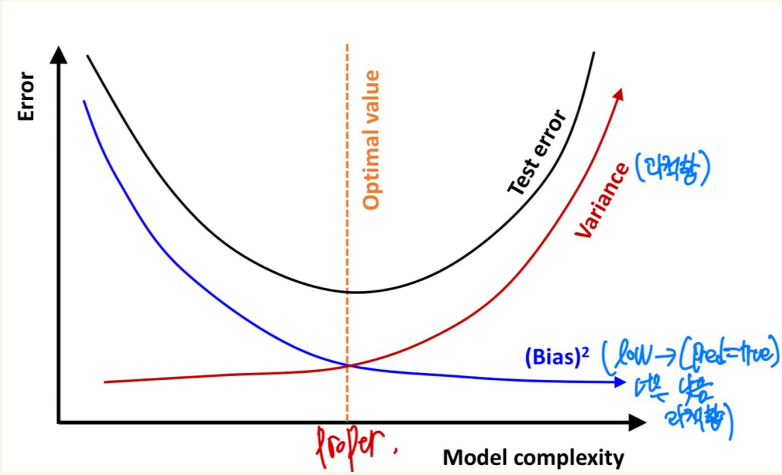
① bias → 예측값들의 평균과 실제값과의 차이.
(정답에 얼마나 가까운지 판단.)
bias ↑ → 정답과 차이 ↑ → 정답 잘 못 맞춤.

② variance → 여러번 값을 시도해 다른 정도.
variance ↑ → 값들의 예측의 차이↑ → 데이터셋 의존.

따라서 bias 높다는 것은 정답 안맞춤 Underfitting.
variance 높다는 것은 특정 데이터셋에 overfitting.



Bias와 variance trade off.

앙상블 Bias와 variance를 줄이기 위한 방법.

↓

앙상블

① Bagging.

→ 복원추출하여 여러 데이터셋 만듦.
(bootstrap)

Aggregating 방법 ① Majority voting
② Weighted voting
③ Stacking.
↳ 예측값을 학습데이터 input으로 활용.

Bagging → 독립성↑ 가정하여 variance가 1/n로 줄어들지만 보통 그렇지 않아서 효과↓

Random forest

Bagging 한 종류로 변수 랜덤하게 선택.

∴ 다른 값들의 Bootstrap이 특징적.

⊕ 변수 중요도까지 산출

② Boosting

Ada boost → 약한 학습기.

틀린것들 데이터 셋 찾고 그것이 분류되게 weight 수정.

$$w_{(i)} \leftarrow w^{(i)} \cdot \exp\left(-\alpha_t y_{(i)} h_t(x^{(i)})\right)$$

→ 모두 -1 곱할 때

맞추면 w↑

틀리면 w↓

GBM

- 경사하강법 + Boosting.

  잔차를 예측하고 그것을 이용하여 보완해 나감 모델에 사용.

  Loss function의 negative gradient = 잔차

XGBoost

- GBM의 상위 버전

  속도 + 성능 향상 (병렬 처리)