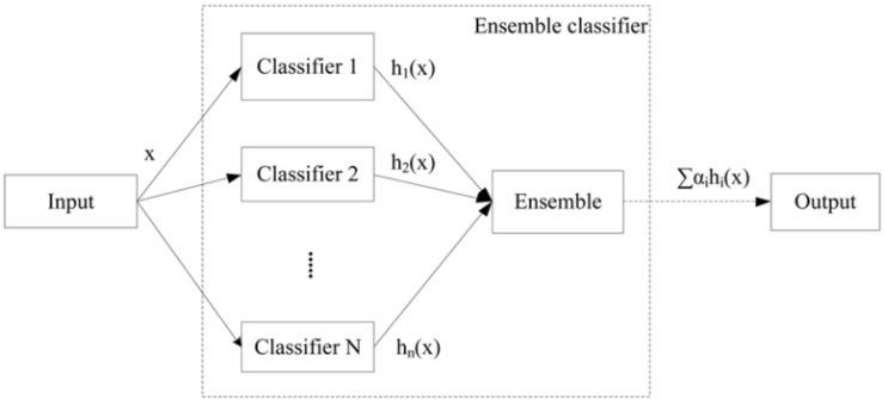
앙상블, 랜덤 포레스트

학부연구생 이현재

앙상블 (Ensemble)

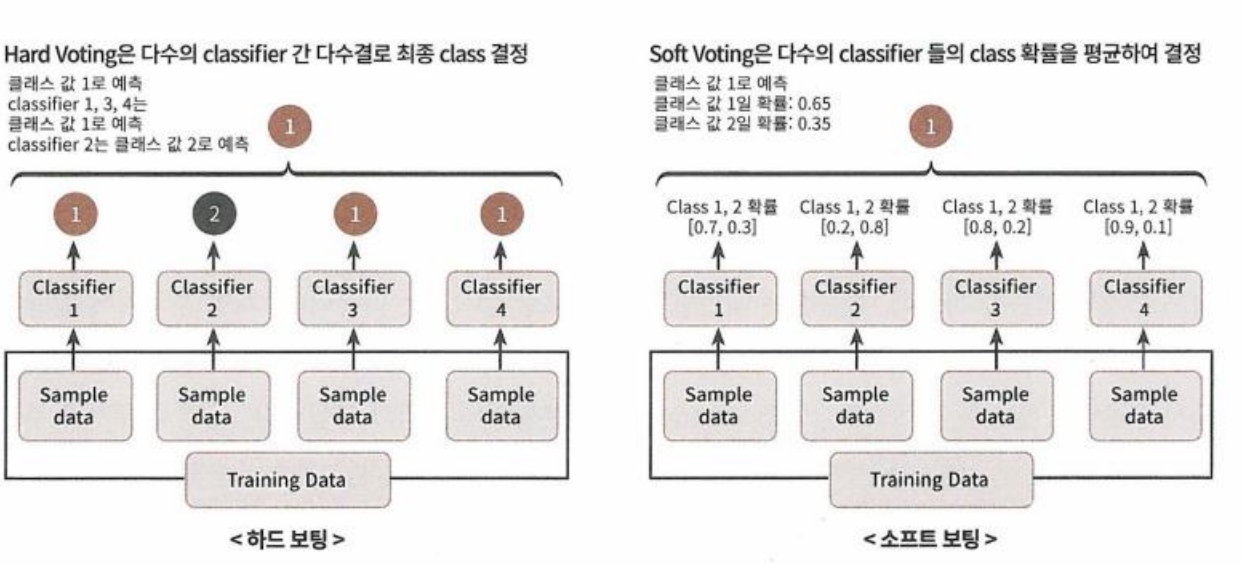
- 정의 : 여러 개의 분류기(Classifier)를 생성하고 그 예측을 결합함으로써 보다 정확한 최종 예측을 도출하는 기법
- 학습의 유형은 보팅(Voting), 배깅(Bagging), 부스팅(Boosting) 그리고 스태킹(Staking)
- 장점
  - 과적합 감소 효과(배깅, 과적합이 쉽다는 결정 트리의 단점을 많은 분류기를 결합해 다양한 상황을 학습하게 함으로써 극복)
  - 다양한 관점을 가진 알고리즘이 서로 결합해 전체 성능을 향상



앙상블 모델 기반의 classifier

보팅 (Voting)

