Ensemble

개념

앙상블은 여러 개별 예측 모델을 합쳐, 성능이 좋고 강건한 모델을 만드는 기법이다.

분산과 편차에 따른 모델 복잡도

- 분산
  - 모델의 분산은 데이터셋의 변동에 얼마나 민감하게 반응하는지를 나타낸다.
  - 분산이 높은 경우 Overfitting 되었을 가능성이 높다.
  - 즉, 훈련 데이터에 대한 예측에 대해 높은 정확도를 가지지만, 새로운 데이터에 대해서는 정확도가 떨어진다.
- 편차
  - 모델의 편차는 실제 값과 모델의 예측 값 사이의 차이를 나타낸다.
  - 편차가 높은 경우 Underfitting 되었을 가능성이 높다.
  - 즉, 모델이 너무 간단하여 실제 데이터 패턴을 잘 포착하지 못한다.

- 앙상블은 분산과 편차 간의 균형을 맞추는데 도움을 주며, 분산과 편차를 감소시키는 역할을 할 수 있다.

배깅과 부스팅

- 배깅
  - 배깅은 훈련 데이터셋을 무작위로 복원 추출하여 각 추출 데이터셋에 대해 독립적으로 모델을 한 후, 해당 모델들의 예측에 대해 평균을 취하거나 다수결로 투표하여 최종 예측을 생성하는 방법이다.
  - 대표적으로 Random Forest 알고리즘이 있다.
- 부스팅

    - 부스팅은 성능이 좋지 못한 모델을 순차적으로 학습시켜 성능을 향상시키는 방법이다. 이전 모델의 오차를 보고 그 오차를 보완하는 새로운 모델을 학습시킨다.
    - 대표적으로 Gradient Bossting, AdaBoost 가 있다.
- 차이점
    - 배깅
        - 모델을 병렬적으로 학습시킬 수 있으므로 학습 시간을 줄일 수 있다.
        - 각 모델은 서로 독립적이다.
    - 부스팅
        - 모델을 순차적으로 학습한다
        - 성능이 좋지 못한 모델의 오차를 보고 학습하므로, 모델간의 연관성이 존재한다.