



GBM 심층 분석 보고서

Gradient Boosting Machine

1. 개요 및 핵심 철학

GBM(Gradient Boosting Machine)은 부스팅 기법에 '**경사 하강법(Gradient Descent)**'이라는 수학적 최적화 개념을 접목한 모델입니다. AdaBoost가 오답에 가중치를 주는 직관적인 방식이었다면, GBM은 "정답과 예측값의 차이(잔차, Residual)"를 수치적으로 미분하여 최소화하는 방향으로 모델을 학습시킵니다.



핵심 컨셉: "잔차(Residual) 학습"

"남은 오차만큼만 더 배운다."

첫 번째 모델이 맞추지 못한 오차(잔차)를 두 번째 모델의 '타겟'으로 설정합니다. 즉, 뒤에 오는 모델은 앞 모델이 남긴 찌꺼기(에러)를 해결하는 데에만 집중합니다.

⚙️ 2. 상세 작동 메커니즘

GBM은 손실 함수(Loss Function)의 기울기(Gradient)가 곧 잔차(Residual)와 같다는 수학적 원리를 이용합니다.

STEP 1. 초기 모델 생성

가장 단순한 예측값(예: 타겟의 평균)으로 초기 모델 F_0 를 만듭니다. 이때 발생하는 오차를 잔차(r_0)라고 합니다.

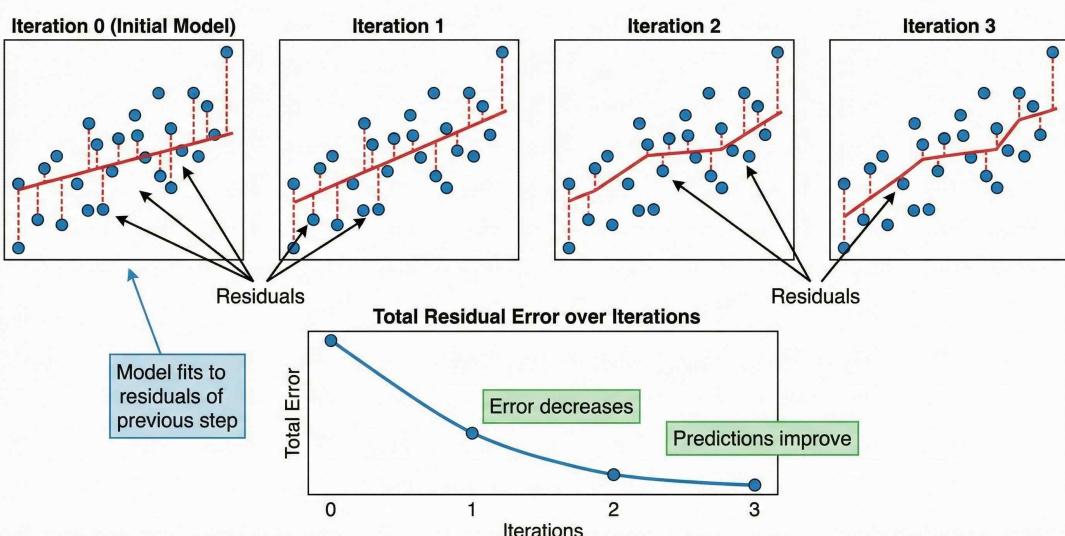
STEP 2. 잔차 예측 모델 학습

새로운 트리는 원본 데이터의 타겟(y)이 아니라, 이전 모델의 잔차($y - \hat{y}$)를 예측하도록 학습됩니다.

STEP 3. 모델 업데이트 및 반복

기존 모델에 (학습률 \times 새 모델)을 더해 업데이트합니다. $F_{\text{new}} = F_{\text{old}} + \eta \times h_{\text{residual}}$. 이 과정을 반복하면 오차가 점진적으로 0에 수렴합니다.

Gradient Boosting Machine (GBM) Residual Reduction Process



▣ 그림 1: 오차(Residual)가 점차 줄어드는 과정

3. AdaBoost vs GBM

두 모델 모두 부스팅 계열이지만, 오차를 처리하는 방식에서 결정적인 차이가 있습니다.

