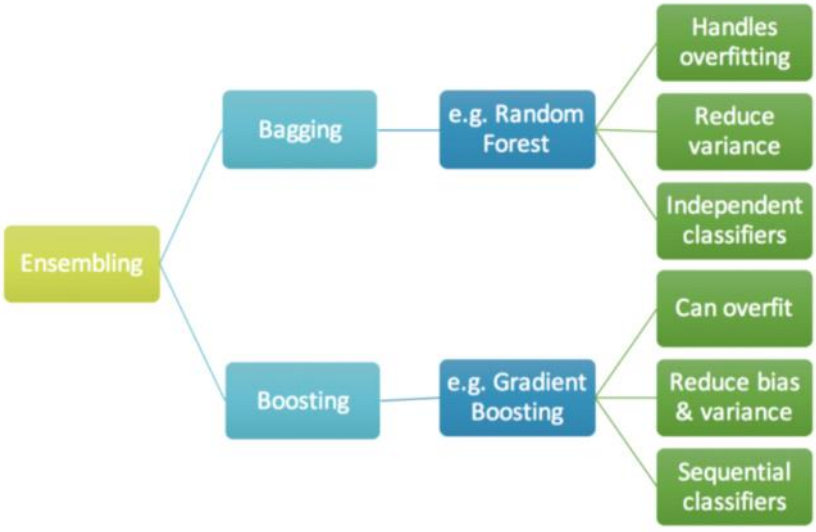
Ensemble Learning (앙상블 기법)

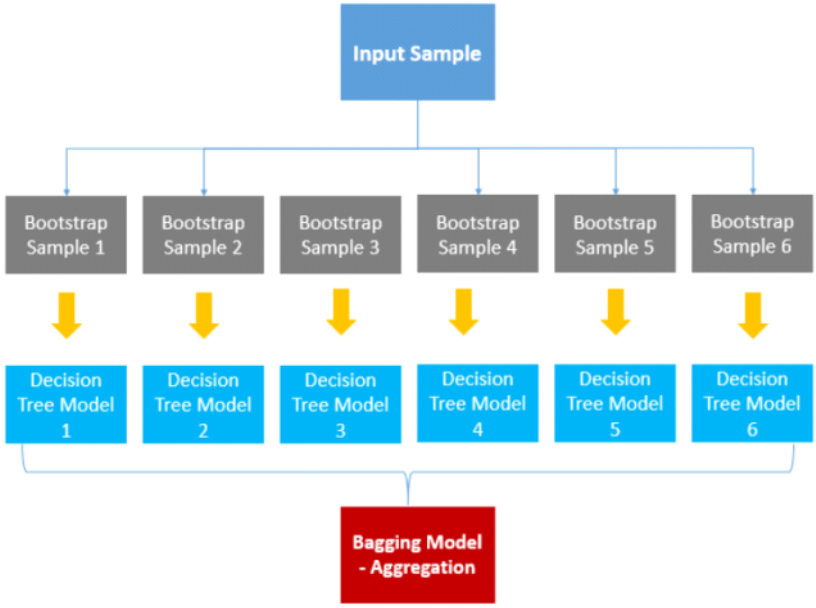
2020년 7월 10일 금요일 오후 2:03

앙상블이란 여러 개의 약한 분류기(weak learners)를 이용해 최적의 답을 찾아내는 기법이다.
주요 앙상블 알고리즘은 bagging과 boosting으로 나눌 수 있다.

