10/1/2019

REGRESSION

5기 이세린

Growth Hackers

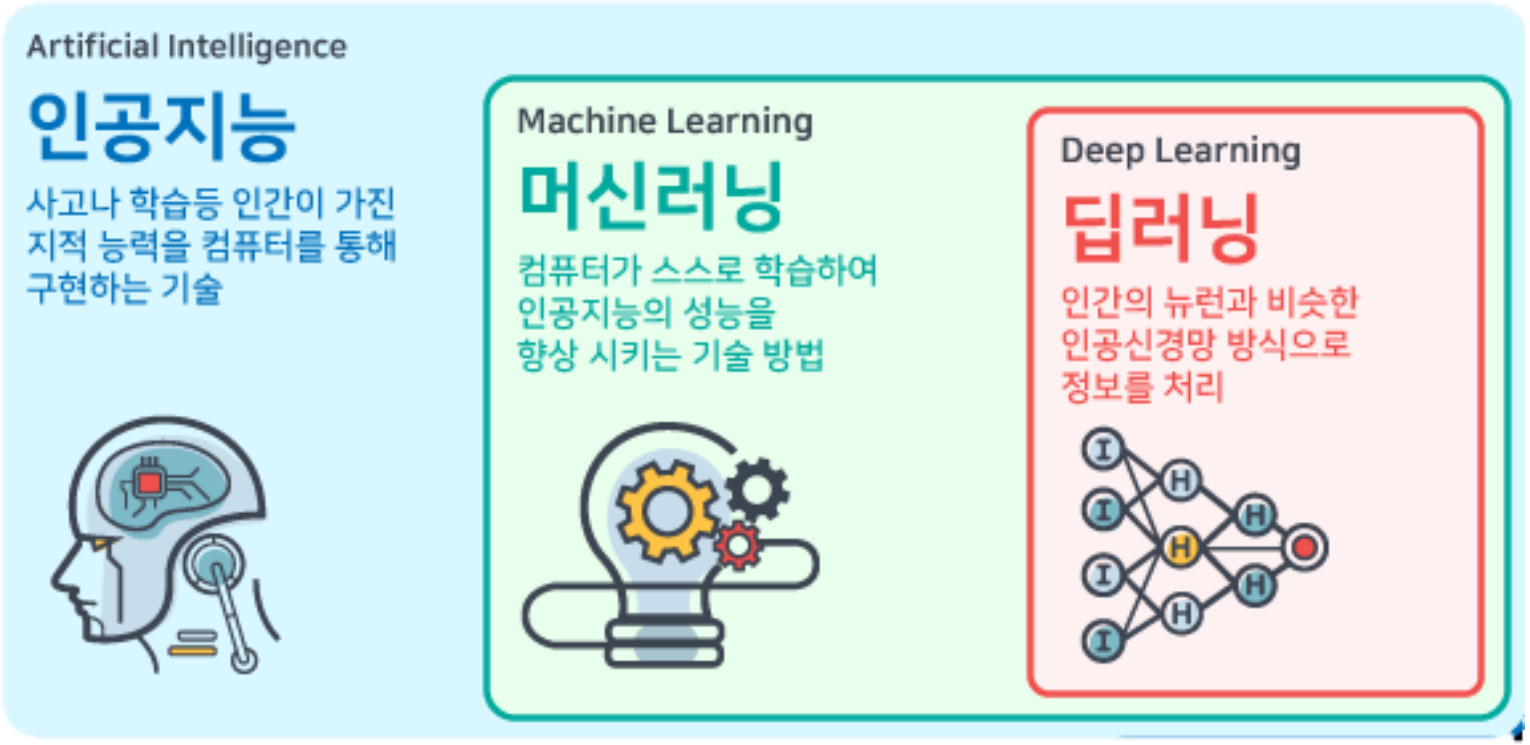
CONTENTS

10/1/2019

0. PREVIEW

Growth Hackers

기계학습(Machine Learning)



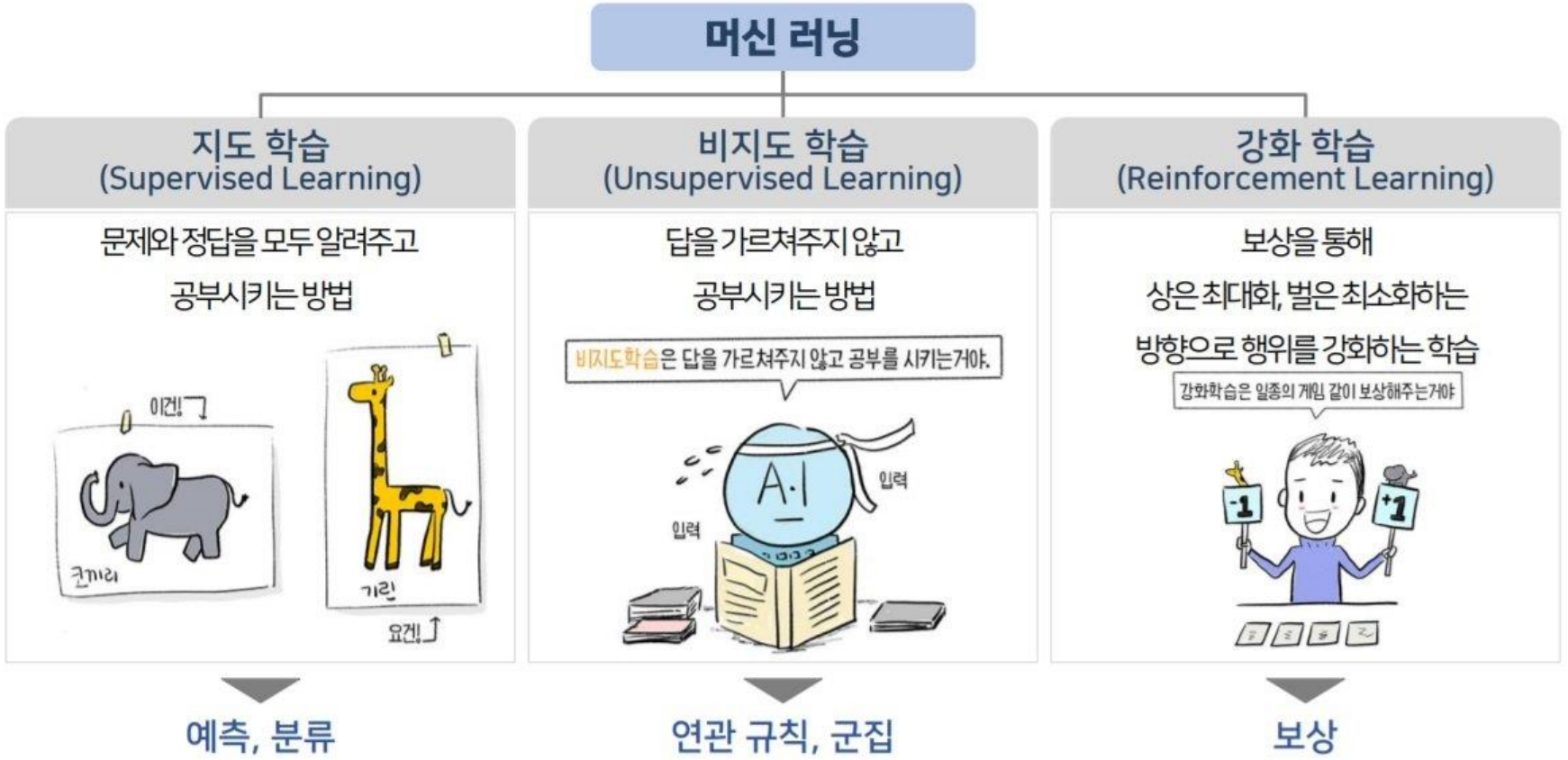
Growth Hackers

기계학습(Machine Learning)

