LGBM과 음주 예측

RoseDev1l

김신용, 이제선, 장주연, 이태윤, 황영진

LGBM(LightGBM) 정의

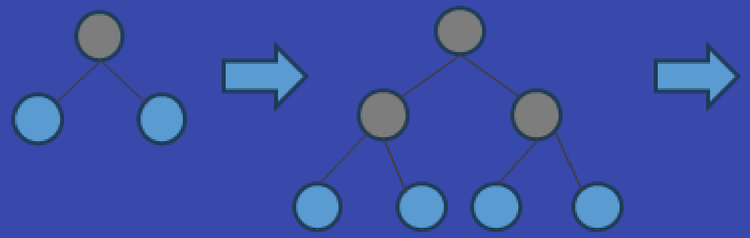
정의

LightGBM = “결정트리(Decision Tree)를 여러 개 쌓아서(Boosting) 오차를 줄이는” Gradient Boosting 모델입니다. 즉, 트리를 타고 내려가서 예측하고(추론), 학습은 손실함수의 기울기(gradient)를 이용해 다음 트리가 무엇을 고쳐야 하는지 정하는 방식(훈련)입니다.

LightGBM이 사용되는 문제 형태

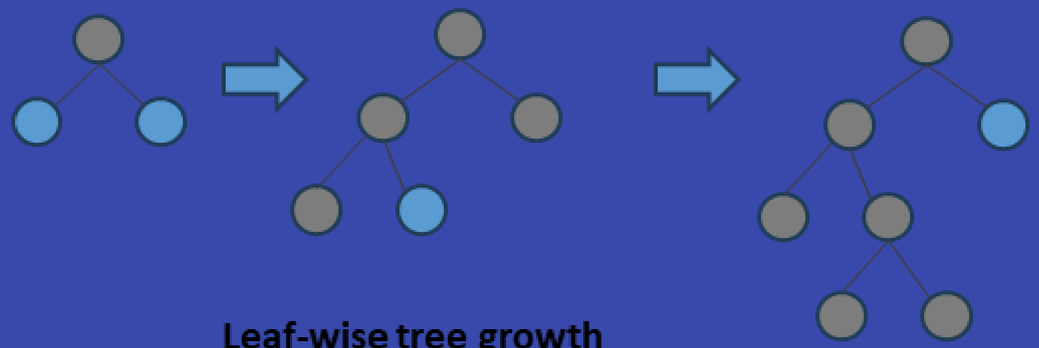
- 회귀 (값 예측)
- 이진 분류 (0 / 1)
- 다중분류 (여러가지의 클래스)
- 랭킹 (검색 / 추천에서 순서 맞추기)
- 이상치 점수처럼 응용가능

XGBoost



Level-wise tree growth

LightGBM



Leaf-wise tree growth

핵심 개념

1. Boosting (부스팅)
- 얕은 트리를 만들면서 이전모델의 실수 보완

2. Gradient Boosting (그래디언트 부스팅)
- 손실함수의 기울기를 확인후 수정해야하는 부분을 정함

3. Decision Tree 의 예측
- 트리 조건에 따라 방향이 나뉘고 leaf의 갑싱 예측 기여분이 됨

4. Additive Model (합산 모델)
- 트리출력 → 트리출력 → 트리출력 → 트리출력 → 방식

5. Learning Rate (학습률)
- 트리 하나가 고치는 양을 줄여서 천천히 학습

LGBM 사용

문제정의

음주 여부

데이터 수집

캐글

데이터 전처리

인포 확인하고 원핫 인코딩 완료,
다중 공선성 : 총콜레스테롤 &
LDL콜레스테롤 수치가 높아서 총
콜레스테롤을 삭제함

LGBM 사용

모델 설정

LightGBM

학습

AUC / Early Stopping 사용

모델 평가

Test Auc

LGBM 시행착오

이상치 제거법으로는 첫 시행에 이상결측치를 모두 제거를 했고,
이후, 그 다음은 아무 결측치를 삭제하지 않고 진행하였고,
마지막으로는 일부 결측치만 삭제했을때 학습률과 결과측이
일부 결측치만 삭제했을때가 제일 높았습니다

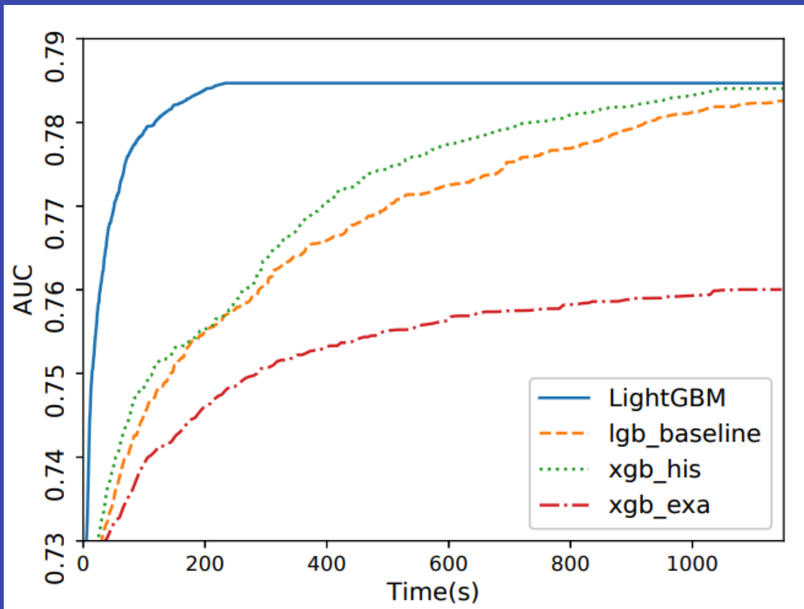


Figure 1: Time-AUC curve on Flight Delay.

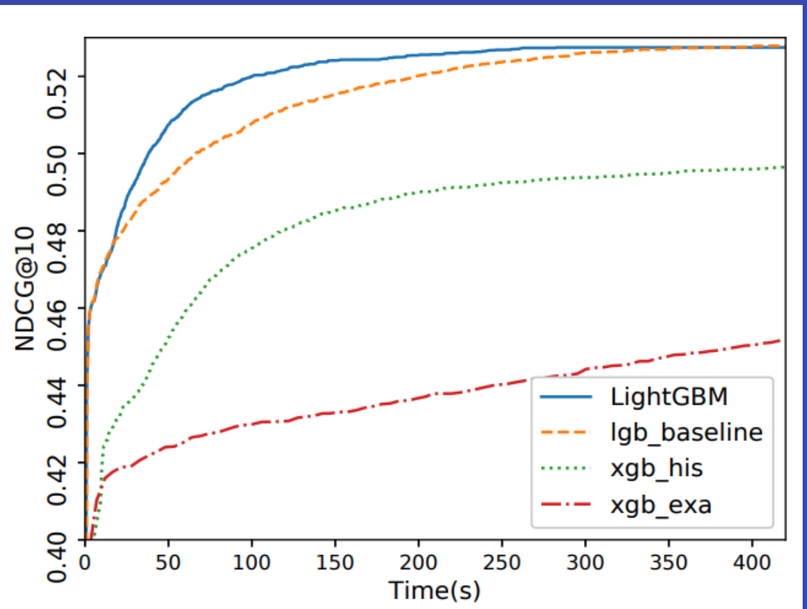


Figure 2: Time-NDCG curve on LETOR.