데이터마이닝팀

4팀

장이준
이선민
김영호
김현우
박시언

CONTENTS

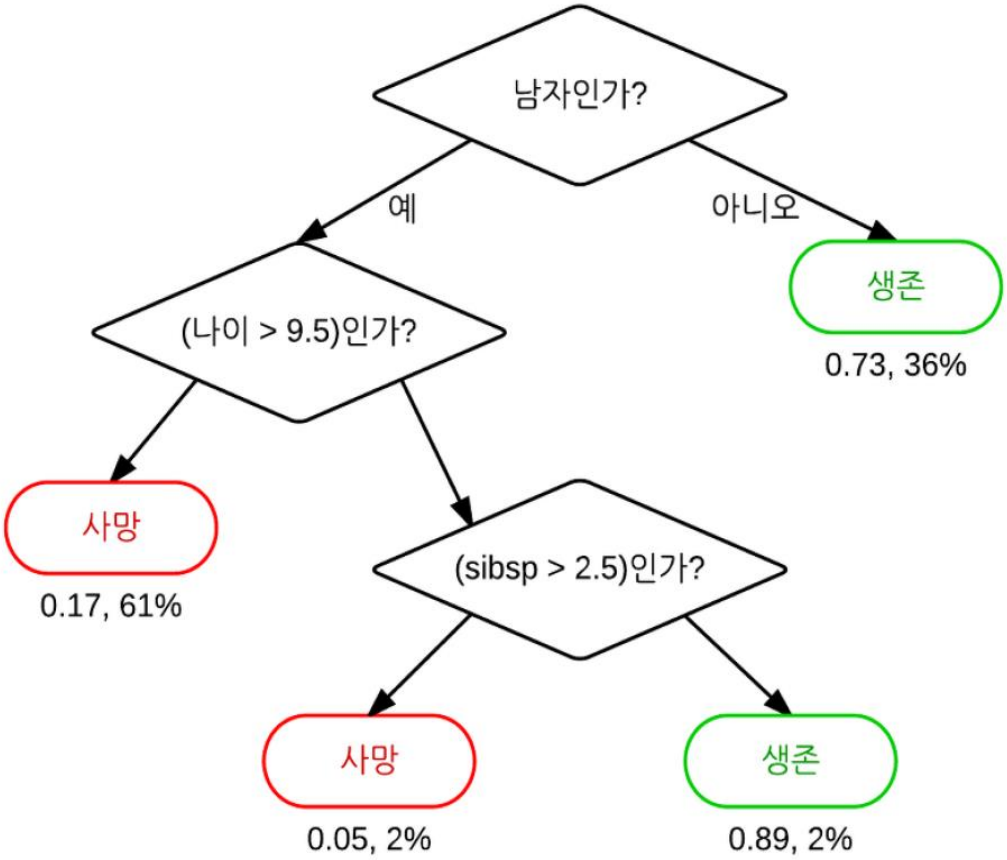
1

트리 기반 모델

트리 기반 모델

트리 기반 모델

의사 결정 나무



여러 정보에 기반하여
생존 여부 예측하기

각 질문에 대한 대답으로
예측값이 결정됨!

Ex) 타이타닉 프로젝트

트리 기반 모델

회귀 모델 | Decisiontree Regressor

Objective Function

$$\min_{c_m} \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 = \min_{c_m} \sum_{i=1}^{N} \left[y_i - \sum_{m=1}^{M} c_m I(X \in R_m) \right]^2$$

- cm 해당 노드의 관측값들의 평균값
- N 전체 관측값의 개수
- M 전체 node 의 개수 해당

트리 기반 모델

분류 모델 | Decisiontree Classifier

“ Impurity(불순도) ”

분기된 영역 (ex. R1, R2) 내에서 각기 다른 다양한 범주(factor)들의 개체들이 얼마나 포함되어 있는가

즉, 한 영역 내에서 불순도가 높을수록 분류가 잘 되지 않음

트리 기반 모델

엔트로피(Entropy)

불순도로서 쓰이는 엔트로피 개념 이해하기 위해

$\hat{P}_{mk}$를 한 영역 내의 특정 클래스의 비율로 생각하고 엔트로피의 수식을 보자!

