차세대융합기술연구원
ADVANCED INSTITUTE OF CONVERGENCE TECHNOLOGY

부스팅 트리 회귀 모델

2021.08.10 여지민

Gradient Boosting Machine(GBM)

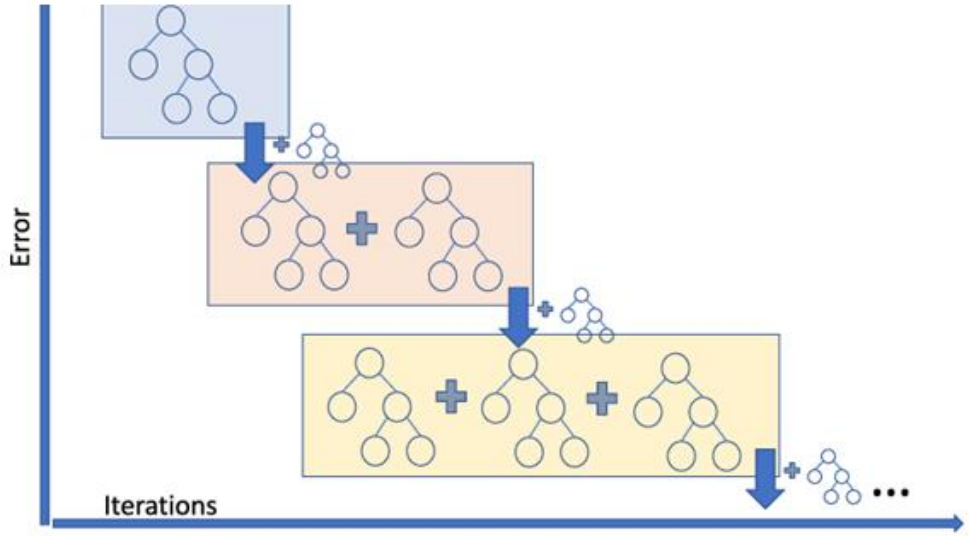
Gradient Boosting Regressor

순차적으로 트리를 생성함으로써 이전 트리의 오차를 보완하여 전체 예측 오류를 최소화

*Boosting: 약한 학습기를 여러 개 연결하여 강한 학습기를 만드는 앙상블 방법

학습률(learning_rate)로 모델의 복잡도 결정

→ 학습률 낮추면 비슷한 복잡도의 모델 만들기 위해서 많은 트리 추가
→ 학습률 높이면 모델 복잡도 및 과대적합 가능성 상승



장점

- 약한 학습기를 사용하여 메모리 적게 사용하고 예측 속도 빠름

단점

- 랜덤포레스트보다 민감한 매개변수 설정
- 긴 훈련시간

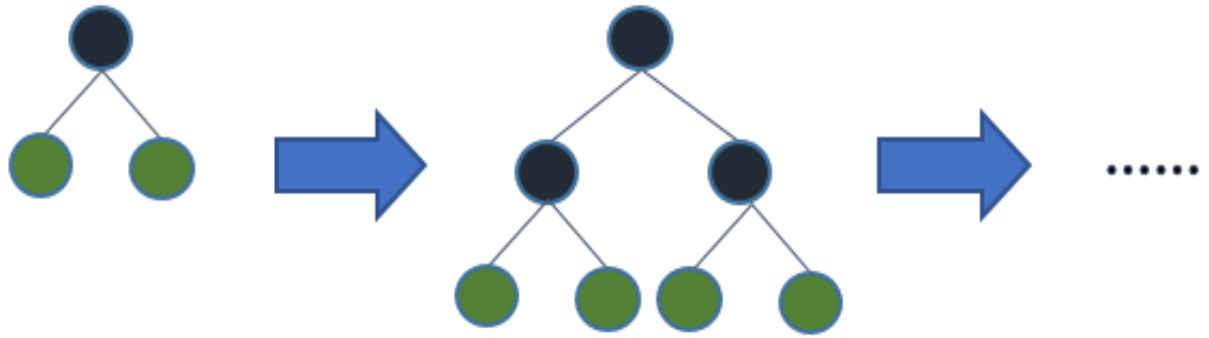
eXtreme Gradient Boosting(XGB)

GBM과의 차이점

- 병렬 처리
- 가지치기 방식 변화

  지정된 최대 깊이(max_depth)까지 분할한 후, 가지치기
  *Information Gain이 없는 가지에 Information Gain이 있는 가지 생성 가능하다 가정
- 과대적합 방지

  Lasso(L1), Ridge(L2) 규제 추가
- 결측치 처리 성능
- Cross-validation



Level-wise tree growth

장점

- 캐쉬 메모리 활용 최대화(계산속도 GBM의 10배)
- 데이터셋을 여러 개의 하드 드라이브에 분산, 병렬 처리

eXtreme Gradient Boosting(XGB)

