Chapter
7

지도학습 – 앙상블기법

1. 앙상블기법이란?
2. 배깅(Bagging)
3. 랜덤 포레스트(Random Forest)
4. 부스팅(Boosting)
5. 분류문제
6. 회귀문제

1

앙상블기법이란?

앙상블 기법(ensemble method)

정의

학습 알고리즘(learning algorithm)들을 따로 쓰는 경우에 비해
더 좋은 예측 성능을 얻기 위해 다수의 학습 알고리즘을 사용하는 방법

목적

앙상블 기법의 **필요성을 이해**하고 여러 가지 기법의 **개념을 소개**한다.

Goal!

1. 앙상블 기법이란 무엇인가?
2. 랜덤 포레스트(Random Forest)
3. 배깅(Bagging)
4. 부스팅(Boosting)

source : https://en.wikipedia.org/wiki/Ensemble_learning

개념 설명을 위해 참고한 사이트들

https://www.quora.com/What-is-the-difference-between-boost-ensemble-bootstrap-and-bagging

https://medium.com/@deepvalidation/title-3b0e263605de

https://blog.naver.com/0325han/221239663065

https://ratsgo.github.io/machine%20learning/2017/03/17/treeensemble/

https://brunch.co.kr/@snobberys/137

https://swalloow.github.io/bagging-boosting

https://m.blog.naver.com/PostView.nhn?blogId=wnswl1119&logNo=221144805300&targetKeyword=&targetRecommendationCode=1&proxyReferer=https%3A%2F%2Fwww.google.co.kr%2F

https://en.wikipedia.org/wiki/Ensemble_learning

https://flonelin.wordpress.com/2016/08/02/stacking%EC%9D%B4%EB%9E%80/

https://www.toptal.com/machine-learning/ensemble-methods-machine-learning

http://blog.hyeongeun.com/21

https://analyticsdefined.com/introduction-random-forests/

http://www.birc.co.kr/2017/01/25/%EC%95%99%EC%83%81%EB%B8%94ensemble%EB%B0%B0%EA%B9%85bagging%EB%B6%80%EC%8A%A4%ED%8C%85boosting/

[핸즈온 머신러닝: 사이킷런과 텐서플로를 활용한 머신러닝, 딥러닝 실무]

[데이터 처리 & 분석 실무]

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method)의 필요성

  알고리즘 예시 – 의사결정나무

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method)의 필요성

  알고리즘 예시 – 의사결정나무

  단순한 모형이기 때문에 예측의 정확도가 낮다는 단점이 있다.

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method)의 필요성

알고리즘 예시 – 의사결정나무

단순한 모형이기 때문에 예측의 정확도가 낮다는 단점이 있다.
-> 이를 개선할 방법은?

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method)의 필요성

알고리즘 예시 – 의사결정나무

단순한 모형이기 때문에 예측의 정확도가 낮다는 단점이 있다.
-> 이를 개선할 방법은?

-> 집단지성: **군중은 똑똑하다.**
여러 개의 모형을 동시에 고려하자.

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method)이란?

- 본래 프랑스어로 '통일, 조화' 등을 나타내는 용어
- 여러가지 동일한 종류의 혹은 서로 상이한 모형들의 예측/분류 결과를 종합하여 최종적인 의사결정에 활용하는 방법론

source : http://www.birc.co.kr/2017/01/25/%EC%95%99%EC%83%81%EB%B8%94ensemble%EB%B0%B0%EA%B9%85bagging%EB%B6%80%EC%8A%A4%ED%8C%85boosting/

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method)이란?

- 본래 프랑스어로 '통일, 조화' 등을 나타내는 용어
- 여러가지 동일한 종류의 혹은 서로 상이한 모형들의 예측/분류 결과를 종합하여 최종적인 의사결정에 활용하는 방법론

-> 목적: 다양한 모형의 예측 결과를 결합함으로써 단일 모형으로 분석했을 때보다 신뢰성 높은 예측값을 얻는 것

source : http://www.birc.co.kr/2017/01/25/%EC%95%99%EC%83%81%EB%B8%94ensemble%EB%B0%B0%EA%B9%85bagging%EB%B6%80%EC%8A%A4%ED%8C%85boosting/

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method) 예시

- 상황: 두 집단을 분류하는 분류기가 5개 있고, 각각의 오분류율이 5%

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method) 예시

- 상황: 두 집단을 분류하는 분류기가 5개 있고, 각각의 오분류율이 5%

  (1) 해당 모형들이 모두 동일한 결정을 내린다면 모형의 오분류율은 5%

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method) 예시

- 상황: 두 집단을 분류하는 분류기가 5개 있고, 각각의 오분류율이 5%

  (1) 해당 모형들이 모두 동일한 결정을 내린다면 모형의 오분류율은 5%

  (2) 5개의 분류기가 상호 독립적이고, 투표를 통해 두 집단을 분류한다면?

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method) 예시

- 상황: 두 집단을 분류하는 분류기가 5개 있고, 각각의 오분류율이 5%

  (1) 해당 모형들이 모두 동일한 결정을 내린다면 모형의 오분류율은 5%

  (2) 5개의 분류기가 상호 독립적이고, 투표를 통해 두 집단을 분류한다면?
  -> 5개의 분류기 중 3개 이상이 오분류를 하는 경우에 최종 결과가 오답

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method) 예시

- 상황: 두 집단을 분류하는 분류기가 5개 있고, 각각의 오분류율이 5%

(1) 해당 모형들이 모두 동일한 결정을 내린다면 모형의 오분류율은 5%

(2) 5개의 분류기가 상호 독립적이고, 투표를 통해 두 집단을 분류한다면?
-> 5개의 분류기 중 3개 이상이 오분류를 하는 경우에 최종 결과가 오답
-> 이론상 전체 모형의 오분류율은 약 0.1%로 훨씬 감소한다.

$$E = \sum_{i=3}^{5} (0.05)^i (1 - 0.05)^{5-i}$$

```
> pbinom(q=2, size=5, prob=0.05, lower.tail=FALSE)
[1] 0.001158125
```

source : http://www.birc.co.kr/2017/01/25/%EC%95%99%EC%83%81%EB%B8%94ensemble%EB%B0%B0%EA%B9%85bagging%EB%B6%80%EC%8A%A4%ED%8C%85boosting/

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method) 요약

조건

– 각각의 분류기는 상호 독립적이어야 한다.
– 각 분류기의 오분류율은 적어도 50%보다는 낮아야 한다.

※ 독립성을 만족하지 않아도 예측력은 높아지는 것으로 알려져 있다.

source : http://www.birc.co.kr/2017/01/25/%EC%95%99%EC%83%81%EB%B8%94ensemble%EB%B0%B0%EA%B9%85bagging%EB%B6%80%EC%8A%A4%ED%8C%85boosting/

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method) 요약

조건

– 각각의 분류기는 상호 독립적이어야 한다.
– 각 분류기의 오분류율은 적어도 50%보다는 낮아야 한다.

※ 독립성을 만족하지 않아도 예측력은 높아지는 것으로 알려져 있다.

장단점

– 이상치에 대한 대응력이 높아지고, 전체적인 분산을 감소시킴으로써 예측의 정확도가 올라간다고 알려져 있다.
– 모형의 투명성이 떨어지게 되어 해석하기 어려워진다.

source : http://www.birc.co.kr/2017/01/25/%EC%95%99%EC%83%81%EB%B8%94ensemble%EB%B0%B0%EA%B9%85bagging%EB%B6%80%EC%8A%A4%ED%8C%85boosting/

Goal 1. 앙상블 기법이란 무엇인가?

- 앙상블 기법(ensemble method) 참고

- 앞으로 다룰 앙상블 기법 내용에서는 모두 단순한 알고리즘인 의사결정나무 기법을 동일하게 사용한다.

- 선형회귀분석처럼 다른 간단한 모형을 사용해도 된다.

