Python Study

Environ. Analysts

<환경을 분석할 줄 아는 사람이 되자!>

2014890021 Theo Kim

Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

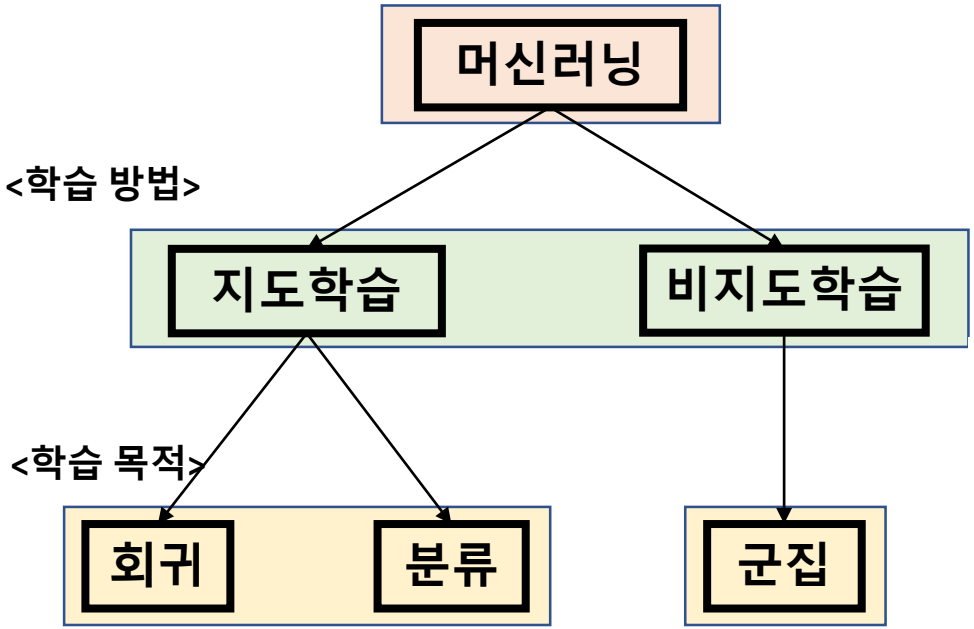
- 머신러닝(Machine Learning)?
- 회귀 분석
- 회귀 분석 실습

Type2 머신 러닝

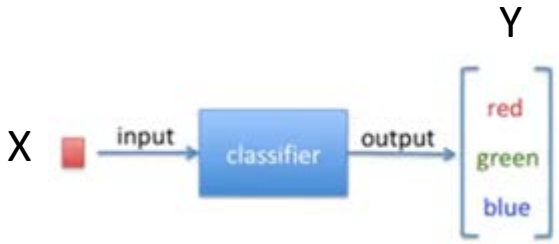
<환경을 분석할 줄 아는 사람이 되자!>

머신러닝(Machine Learning)?

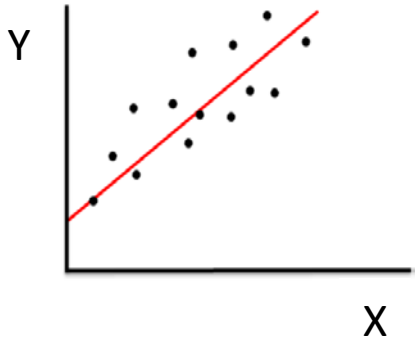
- Machine Learning



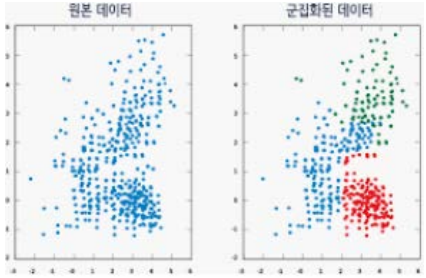
- Classification(이산형)



- Regression(연속형)



- Clustering



- 머신러닝은 기계를 학습시키는 것, 데이터의 숨겨진 패턴을 알아내는 것

- 지도 학습 : 학습을 위해 Y(정답, 종속변수), X(조건, 독립변수)이 둘다 필요함.
- 비지도 학습 : 학습을 위해 X(조건, 독립변수)만 필요함.