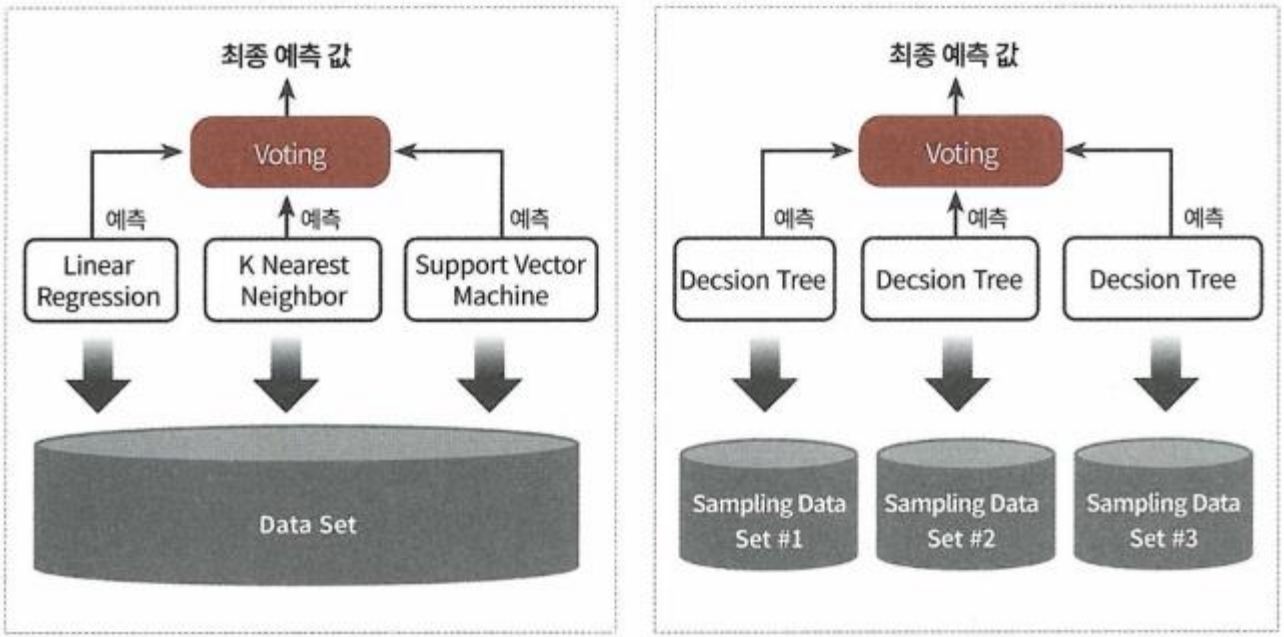
- 정의 : 서로 다른 알고리즘(모델)으로 예측하고 예측한 결과를 가지고 투표를 통해 최종 예측을 선정
- 하드 보팅(Hard Voting) : 예측한 결과값 중 다수의 분류기가 결정한 예측값을 최종 결과값으로 선정 (다수결 원칙)
- 소프트 보팅(Soft Voting) : 분류기들의 레이블 값 결정 확률을 모두 더하고 이를 평균해서 이들 중 확률이 가장 높은 레이블 값을 최종 보팅 결과값으로 선정
- 일반적으로 소프트 보팅이 보팅 방법으로 적용(하드 보팅보다는 소프트 보팅이 예측 성능이 더 좋음)



Prof. Hyerim Bae (hrbae@pusan.ac.kr)

배깅 (Bagging)

- 정의 : 각각의 분류기가 모두 같은 유형의 알고리즘 기반이지만, 데이터 샘플링을 서로 다르게 가져가면서 학습을 수행해 부팅을 수행하는 것
- 대표적인 알고리즘이 랜덤 포레스트
- 분산을 줄여 과적합 억제
- 부트스트래핑(Bootstrapping) : 개별 분류기에 할당된 학습 데이터는 원본 학습 데이터를 샘플링해 추출하는데, 이렇게 개별 분류기에 데이터를 샘플링해서 추출하는 방식(중첩을 허용)