- 한편 서로 다른 알고리즘을 사용하면 예측력이 향상될수도 있으며, 이 또한 앙상블 기법의 일종으로 볼 수 있다.

source : http://www.birc.co.kr/2017/01/25/%EC%95%99%EC%83%81%EB%B8%94ensemble%EB%B0%B0%EA%B9%85bagging%EB%B6%80%EC%8A%A4%ED%8C%85boosting/

Goal 1. 앙상블 기법이란 무엇인가?

- (참고) 통계학적 motivation

- 중심극한정리(CLT; Central Limit Theorem)는 동일한 확률분포를 가진 독립 확률 변수 n개의 '평균'을 고려한다.

$$(\overline{X_n} - \mu) \Rightarrow N(0, \sigma^2/n) \quad as \;\; n \to \infty$$

Goal 1. 앙상블 기법이란 무엇인가?

- (참고) 통계학적 motivation

- 중심극한정리에서 알 수 있는 중요한 정보 중 하나는 극한분포의 분산이다.

- n개의 데이터로 평균 μ 를 추정할 때, n이 커질수록 추정량 $\overline{X_n}$ 가
  μ 에 가까울 확률이 늘어난다고 볼 수 있다.
  (바꿔 말하면 오차에 해당하는 부분의 분산이 줄어든다.)

$$(\overline{X_n} - \mu) \Rightarrow N(0, \sigma^2/n) \quad as \;\; n \to \infty$$

Goal 1. 앙상블 기법이란 무엇인가?

- (참고) 통계학적 motivation

- 사실 동일한 분포를 따르지 않더라도 독립이기만 하면, (적당한 조건 하에서) 중심극한정리가 성립하는 것을 보일 수 있다.

Goal 1. 앙상블 기법이란 무엇인가?

- (참고) 통계학적 motivation

- 사실 동일한 분포를 따르지 않더라도 독립이기만 하면, (적당한 조건 하에서) 중심극한정리가 성립하는 것을 보일 수 있다.

- 비슷하게 앙상블 기법에서도 '**예측을 잘 하는**' 모형을 '**많이**' 모아서 평균적인 값으로 예측을 하면, 분산이 줄어드는 효과가 있을 것이라고 기대할 수 있다. (물론 독립이 아니기 때문에 엄밀하게는 두 상황이 같지 않다.)

Goal 1. 앙상블 기법이란 무엇인가?

- (참고) 통계학적 motivation

- 일반적으로 배깅, 부스팅 등 앙상블 기법을 사용할 때는 단순한 알고리즘인 의사결정나무 기법을 사용하고, 나무의 수를 크게 한다.
  (선형회귀분석처럼 다른 간단한 모형을 사용해도 된다.)

Goal 1. 앙상블 기법이란 무엇인가?

- (참고) 통계학적 motivation

- 일반적으로 배깅, 부스팅 등 앙상블 기법을 사용할 때는 단순한 알고리즘인 의사결정나무 기법을 사용하고, 나무의 수를 크게 한다.
  (선형회귀분석처럼 다른 간단한 모형을 사용해도 된다.)

- 한편 서로 다른 알고리즘을 사용하면 예측력이 향상될수도 있으며,
  이 또한 앙상블 기법의 일종으로 볼 수 있다.

Goal 1. 앙상블 기법이란 무엇인가?

- (참고) 통계학적 motivation

- 일반적으로 배깅, 부스팅 등 앙상블 기법을 사용할 때는 단순한 알고리즘인 의사결정나무 기법을 사용하고, 나무의 수를 크게 한다.
  (선형회귀분석처럼 다른 간단한 모형을 사용해도 된다.)

- 한편 서로 다른 알고리즘을 사용하면 예측력이 향상될수도 있으며,
  이 또한 앙상블 기법의 일종으로 볼 수 있다.

- 하지만 서로 다른 (적은 수의) 모형을 사용하는 경우, (상황에 따라) 단일 모형을 고려할 때보다 훨씬 복잡한 경향성이 있는 오차가 생길 수도 있다.
  따라서 앙상블 기법을 무작정 사용하는 것은 좋지 않을 수도 있다.

2

랜덤 포레스트 (Random Forest)

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

- 랜덤 포레스트(Random Forest)

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

- 랜덤 포레스트(Random Forest)

- 배깅과 유사하지만 각각의 의사결정나무에 사용되는 변수의 수를 제한한다.

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

- 랜덤 포레스트(Random Forest)

- 배깅과 유사하지만 각각의 의사결정나무에 사용되는 변수의 수를 제한한다.

- 각 부트스트랩 샘플에 나무 모형을 적합할 때 매번 가지를 나눌 때마다 p개의 변수 중 **'랜덤하게'** 선택한 m개의 변수만을 고려한다.
  (보통 분류 문제에서는 $m = \sqrt{p}$, 회귀 문제에서는 $m = p/3$를 추천)

source: [따라 하며 배우는 데이터 과학]

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

- 랜덤 포레스트(Random Forest)

- 배깅과 유사하지만 각각의 의사결정나무에 사용되는 변수의 수를 제한한다.

- 각 부트스트랩 샘플에 나무 모형을 적합할 때 매번 가지를 나눌 때마다 p개의 변수 중 **'랜덤하게'** 선택한 m개의 변수만을 고려한다.
  (보통 분류 문제에서는 $m = \sqrt{p}$, 회귀 문제에서는 $m = p/3$를 추천)

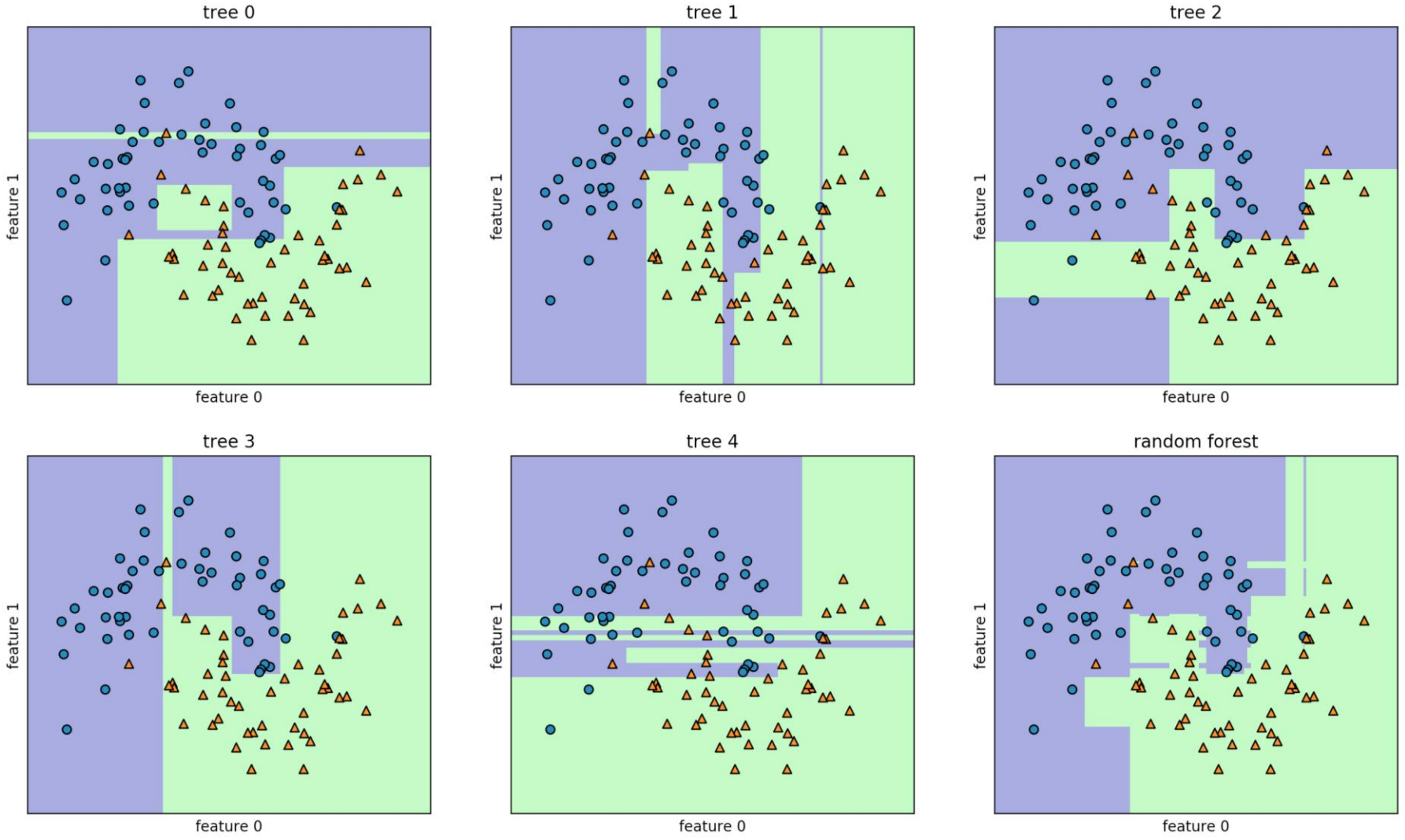
- 배깅과 마찬가지로 B가 크다고 과적합을 하지는 않는다.