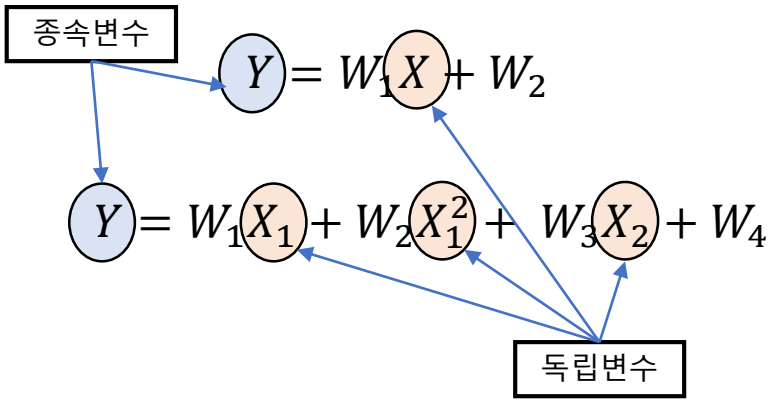
Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

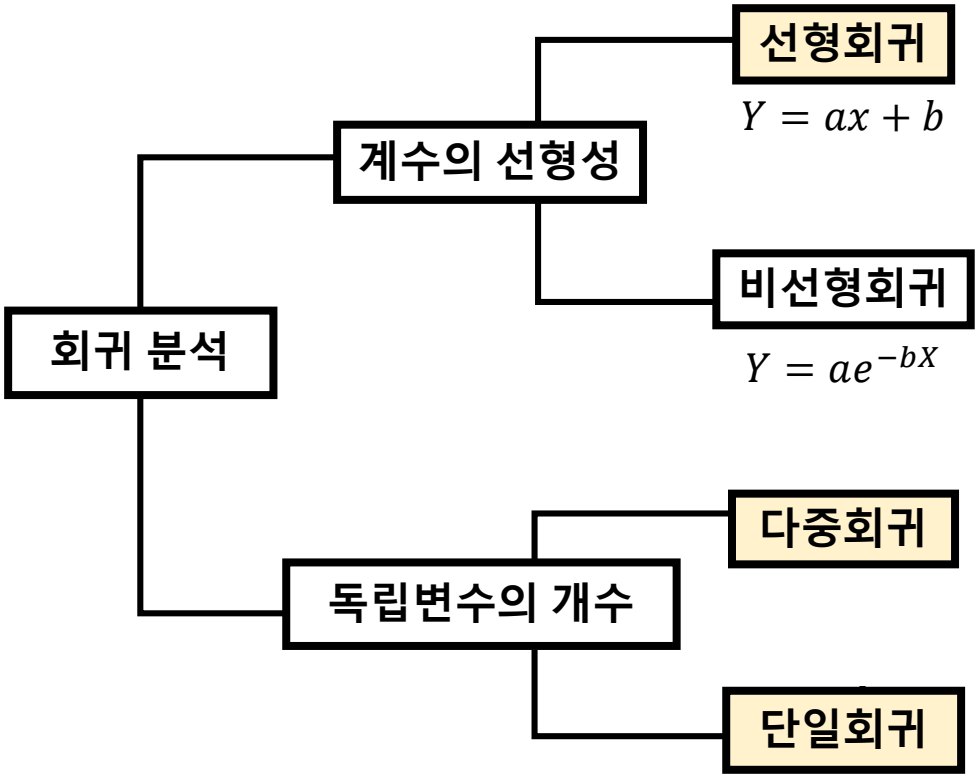
회귀 분석

- What is Regression?

