Data Mining

4팀

김동영
강수경
김재희
유경민
최윤혜

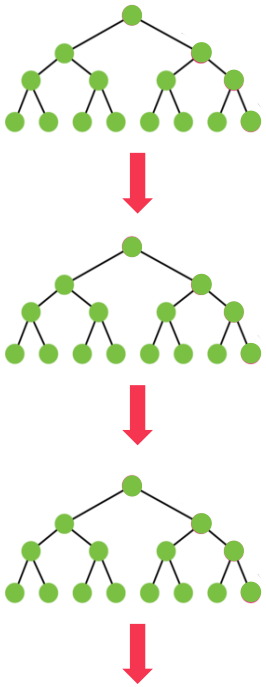
INDEX

Boosting

Bagging

각각의 샘플 데이터에 대하여
병렬적으로 학습

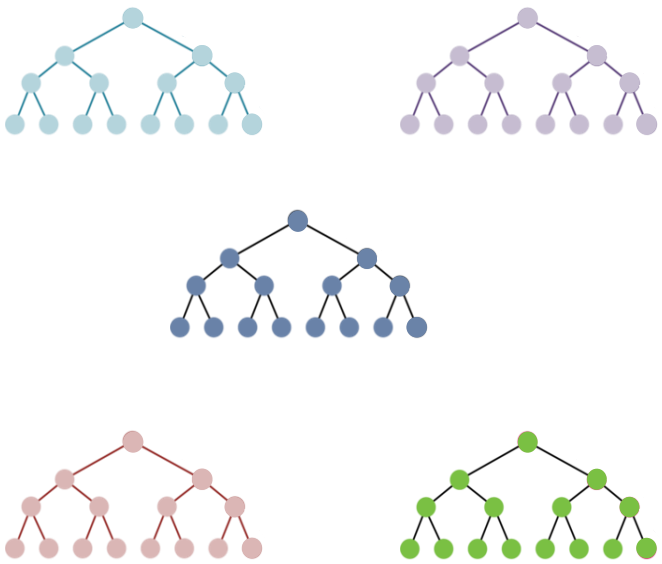
Boosting



Boosting

Boosting

Bagging

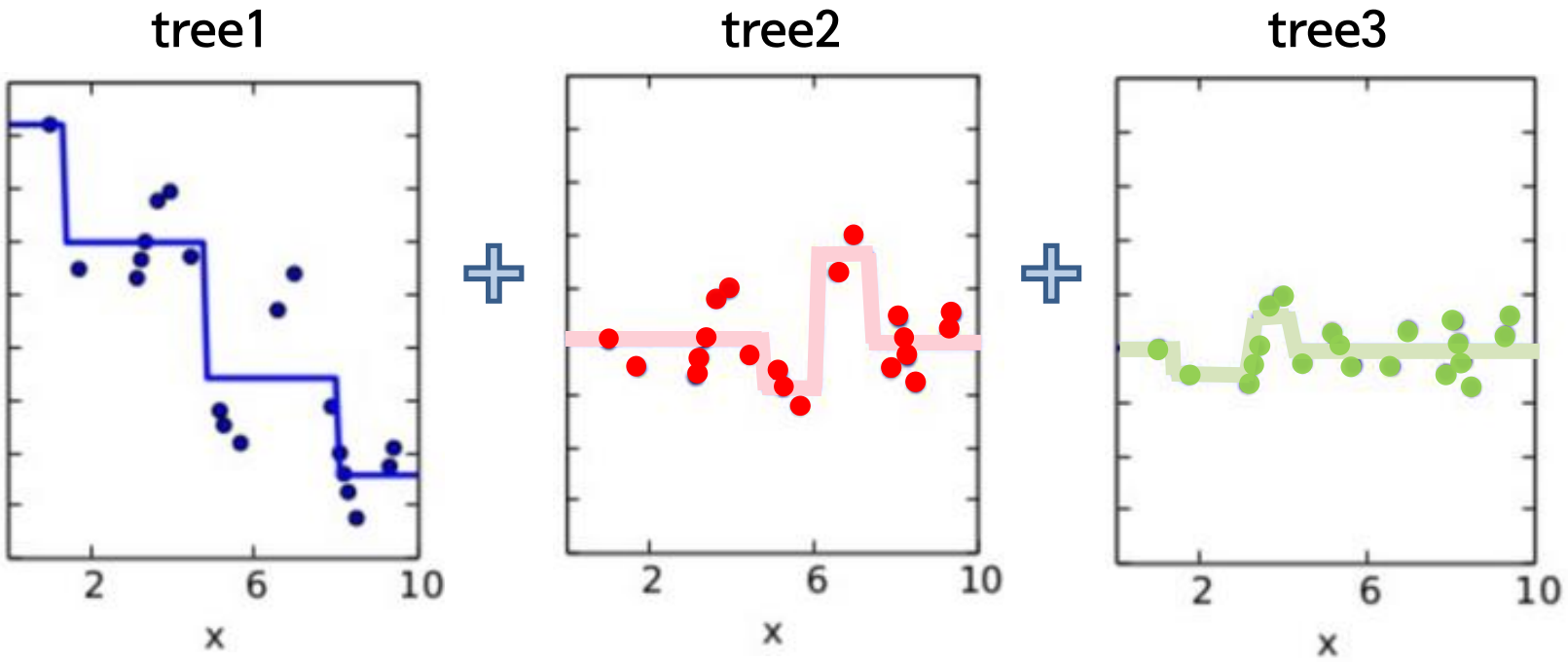


Boosting

약한 트리들을 만들어서
순차적으로 학습

