교육 일지

**2021.10.13 Daily Assignment**

일시: 2021-10-13

장소: YGL 자택

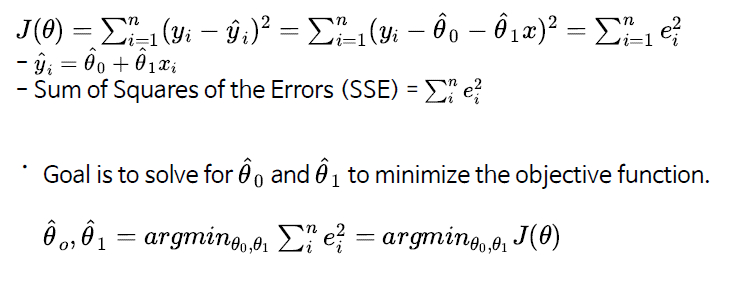
작성자: 강명훈

**오전 교육 내용**

선형과 비선형의 차이

* Parameter 값이 선형적으로 연속해서 연결 되어 있는 경우를 선형이라고 한다.

손실함수



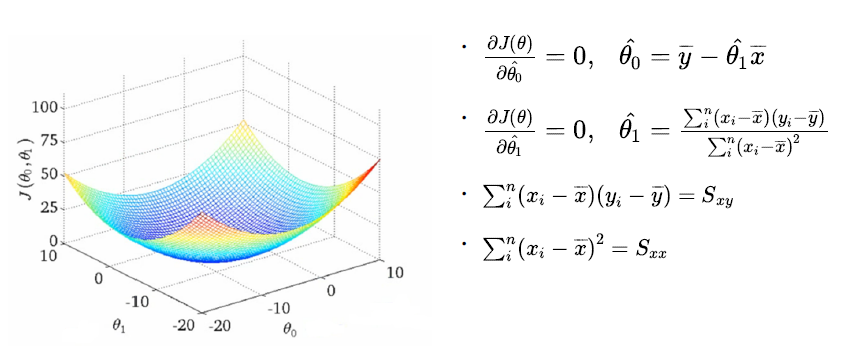
오차의 제곱을 이용해서 얼마나 정확한지를 판단하는 방법이다.

제곱을 사용하는 이유는 분산처럼 부호를 없애기 위해서 이다. 오차의 차이에서 방향은 의미가 없고 그 크기만 의미가 있기 때문에 제곱을 사용한다.

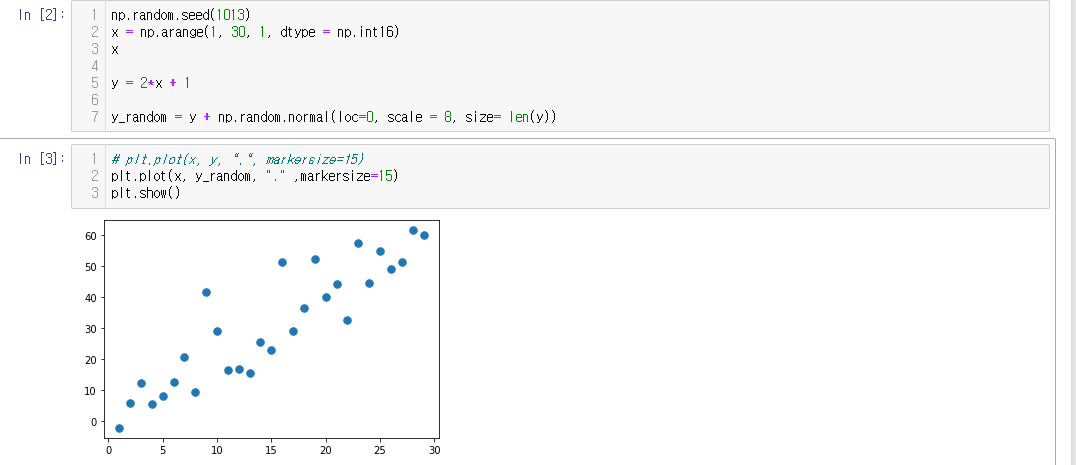
이때 나오는 손실 함수 J는 세타에 대한 함수가 되는데 이 때 이 손실 함수의 최소 값을 찾는 것이 결국 가장 적절한 모델을 찾는 것이 된다.