source: [따라 하며 배우는 데이터 과학]

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

source: https://img.clipartxtras.com/68e19eeae79679431a2016dccfa883d9_trees-tree-clipart-free-clipart-images-3-clipartix-trees-clipart-for-kids_500-417.png
https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcSV19G8BhqhZPhNQ9a2arwtdG1dQu8RUcQ23kcWQbOsz85-dyix

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)



source: http://woolulu.tistory.com/tag/decision%20tree

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

- 모든 요소를 고려하지 않는 이유?

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

- 모든 요소를 고려하지 않는 이유?

- 의사결정나무의 한 단계를 만들 때 모든 변수를 고려한다면,
  모든 의사결정나무가 소수의 강력한 변수만을 가지고 생성될 수 있다.

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

- 모든 요소를 고려하지 않는 이유?

- 의사결정나무의 한 단계를 만들 때 모든 변수를 고려한다면,
  모든 의사결정나무가 소수의 강력한 변수만을 가지고 생성될 수 있다.

- 아무리 소수의 변수가 가장 '강력한' 변수들이어도,
  나머지 '덜 강력한' 변수들까지 고려하는 것이 랜덤 포레스트의 목적이다.

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

- 모든 요소를 고려하지 않는 이유?

- 의사결정나무의 한 단계를 만들 때 모든 변수를 고려한다면,
  모든 의사결정나무가 소수의 강력한 변수만을 가지고 생성될 수 있다.

- 아무리 소수의 변수가 가장 '강력한' 변수들이어도,
  나머지 '덜 강력한' 변수들까지 고려하는 것이 랜덤 포레스트의 목적이다.

- 즉, 단계마다 모든 요소를 고려하지 않는 이유는 역설적으로
  모든 요소를 고려하기 위해서이다.

source: https://medium.com/@deepvalidation/title-3b0e263605de

Goal 3. 앙상블 기법 – 랜덤 포레스트(Random Forest)

- 모든 요소를 고려하지 않는 이유?

- 의사결정나무의 한 단계를 만들 때 모든 변수를 고려한다면,
  모든 의사결정나무가 소수의 강력한 변수만을 가지고 생성될 수 있다.

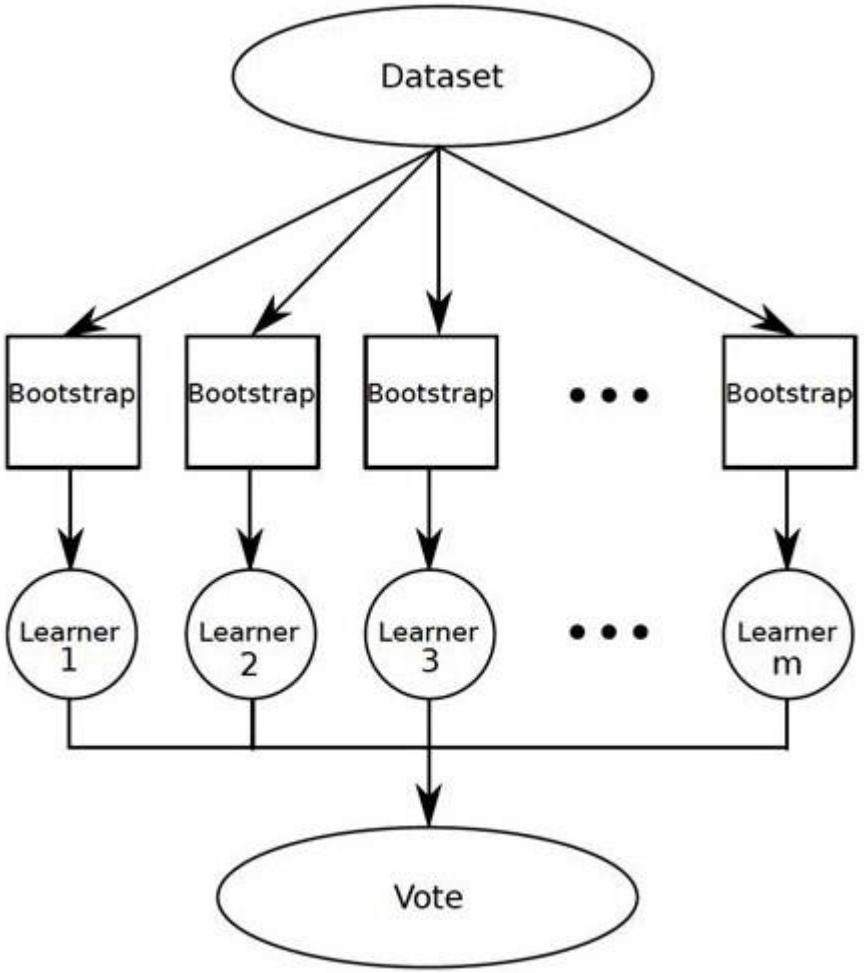
- 아무리 소수의 변수가 가장 '강력한' 변수들이어도,
  나머지 '덜 강력한' 변수들까지 고려하는 것이 랜덤 포레스트의 목적이다.

- 즉, 단계마다 모든 요소를 고려하지 않는 이유는 역설적으로
  모든 요소를 고려하기 위해서이다.

- 변수를 제한하므로 나무끼리의 **상관 관계를 제거**하는 효과가 있다.