$$m\text{번째 영역의 } Entropy = -\sum_{k=1}^{K} \hat{P}_{mk} log_2(\hat{P}_{mk})$$

$$where\ \hat{P}_{mk} = \frac{1}{N_m} \sum_{x_j \in R_m} I(y_i = k)$$

K: 영역 내의 class 종류

$\hat{P}_{mk}$: 분기된 영역 m의 k번째 class 비율

트리 기반 모델

과적합 방지법

Avoid Overfitting in Tree Based Models

사전가지치기

트리의 깊이를
사전에 지정하는 방법

VS

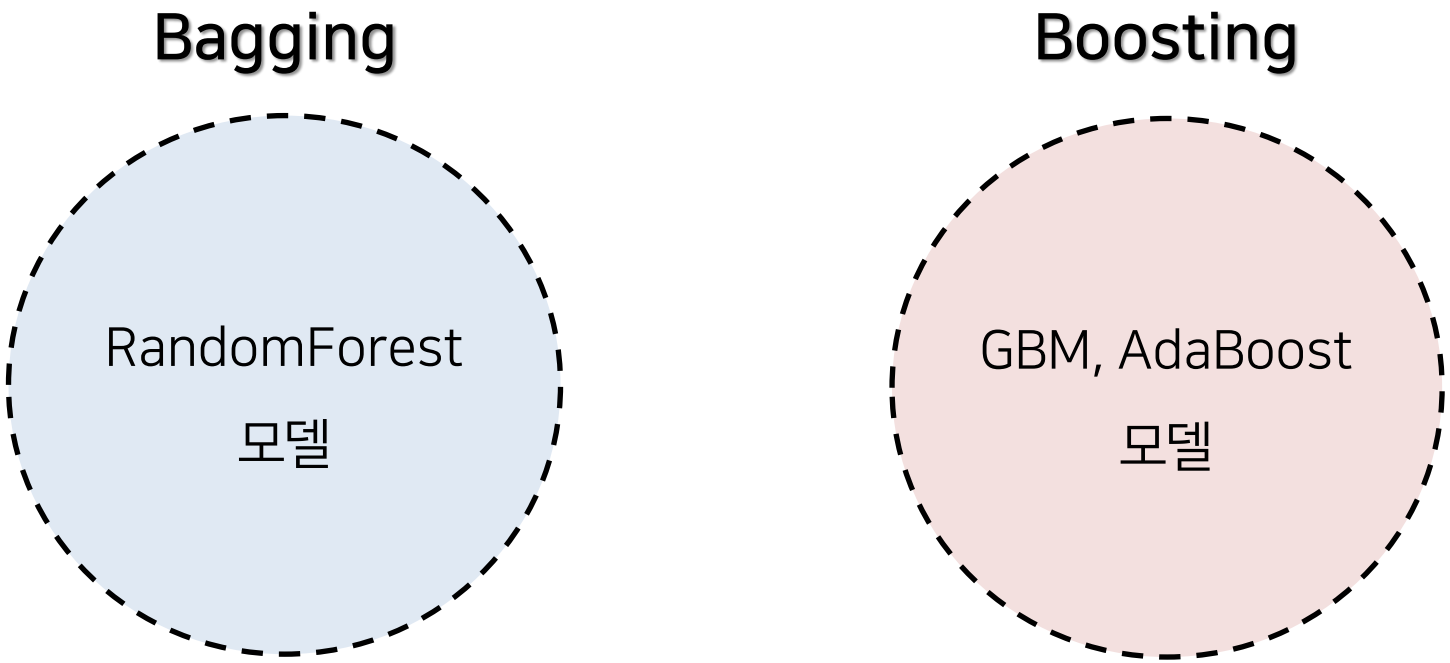
사후가지치기

Full tree를 만든 후
적절한 수준에서 terminal
node를 결합하는 방법

트리 기반 모델

앙상블 기법(Ensemble Method)

