5주차_강의내용 정리

분산과 편차에 따른 모델 복잡도

1. 정의

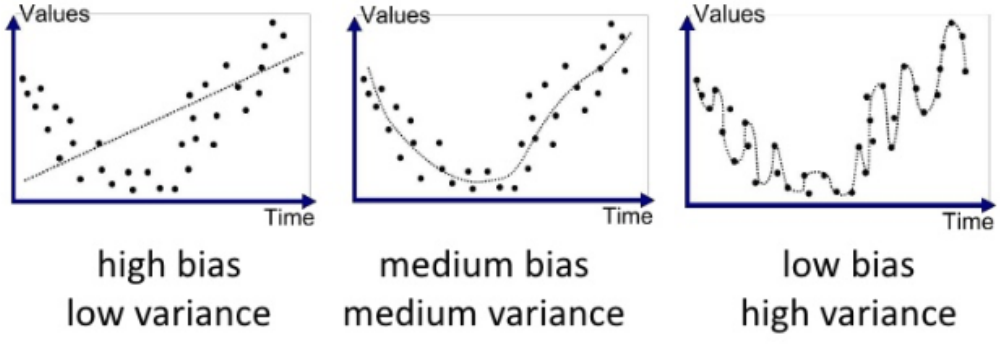
- 분산: 예측치가 예측된 값들의 평균에서 얼마나 떨어져 있는지를 나타내는 정도, 분산이 클수록 정답이 불안정해지는 경우가 많다
- 편차: 예측치가 실제값과 얼마나 떨어져 있는지를 나타내는 정도, 높은 편차는 떨어지는 정확도를 의미

2. 수학적으로 바라보았을 때

$$Err(x_0) = Bias^2(\hat{F}(x_0)) + Var(\hat{F}(x_0)) + \sigma^2$$

실제 값을 예측하기위해 모델 $\hat{F}$를 만들고, 생성된 모델이 실제 값을 얼마나 잘예측하는지를 확인하기 위함.

3. 모델 복잡도



bias가 크고 variance 가 작다면, 첫번 째 사진처럼 단순한 형태의 모델일 것이고, bias 가 작고 variance가 크다면, 3번째 그림처럼 복잡한 모델을 가질것입니다.

첫번째 그림처럼 모델이 너무 단순하다면, underfitting이 일어날 것이고,

세번째 그림처러 모델이 너무 단순하다면, overfitting이 일어날 것입니다.

앙상블

1. 배깅

- 원래 데이터셋에서 복원추출(Boostrapping)을 통해서 데이터셋을 여러개 만들고, 생성된 여러개의 데이터셋을 이용하여 모델 구축.
- 배깅은 여러번 데이터를 활용하므로, 낮은 편차, 높은 분산(낮은 모델 복잡도)를 보여준다
- Training set, 복원추출을 통해 여러개의 데이터셋 그룹을 만들고, 각각의 예측치를 Ensemble하여 최종 예측값 도출
- 부트스트랩 할 경우, 뽑히지 않는 약 1/3개의 데이터를 통해 검증을 진행
- 대표적인 예로 랜덤 포레스트가 존재
- 단점:
  - 복원추출의 한계 → 독립이라는 보장이 없음
  - 비슷한 Tree 가 부트스트래핑을 해도 비슷한 트리가 만들어질 확률이 높음

2. 부스팅

- 특정 데이터 셋에 대한 학습, 모델링을 진행한 후, 잘못 classify된 데이터를 찾고, 해당 데이터를 다음에는 잘 하게 할 수 있도록 sampling 진행
- 대표적으로 GBM, XGboost 등이 존대
- 가장 높은 성능,속도과 뛰어난 해석력
- Overfitting에 강함

- 단점
    - 결측치, 이상치에 취약, 사전에 이상치를 처리해줘야함