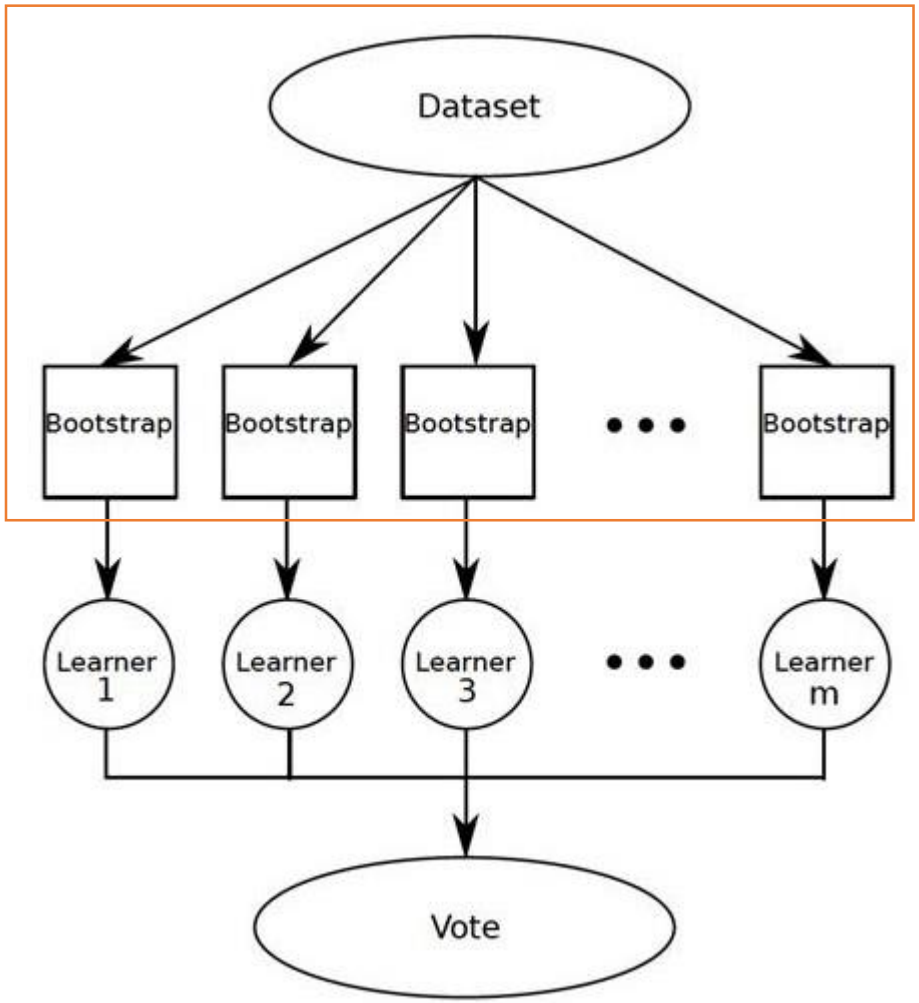
3

배깅(Bagging)



- Bagging: Bootstrap aggregating
- 재추출(Resampling) 방법의 일종인 부트스트랩(Bootstrap)을 이용

Image source : https://www.kdnuggets.com/wp-content/uploads/bagging.jpg

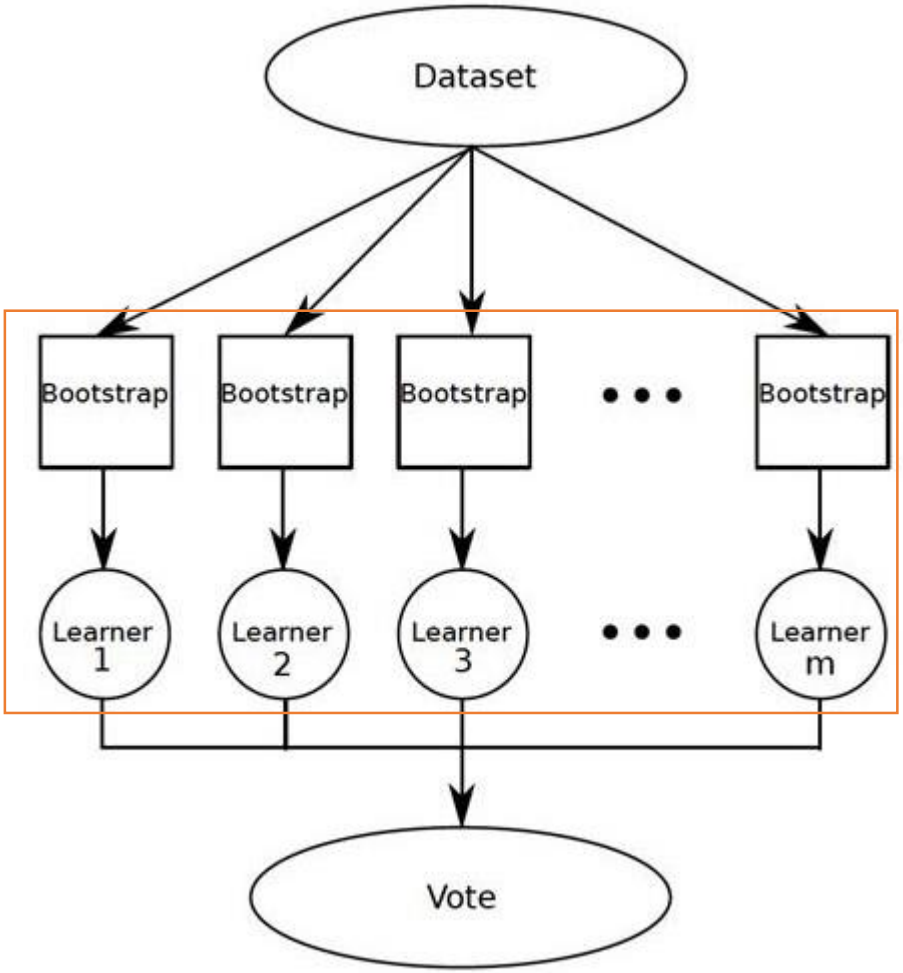


- Bagging: Bootstrap aggregating

- 재추출(Resampling) 방법의 일종인 부트스트랩(Bootstrap)을 이용

절차

1) 원본 훈련 데이터에서 여러 개의 훈련 데이터 집합을 만들어낸다.



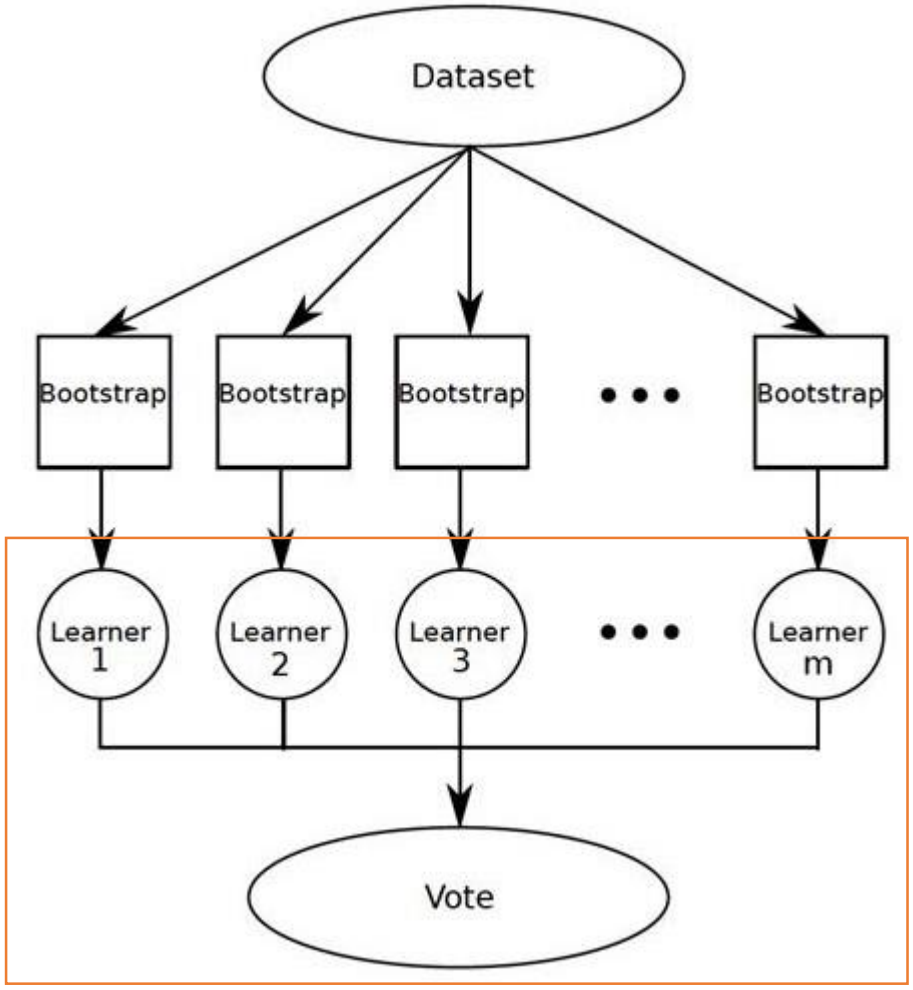
- Bagging: Bootstrap aggregating

- 재추출(Resampling) 방법의 일종인 부트스트랩(Bootstrap)을 이용

절차

1) 원본 훈련 데이터에서 여러 개의 훈련 데이터 집합을 만들어낸다.

