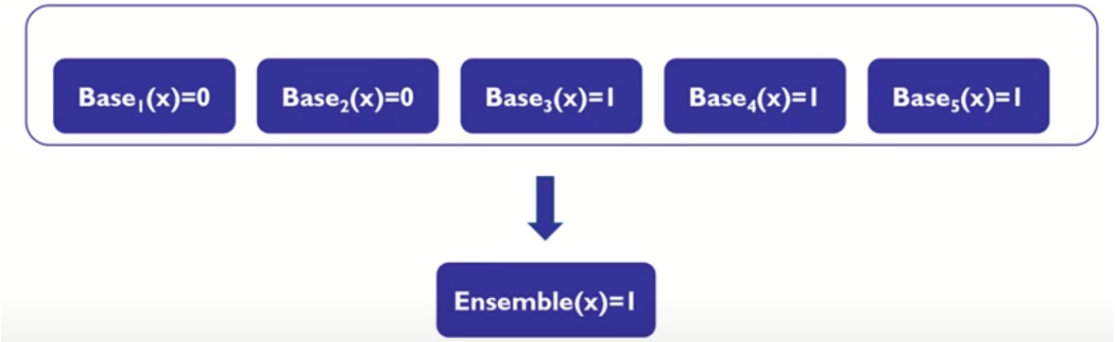
출처가 명시되지 않은 모든 자료(이미지 등)는 조성현 강사님 블로그 및 강의 자료 기반.

<< 머신러닝-앙상블 >>

[앙상블(Ensemble)]

하나의 알고리즘을 사용할 때보다 여러 알고리즘을 사용한 후 결과를 종합하는 것이 더 좋다는 아이디어에서 출발한다.

> *참고* : 김성범 교수님 강의
>
> 
>
>
> 여러 단원들이 모여 하나의 화음을 만들어 내는 오케스트라의 '앙상블'처럼, 여러 모델들의 예측을 다수결 법칙 또는 평균을 이용해 통합하여 예측 정확성을 향상시키는 방법을 말한다.

다만, 앙상블 모델을 통해 하나의 모델보다 우수한 성능을 내기 위해서는 다음의 두 가지 조건이 만족되어야 한다.

- 앙상블의 개별 모델들이 최대한 서로 독립적이어야 한다.
- 앙상블을 이루는 개별 모델들이 무작위 예측을 수행할 때보다 성능이 좋아야 한다.

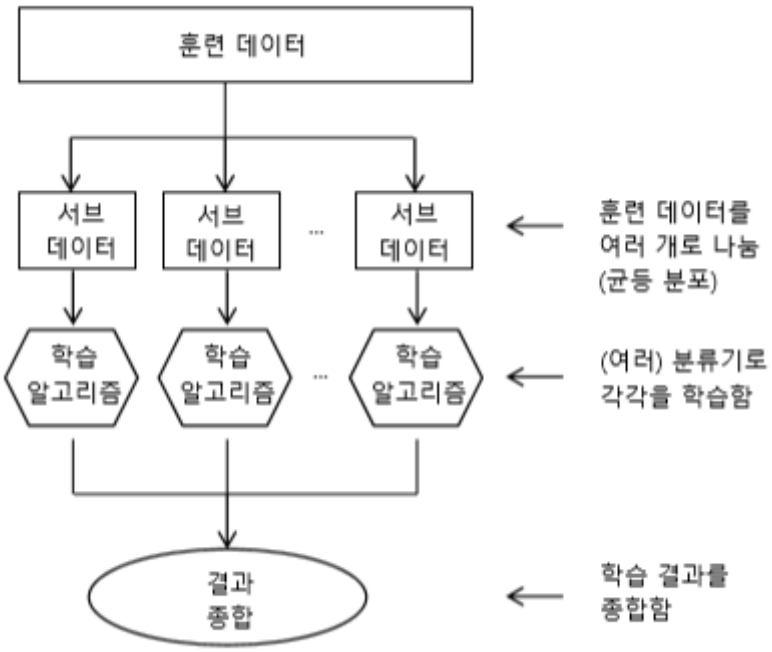
위와 같은 조건을 만족하는 경우, 여러 모델의 예측 결과를 종합하면 그 정확도와 일반화 특성이 향상된 앙상블 모델을 구성할 수 있다.

앙상블 모델을 구축하는 과정은 다음과 같다.

