앙상블모형

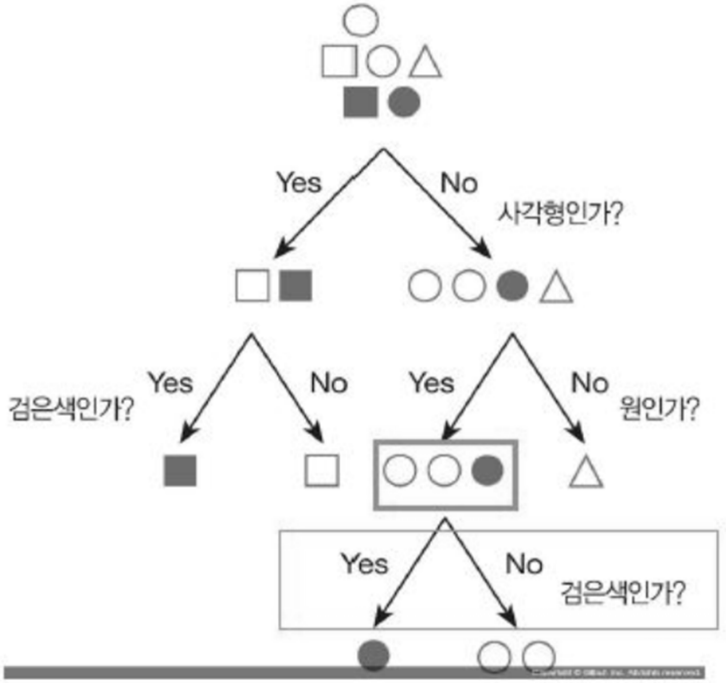
- 의사결정나무
- 앙상블 모형
- **앙상블 (Ensenble) 모형-랜덤 포레스트(Random Forest)**
- **앙상블 (Ensenble) 모형 배깅 (bagging)**
- **앙상블 (Ensenble)-부스팅(Boosting)**

소개

■ 의사결정나무란?
1. 기준 사각형인지 아닌지
2. 원인지 아닌지
3. 마지막으로 검은색인지 아닌지 판별하면 분류 끝남

흰색 삼각형도 똑같은 질문으로 분류 가능

■ 목적
- 세분화: 각 고객이 어떤 집단에 속하는지 파악
- 분　류: 여러변수 근거해 반응 변수의 범주를 분류하고자 하는 경우
- 예　측: 데이터에서 규칙을 찾아내어 미래의 사건을 예측하는 경우

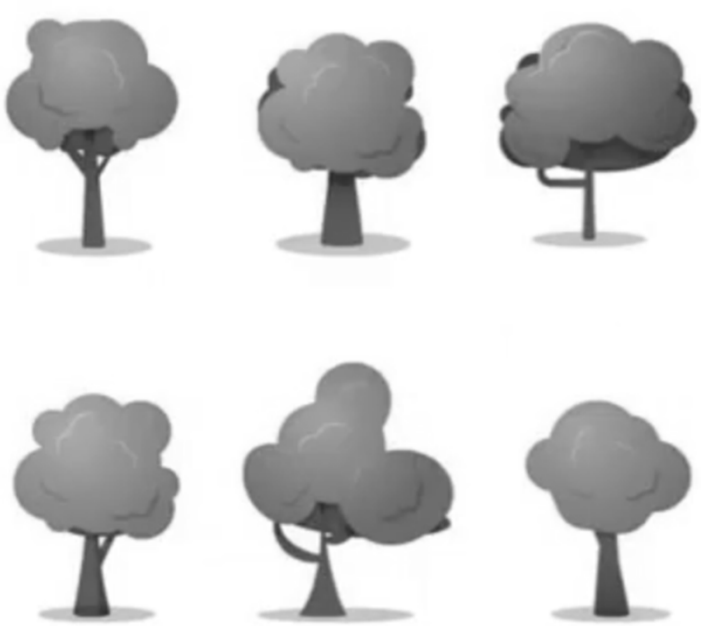


의사결정나무(Decision Tree)

