Part. 01

Machine Learning의 개념과 종류

| 모형의 적합성 평가 및 실험 설계

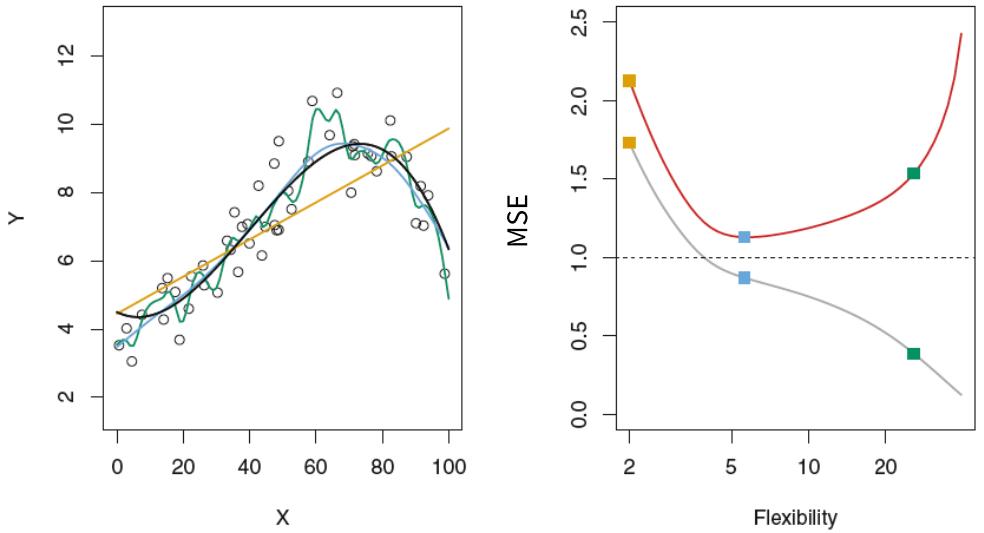
FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

강사. 이경택

I 모형의 적합성 평가 및 실험 설계

▪ 모형의 적합성을 평가하는 방법

- 모형의 복잡도에 따른 학습 집합의 MSE(회색)와 검증 집합의 MSE(빨간색)의 변화는 아래 그림과 같음
- 학습 집합의 MSE는 복잡한 모형일수록 감소하지만, 학습 데이터가 아닌 또 다른 데이터 (검증 데이터)의 MSE는 일정 시점 이후로 증가
- 증가하는 원인은 왼쪽 그림과 같이 모형이 학습 집합에 과적합되기 때문

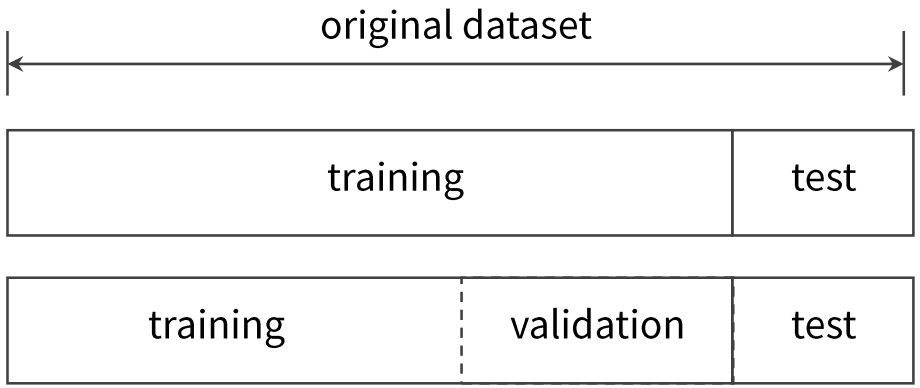


- 검은색: 실제 f를 나타내는 모형
- 노란색: 가장 낮은 복잡도를 가지므로 편파성이 높아져 가장 높은 MSE 값을 가짐
- 초록색: 가장 높은 복잡도를 가지므로 학습 집합에 과적합 되어 분산이 높아짐. 따라서 검증 데이터의 MSE가 하늘색에 비해 상승함
- 하늘색: 검은색 모형과 가장 유사한 형태로, 분산과 편파성이 모두 적절히 낮아져 검증 데이터의 MSE가 가장 낮음

MSE(mean squared error)