- 샘플링: 훈련 데이터를 여러 개의 서브 샘플 데이터로 나눈다.
- 학습: 각각의 서브 샘플 데이터를 서로 다른 알고리즘으로 학습하여 개별 모델을 구축한다.
- 예측: 개별 모델의 예측 결과를 종합 혹은 평균하여 최종 예측 결과를 결정한다.

그리고 위의 과정을 어떻게 수행하는지에 따라 `Bagging`과 `Boosting`의 두 가지로 구분된다.

1. Bagging(*a.k.a* Bootstrap Aggregation)



`Bootstrap`과 `Aggregating`의 단계를 거쳐 모델을 생성한다.

- `Bootstrap` 방법으로 샘플링하여 sub 훈련 데이터를 생성하고,
- 각 데이터를 각각 모델링한 뒤, `Aggregating` 기법에 의해 각 모델의 예측 결과를 집계한다.

샘플링에 의한 결합 방식으로 작동한다고 보면 된다.

BootStrap

`Bootstrap` 이란 샘플링 기법 중 `단순복원 임의추출` 기법을 의미한다. ~~*(응?)*~~

무슨 말인고 하니,

- 한 번 뽑고 다시 집어 넣고(*중복 허용*),
- 원래 데이터의 수만큼 크기를 갖도록,

표본을 추출한다는 의미다.

똑같은 데이터에 대해, 똑같은 수의 크기를 갖도록 서브 샘플을 형성하기 때문에, 각각의 sub training data는 기존의 훈련 데이터와 **균일**한 확률 분포를 따른다. 몇 번의 샘플링을 진행하더라도 동일하다.

실제로 실행해 보면, 위의 그림 중 왼쪽처럼 여러 번 선택되는 데이터가 있을 수 있다. 반대로 생각해 보면, *이론적으로는* 한 데이터가 한 번도 추샘플링되지 않을 수도 있음을 의미한다. `Bootstrapping`의 한계이다.

Aggregating

이제 `Aggregating`의 개념이 들어 온다. `Bootstrap`으로 뽑아 온 데이터를 집계(*aggregate*)한다는 것이다. 이 때 어떤 방식으로 합치는가에 따라 방법이 달라진다.

1. Majority Voting

$$Ensemble(\hat{y}) = argmax_i(\Sigma_{j=1}^{n} I(\hat{y_i}))$$

직관적이다. 여러 모델 중 다수결 투표 원리에 의해 더 많이 예측된 라벨로 분류한다.

2. Weighted Voting

$$Ensemble(\hat{y}) = argmax_i(\frac{\Sigma_{j=1}^{n} TrainAcc_j * I(\hat{y_i})}{\Sigma_{j=1}^{n} TrainAcc_j})$$

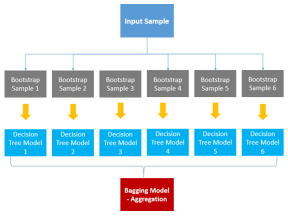
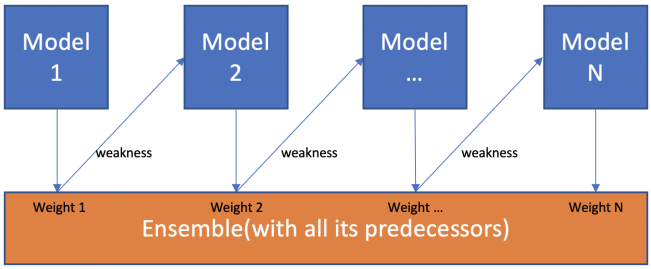
훈련 데이터에 대한 정확도에 가중치를 두어 가중평균을 내자는 원리다. 분류 결과를 맹신하는 것이 아니라, 분류 결과를 도출한 훈련 과정의 정확도까지 고려하자는 아이디어다.