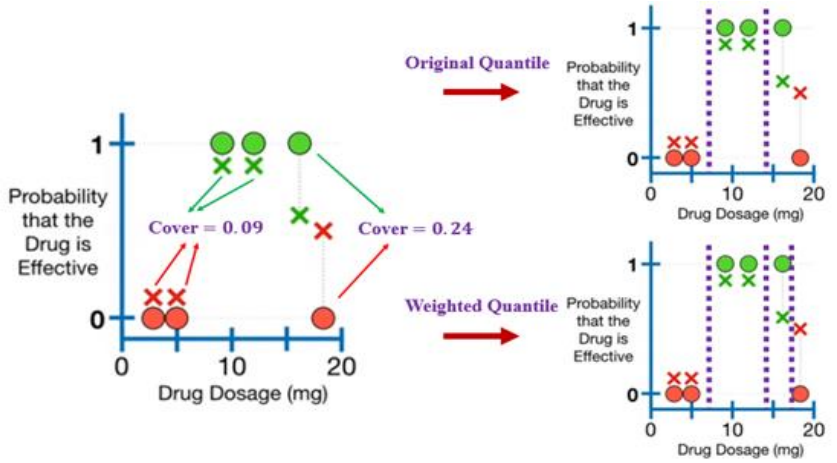
근사적인 트리 학습(Approximate Tree Learning)

- Feature 분포의 백분위수에 따라 후보 분할 지점 제시

→ 연속적인 변수를 이러한 후보지점으로 나뉜 bucket으로 매핑 후, 통계량 집계
집계된 통계량을 바탕으로 가장 좋은 분할 탐색

- Weighted Quantile Sketch 사용

→ 모든 관측치에 weight를 주고, weight의 합을 구간별로 동일하게 유지
모든 관측치의 weight 합 1



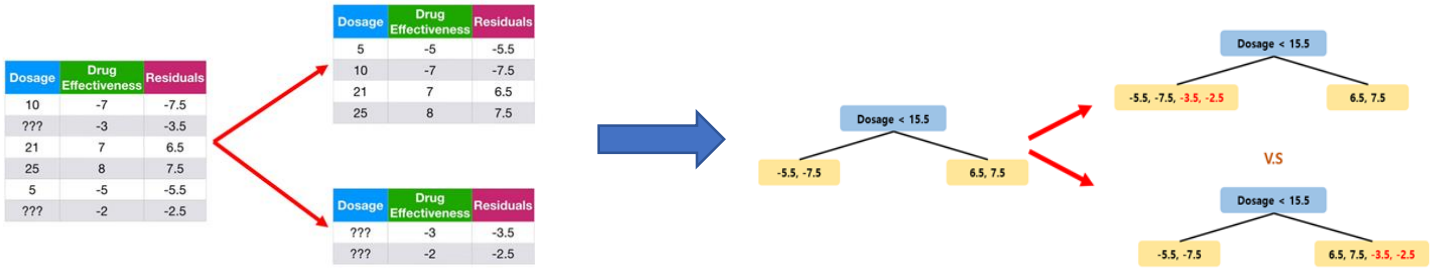
Weighted Quantile Sketch

희소성 인식 알고리즘(Sparsity-Aware Algorithm)

- 결측치에 대해서도 학습 가능

→ 결측치 여부에 따라 데이터 셋 분할하여 처리

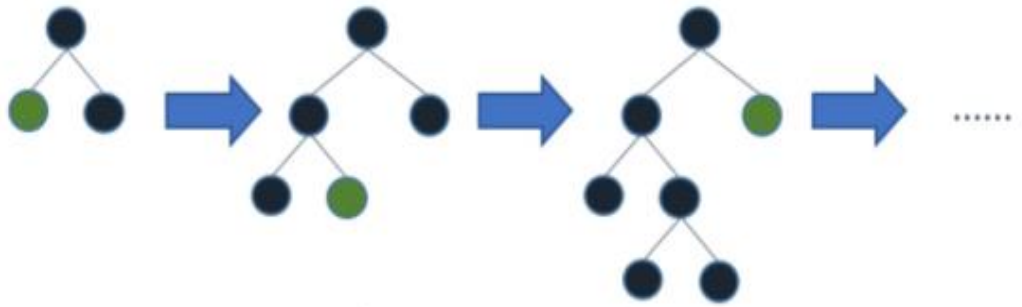
→ 결측치를 각 리프에 추가하여 Information Gain이 큰 경우 선택



LightGBM(LGBM)

LGBM Regressor

Information Gain이 적은 가지의 데이터를 증폭시켜 훈련이 잘 되지 않는 부분에 초점 설정하여 학습



Leaf-wise tree growth

장점

- 트리가 깊어지면서 소요되는 시간 및 메모리 절약 가능
- 동일한 leaf를 생성할 때, level-wise를 사용하는 모델보다 손실 적음

단점

- 데이터 적을 경우, 과대적합 가능성 상승 (여기서 적다는 것은 10000개 기준)

LightGBM(LGBM)

GOSS(Gradient-based One-Side Sampling)

