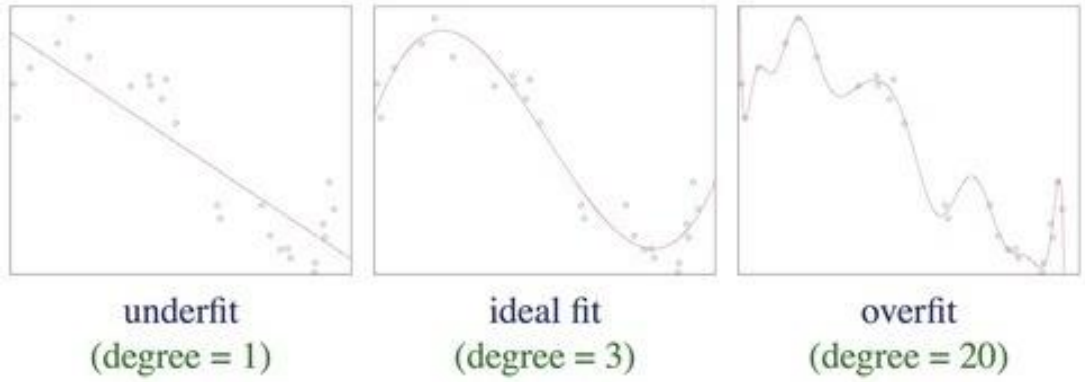
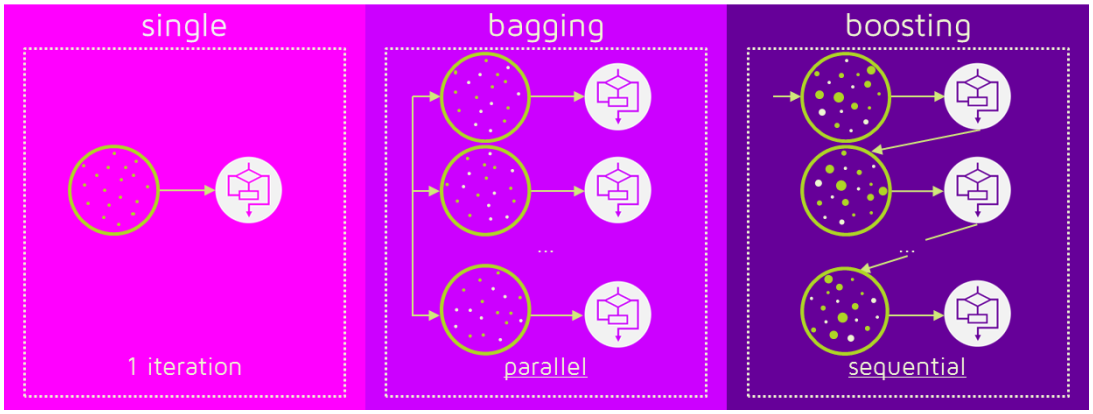
1. 편향과 분산

편향과 분산은 tradeoff 관계에 있다. 편향은 underfitting된 데이터셋에서 높게 나타나는 경향이 있으며, 분산은 overfitting된 데이터셋에서 높게 나타나는 경향이 있다.



2. 배깅과 부스팅



그림에서 나타내는 바와 같이 배깅은 병렬로 학습하는 반면, 부스팅은 순차적으로 학습한다. 한번 학습이 끝난 후 결과에 따라 가중치를 부여합니다. 그렇게 부여된 가중치가 다음 모델의 결과 예측에 영향을 준다. 따라서 순차적 학습만이 가능하다. 부스팅은 오답에 대해서는 높은 가중치를 부여하고, 정답에 대해서는 낮은 가중치를 부여한다. 따라서 오답에 더 집중할 수 있게 된다.

따라서 overfitting의 가능성이 적고, 더욱 확실한 성능 개선을 위해서는 부스팅이 적합하고, overfitting의 가능성을 배제할 수 없다면 배깅이 적합하다고 이야기할 수 있다. 이는 bias, variance의 tradeoff와 함께 고려하여 선택할 수 있다. 두 방식 모두 앙상블의 일종이므로, 단일 모델을 사용할 때보다 정확도를 높일 수 있는데, 이는 hard voting, soft voting, stacking 등의 방식으로 클래스 판단의 오류 가능성을 줄이기 때문이다.

Bagging: random forest

Boosting: adaboost, xgboost, gbm, lgbm