"컴퓨터가 어떤 작업(T)를 하는데 있어서
경험(E)로부터 학습하여
성능에 대한 측정(P)을 향상시키는 학문" – Tom Mitchell

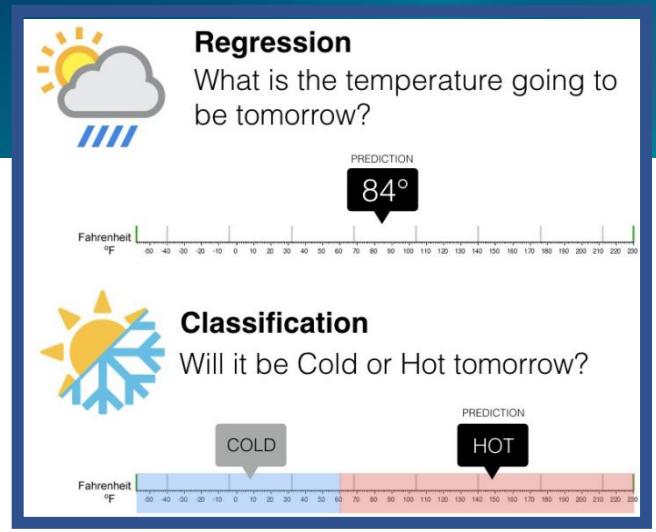
"머신러닝 알고리즘은 데이터를 기반으로 통계적인 신뢰도를 강화하고 예측 오류를 최소화하기 위한 다양한 수학적 기법을 적용해 데이터 내의 패턴을 스스로 인지하고 신뢰도 있는 예측 결과를 도출해 낸다."

지도학습(Supervised Learning)

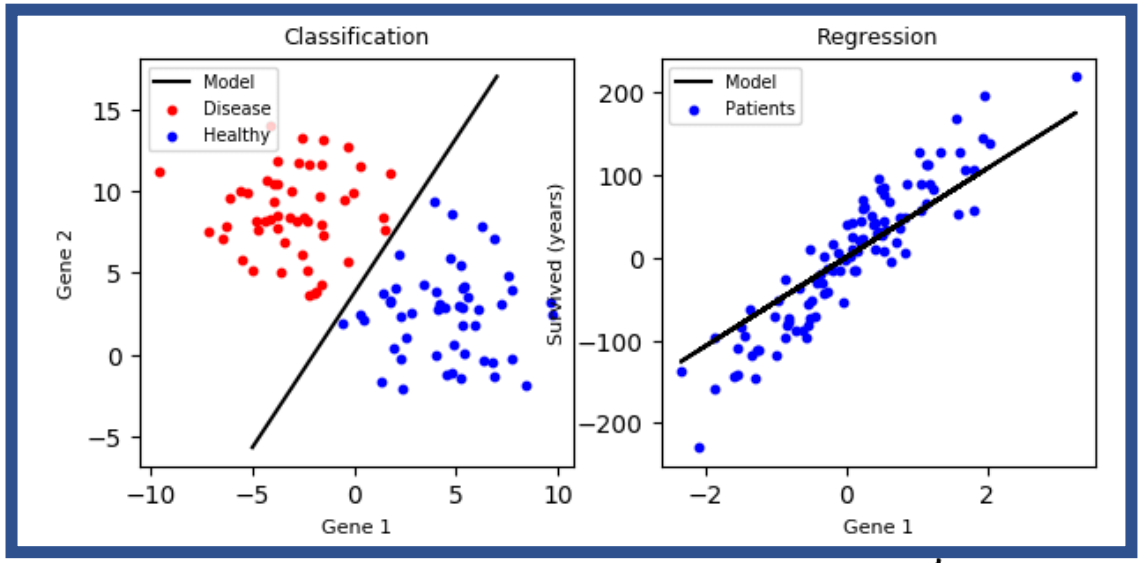


Growth Hackers

Regression VS Classification



- 기계학습 (Machine Learnig)
  - 지도학습 (Supervised Learning)
    - 레이블된 데이터로 학습
    - 미래 데이터 예측
    - y = f(x)
    - 분류 (classification)
      - MNIST 필기체 인식
      - 스팸 메일 분류
    - 회귀 (regression)
      - 주가 예측
  - 비지도학습 (Unsupervised Learning)
    - 레이블 없이 학습
    - 데이터의 숨겨진 구조/특징 발견
    - x ~ p(x), x = f(x)
    - 클러스터링 (clustering)
      - 마케팅 고객 그룹화
  - 강화학습 (Reinforcement Learning)
    - 보상 시스템으로 학습
    - 의사결정을 위한 최적의 액션 선택



Growth Hackers

회귀(Regression)

- 여러 개의 독립 변수와 한 개의 종속변수 간의 상관관계를 모델링하는 기법 통칭
- 머신러닝에서 회귀 예측의 핵심:

  주어진 독립변수 피처와 결정 값 데이터 기반에서 학습

→최적의 회귀 계수(Regression Coefficients)를 찾아내는 것.

