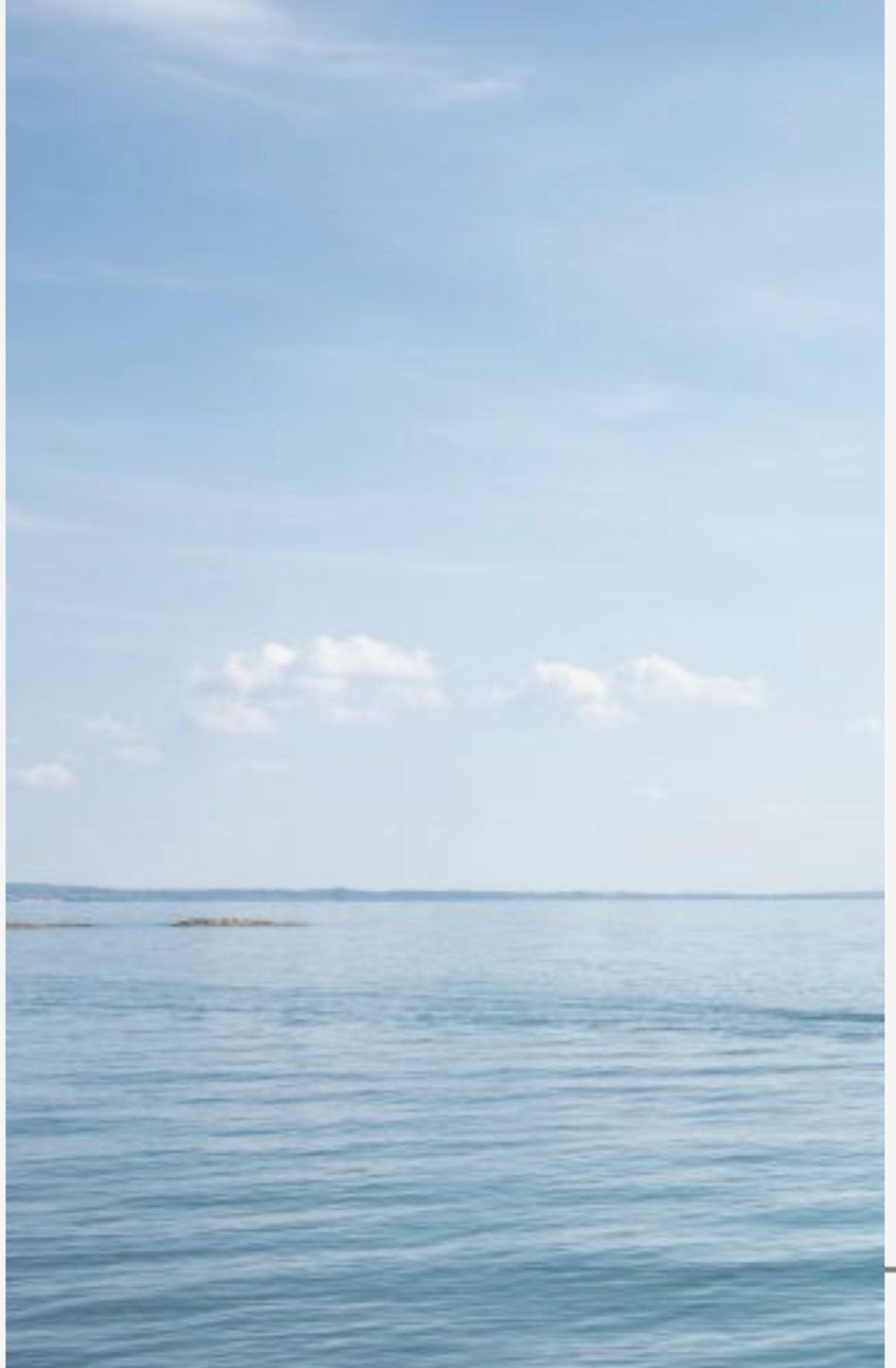
2021 / 01 / 31

Week4
머신러닝 &
topic modeling



목차

01

머신러닝?

머신러닝이란?

데이터를 분석하고, 해당 데이터를 통해 학습한 후,
그 학습된 정보를 바탕으로 목표에 따라 예측/혹은 분류하는
알고리즘.

