Ensemble

2022.01.11

Dongmin Kim (tommy.dm.kim@kaist.ac.kr)

TODO

- 앙상블 (ensemble)
- Random Forest
- Gradient Boosting
- XGBoost, LightGBM
- 여러 모델의 앙상블
- Kaggle Titanic에 Random Forest, XGBoost, LightGBM 적용하기

I. Ensemble

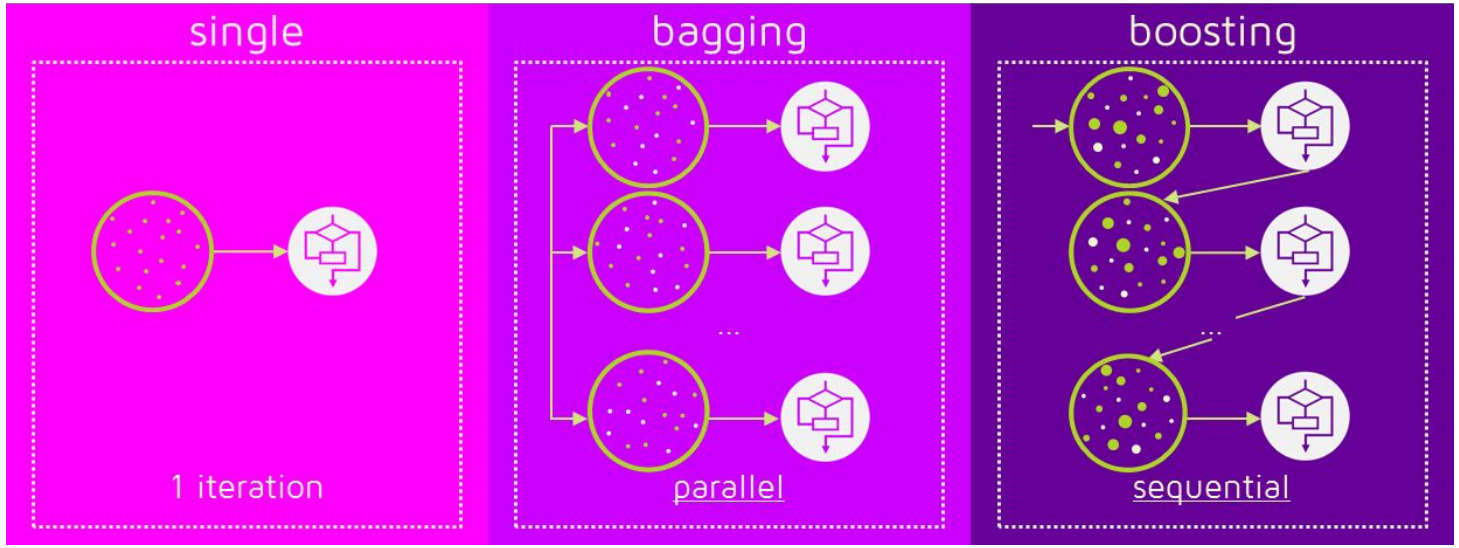
앙상블 (Ensemble)

- 앙상블 (ensemble): 하나의 단일 모델로 예측 하기보다, 여러 모델의 output을 합치는 과정

- Classification
  - Hard voting: one-hot을 결합
  - Soft voting: probability를 결합

- Regression
  - Mean, Median

- 방법론
  - Bagging (random forest)
  - Boosting (boosting algorithms)
  - Stacking