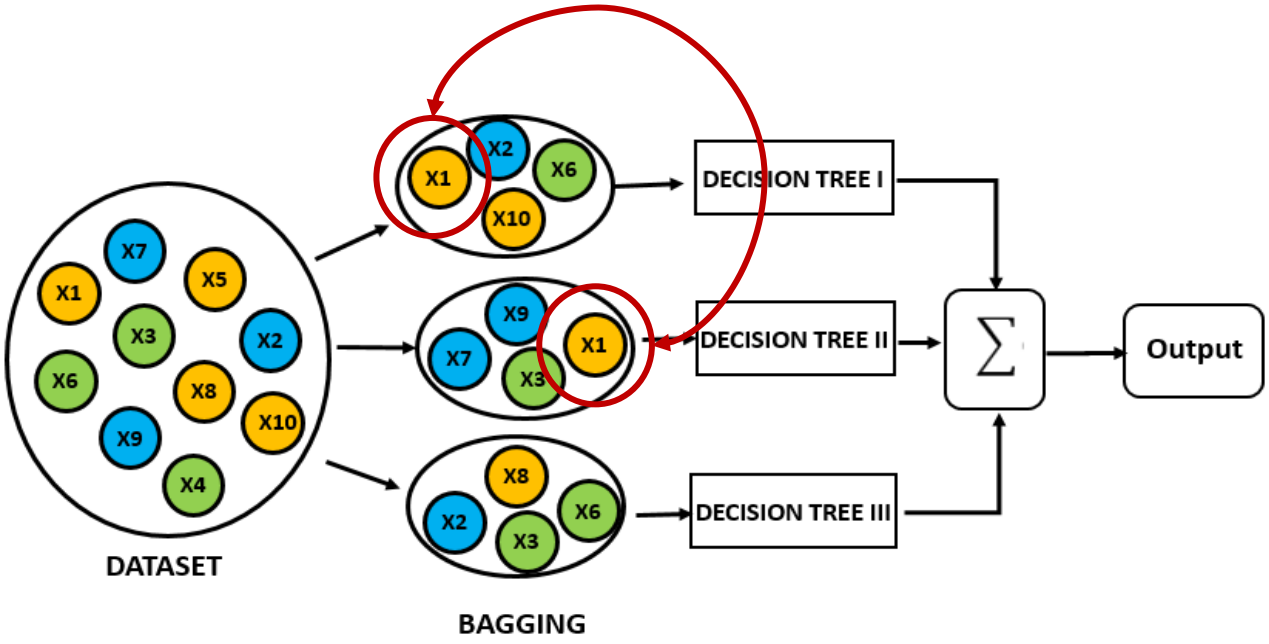
앙상블 기법이란?



트리 기반 모델

배깅 기법(Bagging Method)

배깅(Bagging)이란?



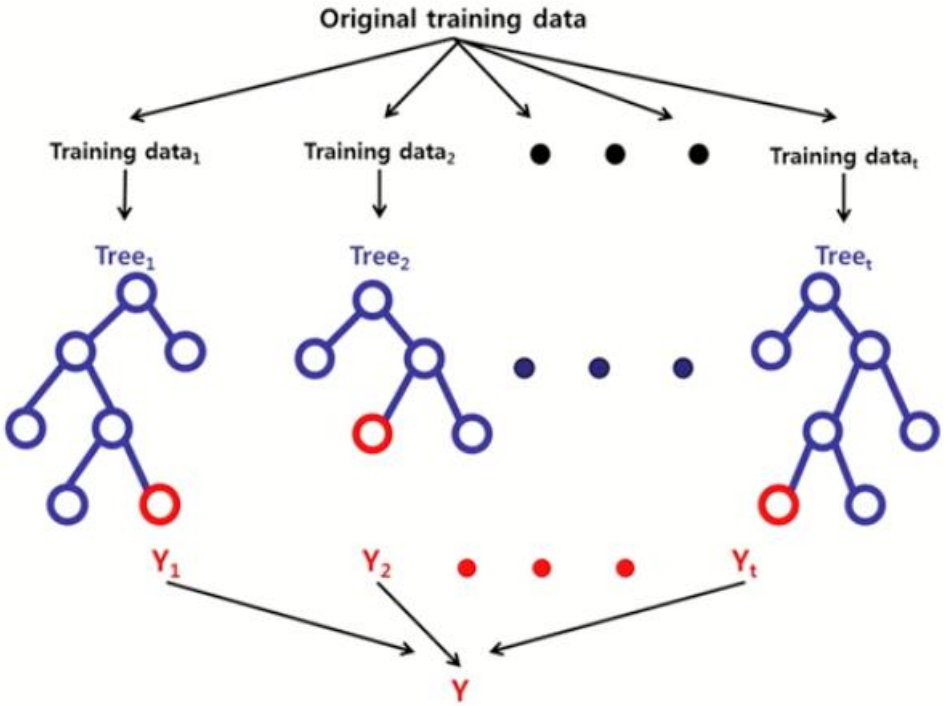
여러 개의 의사결정 나무를 만들되, 각각의 N개의 의사결정 나무를

부트스트랩 기법으로 추출된 N개의 데이터 셋으로 학습을 시키는 기법

트리 기반 모델

배깅 기법(Bagging Method)