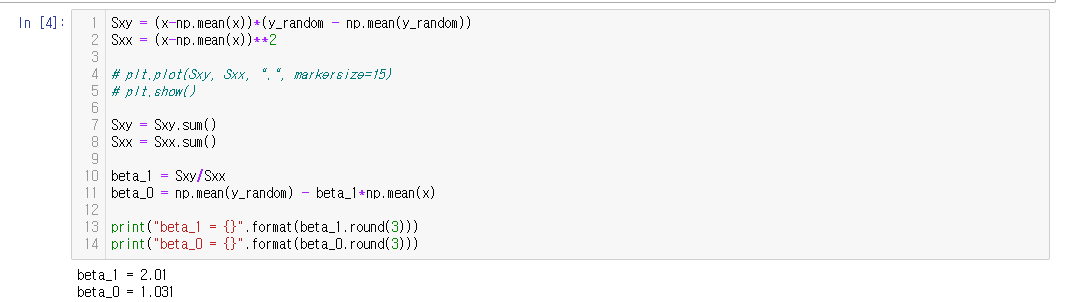
손실함수 J는 세타에 대한 2차 함수인데 왜냐하면 제곱의 합이기 때문이다. 따라서 이 2차 함수에서 최소 값을 찾을 수 있다면 그것이 결국 가장 적절한 모델 값을 찾을 수 있는 방법이다. 이차 함수에서 최소 값을 찾는 방법은 여러 방법이 있다. 그 중 미분을 했을 때 기울기의 값이 0이 되면 최소 값을 찾을 수 있다.

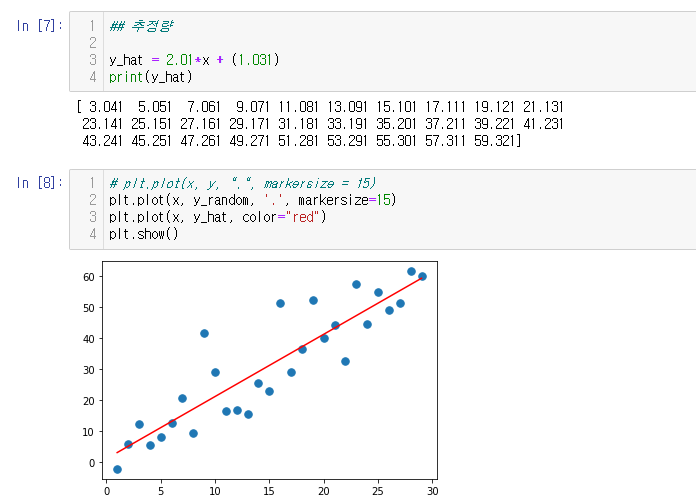
따라서 이를 바탕으로 각 세타에 관한 J의 최소가 되는 지점을 찾으면 그에 따른 세타 값들을 구할 수 있고 이를 바탕으로 가장 적절한 모델을 찾을 수 있다.



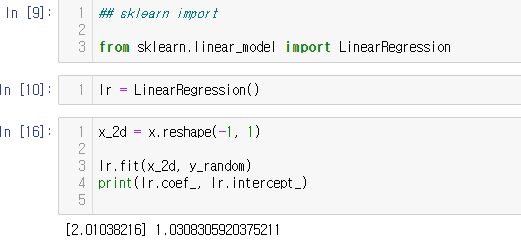
Sklearn 이용 하지 않은 방식과 이용한 방식에 관한 코드 예제







사용



바로 값을 구할 수 있는 것을 확인 할 수 있다.