Boosting

Gradient Boosting



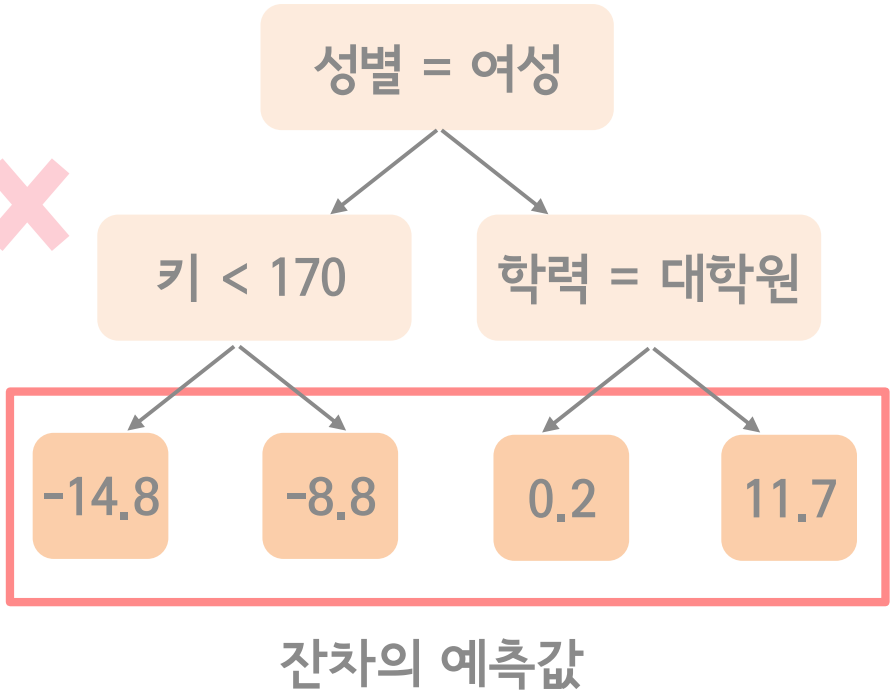
잔차를 학습하는 방식으로 강력한 학습기를 만들어나감

Gradient Boosting

4.5 예측값을 (조금만) 업데이트 해준다

ex) 0.01, 0.001

초기 예측값 64.8 + learning rate ×



Gradient Boosting

5. 앞의 과정을 지정 횟수만큼 반복한다

너무 적으면 underfitting

너무 많으면 overfitting

적절한 횟수를 찾는 것이 중요!