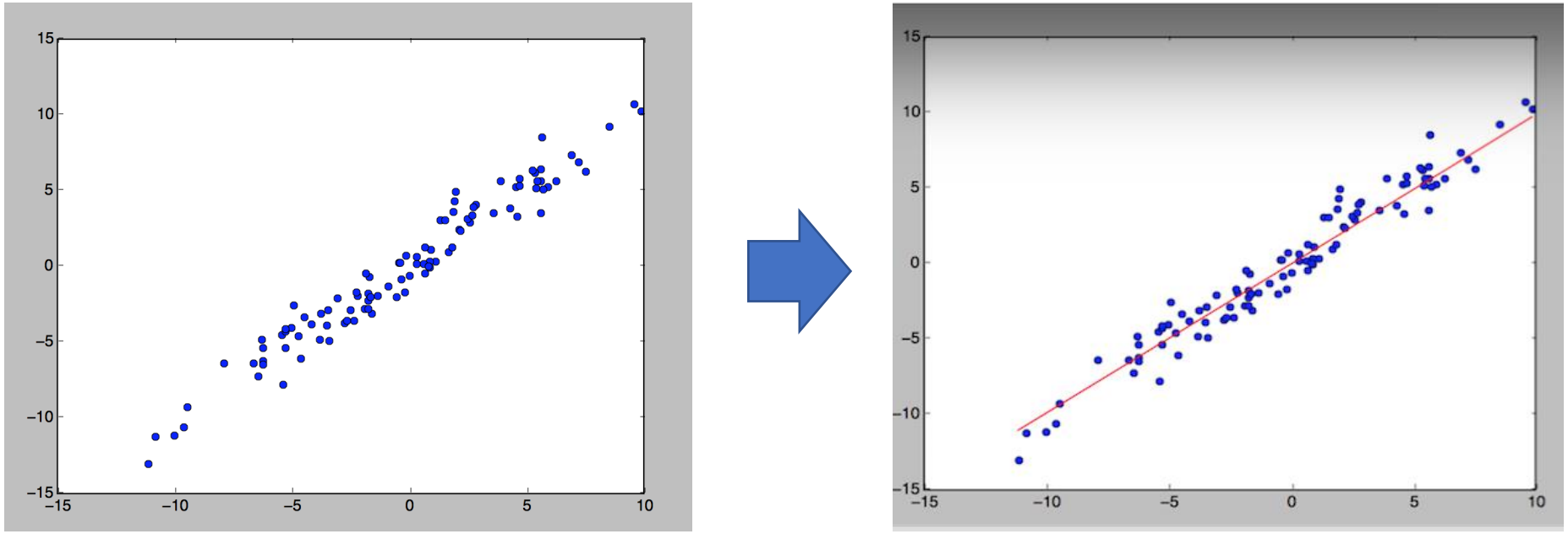
- 단일회귀 VS 다중회귀 / 선형회귀 VS 비선형회귀

10/1/2019

1. SIMPLE & MULTIPLE LINEAR REGRESSION

Growth Hackers

단순선형회귀(Simple Linear Regression)



How do we choose w_0, w_1 for $\hat{y} = w_0 + w_1 x$?

Growth Hackers

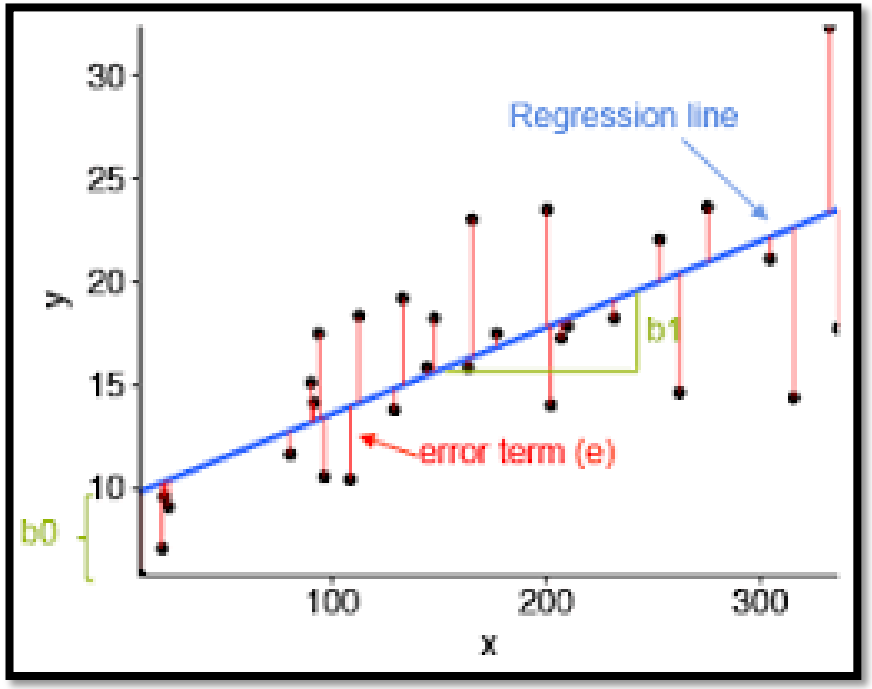
단순선형회귀(Simple Linear Regression)

1. Normal Equation

$$Y_i = w_0 + w_1 X_i + \varepsilon_i$$

$\frac{1}{n}\sum_{i=1}^{n} {\varepsilon_i}^2$ 을 최소화 하는 w_0, w_1 를 찾자!

$$\underset{w_0, w_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - (w_0 + w_1 X_i)\}^2$$



→ 미분값 = 0 일때 최소, 고차원 방정식 이용해 w_0, w_1 도출

Growth Hackers

단순선형회귀(Simple Linear Regression)

2. 경사하강법(Gradient Descent)

- W 파라미터의 개수가 많은 경우, Normal Equation으로 해결이 어려움
- '점진적으로' 반복적인 계산을 통해 W 파라미터 값을 업데이트 하며 오류 값이 최소가 되는 W 파라미터를 구하는 방식.

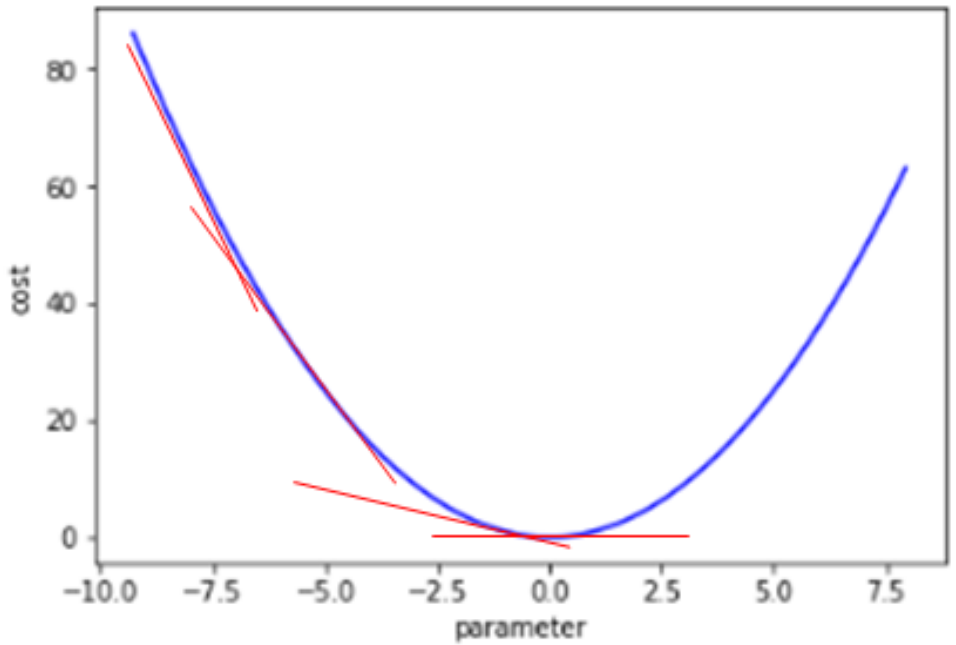
$$W := W - \alpha \frac{\partial}{\partial W} cost(W)$$