지도학습 vs 비지도 학습

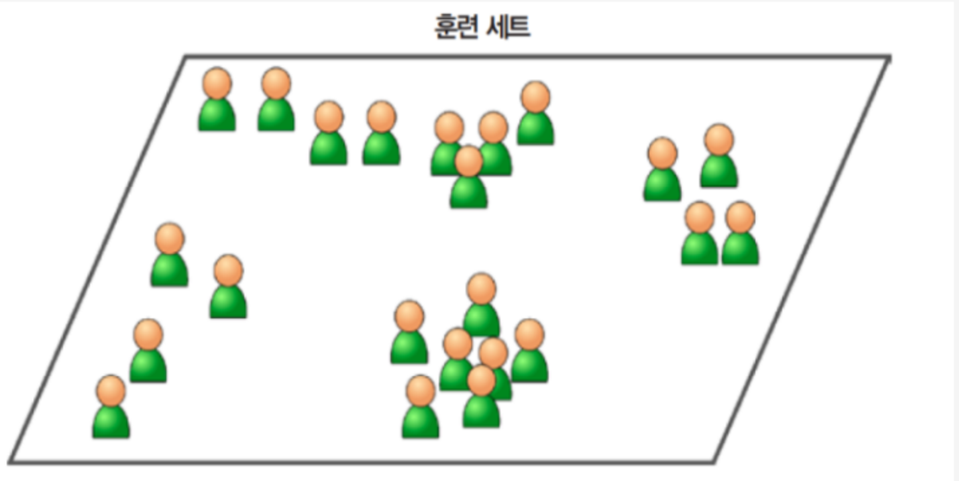
~~뭐 결국... 다 예측하려고 하는 일 아니겠습니까~~

지도학습 vs 비지도학습

지도학습: 답이 제공된, 족보같은 것.
train set과 test set으로 나뉘어 그 훈련된 것이 제대로 맞았는지 확인할 수 있음.

비지도 학습: 답이 제공되지 않은 데이터를 학습시키는 것

그림 1-5 지도 학습에서 레이블된 훈련 세트(예를 들면 스팸 분류)



머신러닝의 과정

1. EDA(탐색적 자료 분석)
2. 전처리
3. train-test-split
4. train data로 모델 하이퍼파라미터 튜닝
5. 4에서 결정된 최종 하이퍼파라미터로
모델을 돌려서 최종 성능 얻기.

머신러닝의 과정

필요한 여러 모듈들 이용해서 진행

01 EDA(탐색적 자료 분석)

데이터를 자유롭게 뒤적뒤적....의미있는 것 찾아보기...

02 전처리

column name 수정, Scaling, type 변환, 외부 데이터 붙이기 등

머신러닝의 과정

필요한 여러 모듈들 이용해서 진행

03 train-test-split

학습을 위한 데이터, 평가를 위한 데이터 나누기.
아 물론 그전에 종속변수, 독립변수 나누어야 합니다.

04 hyper parameter튜닝

hpyer parameter : 내가 직접 설정할 수 있는 파라미터들.

05 성능평가

best score, mean_accuracy, cross_val_score 이용해 성능평가
test 데이터로 성능 평가해야합니다.

