# 5주차 Ensemble 강의 정리

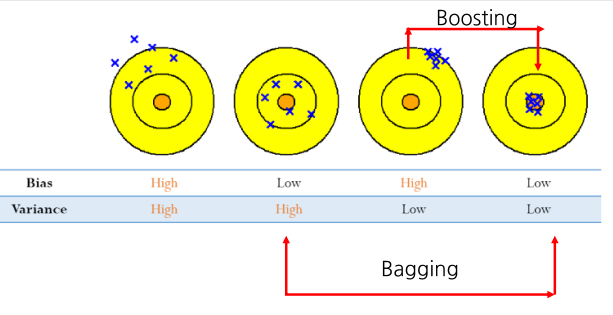
1. 분산과 편차에 따른 모델 복잡도

편차: the amount by which the average estimator differs from the truth

높을수록 정답을 잘 못맞춘다는 특징을 가짐

분산: spread of the individual estimations around their mean

높을수록 퍼져있는 경향이 있음



앙상블에는 분산을 줄이기 위한 배깅, 랜덤포레스트와 편차를 줄이는 AdaBoost가 있다.

편차가 낮고 분산이 높은 경우에는 모델 복잡도가 낮다.

반대로 편차가 높고 분산이 낮은 경우에는 모델 복잡도가 크다.

1. 배깅과 부스팅 모델의 개념과 차이점

배깅은 데이터의 다양성을 확보하자는 목적으로 k-fold data split에서부터 처음으로 시도되었다.

본래 데이터 셋에서 원하는만큼 복원추출한 데이터 셋을 만드는 것이 주된 목적이며 이 때 각각의 샘플을 bootstrap이라고 부른다. 대표적인 방식으로 랜덤포레스트가 있다.

반면 부스팅은 약한 학습기에 집중하여 특정 데이터 셋에 대한 모델링이 끝난 후 잘못 분류된 데이터에 대해 다음 학습에서 뽑힐 확률이 높아지도록 샘플링하는 방식이다. Gradient Boosting, XGBoost 등의 모델이 존재한다.