**오후 교육 내용**

Modeling Process

AI (Artificial Intelligence)

* Decision Tree

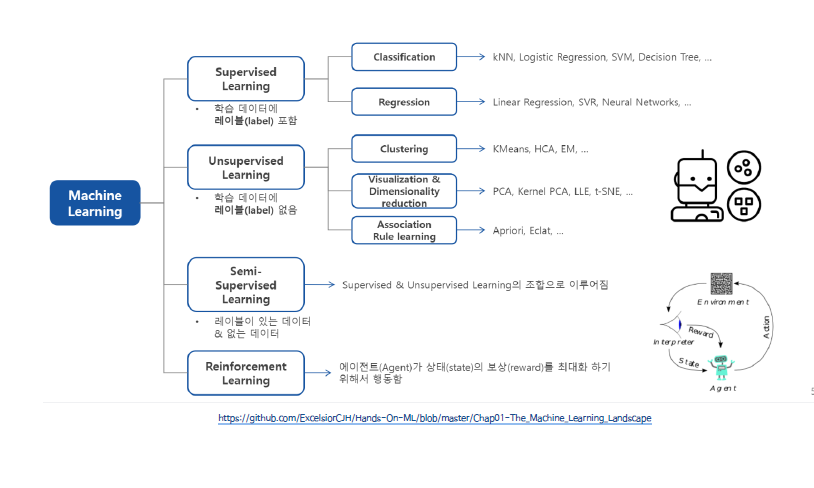
Machine Learning

* Model-based
* Logistic Regression
* SVM (Support Vector Machine)
* Random Forest

Deep Learning

* CNN, RNN

기계 학습의 분류



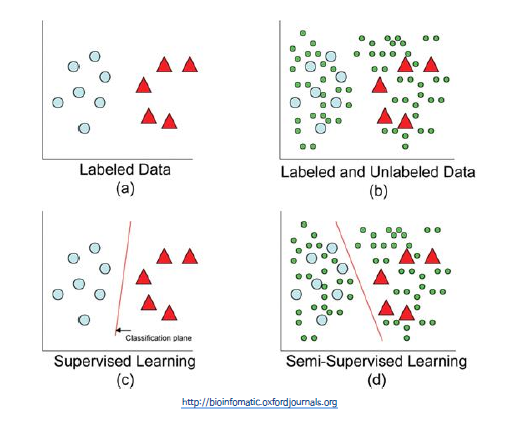
레이블(label): 결과가 가져야 하는 정답 값, 학습 결과에 대한 답이 존재한다.

PCA (Principle Component Analysis)

* 정보화 손실이 최소화 되도록 Feature를 줄이는 방법
* 데이터의 상관성을 고려하여 Feature를 제거한다.

준지도 학습 (지도 학습과 비지도 학습이 섞여 있는 경우)

라벨이 존재하는 데이터와 존재하지 않는 데이터가 섞여 있는 것을 학습해야 하는 경우



ML전처리

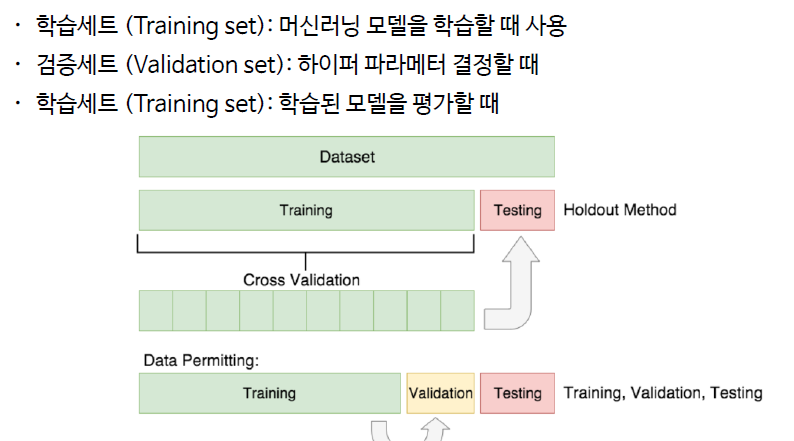
* Feature들에 대한 적절한 전처리 (Pre-processing)
* 초매개변수 조절 (Tuning hyperparameters)
* 모델 성능 평가 (Assessing model performance)