02

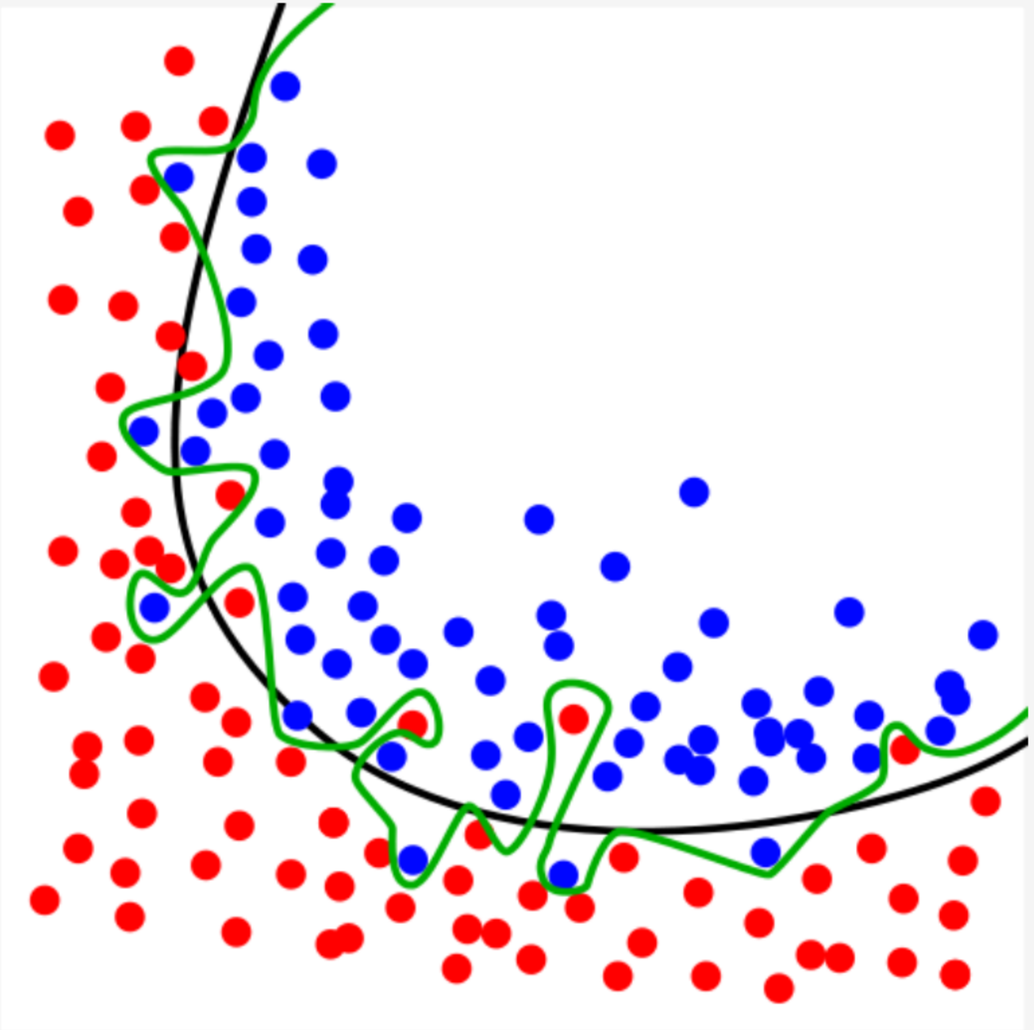
Overfitting &Cross Validation

Overfitting

과적합이란?

학습 데이터에만 너무 지나치게 학습되어
정작 중요한 패턴을 설명할 수 없는 현상.