2) 그 후 각각의 데이터 집합을 활용해 여러 개의 모형을 학습한다.



- Bagging: Bootstrap aggregating

- 재추출(Resampling) 방법의 일종인 부트스트랩(Bootstrap)을 이용

절차

1) 원본 훈련 데이터에서 여러 개의 훈련 데이터 집합을 만들어낸다.

2) 그 후 각각의 데이터 집합을 활용해 여러 개의 모형을 학습한다.

3) 각각의 모형에서 예측을 한 뒤 투표를 통해 최종 예측값을 정한다.

- 부트스트랩(Bootstrap)

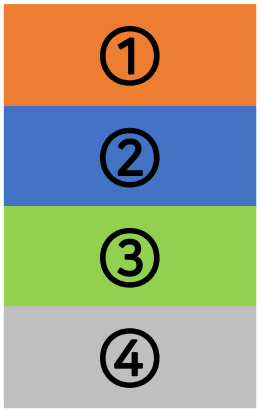
- 추정량의 표준오차, 신뢰구간 등을 이론적으로 구하기 힘든 경우에 사용
- 앙상블 기법인 배깅(Bagging)에도 이용된다.

- 부트스트랩(Bootstrap)

- 추정량의 표준오차, 신뢰구간 등을 이론적으로 구하기 힘든 경우에 사용
- 앙상블 기법인 배깅(Bagging)에도 이용된다.

- 주어진 데이터에서 중복을 허용하여 임의로 n개를 추출하는 것과 같다.