2. Random Forest

Random Forest

x 10000 =

Random Forest

Random Features 1
Random Data 1

0

Random Features 2
Random Data 2

1

Random Features 3
Random Data 3

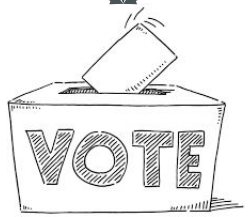
1

Random Features 4
Random Data 4

0

Random Features 5
Random Data 5

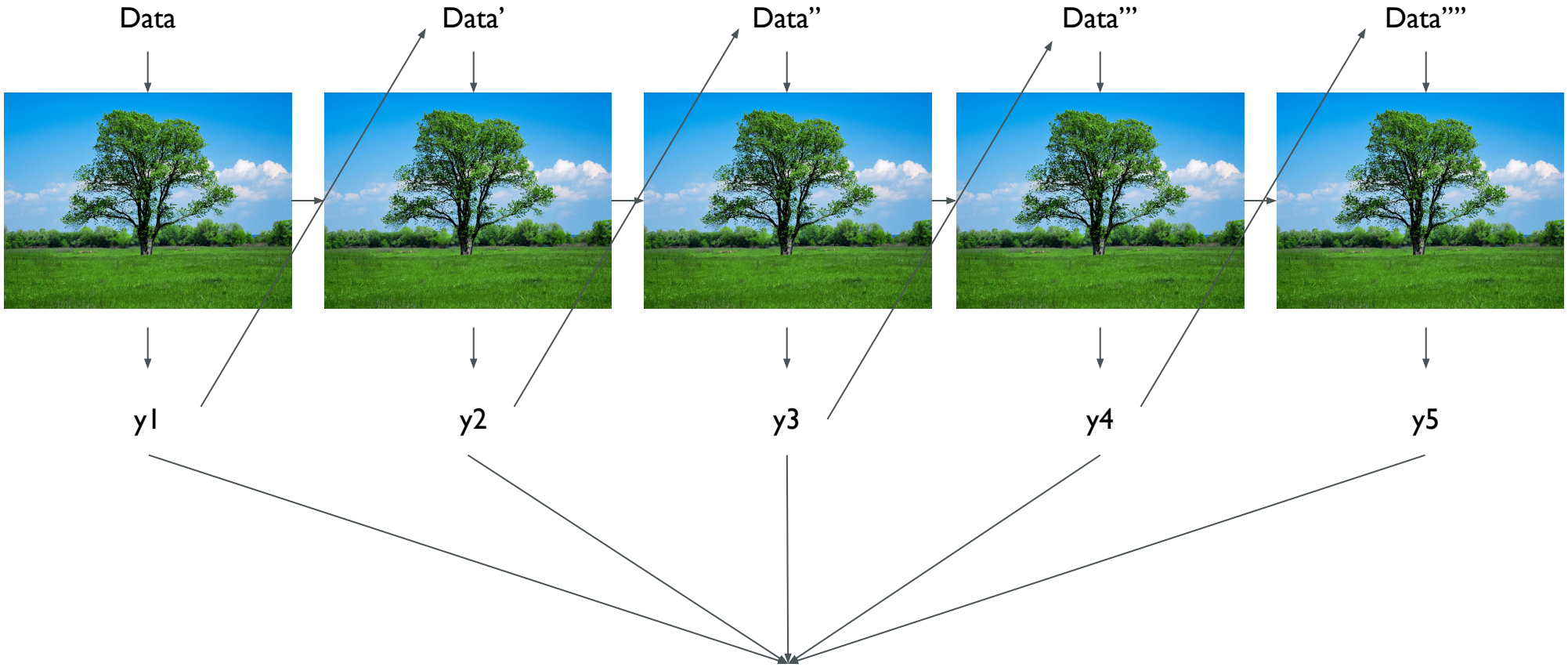
1



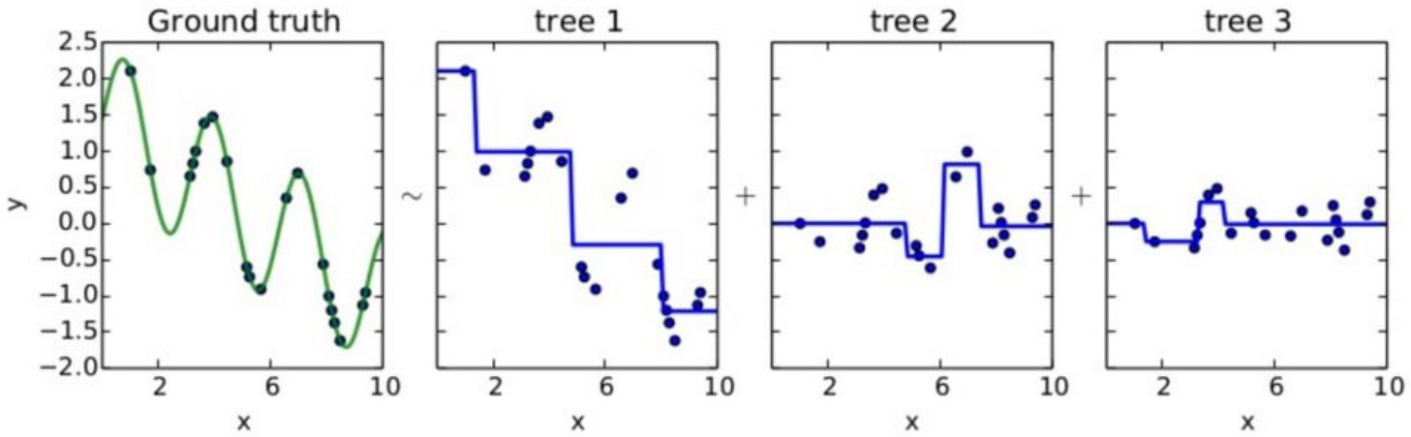
1

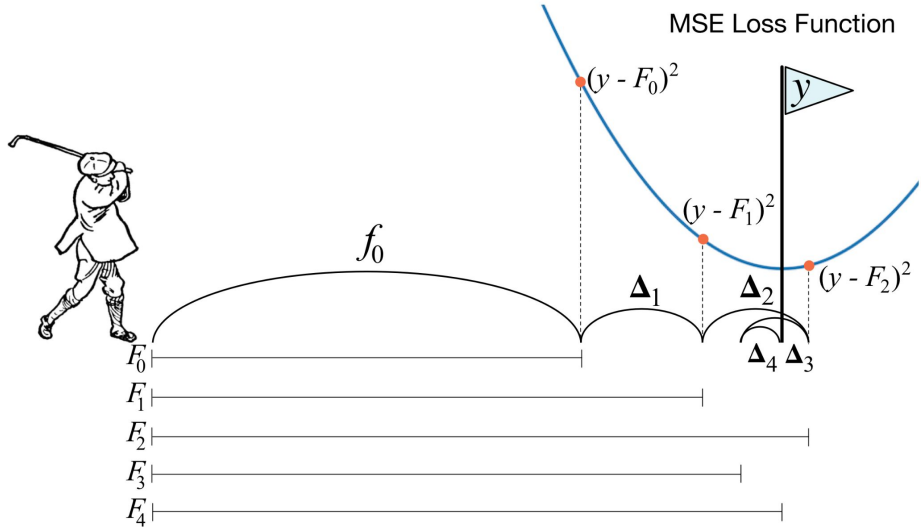
3. Gradient Boosting, Xgboost, LightGBM

Boosting



Gradient Boosting





Gradient Boosting

- Boosting은 앞선 모델의 약점을 보완해가는 과정

- Gradient Boosting은 앞선 모델의 약점을 Gradient를 통해 포착
  - MSE Loss

$$-\frac{\partial L}{\partial y_p} = 2\left(y_i - y_p\right)$$

  - Cross entropy Loss

$$\frac{\partial C}{\partial z_i} = \sum_k \frac{\partial C}{\partial s_k}\frac{\partial s_k}{\partial z_i} = s_i - y_i$$