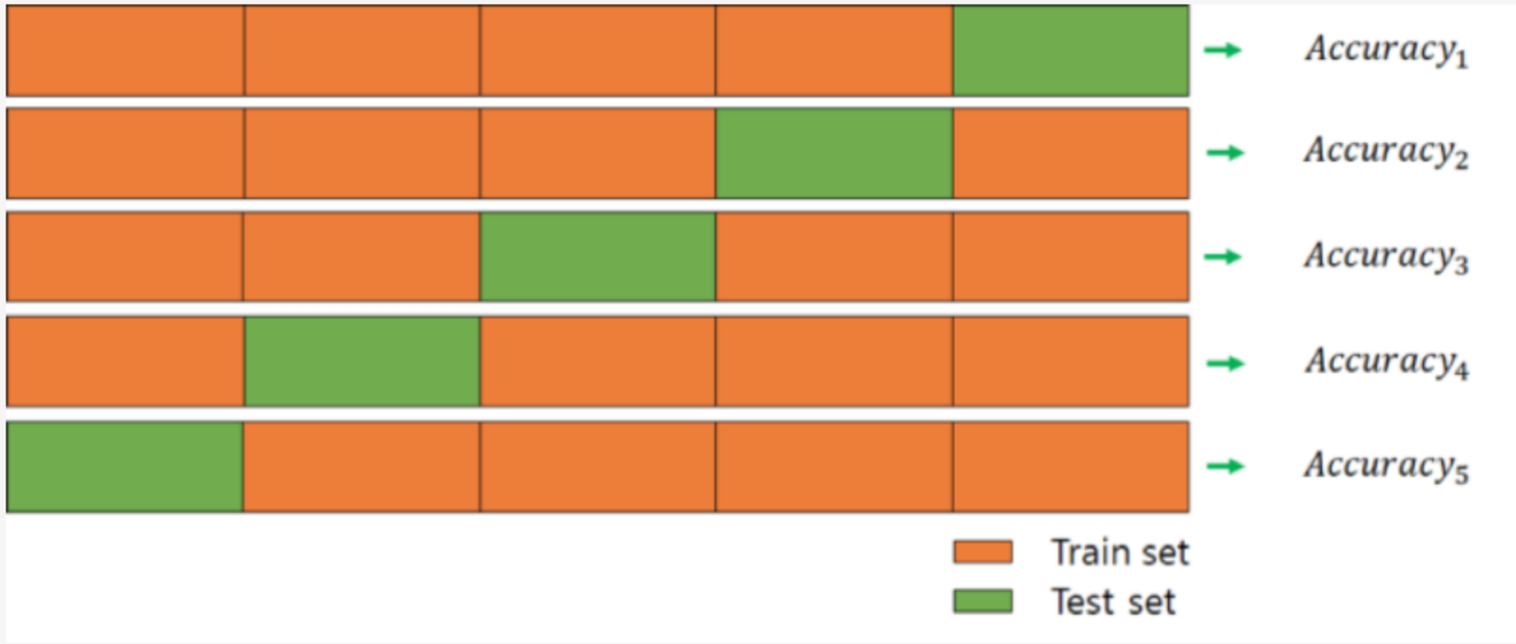
족보만 달달 외웠는데
생판 모르는 문제 나오는 거랑 비슷합니다..



k-fold Crosss validation

전체 데이터 셋을 k개의 subset으로 나누고 k번의 평가를 실행하는데,
이 때 test set을 중복 없이 바꾸어가면서 평가를 진행.

다음으로 k개의 평가 지표(이 경우는 accuracy로 예를 듦)를 평균(때에 따라 평균이 아닌 방법을 사용할 수도 있음) 내어서 최종적으로 모델의 성능을 평가한다.