- 부트스트랩(Bootstrap) 예시



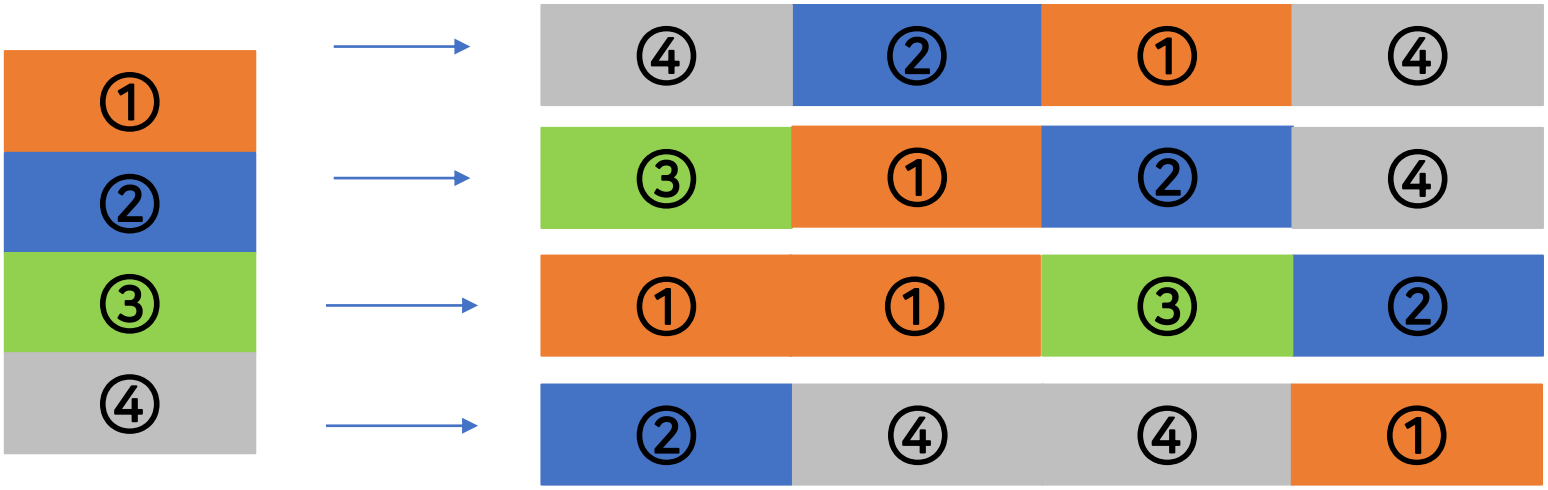
- 데이터의 개수가 4개인 경우(n=4)의 bootstrapping

- 부트스트랩(Bootstrap) 예시

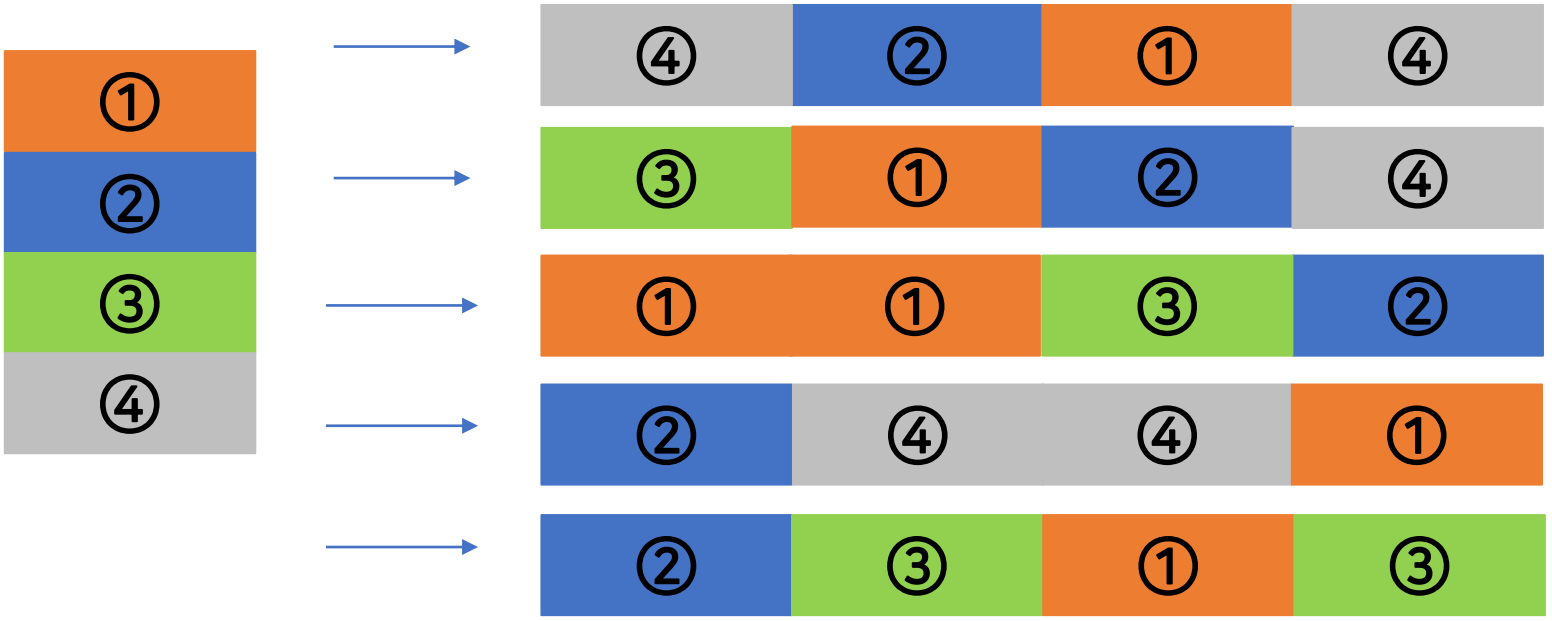
- 부트스트랩(Bootstrap) 예시

- 부트스트랩(Bootstrap) 예시

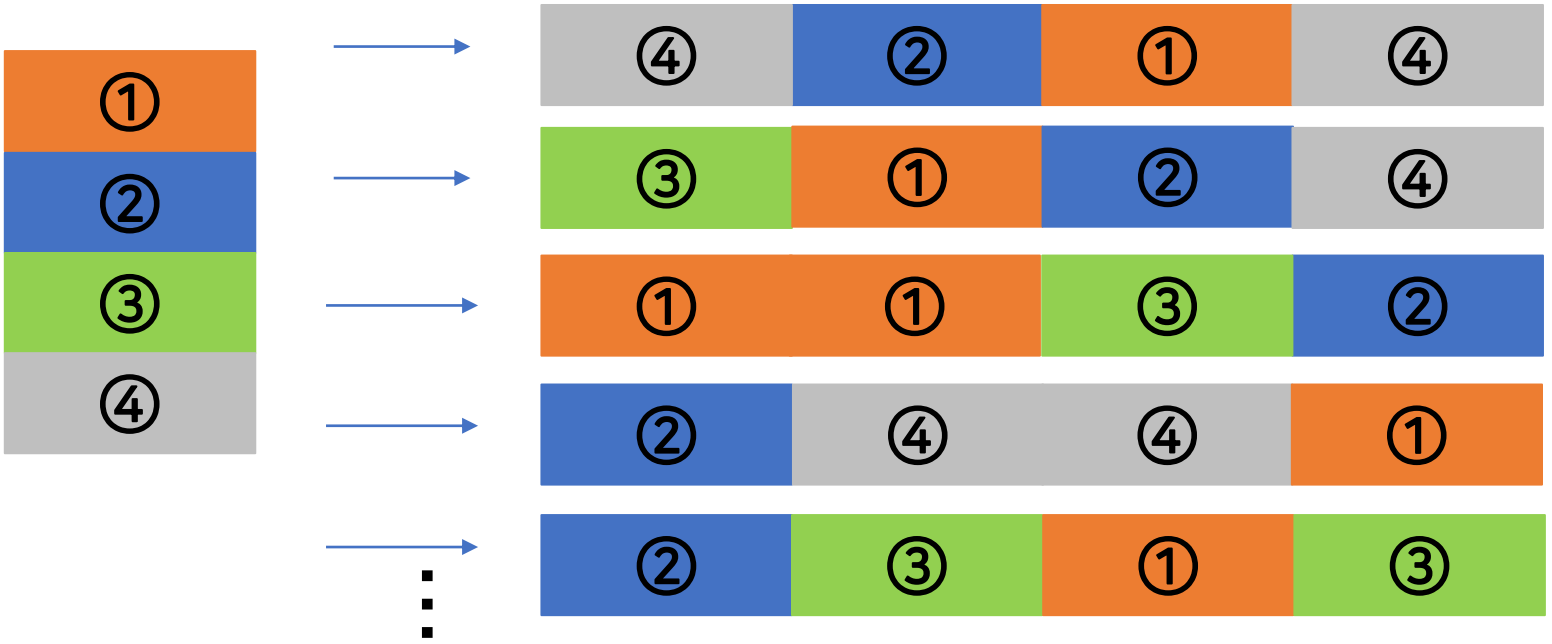
- 부트스트랩(Bootstrap) 예시



- 부트스트랩(Bootstrap) 예시



- 부트스트랩(Bootstrap) 예시



- 데이터의 개수 n이 충분히 큰 경우, n개의 bootstrap sample을 만들 때 한 번 이상 추출되는 데이터의 이론적인 비율은 63.2%임이 알려져 있다.

- $1 - (1 - 1/n)^n \ \rightarrow 1 - 1/e \quad for \quad n \rightarrow \infty$

- 데이터의 개수 n이 충분히 큰 경우, n개의 bootstrap sample을 만들 때 한 번 이상 추출되는 데이터의 이론적인 비율은 63.2%임이 알려져 있다.

- $1 - (1 - 1/n)^n \ \rightarrow 1 - 1/e \quad for \quad n \rightarrow \infty$

- 36.8%의 한 번도 사용되지 않은 데이터는 모형의 성능 평가에 쓸 수 있다 (OOB; out-of-bag).

- 배깅 요약

1) 훈련세트로부터 b = 1, ..., B개의 부트스트랩(bootstrap) 샘플을 얻는다.
2) 각각의 샘플에 비교적 간단한(나무 모형이 많이 쓰인다) 모형을 적합하여 B개의 예측값을 얻는다.
3) 분류에서는 다수결 투표, 회귀에서는 평균값을 최종 예측값으로 사용한다.

source: [따라 하며 배우는 데이터 과학]

Goal 2. 앙상블 기법 – 배깅(Bagging)

- 배깅 요약

1) 훈련세트로부터 b = 1, ..., B개의 부트스트랩(bootstrap) 샘플을 얻는다.
2) 각각의 샘플에 비교적 간단한(나무 모형이 많이 쓰인다) 모형을 적합하여 B개의 예측값을 얻는다.
3) 분류에서는 다수결 투표, 회귀에서는 평균값을 최종 예측값으로 사용한다.

- 배깅을 통하여 의사결정나무보다 일반적인 모형을 만들 수 있다.
- 평균적인 값을 이용하므로 분산을 줄여주는 효과가 있다.
- 일반적으로 B가 커지더라도 과적합이 일어나지 않는다.
- 부트스트랩에 사용되지 않은 데이터로 모형의 성능을 평가할 수 있다.

source: [따라 하며 배우는 데이터 과학]

4

부스팅
(Boosting)

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting)

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting)

- 배깅이 일반적인 모형을 만드는데 집중되어 있다면,
  부스팅은 맞추기 어려운 문제를 맞추는 것이 목적이다.

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting)

- 배깅이 일반적인 모형을 만드는데 집중되어 있다면,
  부스팅은 맞추기 어려운 문제를 맞추는 것이 목적이다.

- 배깅과 유사하게 부스팅은 재추출(resampling)한 데이터에 대한 훈련된
  모델의 앙상블을 사용하고 최종 예측을 결정하게 투표한다.

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting)

- 배깅이 일반적인 모형을 만드는데 집중되어 있다면,
  부스팅은 맞추기 어려운 문제를 맞추는 것이 목적이다.

- 배깅과 유사하게 부스팅은 재추출(resampling)한 데이터에 대한 훈련된
  모델의 앙상블을 사용하고 최종 예측을 결정하게 투표한다.

- 차이점은 리샘플링한 데이터가 보완적인 학습기를 생성하기 위해
  특별하게 구축되고, 투표는 각 모형의 성능을 바탕으로 가중치를 준다.

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting) 원리

- 모래, 자갈, 먼지 등이 섞여 있는 물질에 여러 타입의 체를 가지고 조합해 그것을 분류하는 과정과 유사하다.

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting) 원리
- 예를 들어, 어떤 학습기 M에 대해 Y를 예측할 확률은 아래와 같다.

$$Y = M(X) + error$$

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting) 원리
- 예를 들어, 어떤 학습기 M에 대해 Y를 예측할 확률은 아래와 같다.

$$Y = M(X) + error$$

- 만약 error에 대해 조금 더 상세히 분류할 수 있는 모형 G가 있다면,

$$error = G(X) + error\_2$$

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting) 원리
- 예를 들어, 어떤 학습기 M에 대해 Y를 예측할 확률은 아래와 같다.

$$Y = M(X) + error$$

- 만약 error에 대해 조금 더 상세히 분류할 수 있는 모형 G가 있다면,

$$error = G(X) + error\_2$$

- 'error2'를 더 세밀하게 분리할 수 있는 모형 H가 있다면,

$$error\_2 = H(X) + error\_3$$

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting) 원리

- 앞 과정을 한 번에 표현하면,

$$Y = M(X) + G(X) + H(X) + \ error\_3$$

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting) 원리

- 앞 과정을 한 번에 표현하면,

$$Y = M(X) + G(X) + H(X) + \ error\_3$$

- 여기서 M, G, H 각각 분류기의 성능이 다르기 때문에,
  최적의 가중치 α , β , Υ 를 학습한다면

$$Y = \alpha \times M(X) + \beta \times G(X) + \Upsilon \times H(X) + \ error\_3$$

