GBM(Gradient Boosting Machine)

부스팅 알고리즘

- 여러 개의 약한 학습기를 순차적으로 학습, 예측하면서 잘못 예측한 데이터에 가중치 부여를 통해 오류를 개선해 나가면서 학습하는 방식
- 대표적인 방식으로는 AdaBoost(Adaptive boosting) / 그래디언트 부스트

GBM

- 에이다부스트와 유사, 가중치 업데이트를 경사 하강법을 이용하는 것이 큰 차이
- 랜덤 포레스트보다는 예측 성능이 조금 뛰어난 경우가 많음
- 수행시간이 오래 걸리고 하이퍼 파라미터 튜닝 노력도 필요
- 사이킷런의 GradientBoostingClassifier 클래스
- 약한 학습기의 순차적인 예측 오류 보정을 통해 학습을 수행하므로 멀티 CPU 코어 시스템을 사용하덕라도 병렬 처리가 지원되지 않아서 대용량 데이터의 경우 학습에 매우 많은 시간이 필요

GBM 하이퍼 파라미터

- loss: 경사 하강법에서 사용할 비용 함수를 지정, 기본값은 'deviance'
- learning_rate: GBM이 학습을 진행할 때마다 적용하는 학습률
- n_estimators: weak learner의 개수
- subsample: weak learner가 학습에 사용하는 데이터의 샘플링 비율

XGBoost

XGBoost의 개요

- 일반적으로 다른 머신러닝보다 뛰어난 예측 성능을 가짐
- GBM을 기반으로 하지만 GBM의 단점 보완
- 주요 장점: 뛰어난 예측 성능, GBM 대비 빠른 수행 시간, 과적합 규제, Tree pruning, 자체 내장된 교차 검증, 결손값 자체 처리

LightGBM

XGBoost보다 학습에 걸리는 시간이 훨씬 적고 메모리 사용량도 상대적으로 적음

LightGBM과 XGBoost의 예측 성능은 별다른 차이가 없음

적은 데이터 세트(일반적으로 10,000건 이하)에 적용할 경우 과적합이 발생하기 쉬움

리프 중심 트리 분할 방식을 사용

오버피팅에 보다 강한 구조를 가질 수 있지만 균형을 맞추기 위한 시간이 필요함

카테고리형 피처의 자동 변한과 최적 분할

분류 실습 – 캐글 신용카드 사기 검출

- 언더 샘플링: 많은 데이터 세트를 적은 데이터 세트 수준으로 감소시키는 방식
- 오버 샘플링 이상 데이터와 같이 적은 데이터 세트를 증식하여 학습을 위한 충분한 데이터를 확보하는 방법