- 큰 가중치를 가진 인스턴스 유지하면서 작은 가중치 인스턴스 무작위 다운 샘플링 수행

→ 잘못 예측된 인스턴스: 큰 가중치 부여 / 제대로 예측된 인스턴스: 작은 가중치 부여

- 작은 가중치를 가진 인스턴스에 상수 multiplier를 적용하여 데이터 분포에 미치는 영향 보상

Row id	gradients
4	-5
3	3
2	0.5
6	0.2
5	0.1

상위 100a%개

Select Top 2

Randomly sample 2
From the rest
나머지에서 100b%개

Row id	gradients	weights
4	-5	1
3	3	1
2	0.5	2
5	0.1	2

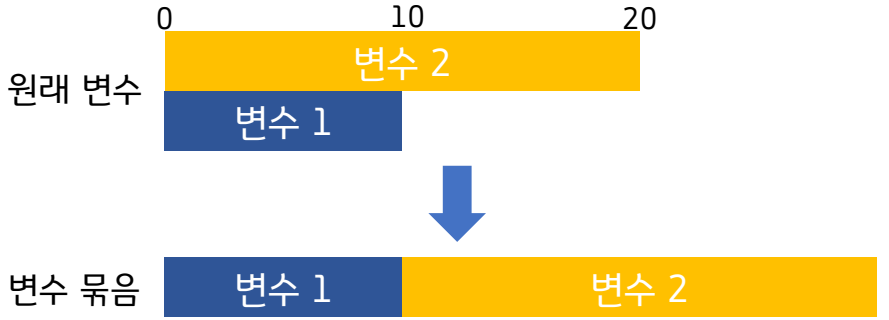
*Information Gain × (1-a)/b

EFB(Exclusive Feature Bundling)

- 히스토그램 작성의 복잡성을 줄이는 것을 목표
- 상호 배타적인 변수들을 하나로 통합하여 계산량 감소

  * 상호 배타적 변수: 컬럼 중 하나만 값이 있고 나머지는 0인 경우

Data수(row) × Feature수(column) → Data수(row) × Bundle수(column)



Categorical Boosting Machine(CATB, 2017)

Catboost Regressor

Target leakage로 인한 과대적합, 범주형 변수 처리 문제 해결

- Ordered Boosting

  : 일부 훈련 데이터 잔차에 대해서 모델 생성 뒤, 해당 데이터 외의 데이터에 대한 잔차 예측

Point	Class label
X1	10
X2	12
X3	9
X4	4

1. t-1까지의 잔차를 기반으로 모델 생성 뒤, t번째 잔차 예측
2. 실제 t번째 y값과 비교하여 잔차 구함

- Random Permutation

  : 과대적합 방지 및 다각적 트리 생성을 위해 매 훈련 동안 데이터 shuffling 및 랜덤 추출

- Ordered Target Encoding

  : Target encoding, Mean encoding, Response encoding 방법 사용

Categorical Boosting Machine(CATB, 2017)

ex) Mean Encoding

Time	Feature1	Class_labels
Mon	Sunny	32
Tues	Cloudy	15
Wed	Cloudy	14
Thurs	Mostly_cloudy	10
Fri	Cloudy	20

현재 데이터의 인코딩하기 위해 이전 데이터들의 인코딩된 값을 사용

ex) Friday : cloudy = (14+15) / 2 = 15.5

Saturday : cloudy = (15+14+20) / 3 =16.3

장점

- 범주형 변수 처리 방법 개선
- Ordered boosting 기법을 활용하여 예측변화 문제 해결
- 모델 파라미터 튜닝 불필요

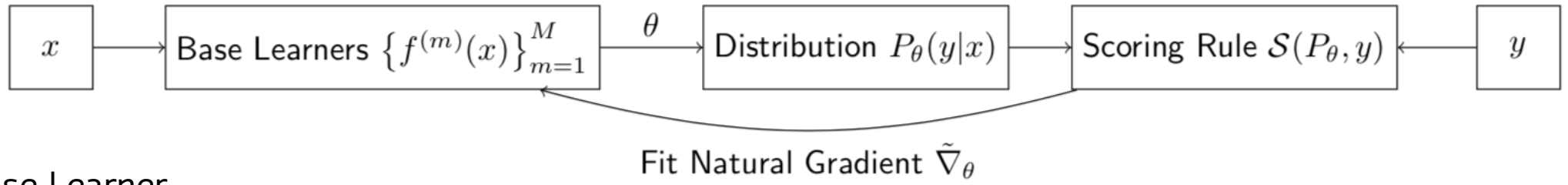
단점

- 희소행렬 처리 어려움

Natural Gradient Boosting(NGB, 2019)

NGB Regressor

목표변수에 대한 예측 및 예측값에 대한 불확실성(확률) 추정



- Base Learner

  : 가장 기본적인 학습기는 Decision Tree(학습기 종류 무관)

- Probability Distribution

  : Point estimate 대신 매개변수를 예측하도록 훈련하여 전체 데이터에 대한 확률 분포 예측 및 생성

  일반적으로는 정규 분포, Binary값에 대해서는 Bernoulli 분포를 따름

- Scoring Rule

  : MLE(Maximum Likelihood Estimation) or CRPS(Continuous Ranked Probability Score)