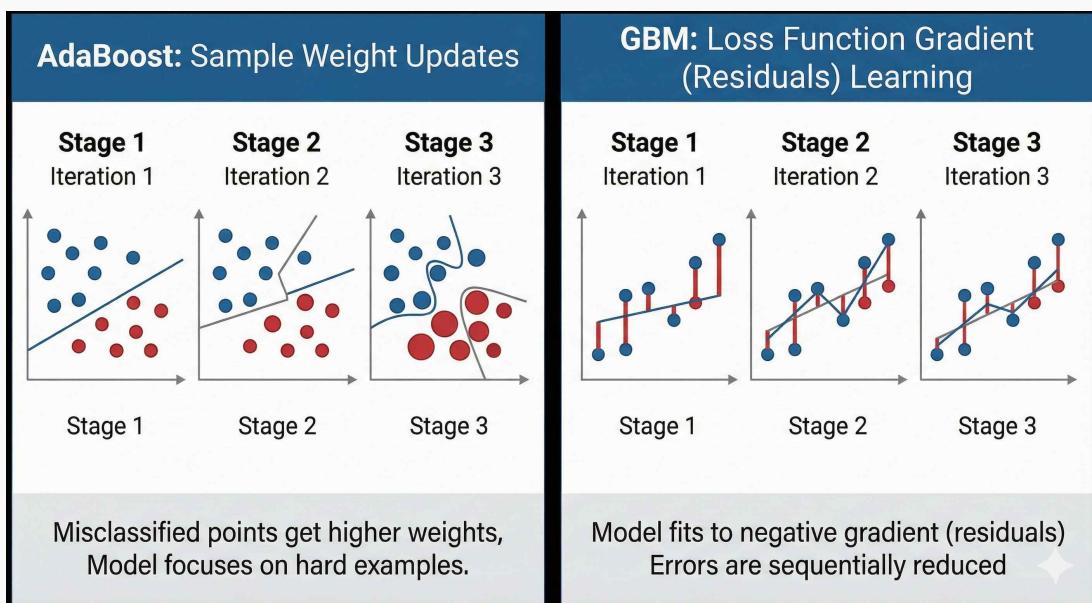


그림 2: 샘플 가중치(AdaBoost) vs 손실 함수 기울기(GBM)

항목	AdaBoost	GBM
오차 처리	잘못 분류된 샘플에 가중치 (Weight) 증가	손실 함수의 음의 기울기 (Gradient/잔차) 학습
약한 학습기	주로 Stump (깊이 1)	일반적으로 깊이 3~5 정도의 트리 (제한적)
민감도	이상치에 매우 민감	이상치에 민감하나, 손실 함수 변경으로 조절 가능
유연성	분류 문제에 특화	다양한 손실 함수 적용 가능 (회귀, 분류, 랭킹 등)

 4. 장단점 심층 분석 장점 (Pros)

- ✓ **높은 예측 성능:** 랜덤 포레스트보다 일반적으로 예측 오차가 적습니다.
- ✓ **다양한 손실 함수 지원:** 목적에 따라 Loss Function(MSE, Log Loss 등)을 자유롭게 교체할 수 있어 확장성이 좋습니다.

 단점 및 주의사항

- ✗ **과적합(Overfitting):** 훈련 데이터에 너무 집중하다 보니 과적합되기 쉽습니다. (트리 개수, 깊이 조절 필요)
- ✗ **긴 학습 시간:** 트리를 순차적으로 하나씩 만들기 때문에 병렬 처리가 어렵고 시간이 오래 걸립니다.

 5. 결론 (Conclusion)

“

GBM은
"수학적 최적화의 정석"을
보여주는 모델입니다.

”

최종 요약

- ✓ 현대 부스팅 알고리즘(XGBoost, LightGBM)의 **모태가 되는 기본 모델**입니다.
- ✓ 순수 GBM을 직접 쓰기보다는 속도와 규제가 보완된 XGBoost나 LightGBM을 사용하는 것이 실무적으로 유리합니다.
- ✓ "잔차를 학습한다"는 개념은 머신러닝의 핵심 아이디어 중 하나입니다.