

GRADIENT BOOSTING

핵심 원리 및 활용

순차 학습, 오차 보정, 앙상블

GBM 개요

GBM(Gradient Boosting Machine)은 여러 개의 약한 결정트리를 순차적으로 학습시키며, 이전 모델이 틀린 부분의 오차를 다음 모델이 보완하도록 설계된 앙상블 모델입니다. 단일 결정트리는 깊게 만들면 과적합, 얕게 만들면 표현력 부족 문제가 있어 이를 보완하기 위해 등장했습니다. RandomForest처럼 트리를 여러 개 쓰지만, 병렬 평균이 아니라 반드시 순서대로 추가한다는 점이 핵심입니다.

핵심 원리

- 데이터를 다시 뽑는 것이 아니라, 오차(잔차 또는 손실의 그래디언트)를 학습합니다.
- 새 트리 $h(x)$ 가 오차를 보정하고 기존 예측에 $learning_rate(\eta)$ 만큼 더해집니다: $F \leftarrow F + \eta \cdot h$
- $learning_rate$ 를 낮추고 트리 수를 늘리되, 검증셋 기반 early stopping으로 과적합을 제어합니다.

주요 장점

잘 못 맞춘 것에 집중

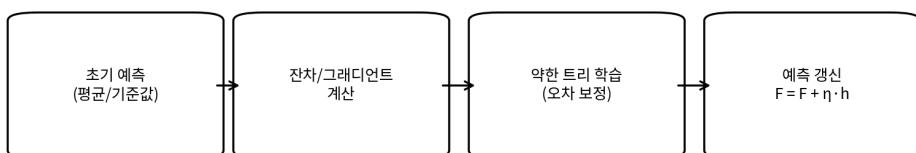
각 단계가 이전 모델의 실수를 보정하도록 학습되기 때문에, 단일 트리보다 높은 예측 성능을 내는 경우가 많습니다.

테이블 데이터에서 강력

트리 기반이라 스케일링이 필수는 아니고, 회귀·분류 모두에서 정밀한 예측이 필요할 때 강력합니다.

GBM learning_rate가 /
learning_rate n_estimators(
early stopping

GBM “
h(x)가 () , F
(learning_rate) : $F = F + \eta \cdot h$.



반복하면서 오차를 점점 줄임

설정/하이퍼파라미터 튜닝

GBM 가 (1)
learning_rate · n_estimators , (2) max_depth/min_samples_leaf
, (3) subsample(GBM)
/ (: RMSE, AUC) 가

n_estimators	100 – 2000		
learning_rate	0.01 – 0.2		
max_depth	2 – 5	,	
subsample	0.5 – 1.0	,	
	1 – 50	,	
max_features	sqrt / log2 / 0.3 – 1.0		

/ (1) learning_rate – n_estimators , (2)
 max_depth/min_samples_leaf , (3) subsample , (4)
 early stopping .

결론

GBM () 가 . learning_rate .
 . , early stopping .

참고 문헌 (추천)

- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*.
- Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.
- Bühlmann, P., & Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD*.
- Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NeurIPS*.