ML알고리즘

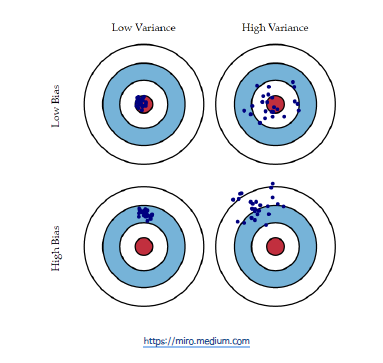
* ML알고리즘 종류
* 모수(매개변수) 학습 (Parameter Learning)
* Python을 이용한 실습

Training set and test set

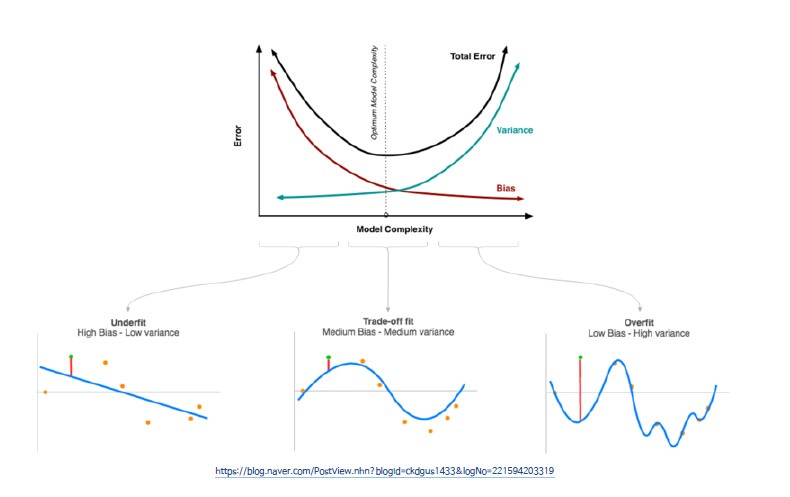
* ML 모델의 성능평가를 위해서 자료를 분할
* Training set: 모델의 알고리즘 learning, 모델에 사용될 FEATURE들을 결정, 초매개변수 조절(약 전체 자료수의 70%로 설정)
  + Training set: 모델의 알고리즘 learning
  + Validation set: 모델에 사용될 featrue들을 결정, 초매개변수 조절, 과적합(Over fitting)방지 -> 초매개변수는 모델을 작성하는 사람이 결정한다.
* Test set: 최종 선택된 모델의 성능평가 (약 전체 자료수의 30%로 설정), 자료의 수가 적을 경우 생략 가능



모델의 치우침 (Bias)과 분산 (Variance)