랜덤포레스트(RandomForest) 모델

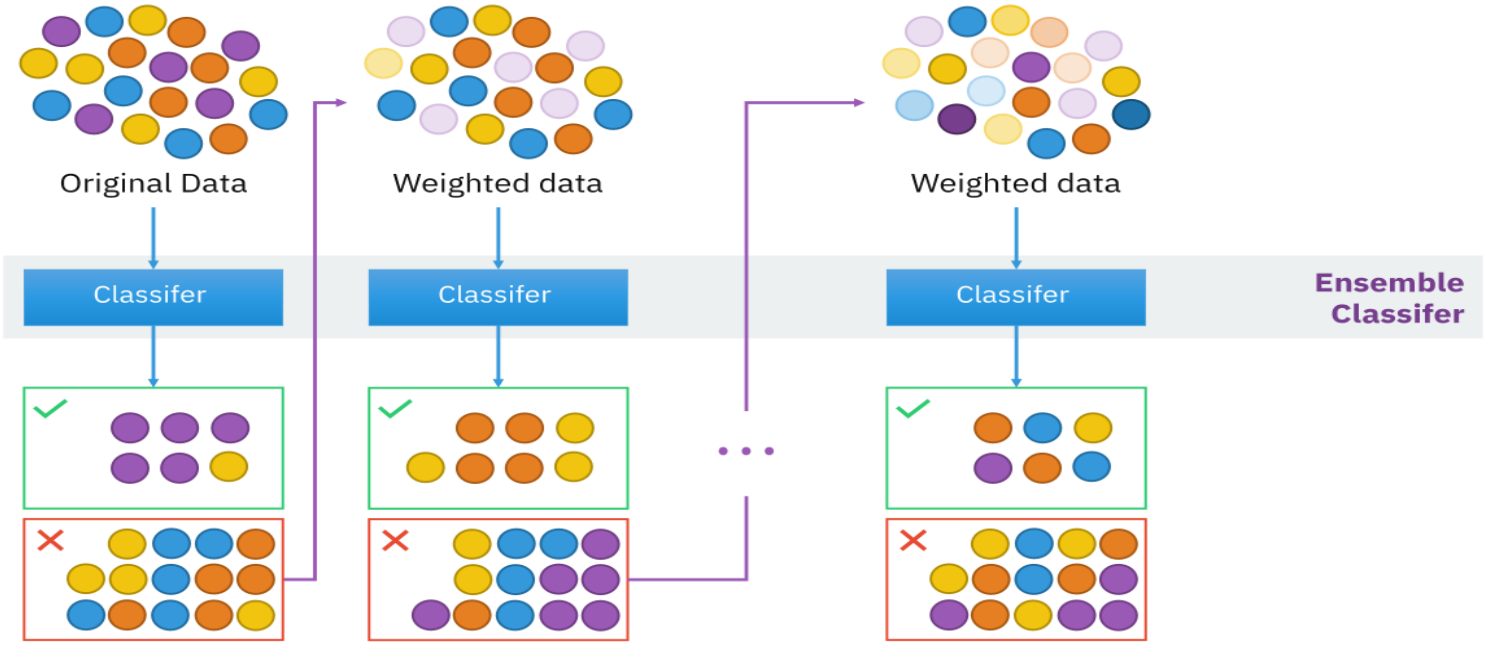


각 모델링마다 사용되는 Feature의 개수를 랜덤하게 선택함

트리 기반 모델

부스팅 기법(Boosting Method)

부스팅(Boosting)이란?



앙상블 기법 중 하나로, 여러 개의 약한 트리(Weak Tree) 모델을 모아서 하나의 강한 모델을 만들어내는 기법

2

Non-linear 모델

Non-linear 모델

Piecewise Polynomials | Piecewise Linear

Piecewise linear

문제점

불연속적 → knot 기준으로 좌극한 ≠ 우극한

knot을 기준으로 각 함수의 좌극한과 우극한이 같도록 하는 제약식 추가

$$f(\xi_i^-) = f(\xi_i^+)$$

Non-linear 모델

Cubic Spline

문제점

불연속적 → 좌극한 ≠ 우극한

미분불가능 → 좌미분계수 ≠ 우미분계수

좌극한 = 우극한

제약식 추가

$$f(\xi_i^-) = f(\xi_i^+)$$

좌미분계수 = 우미분계수

제약식 추가

$$f'(\xi_i^-) = f'(\xi_i^+)$$

$$f''(\xi_i^-) = f''(\xi_i^+)$$