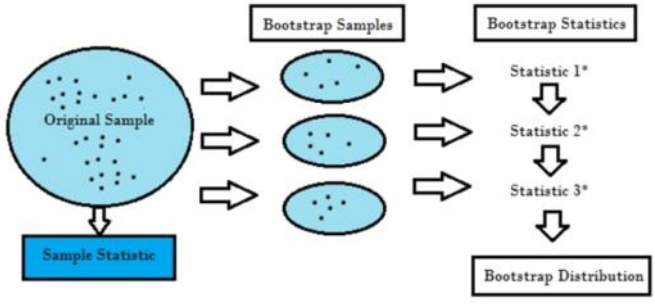


Bagging (배깅)

Bagging은 Bootstrap Aggregation의 약자로, bootstrap sampling을 통해 여러 개의 샘플을 추출한 후, 각 모델을 학습시켜 결과물을 집계(aggregate)하는 방법이다.
가장 대표적인 배깅 기법은 랜덤 포레스트(Random Forest)이다.



Bootstrap sampling은 랜덤 복원 추출기법으로 데이터가 많지 않은 경우에 유용하다.
Categorical data -> Voting
Numerical data -> Average

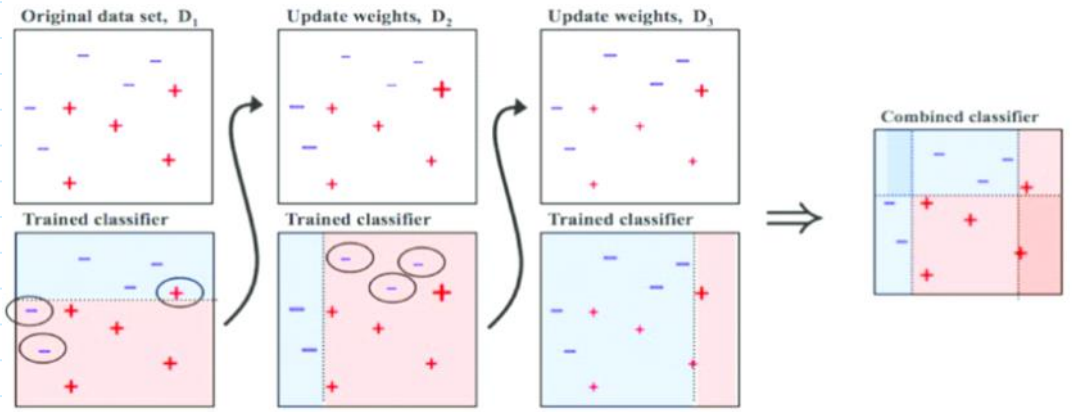


Boosting (부스팅)

Boosting은 가중치를 활용하여 약한 분류기를 강한 분류기로 만드는 방법이다.

여러 개의 독립적인 모형을 집계하여 최종 결과를 예측하는 Bagging 방법과 달리, Boosting은 모델 간 팀워크를 통해 최종 결과가 도출된다.

처음 모델이 예측을 하면 그 결과에 따라 데이터에 가중치가 부여된다. 이렇게 부여된 가중치가 다음 모델에 영향을 주게 되고, 잘못 분류된 데이터에 집중하여 새로운 분류 규칙을 만드는 과정을 반복한다. 이런 과정을 거쳐 만들어진 분류기들을 모두 결합하여 최종 모형을 만드는 것이 Boosting 기법이다.



Bagging (배깅)과 Boosting (부스팅)의 차이

Bagging은 병렬로 학습하는 반면, Boosting은 순차적으로 학습한다.

Bagging 기법을 활용하면 단일 모형을 사용했을 때보다 분산을 줄이고 과적합을 피할 수 있다.

Boosting은 한번 학습이 끝난 후 결과에 따라 가중치를 부여하고, 이는 다음 모델의 결과에도 영향을 준다. 오답에 대해서는 높은 가중치를 부여하고, 정답에 대해서는 낮은 가중치를 부여한다. 따라서 오답을 줄이는 데에 집중하게 된다. 하지만, 그 만큼 outlier에 취약하다는 단점이 있다.

Boosting은 Bagging에 비해 error가 적다. 즉, 성능이 좋다. 하지만 계산량이 많으므로 속도가 느리고 overfitting(과적합)이 될 가능성이 높다.