여러 개의 독립변수와 한 개의 종속변수 간의 상관관계를 모델링하는 기법



- 회귀분석의 계통



Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석

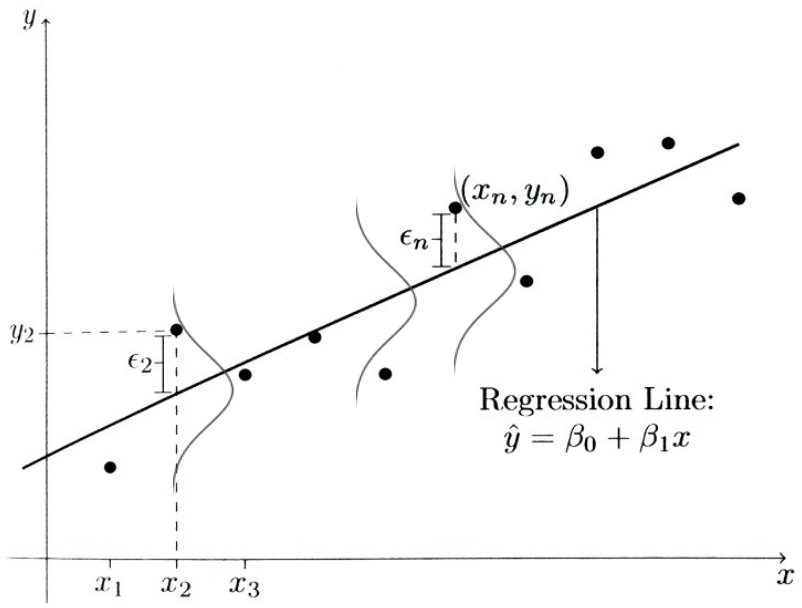
- 단순 선형 회귀 분석

$$\hat{Y}_{regression} = W_1 X + W_2$$

$$Y_{true} = W_1 X + W_2 + error$$

변수는 회귀 계수이다!
즉, W_1, W_2 가 변수 이다.

물음 : 어떻게 W_1, W_2를 결정해 나갈 것인가



예) 호수 내 **녹조의 농도와 TP와의 상관관계가 있다고 생각**했다. 대충 산포도나 박스 플랏을 그리다 보니 이 관계가 선형적인 관계라고 판단하게 되었다. 배를 타고 나가 종속변수인 녹조의 농도와 독립변수인 TP를 측정하였다. 이 실험을 통한 자료의 일부를 단순 **선형 회귀 분석을 통해 훈련 시키고 훈련시킨 후 나온 회귀 계수**를 통해 실험을 통한 자료 일부에 잘 맞는지 검증을 해보았다.

Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석

- 회귀계수 최적화 방법

$$\hat{Y}_{regression} = W_1 X + W_2$$

$$\hat{W}$$

회귀 계수들의 최적화 <Optimization>

Option1. 정규방정식

수학적으로 풀이가 가능함. But Parameter가 많으면 시간이 오래 걸리고 데이터의 완결성이 떨어지는 경우는 성립이 안되는 경우도 많음.

일반식 : $\hat{W} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$

Option2. 머신 러닝

엄청나게 여러 번 계수들을 갱신하면서 시행착오를 통해 오차를 줄이는 최적의 회귀 계수 조합을 찾게 됨. 머신러닝 결과 결국엔 정규방정식의 식으로 수렴하게 된다.

Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석

- 회귀 분석 성능평가를 위한 목적함수

목적함수(Objective Fuction)

비용함수 (Cost Function)
손실함수(Loss Function)

- **MAE(Mean Absolute Error)**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

- **MSE(Mean Squared Error)**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- **R^2(R-Squared)**

$$R^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- **RMSE(Root Mean Squared Error)**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

- 회귀식의 성능을 평가하는 것은 목적함수로 오차의 경우 0에 가까워 질수록 좋고 R^2의 경우 1에 가까워질수록 '모델이 잘 맞다'라고 판단 가능하다.

Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석

- 휘귀 계수 최적화를 위한 알고리즘(경사하강법, Gradient Descent, GD)

목적함수(R) : MSE

$$\hat{Y}_i = W_1 X_i + W_2$$

$$\frac{\partial R}{\partial W_1} = \frac{\partial \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\partial W_1} = \frac{2}{n}\sum_{i=1}^{n}(-xi * (yi - (W1xi + W2))$$

$$\frac{\partial R}{\partial W_2} = \frac{\partial \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\partial W_2} = \frac{2}{n}\sum_{i=1}^{n}(yi - (W1xi + W2))$$

$$W_1' = W_1 - \eta \frac{\partial R}{\partial W_1}$$

$$W_2' = W_2 - \eta \frac{\partial R}{\partial W_2}$$

갱신값(Renewal Value)

학습방향(Direction)