■ 의사결정나무(Decision Tree)
- 거대한 나무

■ Randomforest
- 여러 개의 작은 나무
- 각자 모양이 다름

decision tree

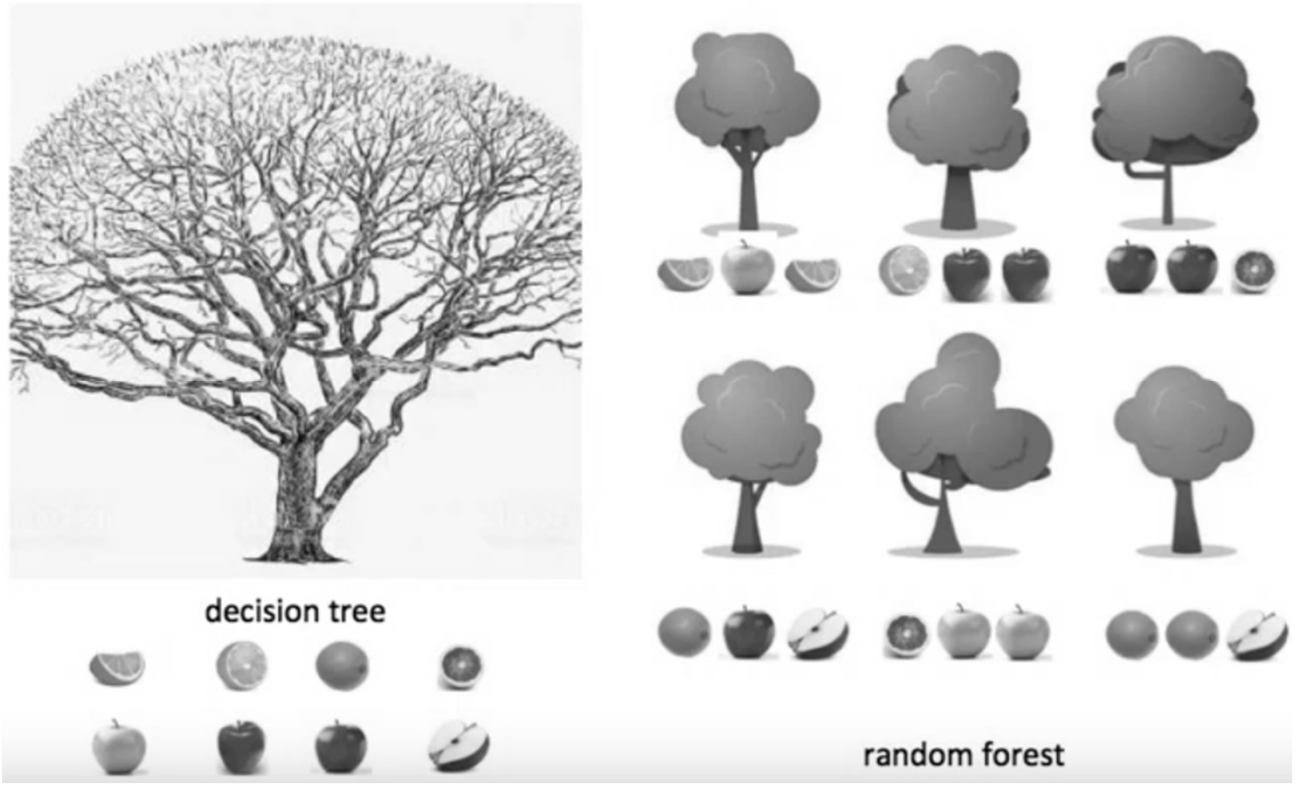


random forest

Decision Tree VS Randomforest

■ 의사결정나무(Decision Tree)
- 가지고 있는 데이터 모두 사용
- Overfiting 발생
- 사과와 나머지로 구분

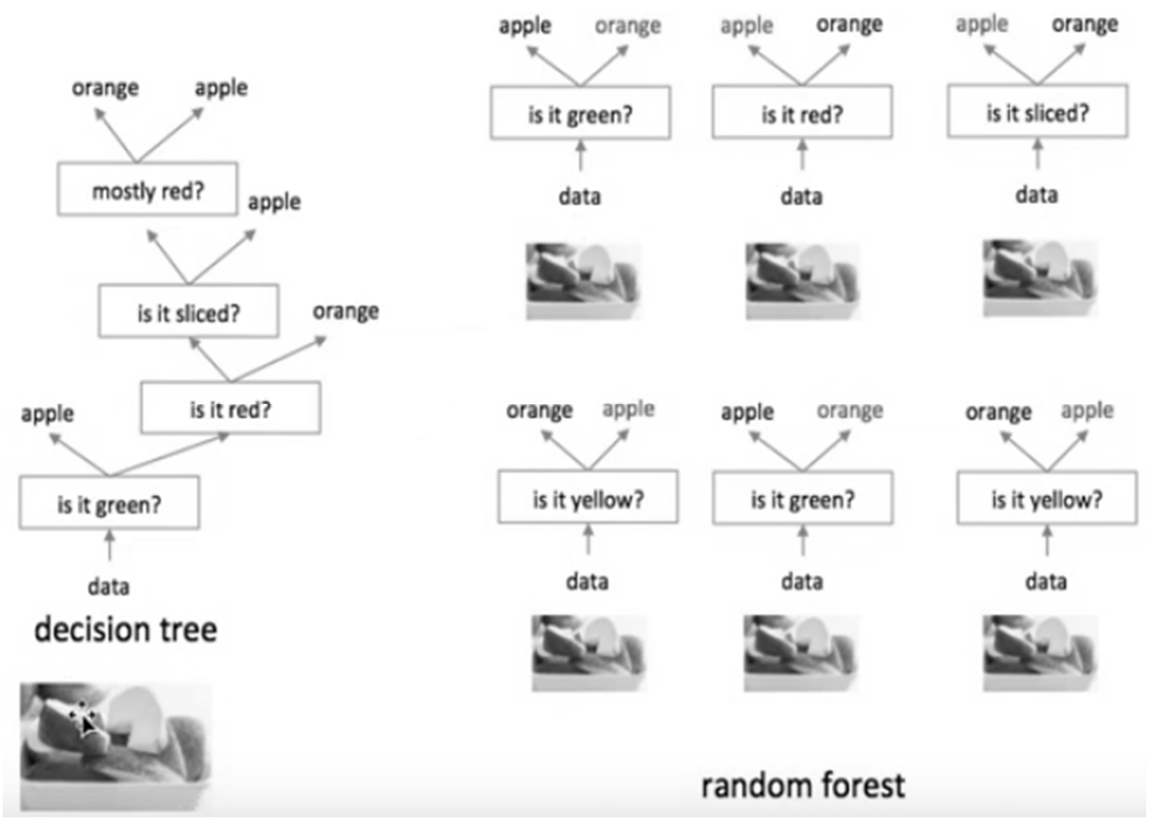
■ Randomforest
- 6개의 작은 Tree
- 중복된 데이터 허용(boosting)
- 질문이 랜덤함



Decision Tree VS Randomforest

■ 의사결정나무(Decision Tree)
- 녹색?
- 과일이 잘렸는가?
- 틀린 답

■ Randomforest
- 질문이 랜덤
- Aggregating Result →투표 통해 선택
- boosting과 Aggregating →Bagging이라 표현



Decision Tree VS Randomforest

■ 앙상블(ensemble) 모형이란 여러 개의 분류 모형에 대한 결과를 종합하여 한 데이터를 분류하는 모형

■ 결과의 종합은 다수결이나 기타 방법을 사용

■ 한 모형으로 데이터를 분류할 때보다 오분류를 많이 줄일 수 잇다.
- 만일 5개의 독립적인 분류모형이 각각 5%의 오분류를 가질 때 앙상블 모형의 오분류율은 다음과 같다.

$$e_{ensemble} = \sum_{i=3}^{5} \binom{5}{i} (0.05)^i (1-0.05)^{5-i} = 0.0001$$

앙상블 (Ensenble) 모형

■ 각각의 분류기로 어떠한 분류모형도 앙상블 모형에 적용 가능

■ 한 가지 분류모형에서도 여러 개의 분류기를 만들 수 있다.

