D&A

ML Session 6차시

Bagging .

2022 / 01 / 31

D&A 운영진 윤경서

CONTENTS.

01 Bagging

02 Random Forest

03 ExtraTrees

01. Bagging(Bootstrap Aggregating)

앙상블이란?

여러 개의 단순한 모델을 결합하여 보다 정확한 모델을 만드는 방법으로 Bagging, Boosting, Voting, Stacking이 있다.

- 단일 모델로는 overfitting의 위험이 있을 수 있기 때문에 깊고 성능이 매우 뛰어난 단일 모델보다
  가볍고 성능이 조금 떨어지더라도 여러 개의 모델을 결합하여 사용하는 것이 더 좋은 일반화 성능을 낼 것이라는 Idea

Bagging이란?

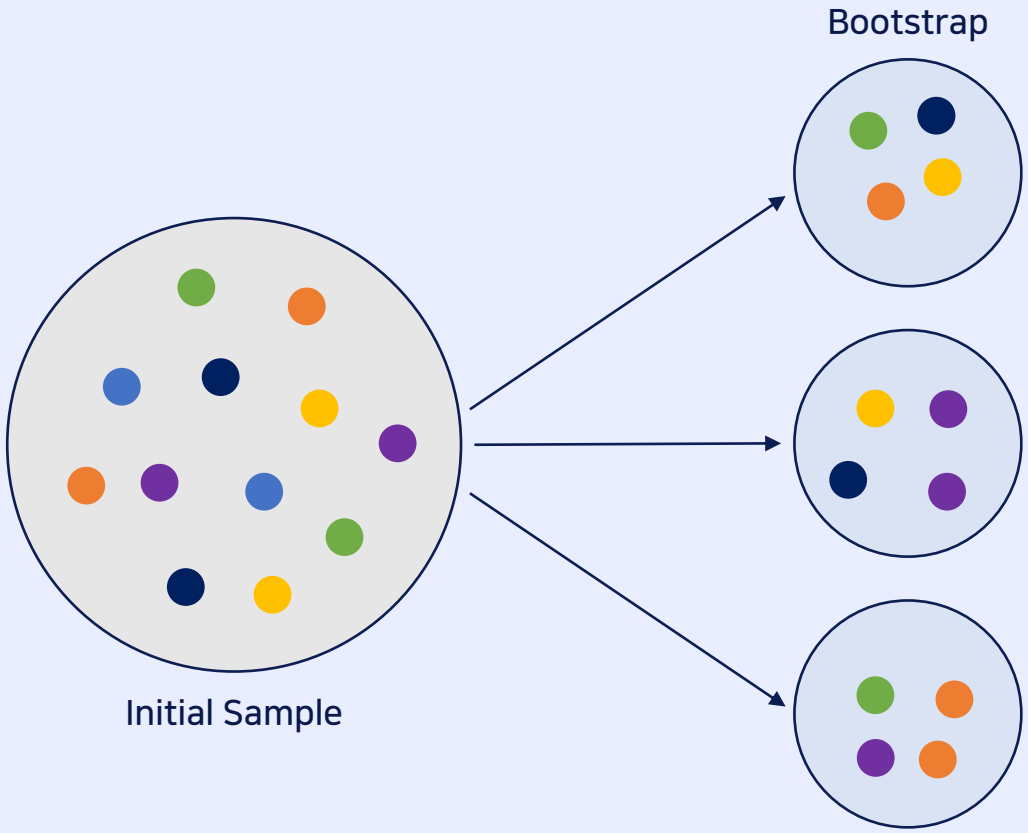
각 모델별로 기존 데이터 셋에서 중복을 허용하여 무작위로 N개의 feature를 선택한 후, 선택한 feature를 통해 만들어진 각 모델의 결과를 취합하는 방법이다. 이때, 각 모델은 서로 독립적이다.

※ 각 모델별로 중복을 허용하여 데이터를 무작위로 선택하는 것을 Bootstrap이라 한다.

Bagging 방식으로 만들어진 앙상블 모델은 각 모델 들의 결과를 취합하여
분류일 경우 voting(투표), 회귀일 경우 평균값으로 결과를 도출한다.

01. Bagging

Bagging의 Bootstrap



그림과 같이 어떤 샘플은 여러 번 샘플링이 되고, 어떤 것은 전혀 선택되지 않을 수 있다.