Boosting Model

XG boost

LightGBM

Catboost

우리가 지금까지 배운 트리의 구조를 통해
부스팅 3인방의 특징을 알아보자!!

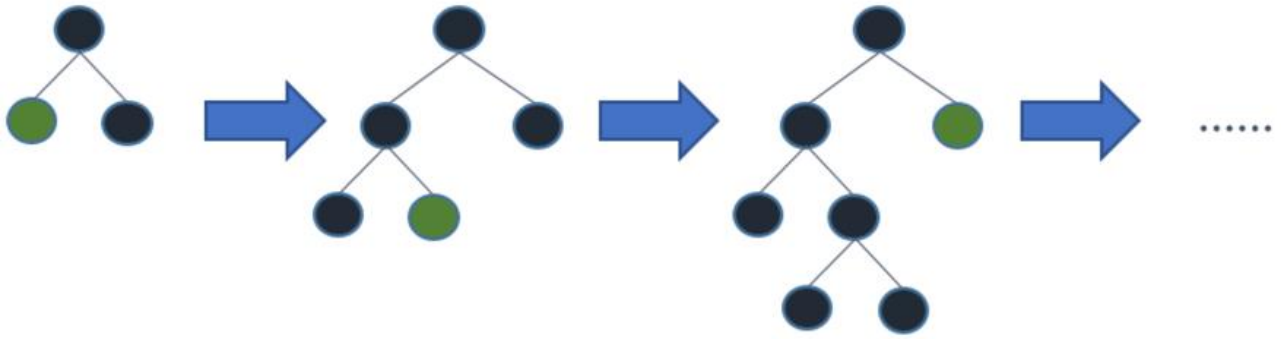
Boosting Model

XG Boost

1. 병렬 처리를 사용 가능
   - 속도를 높일 수 있기 때문!
2. 가지치기를 통한 분할 수 줄이기 가능
   - 미리 세팅해둔 max_depth까지만 split하고 pruning 하기
3. 잔차를 학습할 때 regularization (패널티)를 통해 과적합을 방지

Boosting Model

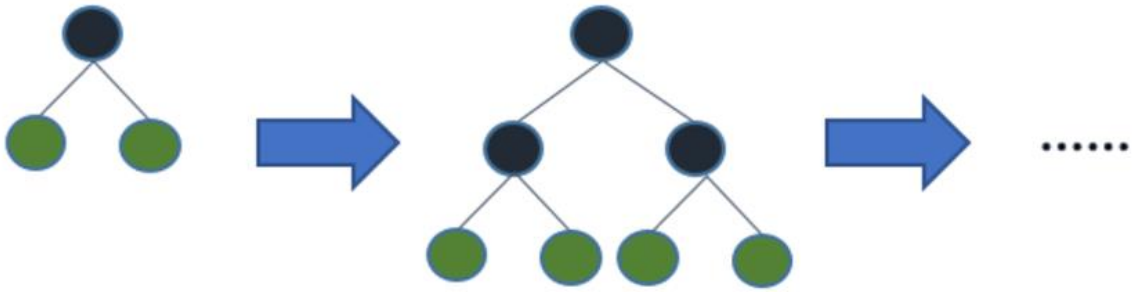
Light GBM



Light GBM은 Leaf-wise 방식을 사용!