I 모형의 적합성 평가 및 실험 설계

- 데이터 분할



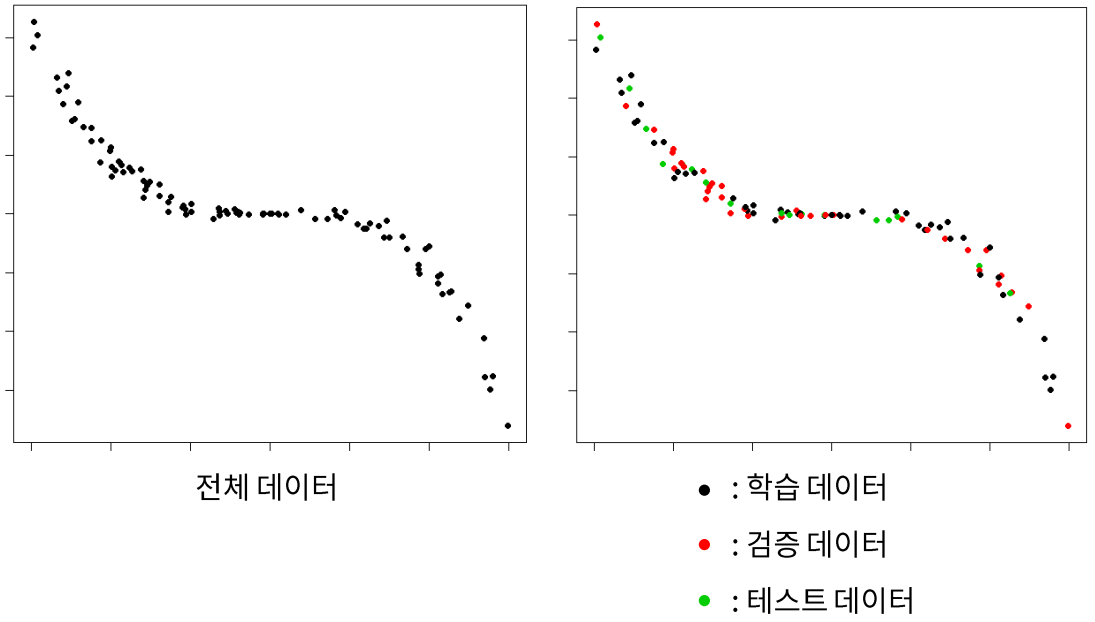
- 과적합을 방지하기 위해 전체 데이터를 학습 데이터, 검증 데이터, 테스트 데이터로 나누며 보통 비율은 5:3:2로 정함
  - 학습 데이터(training data): 모형 f를 추정하는데 필요
  - 검증 데이터(validation data): 추정한 모형 f 가 적합한지 검증함
  - 테스트 데이터(test data): 최종적으로 선택한 모형의 성능을 평가

I 모형의 적합성 평가 및 실험 설계

- 데이터 분할

① 데이터 분할

- 전체 데이터를 학습 데이터, 검증 데이터, 테스트 데이터로 나눔



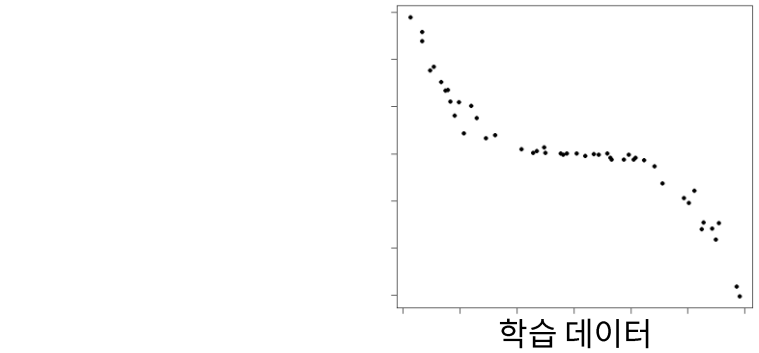
② ③ ④

I 모형의 적합성 평가 및 실험 설계

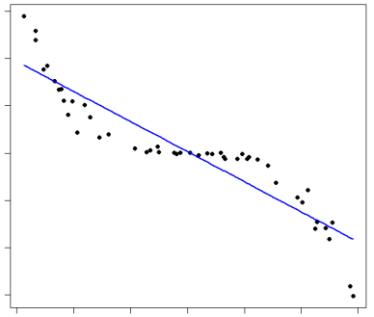
- 데이터 분할

① | ② 모형 학습 | ③ | ④

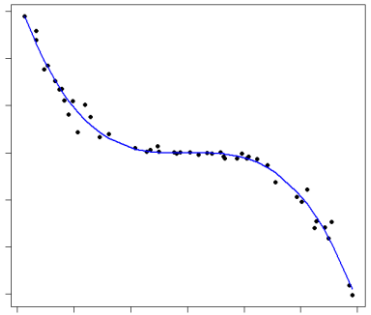
- 학습 데이터를 사용하여 각 모형을 학습함



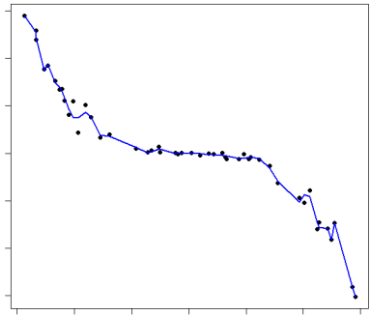
학습 데이터



모형 A (underfitting)