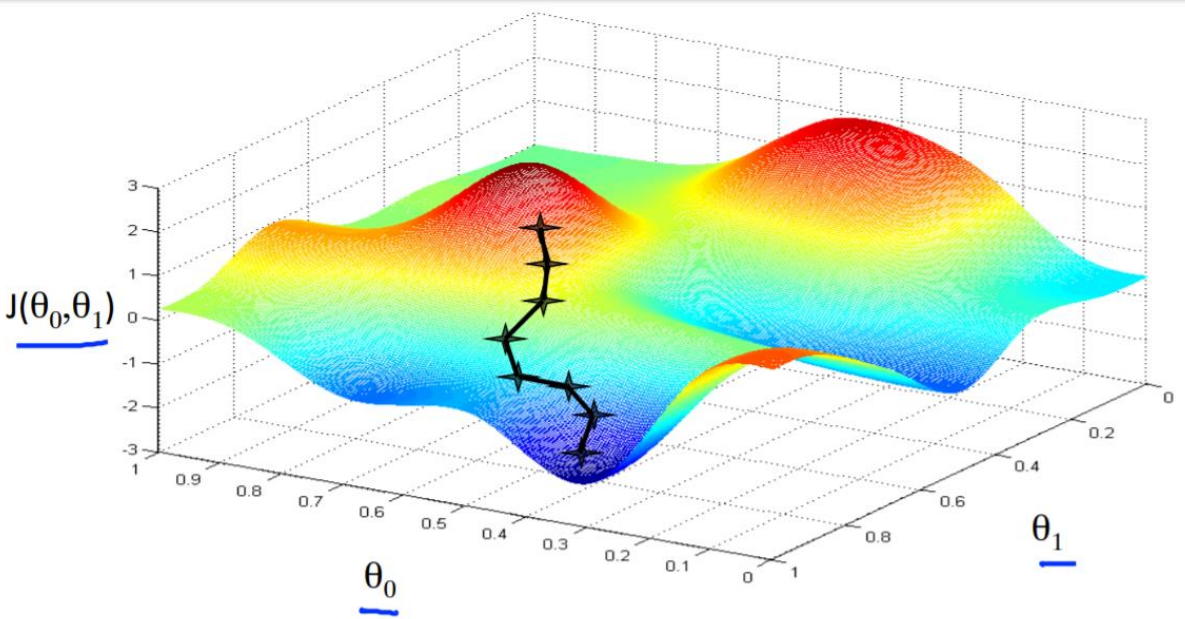
$$(cost(W) = \frac{1}{2n} \sum_{i=1}^{n} (Wx^{(i)} - y^{(i)})^2)$$

- Local minimum(O) Global minimum(X)

단순선형회귀(Simple Linear Regression)

2. 경사하강법(Gradient Descent)