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting) 원리

- 앞 과정을 한 번에 표현하면,

$$Y = M(X) + G(X) + H(X) + \ error\_3$$

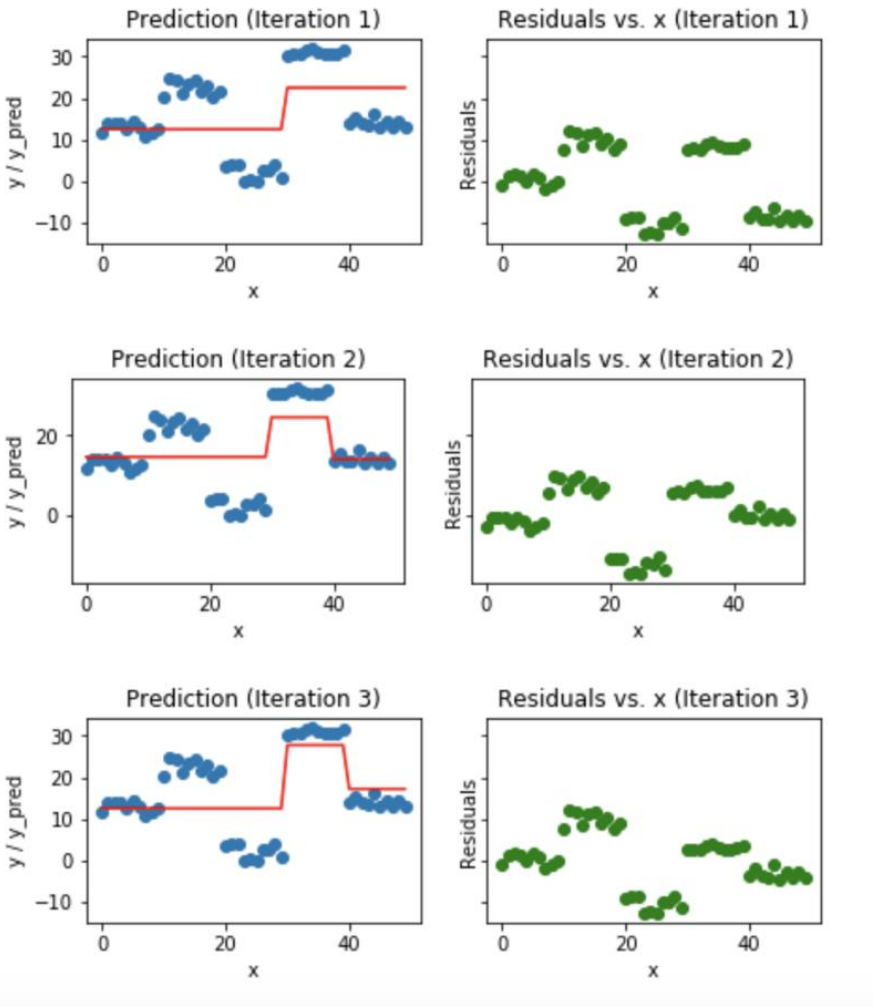
- 여기서 M, G, H 각각 분류기의 성능이 다르기 때문에,
  최적의 가중치 α , β , ϒ 를 학습한다면

$$Y = \alpha \times M(X) + \beta \times G(X) + \Upsilon \times H(X) + \ error\_3$$

- 부스팅은 배깅과 다르게 학습이 순차적으로 진행된다.

-> 시간이 상대적으로 오래 걸리며, 반복 횟수가 커지면 과적합 위험이 있다.

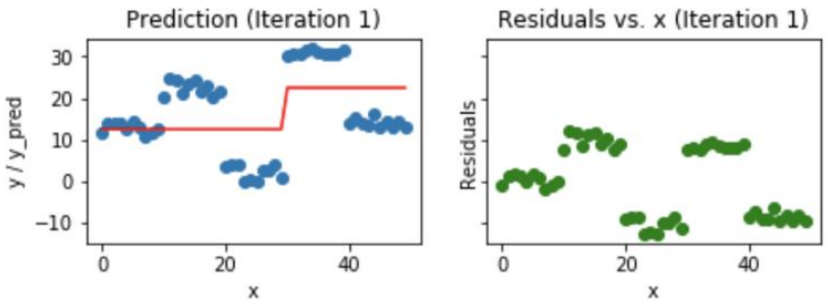
Goal 4. 앙상블 기법 – 부스팅(Boosting)

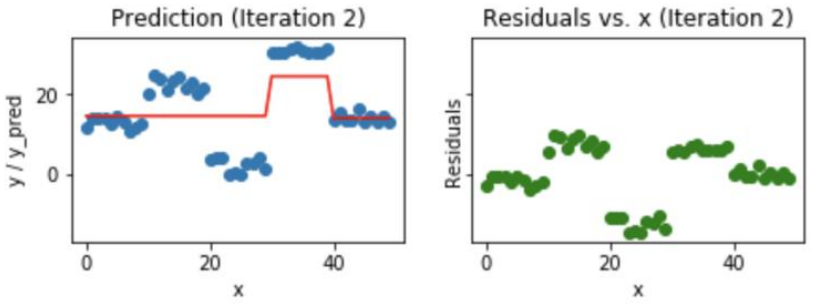


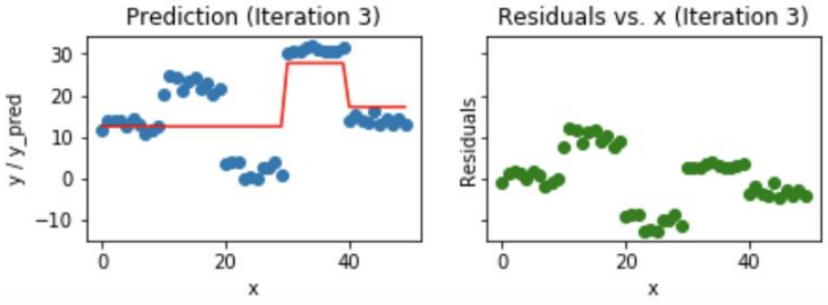
...

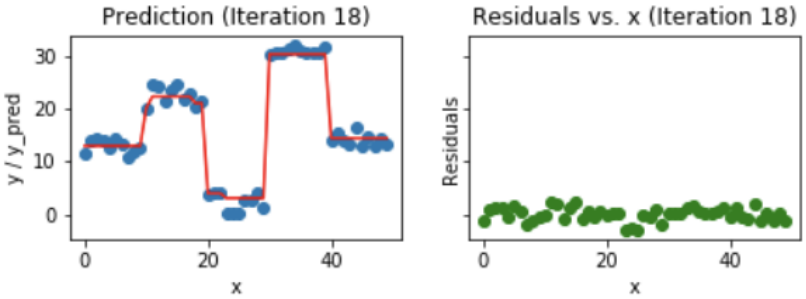
source: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d

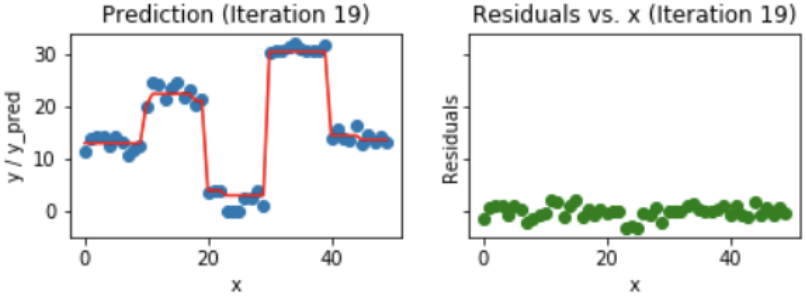
Goal 4. 앙상블 기법 – 부스팅(Boosting)