- 최대 손실값을 가지는 리프 노드를 지속적으로 분할
- 트리의 깊이가 깊어지고 비대칭적인 트리가 생성
- 균형 트리보다 예측오류를 최소화 하고 속도가 향상

Catboost

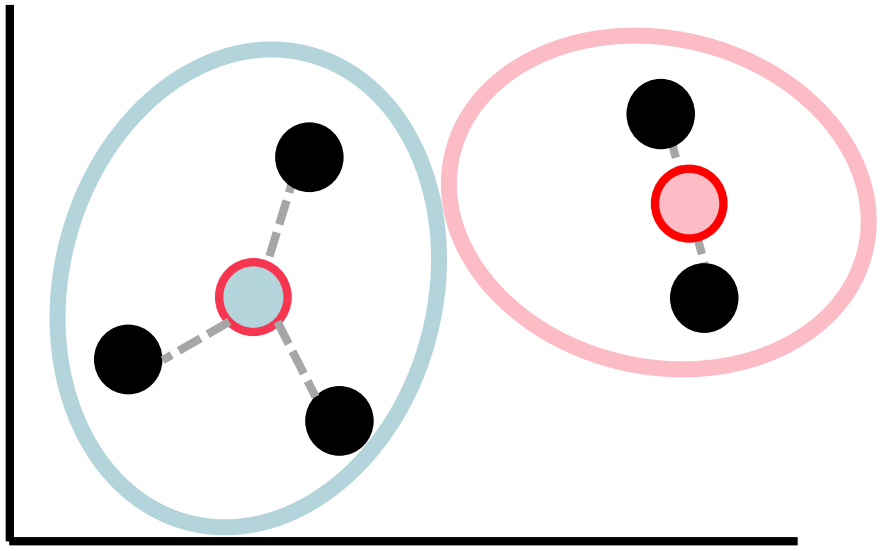


Light GBM과 다르게 Level-wise (symmetric tree구조)

- 균형 잡힌 트리이므로 overfitting을 방지할 수 있다
- 구조가 같으므로 leaf node의 값들을 벡터에 저장해 tree structure을 메모리에 저장할 필요 없이 leaf value를 저장하고 불러올 수 있음!

  ---> 효율적인 테스트 가능

K-Means Algorithm



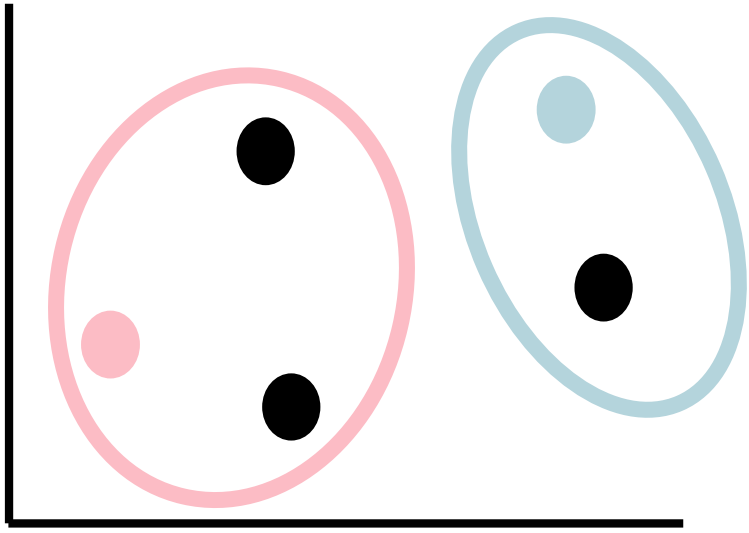
K

중심점의 개수

Means

평균 좌표가 중심점이 됨

K-Medoids Algorithm



K

중심점의 개수

Medoids

클러스터 내 관측치 중
이를 대표하는 관측치가 중심점이 됨

K-Means VS K-Medoids

K-Means

계산량 ↓

이상치 영향 ↑

K-Medoids

이상치 영향 ↓

계산량 ↑

K 정하기

클러스터 개수는 내 맘대로 뇌피셜인가요 ?

- **도메인 지식** ex.피셋 팀원 분류: 6개 팀이므로 K=6
- **Rule of Thumbs** $k = \sqrt{\frac{n}{2}}$
- **WCSS** 클러스터 내 중심점과의 거리제곱합
- **Silhoette** 데이터의 클러스터 내외 거리 비교