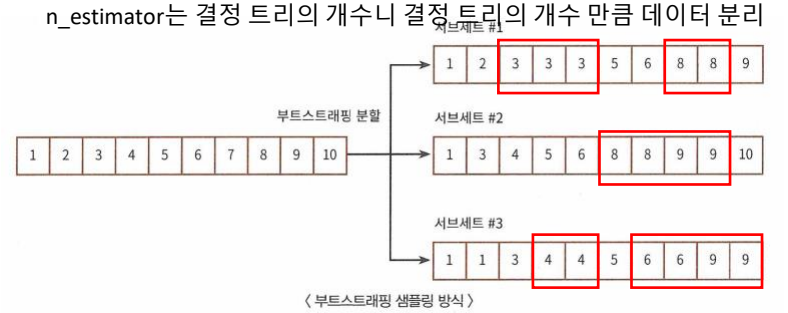


Prof. Hyerim Bae (hrbae@pusan.ac.kr)

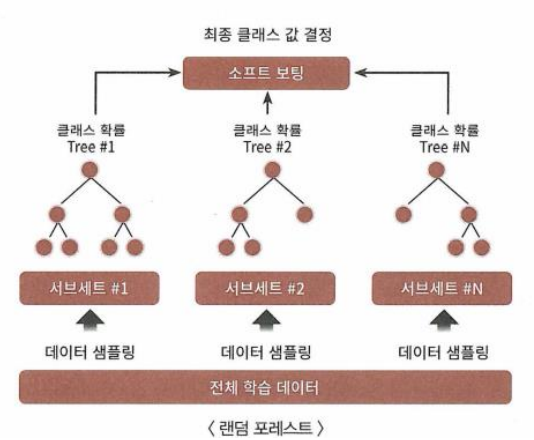
랜덤 포레스트 (Random Forest)

- 결정 트리(Decision Tree)가 모여 랜덤 포레스트(Random Forest)를 구성(여러 개의 결정 트리를 통해 랜덤 포레스트를 만들면 오버피팅 되는 단점을 해결)
- 앙상블 알고리즘 중 비교적 빠른 수행 속도를 가지고 있으며, 다양한 영역에서 높은 예측 성능을 보임
- 여러 개의 결정 트리 분류기가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링해 개별적으로 학습을 수행한 뒤 최종적으로 모든 분류기가 보팅을 통해 예측 결정
- 부트 스트래핑(Bootstrapping) : 전체 데이터에서 일부가 중첩되게 분리하는 방법
  - : 중첩된 부분

n_estimator=3으로 하면 다음과 같이 3개의 데이터 서브세트가 만들어진다.
n_estimator는 결정 트리의 개수니 결정 트리의 개수 만큼 데이터 분리



〈부트스트래핑 샘플링 방식〉



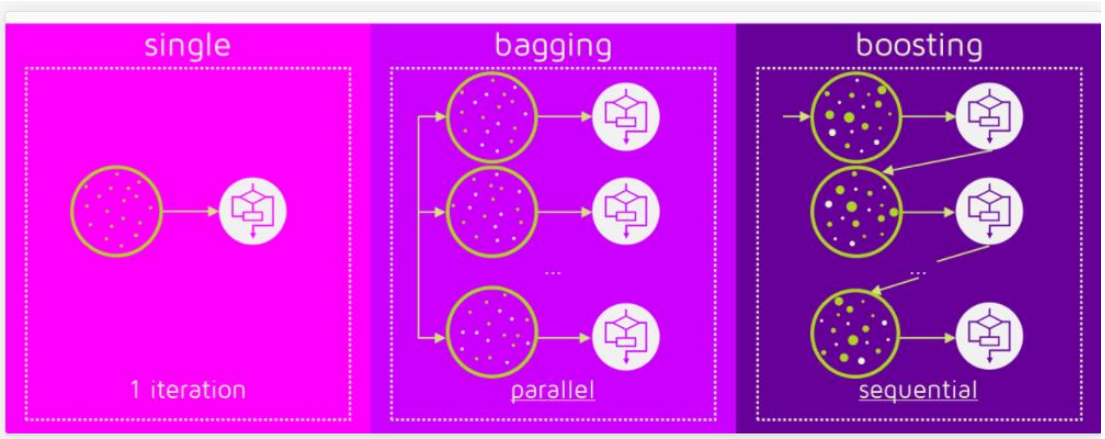
〈랜덤 포레스트〉

- 랜덤 포레스트 하이퍼 파라미터(sklearn)
  - n_estimator : 랜덤 포레스트에서 결정 트리의 개수
  - max_features : 최적의 분할을 위해 고려할 최대 피처 개수

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

부스팅 (Boosting)