1) 데이터를 조절하여 여러 분류기를 만드는 방법
 - 배깅 (Bagging)
 - 부스팅 (Boosting)

2) 변수의 수를 조절하는 방법
 - 랜덤 포레스트 (Random Forest)

3) 집단의 종류가 많은 경우 소집단으로 묶는 방법

4) 분류모형의 가정을 조절하는 방법
 - 모수에 대한 가정 도는 알고리즘의 변화

앙상블 (Ensenble) 모형

– 2001년에 Leo Breiman에 의해 처음 소개된 기법/의사결정 트리 단점 개선 위한 알고리즘 중 가장 지배적인 알고리즘

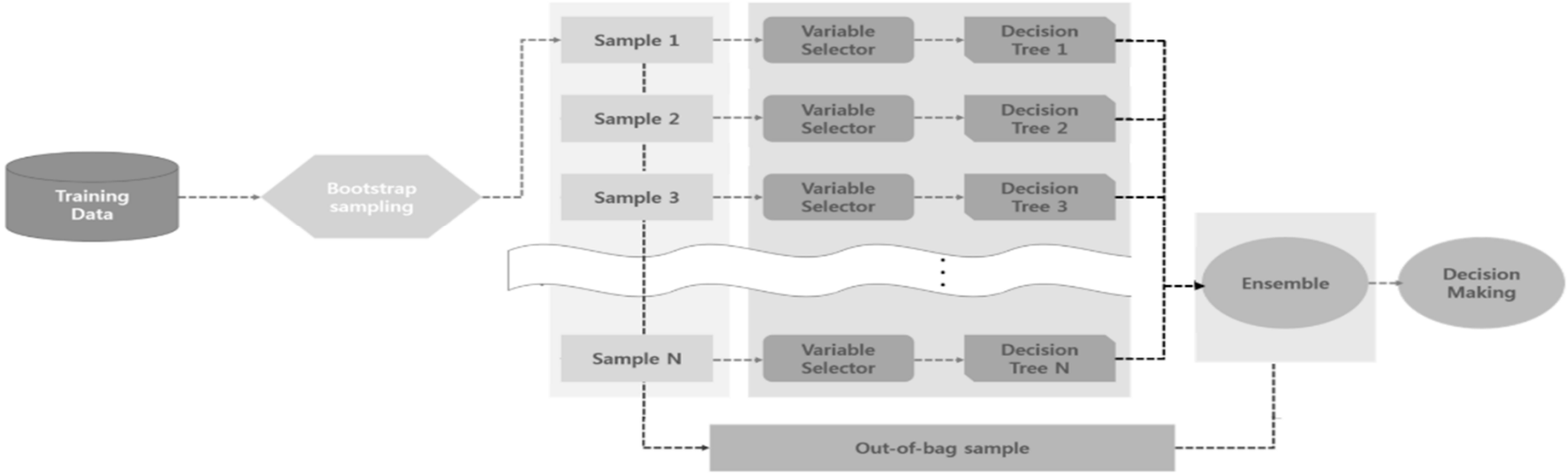
– 부트스트랩 (bootstrap)표본을 이용할 뿐 아니라
변수에 임의성을 더한 방법을 말하며 앙상블 이론이 갖는 장점을 극대화하여
예측 및 분류 정확도를 기존의 방법보다 개선하며 안정성을 얻음.

– 랜덤포레스트의 성능평가는 OOB(Out-Of-Bag) error라는 수치로 파악함.
- Out-of-Bag 샘플은 부트스트랩 샘플링 과정에서 추출되지 않은 데이터들을 말하며
이 데이터들을 따로 랜덤포레스트로 학습을 시켜 나온 결과가 OOB error임.

* OOB예측방법이 테스트셋을 사용하여 검증하는 것 만큼 정확하므로 따로 테스트 셋을 구성할 필요가 없어졌고
OOB샘플들은 주로 평가용 데이터에서의 오분류율을 예측하는 용도 및 변수 중요도를 추정하는 용도로 많이 이용됨.