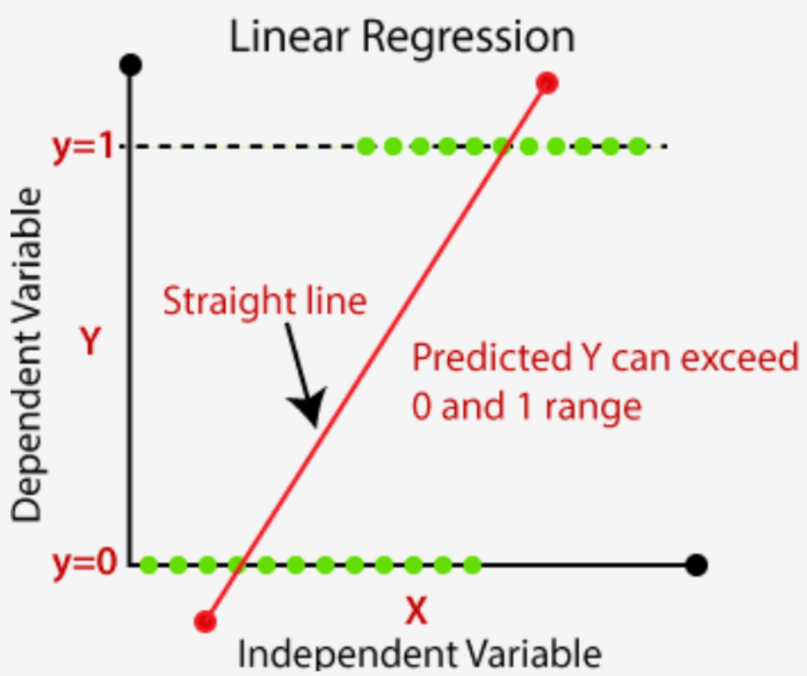


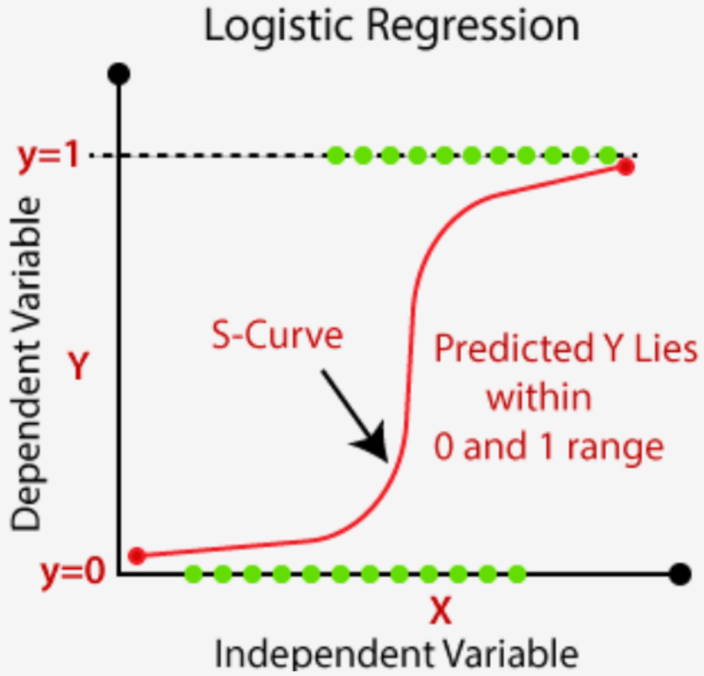
03

Machine Learning Models

Linear Regression
Logistic Regression

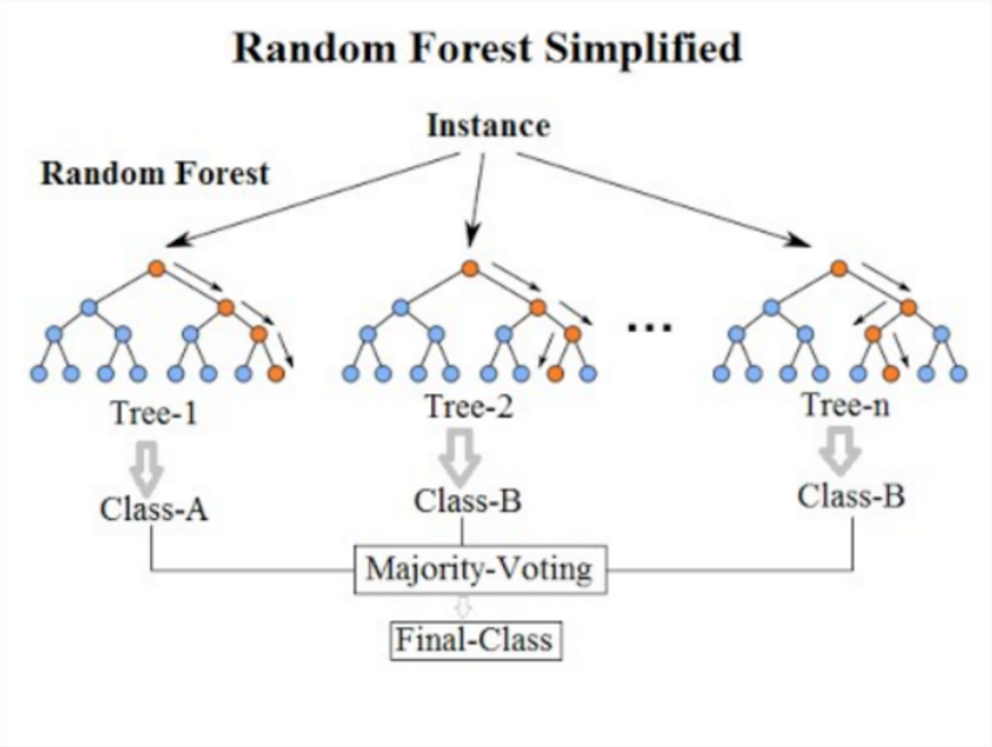
선형회귀/ 로지스틱 회귀





RandomForest Classifier

여러개의 decision tree를 형성하고
새로운 데이터 포인트를 각 트리에 동시에 통과시키며,
각 트리가 분류한 결과에서 투표를 실시하여 가장 많이 득표한 결과를 최종 분류 결과로 선택



실습하러 가봅시다!