초기값(Initial Value)

학습률(Learning Rate)

<다변수에 일반화된 식>

$$W^{next} = W - \eta \nabla_w R(W)$$

Type2 머신 러닝

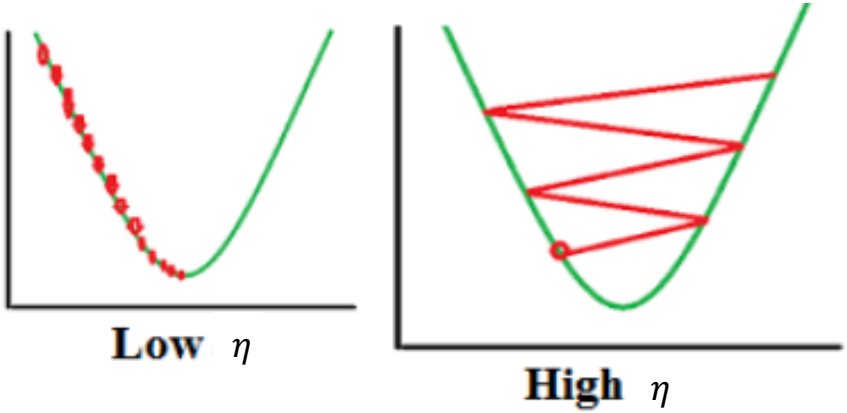
<환경을 분석할 줄 아는 사람이 되자!>

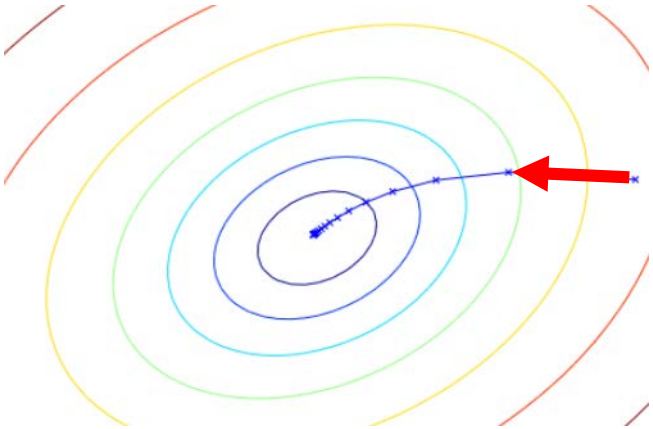
회귀 분석

<다변수에 일반화된 식>

$$W^{next} = W - \eta \nabla_w R(W)$$

- Learning Rate
- Direction





$W^{next} = W - \eta \nabla_w R(W)$

어느정도로 학습을 진행시킬 것인가

Learning Rate = 0.01, 0.001, 0.0005등

$W^{next} = W - \eta \nabla_w R(W)$

목적함수에 대한 각 회귀계수의 미분값의 벡터 방향

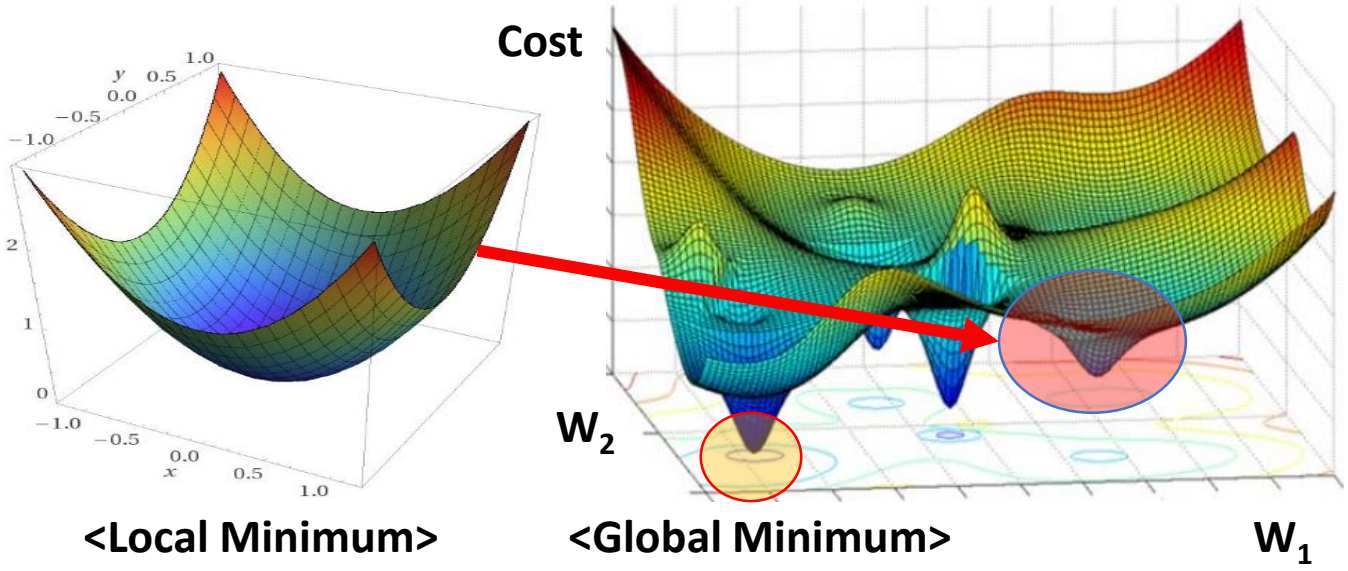
Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석

