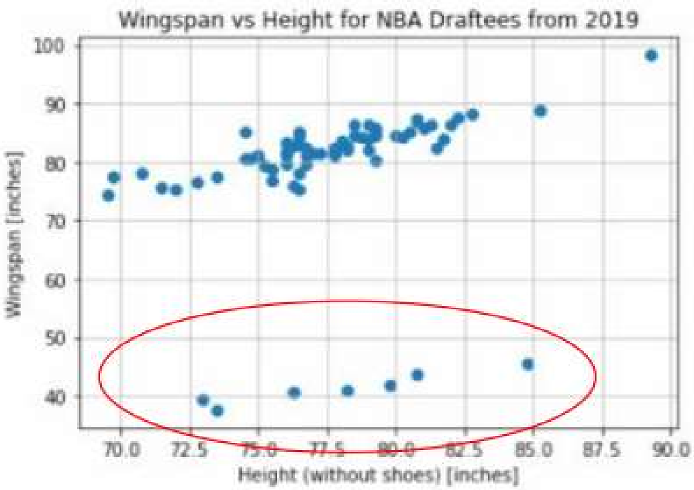
선형 모델

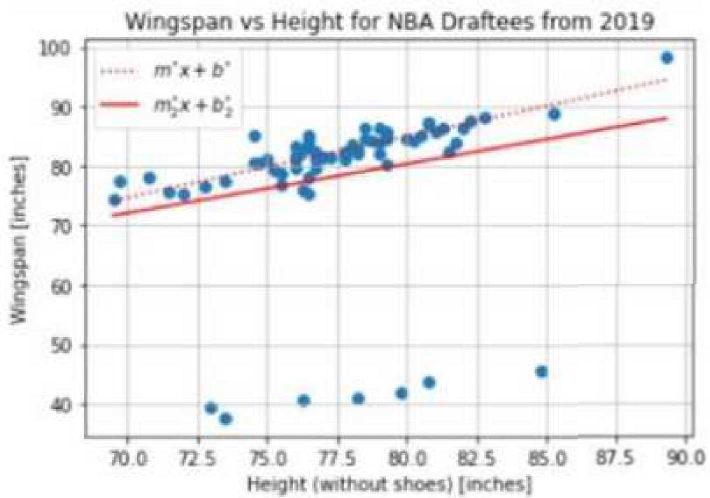
• **최소 제곱법**

선형 회귀

최소 제곱법의 한계

- **Outlier**가 많이 존재하는 데이터에서는 최소제곱법을 적용할 수 없다.
- **Dataset**의 **Data point** 개수 ≤ 1
- 모든 **Data point**가 같은 x_i 값을 가지는 경우





OUTTA

선형 회귀

OUTTA

회귀 분석

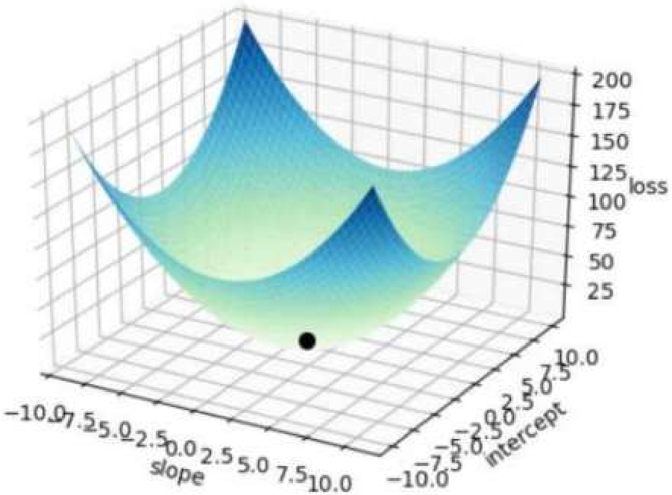
선형 모델

최소 제곱법

• **선형 회귀**

m^*, b^* 구하기

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



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

OUTTA

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, b에 대해 편미분한다.

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

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

m^*, b^* 구하기

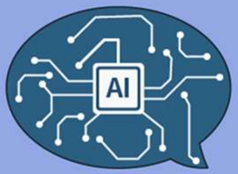
$$\frac{\partial \mathcal{L}(m,b)}{\partial m} = \sum_{n=0}^{N-1} (2mx_n^2 + 2bx_n - 2x_n y_n) = 0 \quad - \quad (1)$$

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

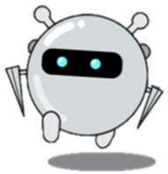
1과 2를 연립하면..

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

이렇게 구한 m^*, b^* 를 '최소제곱추정량' 이라고 한다.



감사합니다



OUTTA