이 때, 한번도 선택되지 않은 샘플을 oob(out-of-bag) 샘플이라고 부른다.

Bagging에서 oob샘플은 훈련 시에 사용되지 않는다.
➡ oob 샘플을 사용하여 모델을 평가할 수 있다.
➡ 앙상블의 평가는 각 예측기의 oob 평가를 평균하여 얻게 된다.

Bagging의 한계점

1) 각 서브샘플을 구성할 때 복원 추출을 하기 때문에 샘플의 특징이 유사
   - 일반적으로 모집단의 1/3 정도가 oob가 된다.

2) 각 dicision node를 분리할 때 모든 feature를 고려해서 에러를 계산
   - Decision Tree에서 node를 분리할 때 모든 feature의 정보 이득 계산 후, 정보 이득이 높은 feature를 기준으로 분리

Bagging 방식의 대표적인 앙상블 모델
- RandomForest

02. Random Forest

RandomForest란?

수많은 Decision Tree가 합쳐져 만들어진 앙상블 모델

- 여러 개의 Decision Tree가 생성되고, 각 모델은 각자의 방식으로 데이터를 샘플링하여 개별적으로 학습이 된다.
- 분류일 경우 voting(투표)으로, 회귀일 경우 평균값으로 결과를 도출한다.

➡ 이렇게 랜덤으로 데이터를 가져오고, 여러 개의 tree가 모여 숲을 만들기 때문에 Random Forest라는 모델명이 붙게 됨

RandomForest의 장점

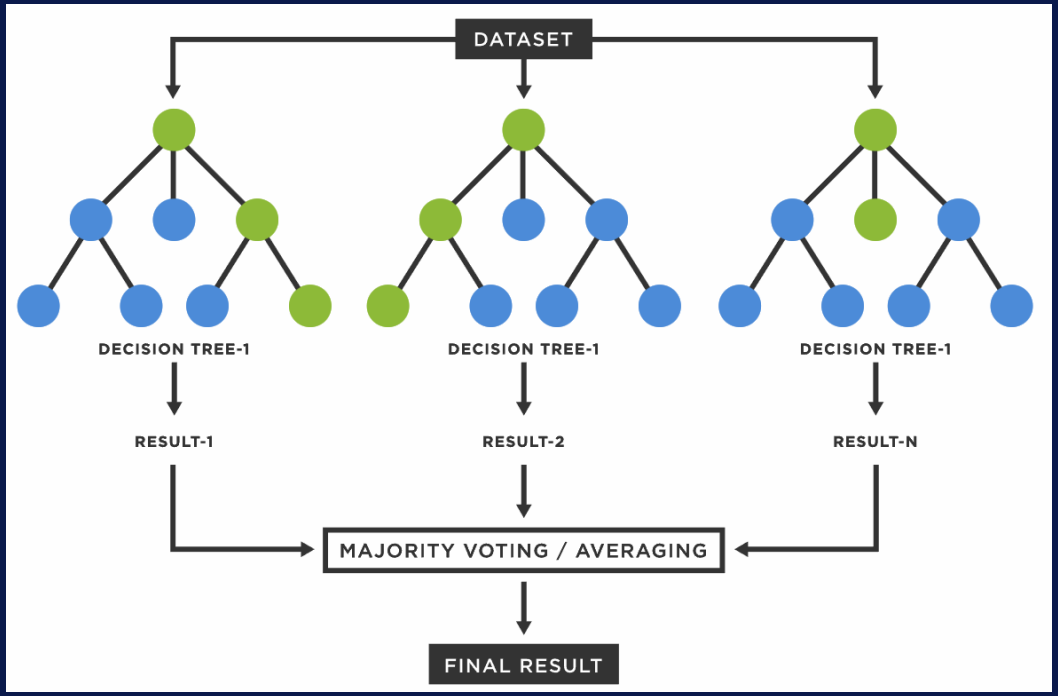
- 분류와 회귀에 모두 사용 가능하다.
- 대용량 처리에 효과적이고, decision tree의 훈련데이터에 overfitting 되는 단점을 해결할 수 있다.
  ➡ 다수의 나무를 기반으로 예측하기 때문에 각 분류기의 영향력이 줄어들게 되어 좋은 일반화 성능을 보인다.
- Version 1.1 이후 각 Decision Tree에서 node를 분할할 때 모든 feature가 아니라 랜덤하게 일부의 feature만을 사용하였다.
  ➡ Bagging의 한계점을 보완하였다.