모형 B



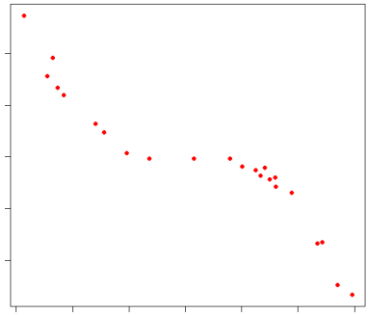
모형 C
(overfitting)

I 모형의 적합성 평가 및 실험 설계

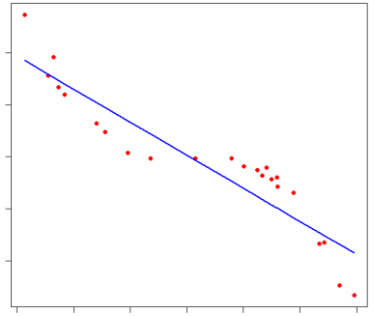
- 데이터 분할

① ② ③ 모형 선택 ④

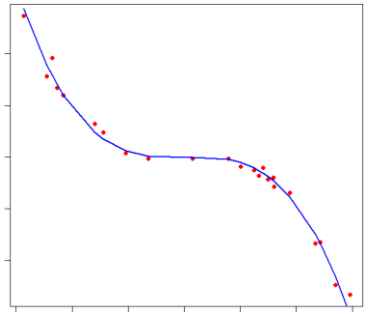
- 검증 데이터를 사용하여 각 모형의 성능을 비교하고 모형 선택



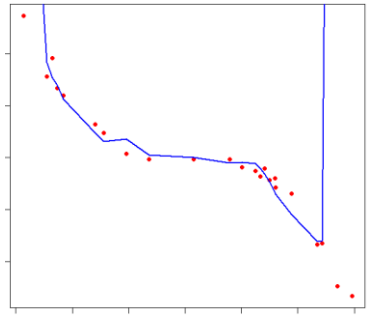
검증 데이터



모형 A (MSE=24.42)



모형 B (MSE=2.14)



모형 C
(MSE=1854.05)

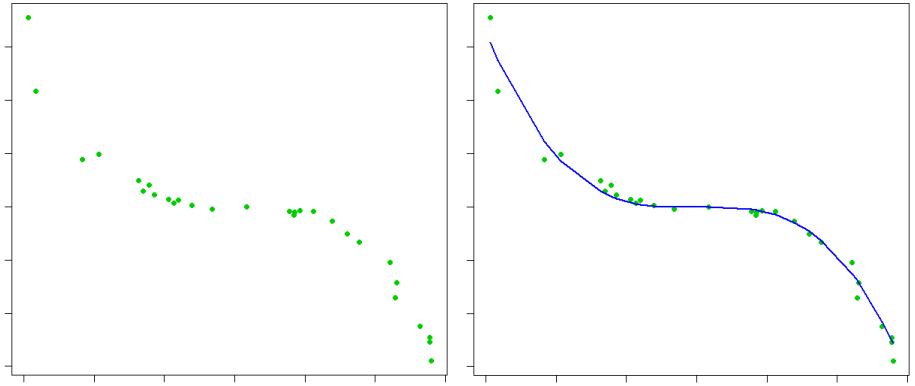
MSE (mean squared error)