앙상블 (Ensenble) 모형-랜덤 포레스트(Random Forest)

– 랜덤포레스트는 데이터에서 부트스트래핑 과정을 통해 N개의 샘플링 데이터 셋을 생성함.
– 각 샘플링된 데이터 셋에서 임의의 변수를 선택하는 과정을 진행함. 변수의 갯수를 선택하는 방법은 M개의 총 변수들 중에서 sqrt(M) 또는 M/3개의 개수만큼 변수들을 랜덤하게 선택하고 나머지 변수는 모두 제거하는 과정 반복함.
– 변수선택이 진행된 의사결정트리들을 종합하여 앙상블 모델을 만들고 OOB error를 통해 오분류율을 평가함.
– 랜덤 포레스트(Random forest)의 흐름도는 아래 그림과 같음.

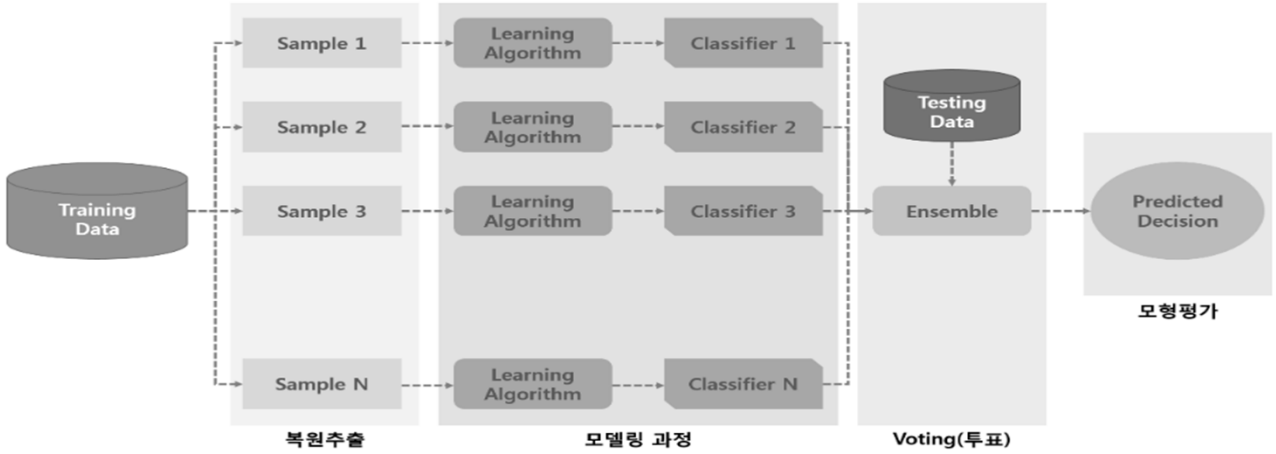


앙상블 (Ensenble) 모형–랜덤 포레스트(Random Forest)

– 의사결정트리는 학습데이터에 따라 매우 달라지기 때문에 일반화하여 사용하는데 어려움이 있음.

– 랜덤포레스트는 변수선택의 임의성과 배깅을 통해 각 트리들의 예측들이 비상관화 되게하여 일반화 성능을 향상시켜 이러한 문제점을 해결하였으며 노이즈가 포함된 데이터에 이용하기 좋음.
– 또한 데이터 셋 내의 데이터 분포가 고르지 않은 경우에 사용됨.

■ 배깅과 공통점
– 배깅의 경우 전체 데이터에서 여러 샘플링 데이터를 추출하여 서로 다른 학습 분류기를 통합하는 방법을 말함.



앙상블 (Ensenble) 모형-랜덤 포레스트(Random Forest)

– 배깅과 랜덤포레스트는 부트스트랩(Bootstrap)과정이 진행된다는 점에서 공통점을 가짐.

* 부트스트랩이란 가설검증(test)을 하거나 계산하기 전에 random sampling을 적용하는 방법을 말함.