2. Boosting

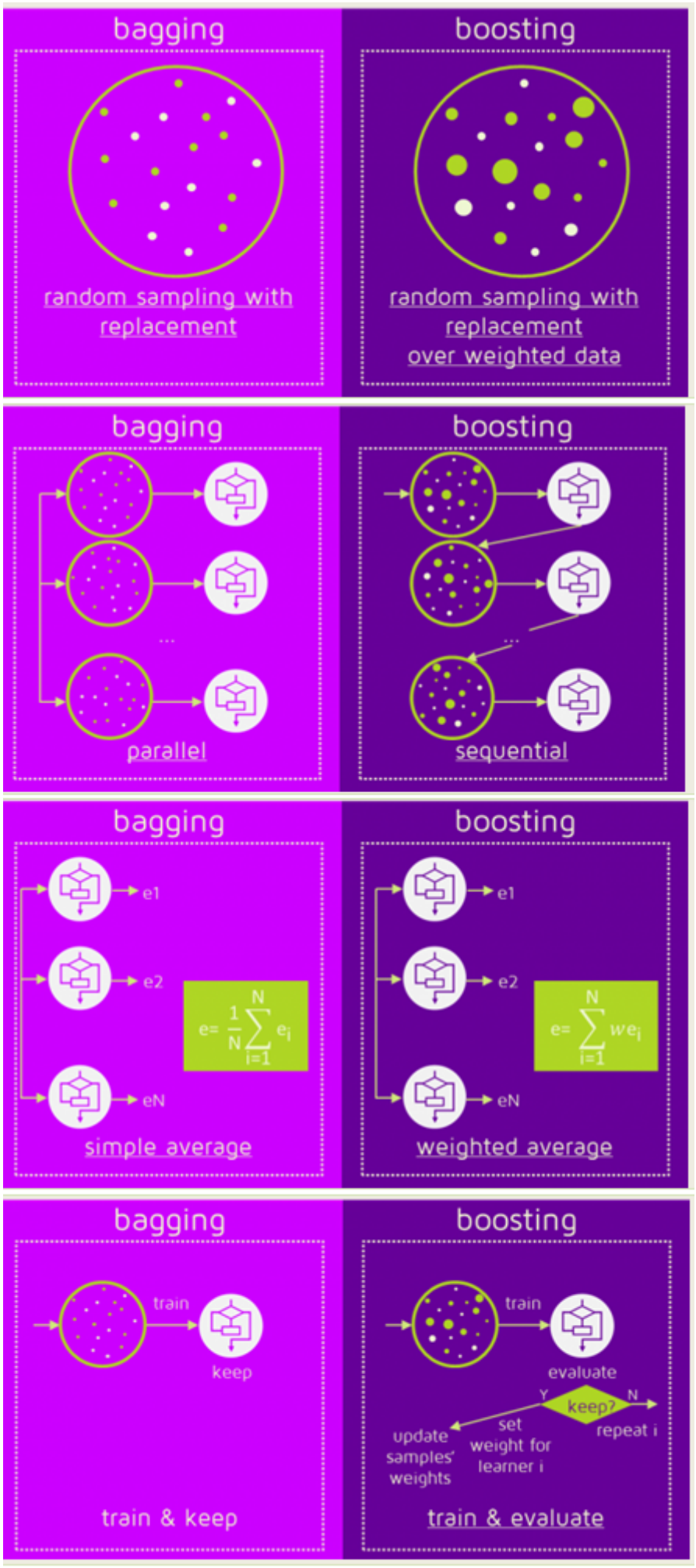
부스팅 기법은 배깅과 달리 예측이 어려운 데이터에 더욱 집중하며 정확도를 만드는 방법이다. 말하자면 약한 모델을 점점 강화시켜 나가는 방식이다.

배깅과 마찬가지로 `Bootstrap` 기법을 활용하여 서브 훈련 데이터 샘플을 만든다. 그러나 배깅과는 달리, 이후 학습을 진행해 나가며, 잘못 분류된 데이터의 샘플링 가중치가 높아진다. 동일한 샘플링 기법이더라도, 잘못 분류된 샘플일수록 이후 학습의 샘플링 과정에 선택될 확률이 높아진다. 이렇게 새롭게 샘플링된 데이터에 대해 학습을 반복하며, 예측이 어려운 패턴들이 점점 더 많이 선택된다.

3. 두 기법 비교

	Bagging	Boosting
학습 순서	병렬 	순차적 Model 1,2,..., N are individual models (e.g. decision tree) 
모델 간 관계	독립적 각 모델의 결과를 집계하여 예측한다.	영향 있음 현재 모델의 가중치가 다음 모델의 샘플링에 전달된다.
샘플링	균일한 확률분포	예측이 어려운 데이터에 집중
강점	과적합에 강함	정확도가 높음
약점	특정 영역에서 정확도 낮음	Outlier(이상치, 결측치)에 취약
대표 모델	랜덤 포레스트	AdaBoost, GradientBoost, xgBoost

출처: https://joyfuls.tistory.com/61



출처: [xwMOOC 기계학습](https://statkclee.github.io/ml/ml-tree-model.html#1_나무기반_모형_1_2)