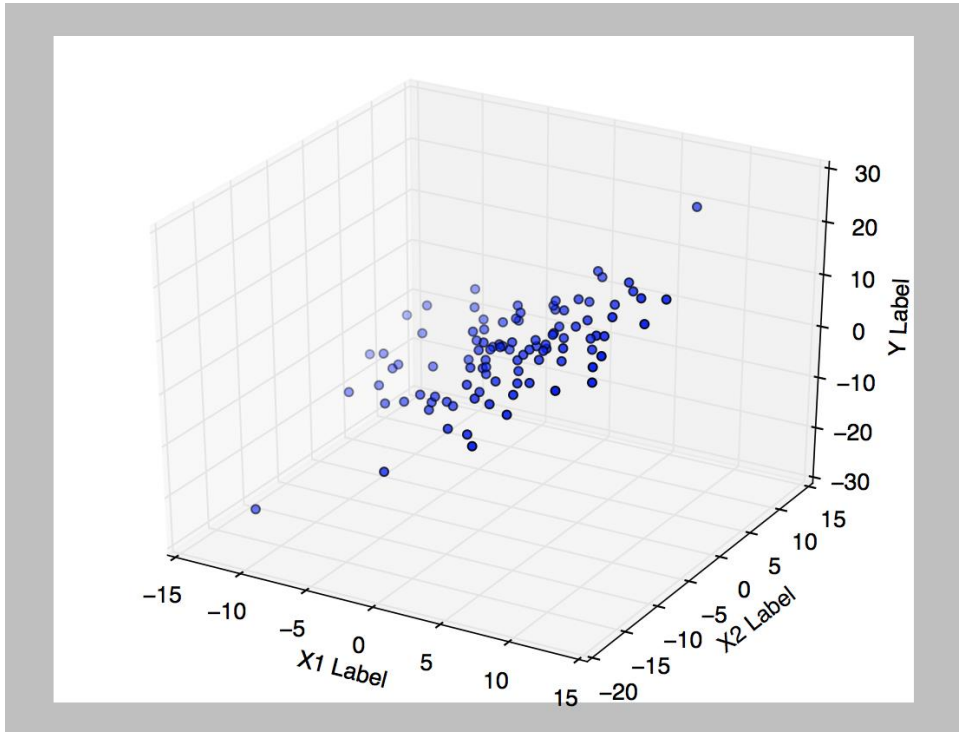


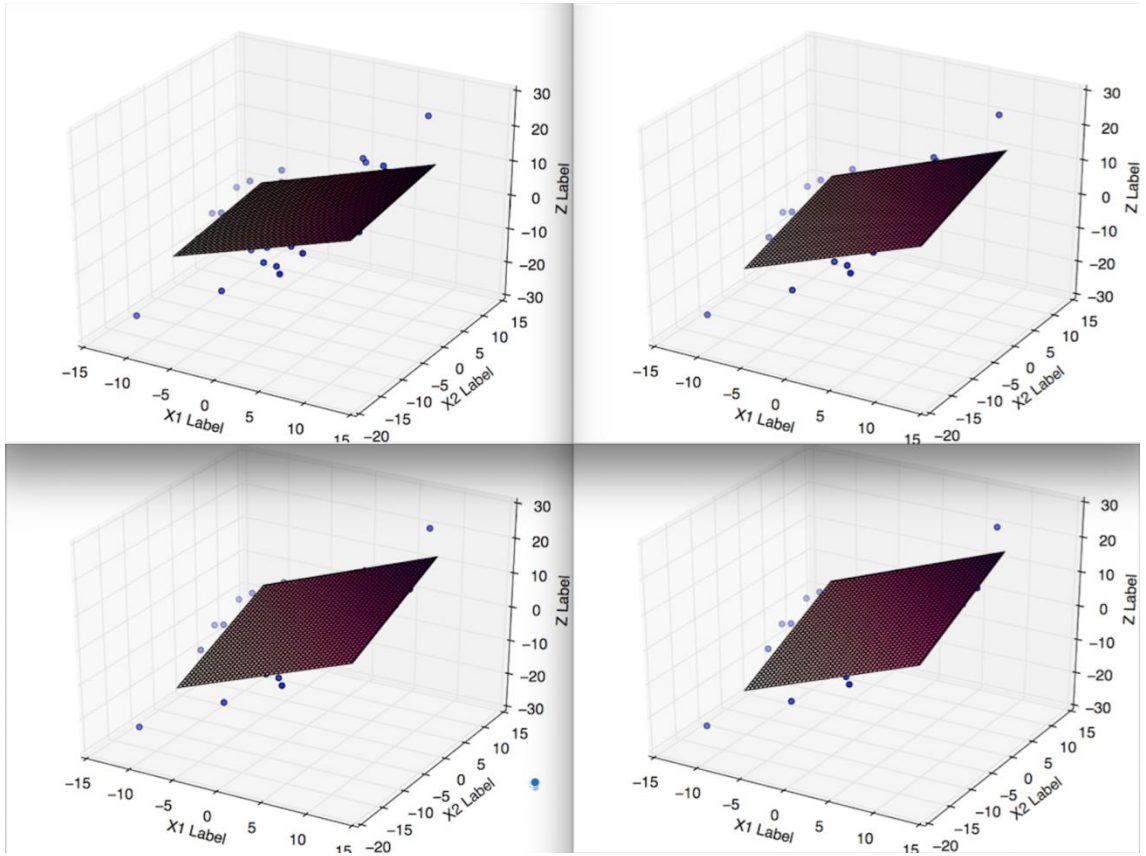


Growth Hackers

다중선형회귀

'독립변수가 하나가 아니라면?'
'변수가 추가되면 더 좋은 모델을 만들 수 있지 않을까?'





Growth Hackers

다중선형회귀

- $Y = XW + \epsilon$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y_i = W_0 + W_1 x_i + \varepsilon_i$$

<단순선형회귀>

- $Y = XW + \epsilon$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} W_0 \\ \vdots \\ W_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y_i = W_0 + W_1 x_1 + \cdots + W_n x_n + \varepsilon_i$$

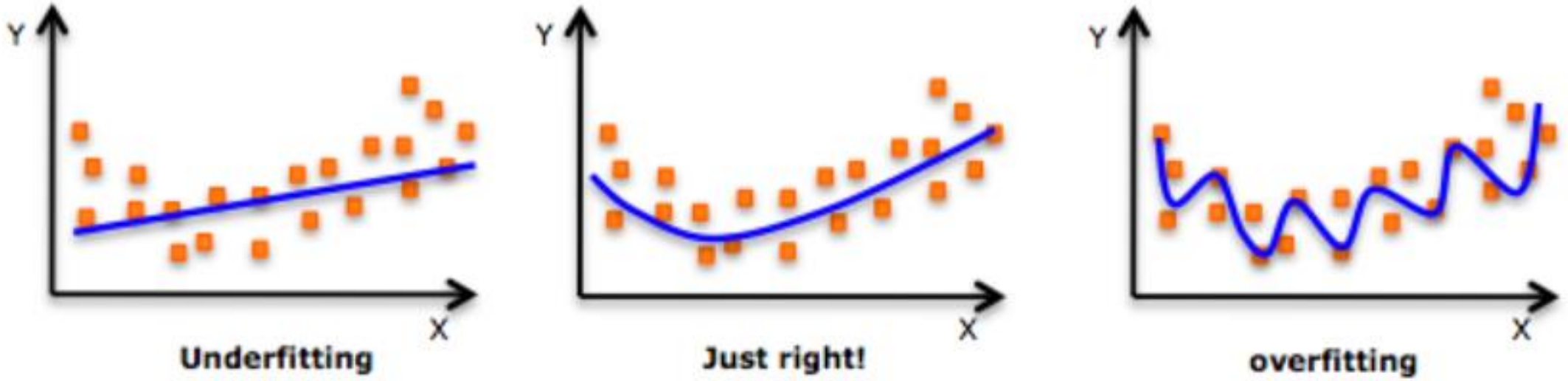
<다중선형회귀>

Same Form!