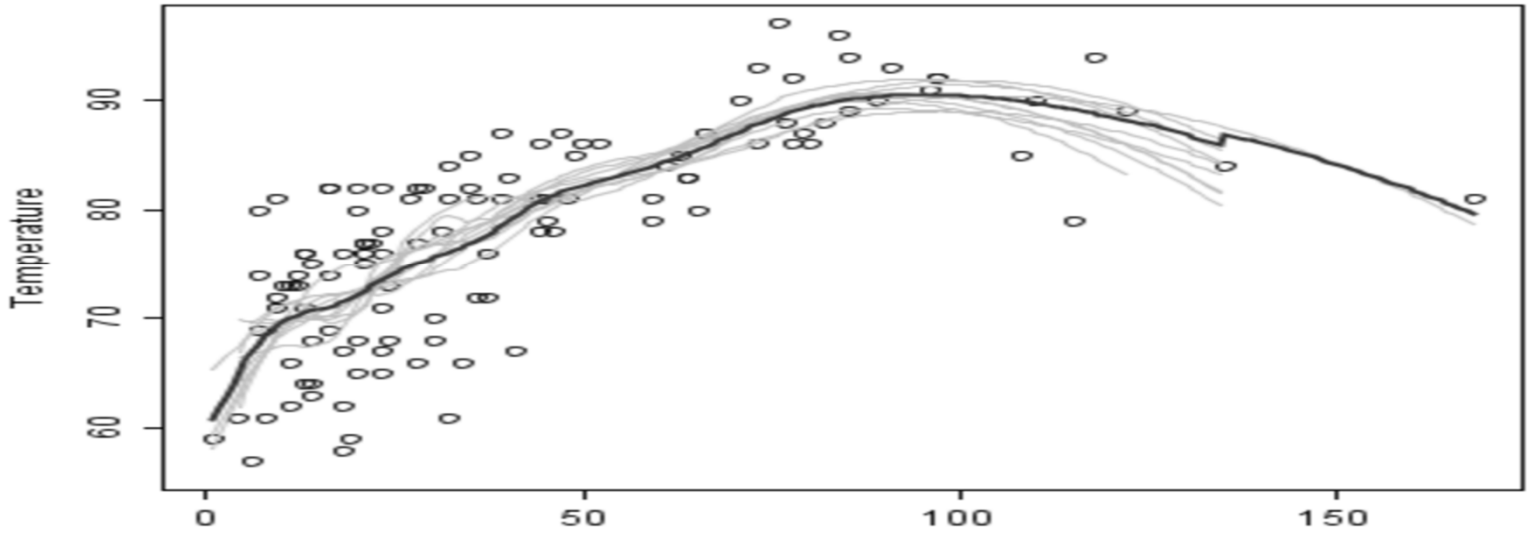
* 배깅과 랜덤포레스트에서 부트스트랩은 두 가지의 이유로 사용되며
  데이터 셋 내의 데이터 분포가 고르지 않은 경우
  과적합을 줄이는 경우를 말함.

* **데이터 셋 내의 데이터 분포가 고르지 않은 경우**
  개와 고양이를 구분하는 분류기를 학습시키는데
  100장의 개 이미지와 1장의 고양이 이미지가 있다면
  분류기는 고양이의 error는 무시하는 방향으로 학습되기 쉬움.
  이런 경우에 부트스트랩을 이용하여 고양이의 데이터 수를 늘려 학습을 시킴.

* **과적합을 줄여 준다는 것**
  부트스트랩을 이용한다며 충분하지 않은 데이터를 이용하여
  여러 개의 앙상블 모델을 만들 수 있고
  이를 통해 각 모델들이 과적합 되어있더라도
  그들을 평균 내어 과적합을 상쇄시켜 일반화된 모델을 만들 수 있음.

앙상블 (Ensenble) 모형-랜덤 포레스트(Random Forest)

– 랜덤 포레스트는 변수선택 과정을 적용했다는 것이 큰 차이점임.
– 랜덤 포레스트가 랜덤하게 변수를 추출하는 이유는 배깅을 통해 얻은 트리들간의 상관성 때문인데, 소수의 변수가 결과에 강학 예측 성능을 보인다면 여러 개의 분류기를 이용하여 학습을 시켜도 결국 소수의 변수가 중복되어 선택되어 결과적으로 트리들이 상관화가 되는 결과를 얻음.
– 즉, 변수선택 과정이 없는 배깅의 경우 지배적인 변수가 포함 되어 있다면 뿌리 노드가 변하지 않아 트리가 유사한 형태로 학습됨.
– 따라서 분류기준의 대표성을 띄는 변수를 찾기 위해 변수선택 과정을 진행함.