- 정의 : 여러 개의 분류기가 순차적으로 학습을 수행하되, 앞에서 학습한 분류기가 예측이 틀린 데이터에 대해서는 올바르게 예측할 수 있도록 다음 분류기에게는 가중치를 부여하면서 학습과 예측을 진행(편향 감소)
- 예측 성능이 뛰어나 앙상블 학습을 주도(XGBoost, LightGBM)
- 오답에 대해서는 높은 가중치를 부여하고, 정답에 대해서는 낮은 가중치 부여(오답을 정답으로 맞추기 위해 더 집중)
- 배깅과 부스팅 차이



  - 배깅은 병렬로 학습하는 반면, 부스팅은 순차적으로 학습
  - 부스팅은 배깅에 비해 error가 적음. 하지만 속도가 느리고 오버피팅될 가능성 있음.

    -> 개별 결정 트리의 낮은 성능이 문제라면 부스팅이 적합하고, 과적합 문제라면 배깅이 적합

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

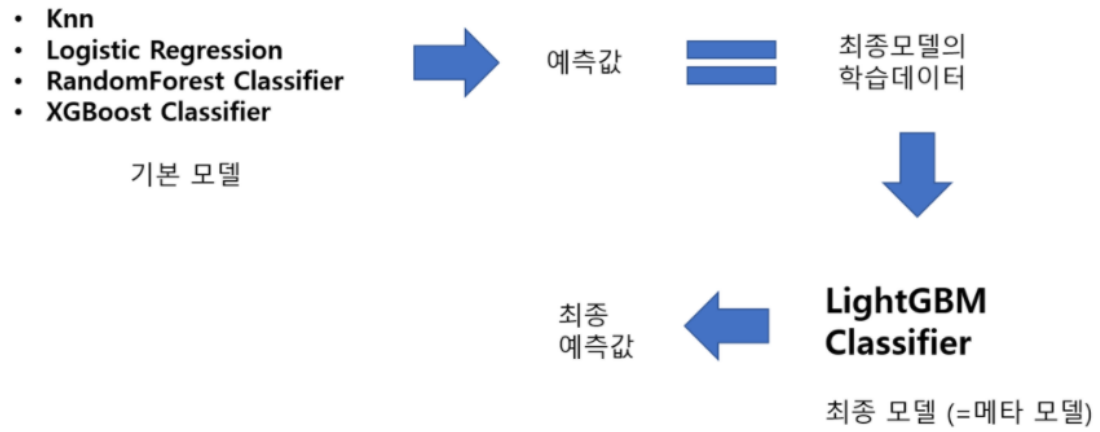
부스팅 모델 (Boosting Model)

- 에이다부스트(AdaBoost)
  - 오류 데이터에 가중치 부여를 통해 오류를 개선해 나가면서 학습하는 방식
- GBM(Gradient Boost Machine)
  - 에이다부스트와 유사하나, 가중치 업데이트를 경사 하강법을 이용
- XGBoost(eXtra Gradient Boost)
  - GBM에 기반하고 있지만, GBM의 단점인 느린 수행 시간 및 과적합 규제 부재 등의 문제를 해결
  - 최고의 분류 모델
  - 빠른 속도와 높은 효율
- LightGBM
  - XGBoost보다 학습에 걸리는 시간이 훨씬 적고 메모리 사용량도 적음
  - 적은 데이터 세트에 적용할 경우 과적합이 쉽게 발생한다는 단점
  - 리프 중심 트리 분할

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

스태킹 (Staking)

- 정의 : 여러 가지 다른 모델의 예측 결과값을 다시 학습 데이터로 만들어서 다른 모델(메타 모델)로 재학습 시켜 결과를 예측하는 방법



스태킹 예시

- 현실 모델로 많이 사용되지 않지만 성능이 올라가는 경우가 있기 때문에 캐글이나 데이콘에서 주로 사용하는 모델

앙상블 기반 모델 코드 구현

https://github.com/LeeYunseol/Lab_study/blob/main/%EC%95%99%EC%83%81%EB%B8%94%2C%EB%9E%9C%EB%8D%A4%ED%8F%AC%EB%A0%88%EC%8A%A4%ED%8A%B8/%EC%95%99%EC%83%81%EB%B8%94%2C%20%EB%9E%9C%EB%8D%A4%ED%8F%AC%EB%A0%88%EC%8A%A4%ED%8A%B8.ipynb

참고 자료

- Pytorch로 시작하는 딥러닝 입문
- 귀퉁이 서재 블로그(https://bkshin.tistory.com/)