Growth Hackers

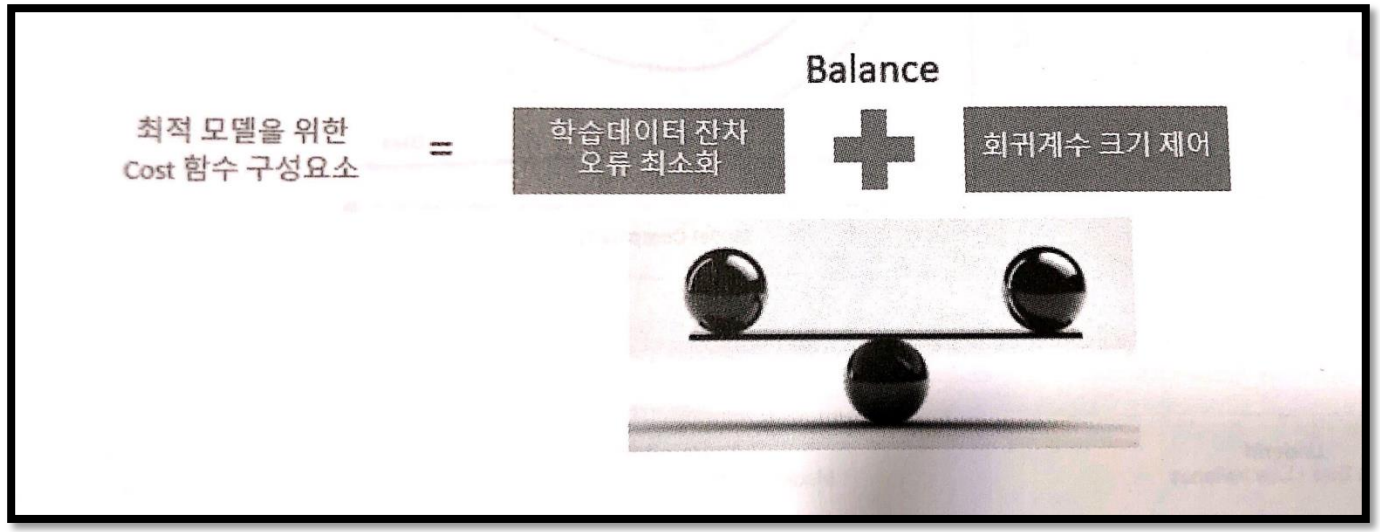
2. REGRESSION UNDER REGULATION

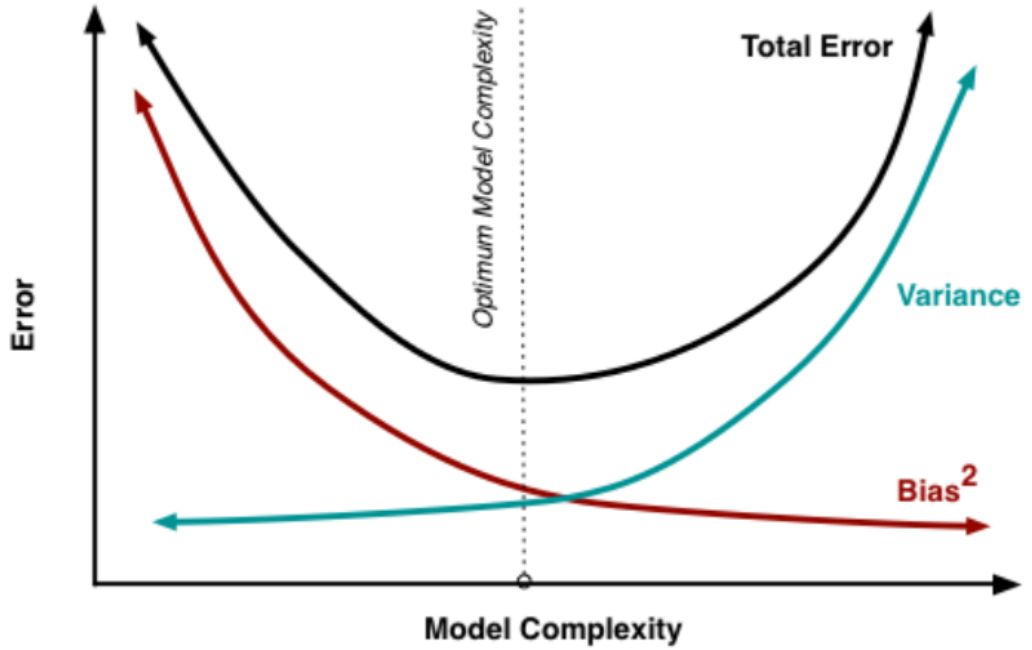
과적합/과소적합 이해



Growth Hackers

최적 모델을 위한 cost함수 구성요소





모델이 복잡해진다? → 모델의 Bias 감소/ 모델의 Variance 증가

모델이 단순해진다? → 모델의 Variance 감소/ 모델의 Bias 증가

Growth Hackers

릿지 회귀(Ridge Regression)

$$\text{cost(W)} = \frac{1}{n}\sum_{i=1}^{n}(Wx^{(i)} - y^{(i)})^2 + \alpha\sum_{i=1}^{n} {w_i}^2$$

→ (weight)^2 을 가능한 한 0에 가깝게 만든다.

$\alpha \uparrow \rightarrow$ 패널티 $\uparrow$ $\rightarrow$ weight $\downarrow$

$\alpha \downarrow \rightarrow$ 패널티 $\downarrow$ $\rightarrow$ weight $\uparrow$

라쏘 회귀(Lasso Regression)

$$\text{cost(W)} = \frac{1}{n}\sum_{i=1}^{n}(Wx^{(i)} - y^{(i)})^2 + \alpha\sum_{i=1}^{n}|w_i|$$

→ |weight|을 가능한 한 0에 가깝게 만든다.