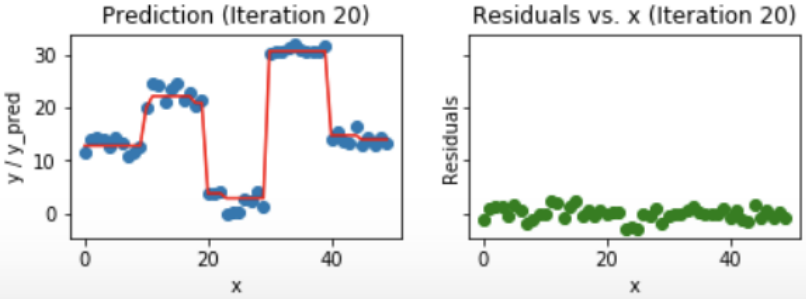








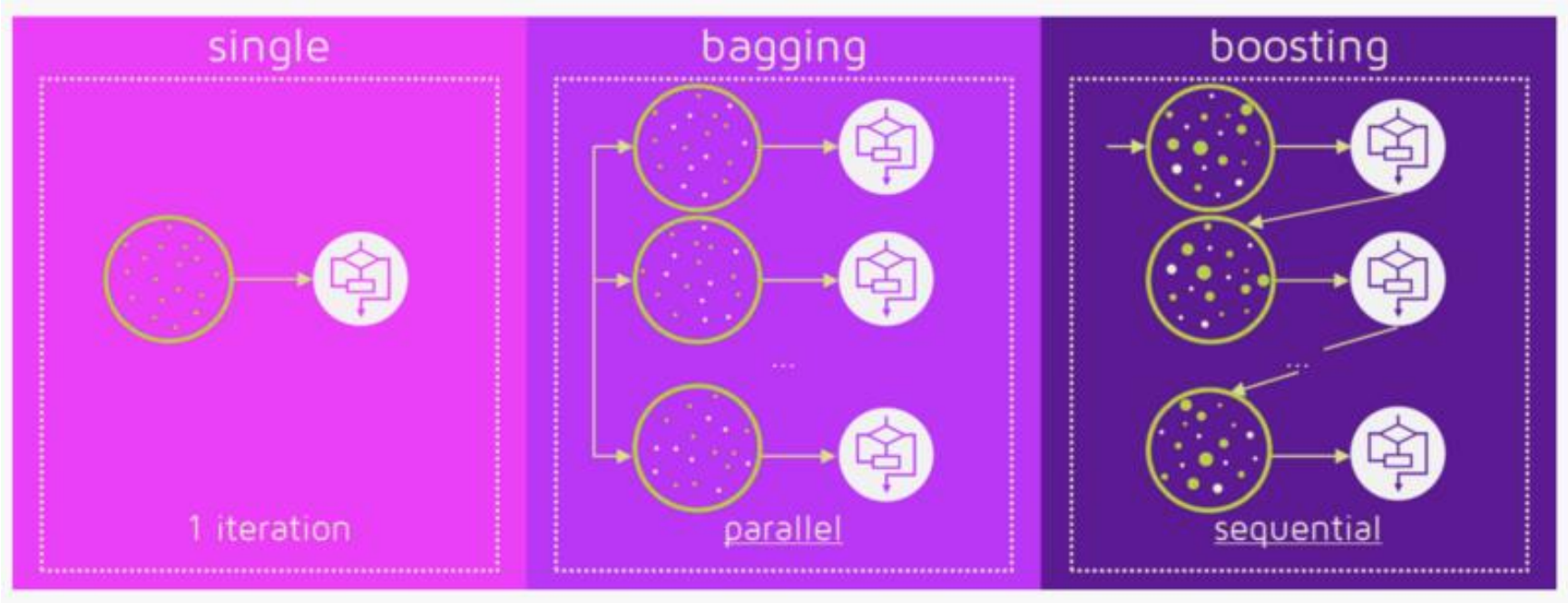




source: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 단일 알고리즘, 배깅, 부스팅의 차이



source: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting) 알고리즘

- 부스팅을 위한 여러 가지 알고리즘이 개발되어 있다.
  (AdaBoost, Gradient Boosting, Xgboost, CatBoost, Light GBM…)

source: https://brunch.co.kr/@snobberys/137

Goal 4. 앙상블 기법 – 부스팅(Boosting)

- 부스팅(Boosting) 알고리즘

- 부스팅을 위한 여러 가지 알고리즘이 개발되어 있다.
  (AdaBoost, Gradient Boosting, Xgboost, CatBoost, Light GBM…)

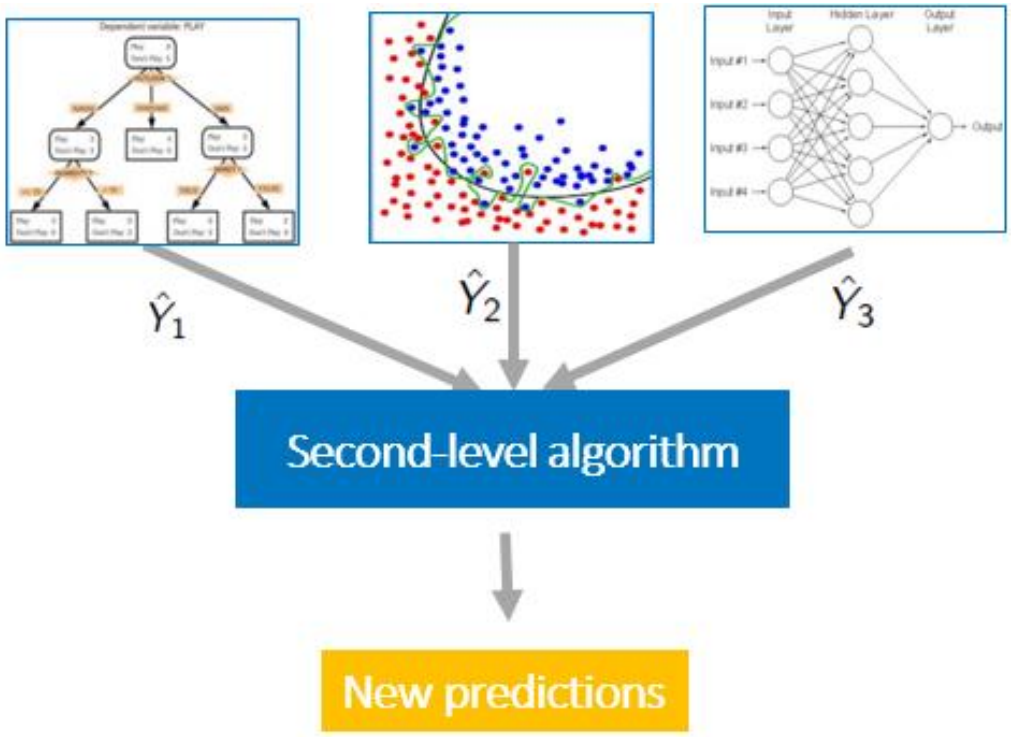
- AdaBoost: 잘못 예측한 데이터에 가중치를 부여하자.
- GBM(Gradient Boosting): 가중치를 계산할때 경사 하강법을 이용하자.
- Xgboost(eXtreme Gradient Boosting): 'CART' 알고리즘 사용
  (의사결정나무의 일종, 의사결정나무도 실제로는 여러 가지가 존재한다)

source: https://brunch.co.kr/@snobberys/137

(참고) 스택킹(Stacking)

- 스택킹(Stacking); Meta Ensembling

- 스택킹은 여러 가지 알고리즘에서 예측한 값을 입력값으로(input) 하여 그 위에 추가적인 모형(주로 로지스틱 회귀분석)을 쌓는다(stacking).



source: https://blogs.sas.com/content/subconsciousmusings/files/2017/05/modelstacking.png

(참고) 스택킹(Stacking)

- 스택킹(Stacking); Meta Ensembling

- 필요한 연산량이 훨씬 증가하지만 정확도는 상대적으로 소량 증가한다.

-> 실무보다는 Kaggle과 같은 대회에서 자주 사용된다.

•Jeong-Yoon Lee, Winning Data Science Competitions

source: https://blogs.sas.com/content/subconsciousmusings/files/2017/05/stackedapproach.png