K-fold 교차 검증

교차검증

표본내 성능과 표본외 성능

- 표본내 성능 검증 (in-sample testing) : 학습 데이터 집합의 종속 변수값을 얼마나 잘 예측 하였는지를 나타내는 성능
- 표본외 성능 (out-of-sample testing) / 교차검증(cross validation) : 학습에 쓰이지 않는 표본 데이터 집합의 종속 변수값을 얼마나 잘 예측 하였는지를 나타내는 성능

교차검증을 통한 과최적화 파악

과최적화

표본내 성능은 좋으나, 표본외 성능이 상대적으로 많이 떨어짐 → 과최적화 파악.

과최적화가 발생하면 학습에 쓰였던 표본 데이터에 대해서는 종속 변수값을 잘 추정하지만, 새로운 데이터를 예측하지 못하기 때문에 예측 목적으로 쓸 수 없는 모델/모형이 됨

검증용 데이터 집합

- training data set : 모형 추적, 즉 학습을 위한 데이터 집합
- test data set : 성능 검증을 위한 데이터 집합

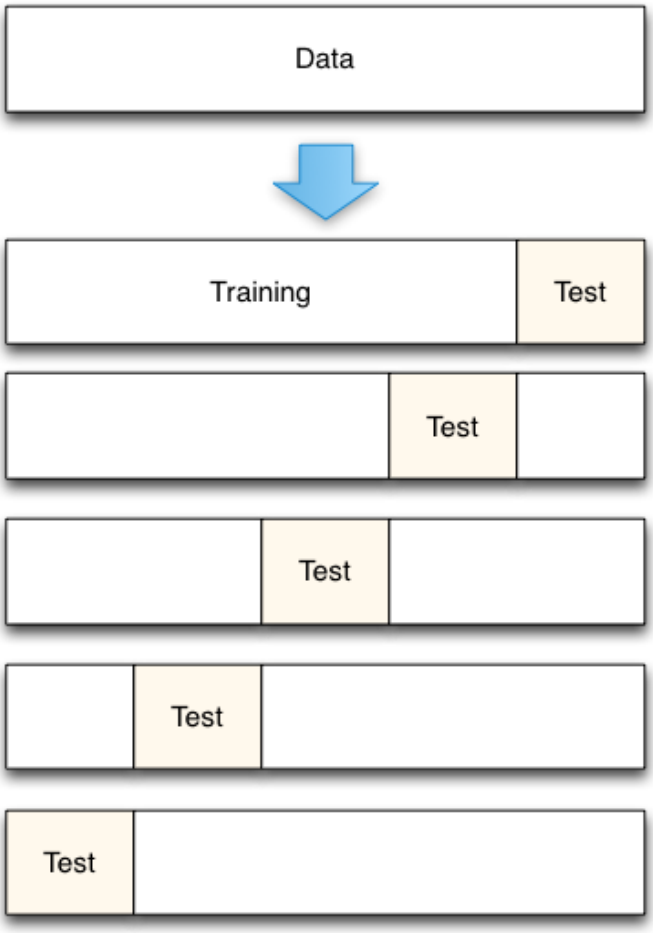
보통 가지고 있는 데이터 집합을 학습용과 검증용으로 나눔 → 학습 / 검증 데이터 분리 (train-test split)

K-fold 교차 검증

데이터의 수가 적은 경우, 이 데이터 중의 일부인 검증 데이터의 수도 적음 → 검증 성능의 신뢰도가 떨어짐.

그렇다고 검증 데이터를 늘리면 학습용 데이터의 수가 상대적으로 작아짐 → 이러한 딜레마를 해결하기 위해 **모든 데이터를 훈련 및 검증에 사용하는 K-fold 교차 검증**을 사용함

또한 K-fold 교차검증은 데이터가 충분하더라도 검증 데이터가 쏠리는 문제를 해결할 수 있음. 즉, 성능 평가 일반화.



단계

1. 전체 데이터를 K개의 부분집합($\{D_1, D_2, ..., D_K\}$)으로 나눔.
2. 데이터 ($\{D_1, D_2, ..., D_{K-1}\}$)를 학습용 데이터로 사용하여 회귀분석 모형을 만들고, 데이터 ($\{ D_k \}$)로 교차 검증 함
3. 데이터 ($\{D_1, D_2, ..., D_{k-2}, D_K\}$)를 학습용 데이터로 사용하여 회귀분석 모형을 만들고, 데이터 ($\{ D_{k-1} \}$)로 교차 검증 함

.

.

.

4. 데이터 ($\{D_2, ..., D_{k-1}, D_K\}$)를 학습용 데이터로 사용하여 회귀분석 모형을 만들고, 데이터 ($\{ D_1 \}$)로 교차 검증 함
5. 총 K개의 모형과 K개의 교차성능 검증이 나오면, 이 K개의 교차검증 성능을 평균하여 최종 교차검증 성능을 계산

참조 자료

- https://datascienceschool.net/view-notebook/266d699d748847b3a3aa7b9805b846ae/
- https://devkor.tistory.com/entry/머신-러닝-입문자를-위한-설명-교차-검증K-Fold-Cross-Validation