$\alpha \uparrow \rightarrow$ 패널티 ↑ → weight ↓

$\alpha \downarrow \rightarrow$ 패널티 ↓ → weight ↑

3. LOGISTIC REGRESSION

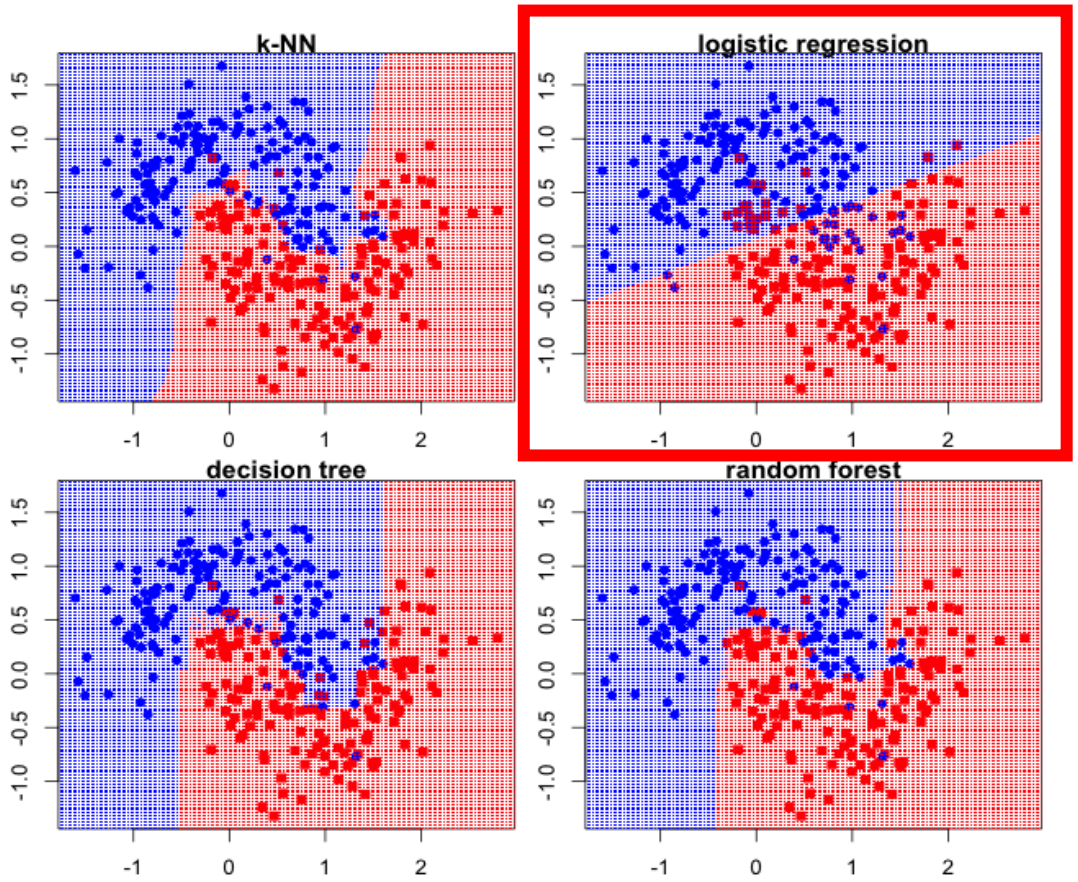
로지스틱 회귀

- 회귀
  - 1. 단순회귀
  - 2. 다중회귀
  - 3. 릿지회귀
  - 4. 라쏘회귀

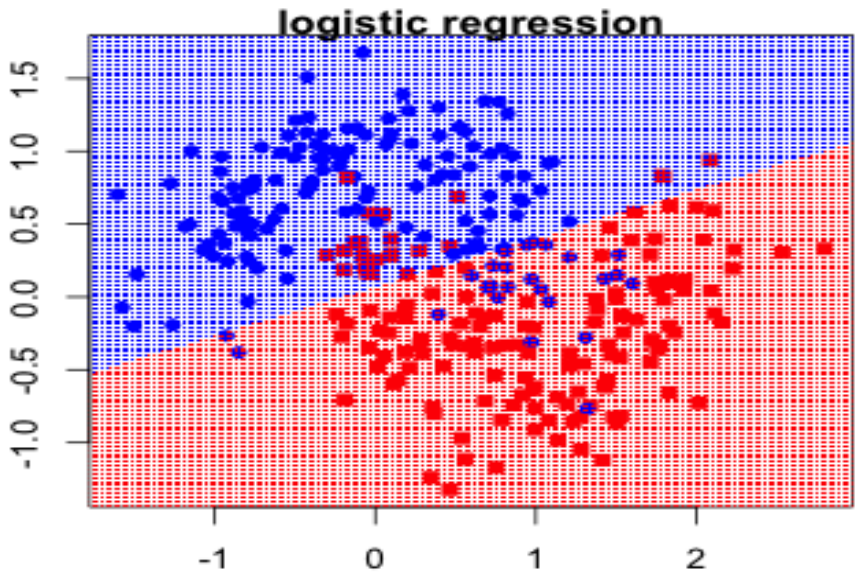
Target이 Quantitative

- 로지스틱회귀 — Target이 Qualitative
  - 선형회귀 방식을 분류(classification) 에 적용한 알고리즘.
  - 이진 분류에 사용 ex) Spam E-mail detection: Spam or Ham
  - 분류 알고리즘들 중 굉장히 정확도가 높은 알고리즘으로 알려져 있음

로지스틱 회귀

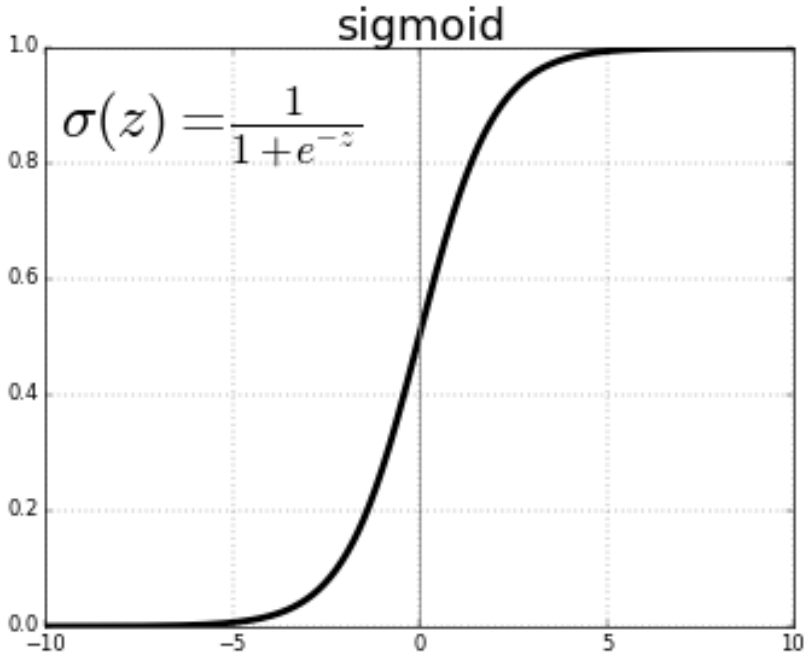


- 결정 경계(Decision Boundary)
  - 두 클래스의 영역을 나누는 경계



Growth Hackers

로지스틱 회귀



- Sigmoid Function
  (= Logistic Function)

특징: 무슨 일이 있어도 출력값이 0과 1 사이에 놓인다.

→ 회귀식을 Sigmoid Function에 대입하면,
출력 값이 반드시 0과 1 사이의 값이 나옴

Growth Hackers

로지스틱 회귀

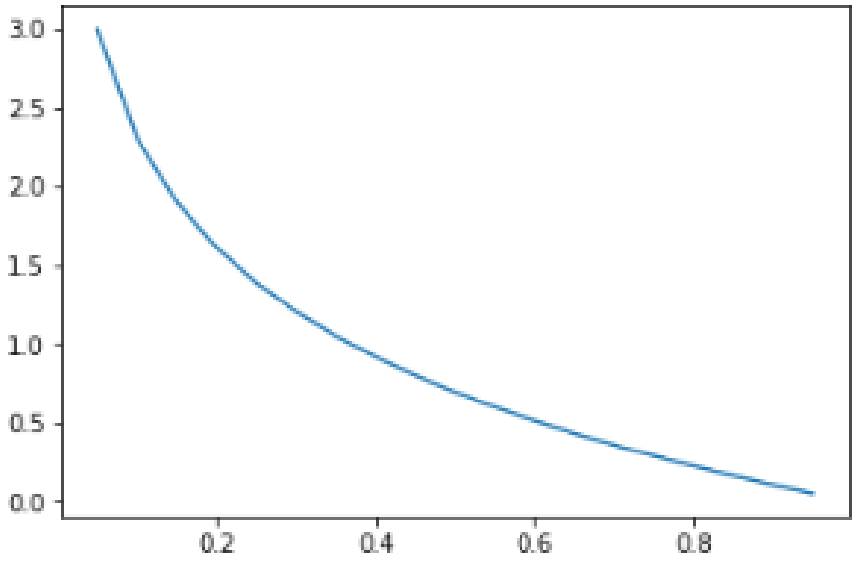
경사하강법(gradient descent) 적용을 위해 cost 함수 변환

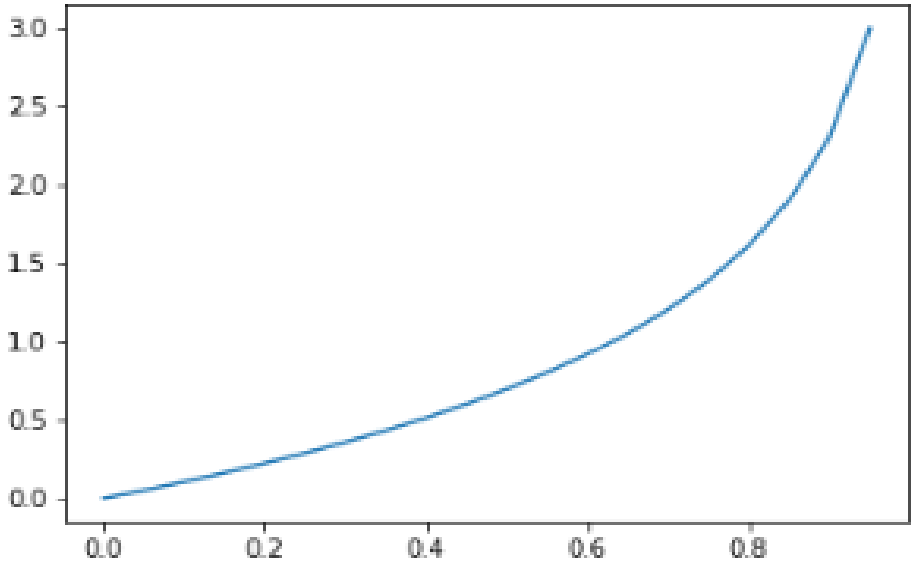
- cost(W) $= \frac{1}{n}\sum_{i=1}^{n} c(H(x), y)$