앙상블 (Ensenble) 모형–랜덤 포레스트(Random Forest)

■ 배깅(bagging)은 boot strap aggregation의 준말

■ 훈련용 데이터로부터 크기가 같은 표본을 여러 번 단순확률 반복추출 (with replacement)로 분류기를 생성하여 앙상블하는 방법

■ 반복추출법을 사용하기 때문에 한 표본에 같은 데이터가 여러 번 추출될 수도 있고, 어떤 데이터는 추출되지 않을 수도 있다.

■ n개의 데이터가 있을 때 n개의 표본을 단순 확률 반복 추출할 경우 각 데이터가 다시 뽑힐 확률은 1 – (1 – 1 / n)^n → 1 – 1/e = 0632

앙상블 (Ensenble) 모형 배깅 (bagging)

알고리즘 6.4 배깅(Bagging) 알고리즘

단계 1: R = 붓스트랩 표본의 수, n = 표본의 크기

2: for k = 1 to R do

3: 크기가 n인 붓스트랩 표본 D_k를 생성

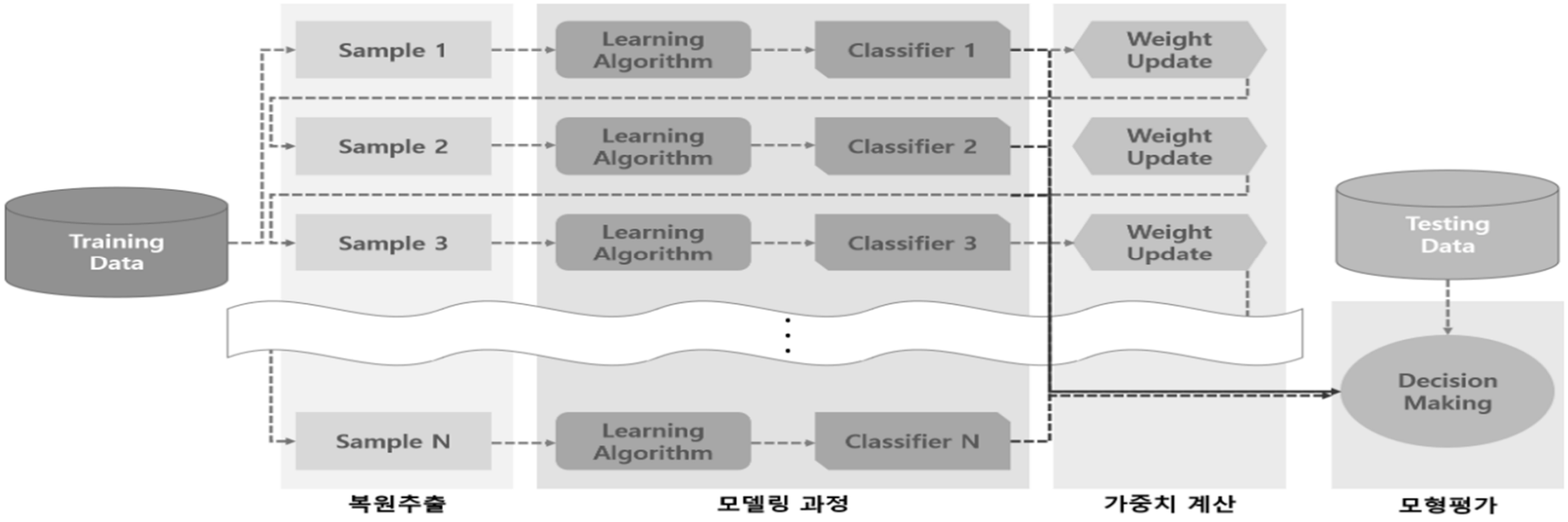
4: 붓스트랩 표본 D_k를 이용하여 분류기 C_k를 만듦

5: end for

6: $C^*(x) = \underset{y}{\operatorname{argmax}} \sum_{k=1}^{R} I(C_k(x) = y)$ {x를 각 분류기가 다수 분류한 집단으로 분류}

앙상블 (Ensenble) 모형 배깅 (bagging)

– Bagging과 유사하게 초기 샘플 데이터를 조작하여 다수의 분류기를 생성하는 기법 중 하나
– 일반적으로 부스팅은 약검출기(weak classifier)들을 여러 개 모아 강검출기(strong clssifier)를 생성하는 방법을 말하고 주로 의사결정나무 모형을 사용함.
– 부스팅(Boosting)의 흐름도는 아래 그림과 같음..