I 모형의 적합성 평가 및 실험 설계

- 데이터 분할

①	②	③	④ 최종 성능 지표 도출

- 테스트 데이터를 사용하여 검증 데이터로 도출한 최종 모델의 성능 지표를 계산함
- 모형 B의 최종 성능 지표는 3.77이라고 할 수 있음



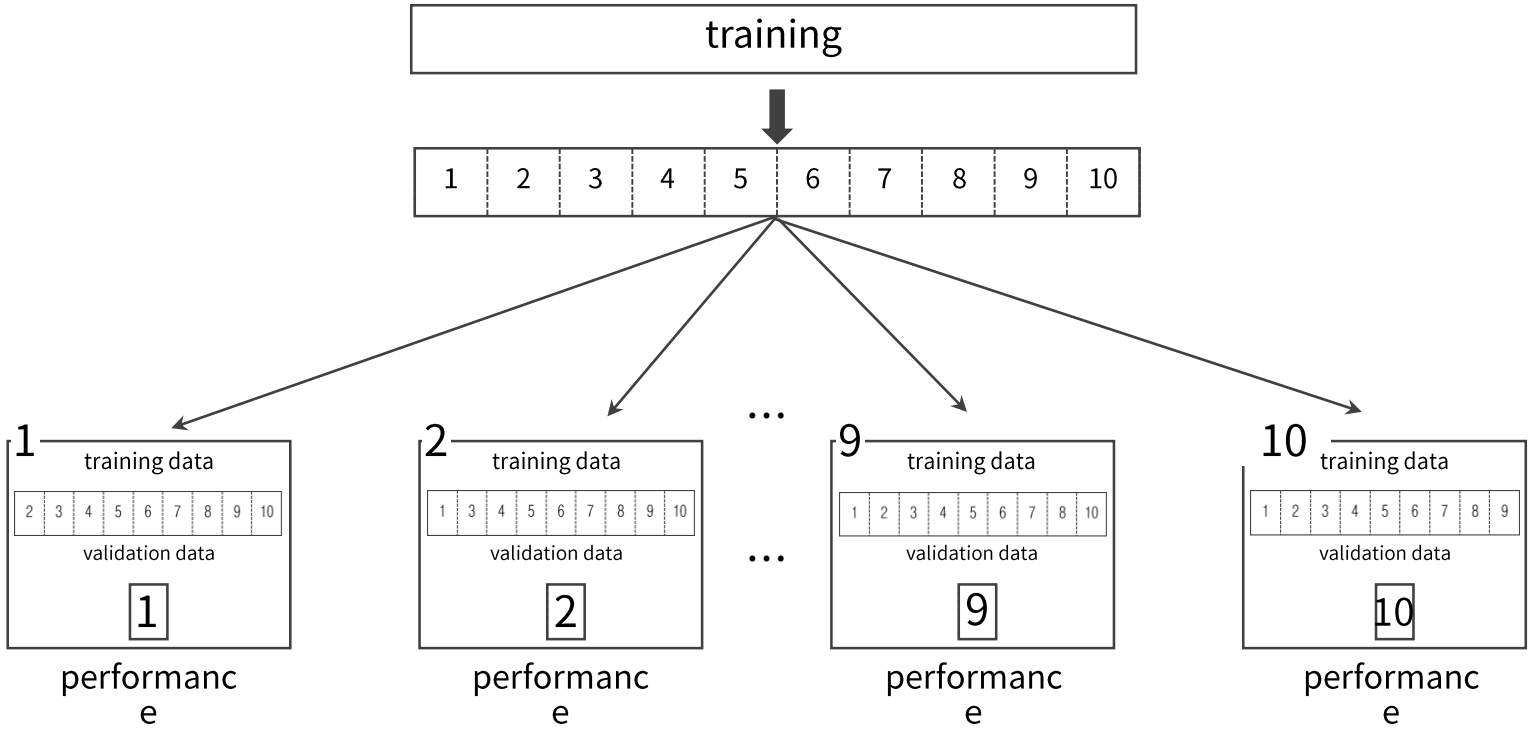
테스트 데이터 모형 B (MSE=3.77)

- 데이터를 어떻게 분리하느냐에 따라 성능에 영향이 달라지며, 이는 최종 모형 선택에 영향을 끼칠 수 있음 (k-fold cross validation)

MSE (mean squared error)

I 모형의 적합성 평가 및 실험 설계

- k-Fold 교차검증(k-Fold Cross Validation)
  - 모형의 적합성을 보다 객관적으로 평가하기 위한 방법
  - 데이터를 k(주로 5 또는 10)개 부분으로 나눈 뒤, 그 중 하나를 검증 집합, 나머지를 학습 집합으로 분류
  - 위 과정을 k번 반복하고 k개의 성능 지표를 **평균하여** 모형의 적합성을 평가



I 모형의 적합성 평가 및 실험 설계

- LOOCV(Leave-One-Out Cross Validation)
  - 데이터의 수가 적을 때 사용하는 교차검증 방법
  - 총 n(데이터 수 만큼)개의 모델을 만드는데, 각 모델은 하나의 샘플만 제외하면서 모델을 만들고 제외한 샘플로 성능 지표를 계산함. 이렇게 도출된 n개의 성능 지표를 평균 내어 최종 성능 지표를 도출