- Useful References
  - statquest: https://youtu.be/3CC4N4z3GJc

XGBoost, LightGBM

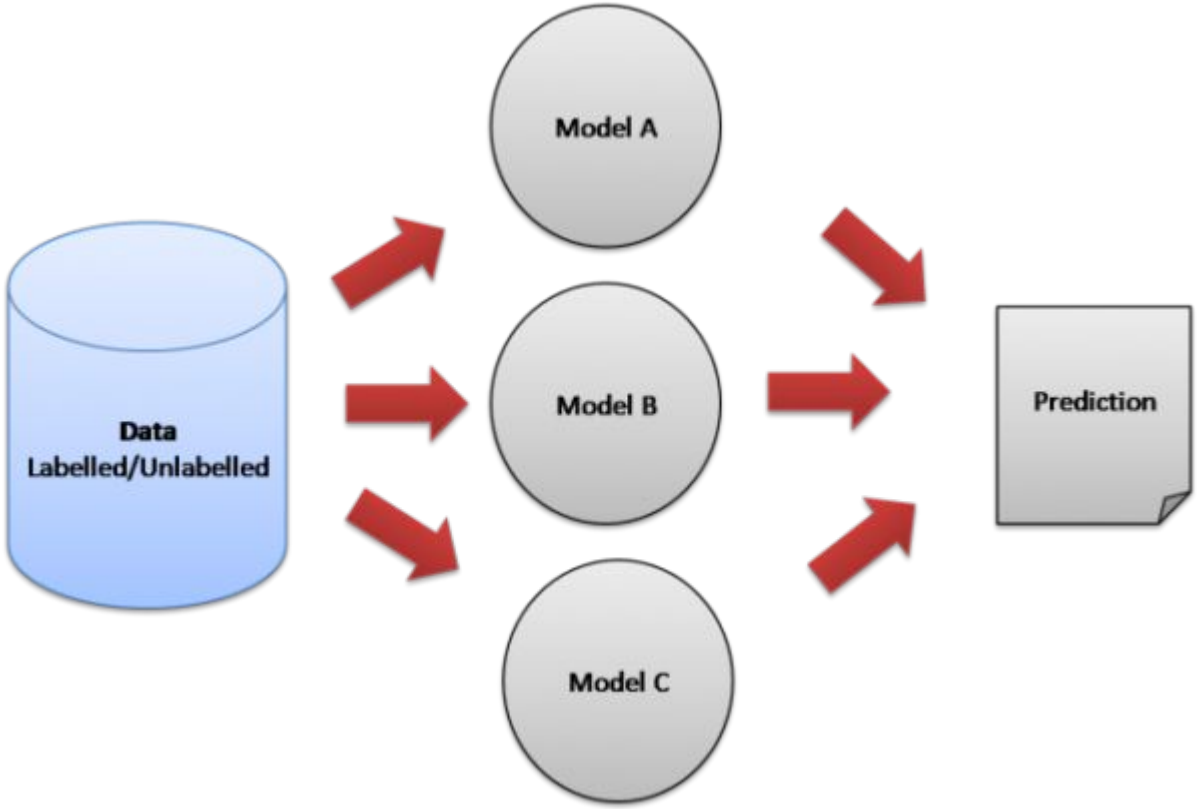
기존 Gradient Boosting 방법론을 발전시킨 알고리즘들

- XGBoost (eXtreme Gradient Boosting)
  - https://youtu.be/OtD8wVaFm6E
  - learning rate (eta) : step size
  - gamma: tree의 minimum impurity decrease와 같은 역할. gamma가 커지면 조금 더 보수적으로 가지치기를 함 (regularization)
  - n_estimators: tree 개수

- LightGBM (Light Gradient Boosted Machine)
  - https://youtu.be/4C8SUZJPlMY
  - learning rate (eta) : step size
  - lambda_l1, lambda_l2 : regularization
  - n_estimators: tree 개수

4. 여러 모델 앙상블하기

Voting


Model A
Data
Labelled/Unlabelled
Model B
Prediction
Model C

5. Kaggle

Kaggle Titanic에 Random Forest, XGBoost, LightGBM 적용하기

0. Kaggle 노트북 준비하고, decision tree 부분까지 실행시키기

6. Random Forest, XGBoost, LightGBM 모델 써보기
   6.1. 모델 트레이닝 시키기
   6.2. accuracy 출력, confusion matrix 그리기, classification report 생성해보기, feature importance 뽑아보기

7. Hyperparameter Tuning
   7.1. sklearn의 gridsearchCV를 사용하여, 튜닝을 해봅시다.
   7.2. 튜닝 된 모델들에 대하여, 6.1, 6.2 에서 했던 모든 과정들을 반복해주세요.

8. Voting
   8.1. sklearn의 soft voting, hard voting을 이용하여,
   decision tree, random forest, xgboost, lightgbm, logistic regression 총 5개의 모델을 앙상블 해주세요.

[Submission 3]

- 우리의 세번째 모델에 대한 결과를 제출해봅시다.
- 오늘 다룬 decision tree, random forest, boosting 모델들, 튜닝, 앙상블 등을 적절히 잘 활용해서, 가능한 최고의 결과를 뽑아보세요!