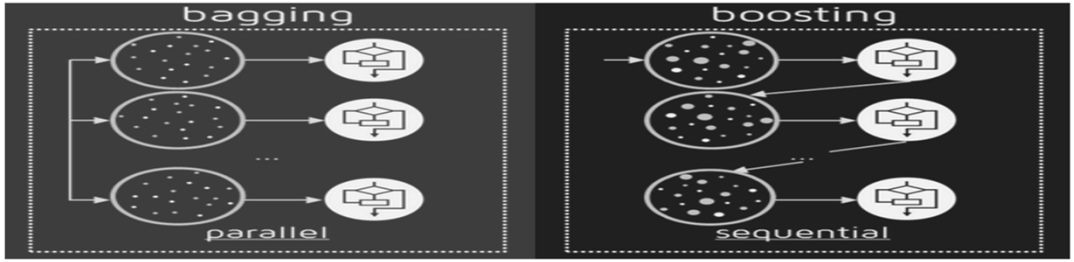
앙상블 (Ensenble)-부스팅(Boosting)

- 전체 데이터에서 여러 샘플링 데이터를 추출
- 순차적으로 이전 학습 분류기의 결과를 토대로 다음 학습 데이터의 샘플 가중치를 조정하면서 학습을 진행
- 특징
  다음 단계의 weak classifier가 이전 단계의 weak classifier의 영향을 받음
  이전의 classifier의 양상을 보고 보다 잘 맞출 수 있는 방향으로 다음 단계를 진행
  각 classifier의 weight를 업데이트 함.
  최종적으로 서로 영향을 받아 만들어진 여러 weak classifier와 서로 다른 weight를 통해 strong classifier를 생성하게 됨.

■ 부스팅(Boosting) 의 목적

– 일반적인 분류 문제는 잘못 분류된 개체들을 더 잘 분류하는 것이 목적
  부스팅은 잘못 분류된 개체들에 집중하여 새로운 분류 규칙을 만드는 것이 목적
– 배깅이 일반적인 모델을 만드는데 집중
  → 부스팅은 맞추기 어려운 문제를 맞추는데 초점이 맞춰져 있음

■ 배깅(Bagging) VS 부스팅(Boosting)



출처 : https://swalloow.github.io

앙상블 (Ensenble)–부스팅(Boosting)

https://www.youtube.com/watch?v=Dhwmd_IyW3g&feature=youtu.be

- **Raw data의 가중치의 유무 차이**

- 배깅

단순 복원 임의 추출법을 통해
raw data로 부터 크기가 동일한 여러 개의 표본 자료를 이용하여 학습

- 부스팅(Boosting)

raw data의 객체들이 동일한 가중치에서 시작
모델링을 통한 예측변수에 의해 오 분류된 객체들에게는 높은 가중치를 부여
가중치를 조정하여 오 분류된 객체들이 더 잘 분류되도록 학습 시킴

- **모델의 학습 방법차이(병렬적 OR 순차적 방법)**

- Bagging

병렬적 결합 방법 → 각각의 분류기 들이 학습 시에 상호 영향을 주지 않고 그 결과를 종합
각 분류기로 부터 얻어진 결과를 한 번에 모두 함께 고려하여 하나의 최종 결과를 얻는 것

- Boosting

순차적 결합 방법으로 학습시킴 → 이전 분류기의 학습 결과 토대 다음 분류기의 학습 데이터의 샘플 가중치 조정 학습 진행
각 분류기의 결과를 단계별로 나누어, 앞 단계에 배치된 분류기의 결과가 뒤에 배치된 분류기의 학습과 분류에 영향을 줌

앙상블 (Ensenble)-부스팅(Boosting)