- c(H(x), y) $= \begin{cases} -\log\left(H(x)\right), & if\ y = 1 \\ -\log\left(1 - H(x)\right), & if\ y = 0 \end{cases}$

로지스틱 회귀

- $c(H(x), y) = \begin{cases} -\log(H(x)) , & if\ y = 1 \\ -\log(1 - H(x)) , & if\ y = 0 \end{cases}$

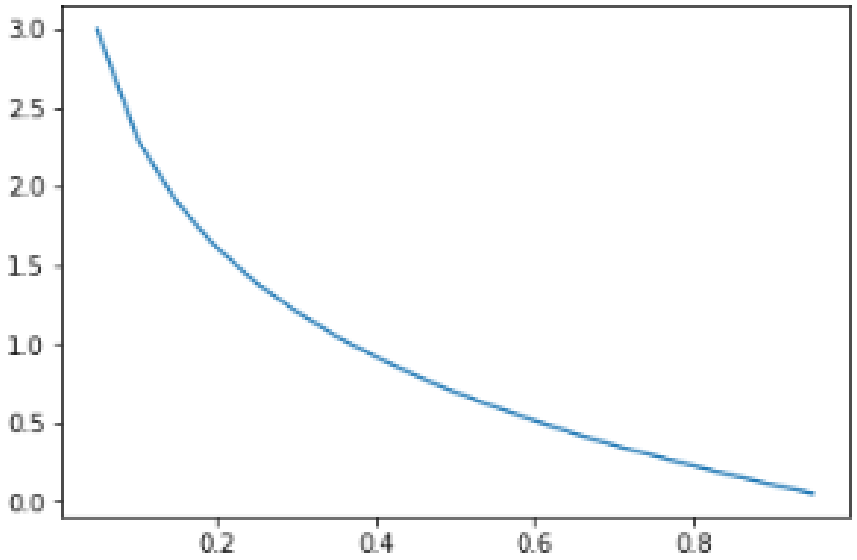


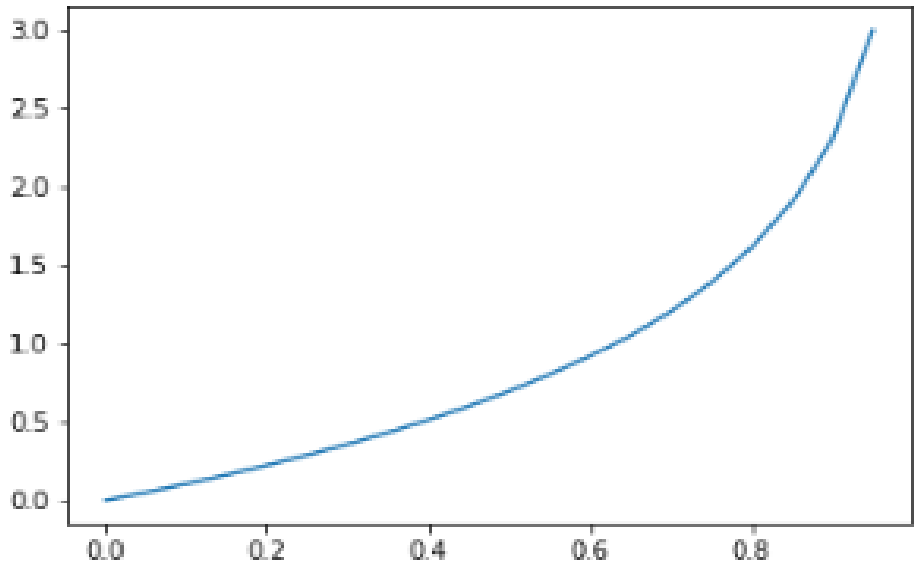


Growth Hackers

로지스틱 회귀

- cost(W) = $-\frac{1}{n}\sum_i^n ylog\big(H(x)\big) + (1-y)\log(1-H(x))$





- W := W - $\alpha\frac{\partial}{\partial W}cost(W)$

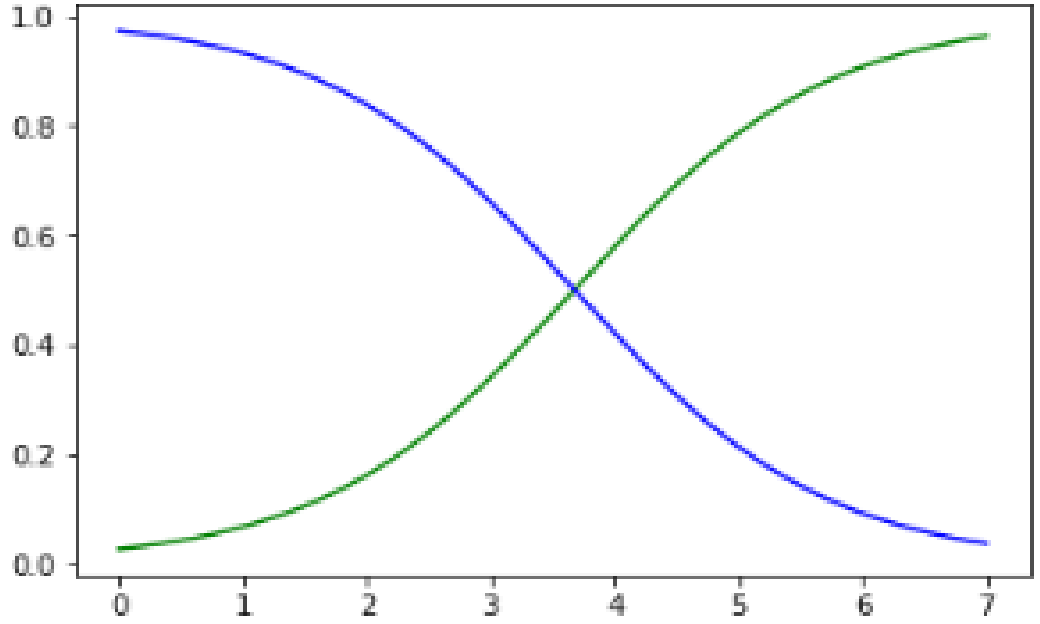
Growth Hackers

로지스틱 회귀

- 고전적인 Iris data

- 꽃받침, 꽃잎의 길이와 너비를 바탕으로 꽃잎의 종을 예측 (분류)

로지스틱 회귀



- 목적:
  '꽃받침 길이를 기준으로
  Sentosa/Not Sentosa를 분류하고 싶다.'

- 결과 해석:
  꽃받침 길이가 2이면 Not Sentosa
  꽃받침 길이가 4.5이면 Sentosa

Growth Hackers

Quest

- ‘breast cancer’ 데이터 셋을 사용

Radius 변수를 기준으로 breast cancer 양성/음성을 분류하는 로지스틱 회귀분석 모델을 만들고, 이를 시각화하고, Radius 길이가 20, 0.1일 때의 결과를 해석해주세요.

- 파일 불러오기:

```
from sklearn.datasets import load_breast_cancer
```