02. Random Forest

Random Forest의 트리 형성 과정

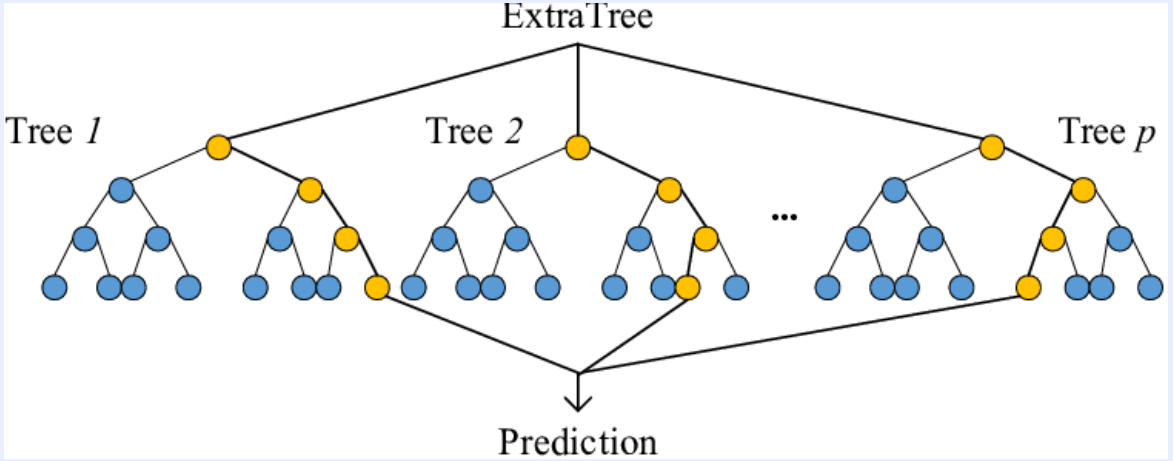
1) 각 모델이 무작위로 feature를 샘플링한다. 이때, 복원 추출을 사용한다.
2) 샘플링된 feature N개로 분할될 수 있는 모든 경우의 수 중 최적의 분할 방법을 찾는다. (= 정보 이득이 가장 높은 분할을 찾는다.)
3) 2번과 3번을 반복하며 선택된 최적의 분할 방법에 따라 각 모델이 트리를 생성하고 학습시킨다.

※ version 1.1 이후 ExtraTrees와 비슷한 방식으로 트리를 형성하도록 업데이트됨.



03. Extra Trees

(Extremely Randomized Trees)



Extra Trees란?

Random Forest와 매우 비슷하게 동작하지만 Random Forest 보다 조금 더 극단적으로 Random 하게 만든 모델

- 비복원 추출로 Bootstrap을 사용하지 않고 Bagging 모델이 아니다.

Extra Trees의 트리 형성 과정

1) 각 모델이 무작위로 feature를 샘플링한다. 이때, 비복원 추출을 사용한다.
2) 샘플링된 feature를 N개라고 할 때, 그 중 랜덤으로 $\sqrt{N}$개씩 분할한다.
3) $\sqrt{N}$개씩 분할된 것 중 최적의 분할 방법을 갖는 트리를 찾는다.
4) 선택된 최적의 분할 방법에 따라 각 모델이 트리를 생성하고 학습시킨다.

ExtraTrees의 장점

➡ 개별 Tree의 성능은 Random Forest보다 낮지만 많은 트리를 앙상블하기 때문에 과대적합을 막고 일반화 성능을 높일 수 있다

➡ Random Forest 보다 무작위성을 더 부여하여 연산량이 적고 속도가 빠르다.

Reference

02. Random Forest

RandomForest 이미지 출처
: https://www.tibco.com/sites/tibco/files/media_entity/2021-05/random-forest-diagram.svg

03. Extra Trees

ExtraTrees 이미지 출처
: https://www.researchgate.net/publication/346995264/figure/fig1/AS:969705405812741@1608207193473/The-structure-of-ExtraTree.png

---

폰트

네이버 글꼴 모음 _ 나눔 스퀘어 사용
출처 : https://hangeul.naver.com/font

D&A
ML Session 6차시 Bagging
Thank You.

2022 / 01 / 31
D&A 운영진 윤경서