- 경사하강법의 문제

<W1, W2에 대한 비용함수 곡선>



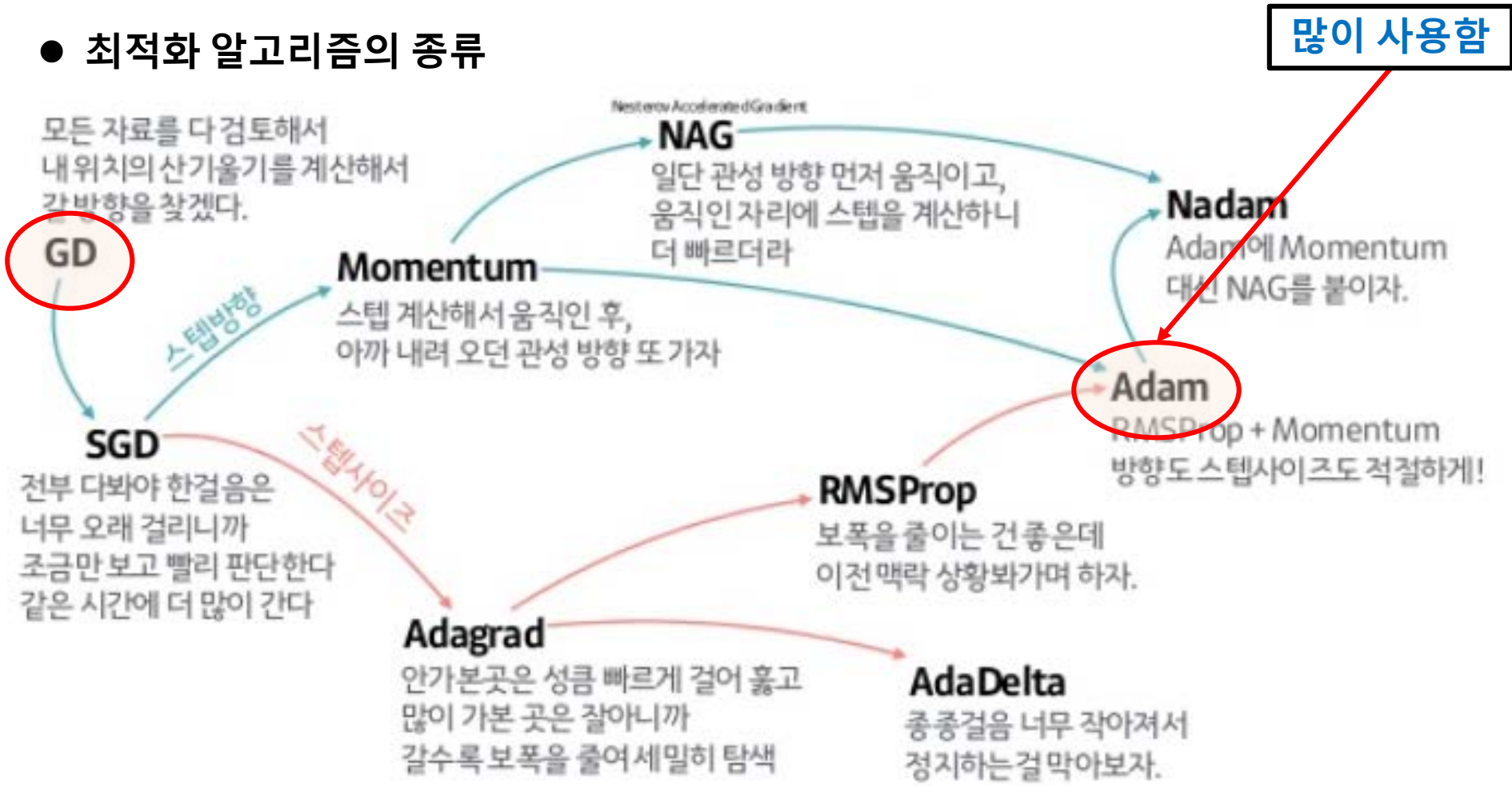
- 이 문제를 개선하기 위한 여러가지 개선된 알고리즘이 존재함.

Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석

- 최적화 알고리즘의 종류



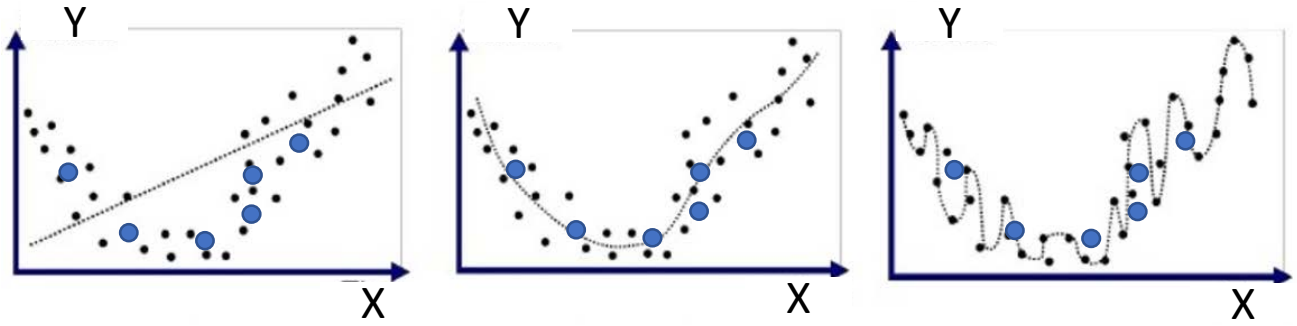
https://sacko.tistory.com/42

Type2 머신 러닝

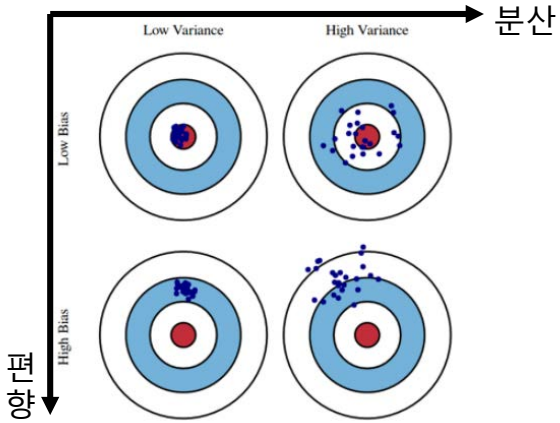
<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석

- 과적합 & 과소 적합 : 예측 시 잘 맞는 다항회귀식의 최대 차수를 정해야 한다.



- 분산-편향 트레이드 오프



- 아래의 원은 Train데이터 통해 만든 최적화된 도형이며 점은 Test데이터임.
- Low Variance(분산), Low Bias(편향)을 가지게 하는 모델을 만드는 것이 좋다.
- Underfitted는 고편향, Overfitted는 고분산으로 중간에 잘 Fitted된 모델을 찾아야 한다.

Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석

- 라쏘 회귀, 릿지 회귀, 엘라스틱넷 회귀

목적함수(Objective Fuction)
비용함수 (Cost Function)
손실함수(Loss Function)
= **학습데이터 잔차 오류 최소화** + **회귀계수의 크기 제어**

라쏘회귀(Lasso) $R(W) = MSE(W) + \alpha\|W\|_1$

릿지회귀(Ridge) $R(W) = MSE(W) + \alpha\|W\|_2{}^2$

엘라스틱넷회귀(ElasticNet) $R(W) = MSE(W) + r\alpha\|W\|_1 + \frac{1-r}{2}\alpha\|W\|_2{}^2$

- α, r 은 직접 사용자가 설정해 주어야 하는 **Hyperparameter**이다. 앞서 Learning rate도 사용자가 지정해 주어야 한다.

Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석

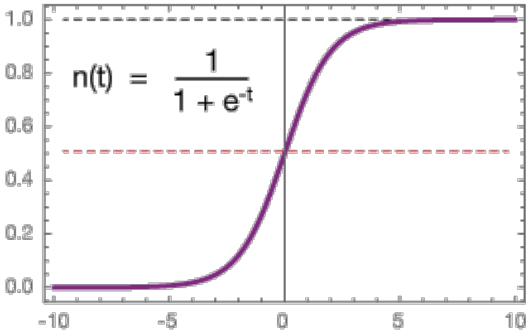
- **로지스틱 회귀분석**
  - 샘플이 특정 클래스에 속할 확률을 추정하는데 널리 사용됨.
  - 그 샘플이 해당 클래스에 속할 때, '1'로 표현하고 그렇지 않을 때 '0'으로 표현한다. -> 분류를 위한 회귀 분석이라고 생각하면 편함.

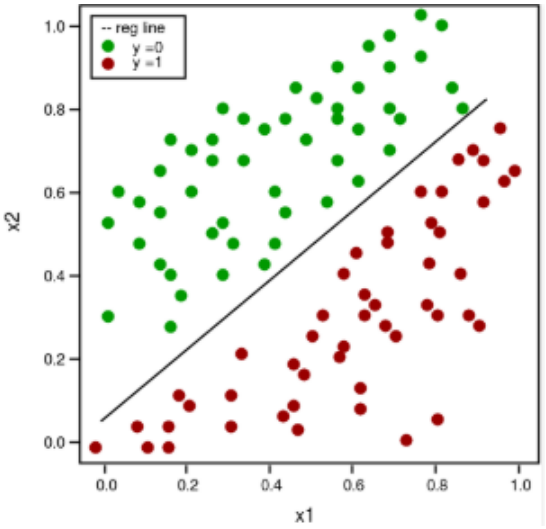
n(t) = $\frac{1}{1+e^{-t}}$ (Sigmoid 함수)

$\hat{P} = n(W_1 X_1 + W_2)$ (0~1확률로 반환)

$$\hat{y}_{logistic} = \begin{cases} 1\ (\hat{P} < 0.5) \\ 0\ (\hat{P} \geq 0.5) \end{cases}$$

$$R(W) = \frac{1}{m}\sum_{i=1}^{m}\left[Y_i log\left(\hat{P}_i\right) + (1 - Y_i)\log(1 - \hat{P}_i)\right]$$





두가지 독립변수를 통해 Y분류 예측

길이	너비	Y
12.3	100	0(소나무)
15	121	1(소나무아님)

Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

회귀 분석 실습

- Python 실습!

  머신러닝 분석 툴 – 사이킷런 다운로드
  pip install scikit-learn

Type2 머신 러닝

<환경을 분석할 줄 아는 사람이 되자!>

보충 자료

- 회귀 분석의 가정

1. 선형성 - X, Y가 직선 처럼 보여야 한다.
2. 비상관성 – 다중공선성을 제거해주어야한다. 적절한 독립변수의 개수를 파악해야한다.
3. 정상성- 오차가 정규분포?
4. 등분산성 – 분산의 정도가 비슷
5. 독립성 – 자기상관성

- 데이터 전처리시

- 정규성 검증 해주어야 함, Shapiro test, Kolmogrov-Sminrnov test, Bartlett test 등
- Log-transform
- Root
- 이상치 제거 등 해결해야함.