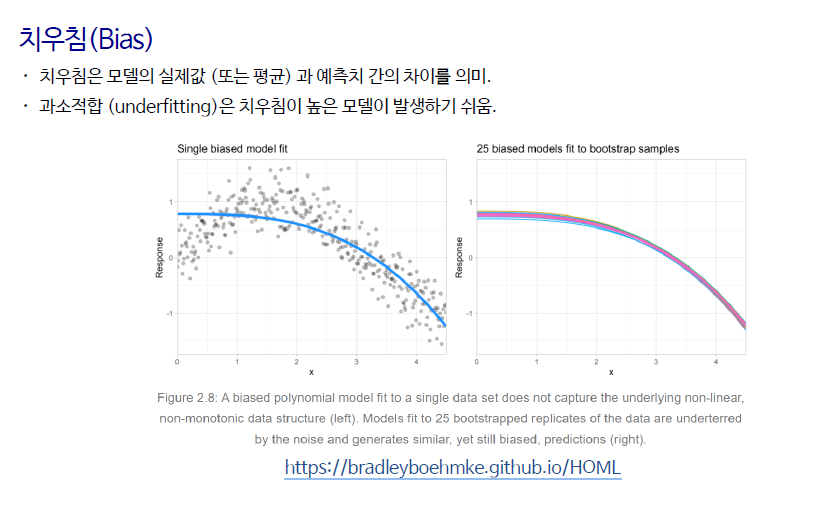


바이어스와 분산은 Trade-off 관계이다. -> 반비례 관계



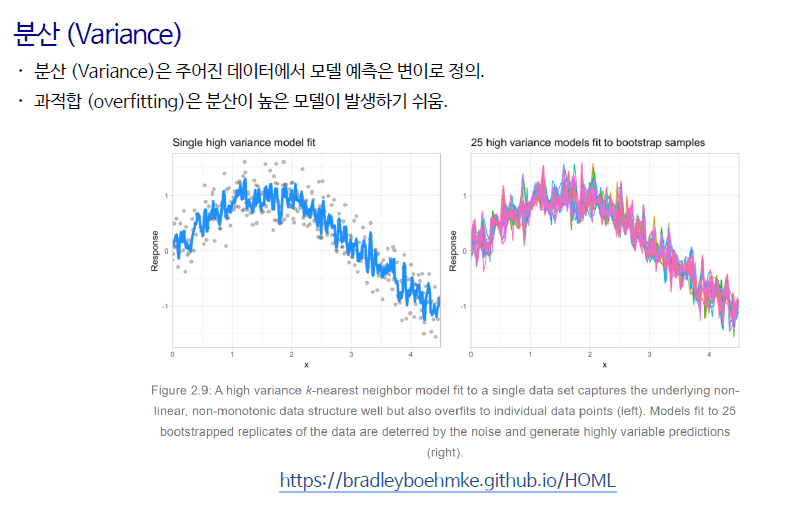
Underfit(과소적합)

* 모델이 너무 단순해서 에러가 너무 많이 발생한다.
* 바이아스가 너무 높다.
* 실제로는 잘 안생김 -> 현실에서는 모델을 복잡하게 구현하는 경우가 많기 때문

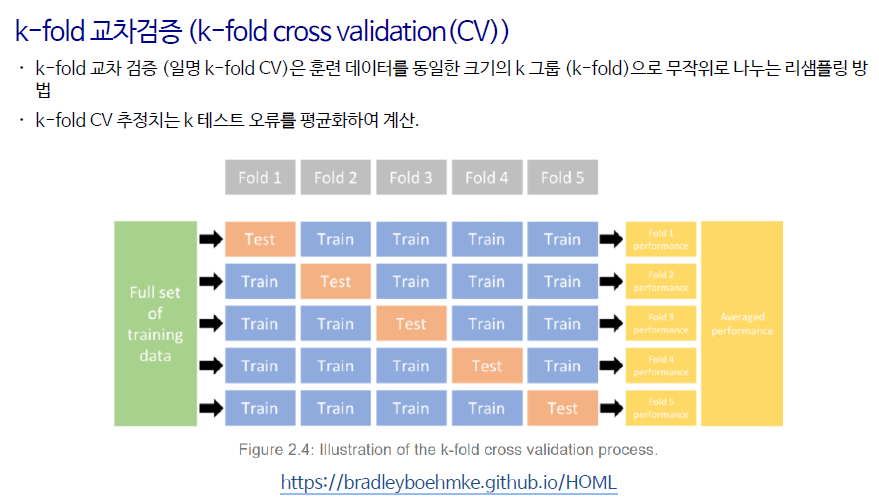


Overfit(과적합)

* 모델이 학습 데이터에 지나치게 치우쳐서 새로운 데이터에 대해서는 정확한 값을 내놓지 못한다.
* 모델이 너무 복잡하며 학습 데이터에 대한 에러는 없다.
* 바이아스가 너무 낮다. -> 한쪽으로 너무 쏠려 있다



기계학습 모델 평가



훈련 데이터를 K개의 집단으로 나눈 다음에 나눈 집단에서 하나만 Test로 두고 나머지를 학습 시켜서 Test로 둔 데이터를 통해 학습을 한다. 전체 데이터가 K개의 집단으로 구분되어지기 때문에 K번이를 반복하고 이를 바탕으로 성능을 테스트 한다.