Part. 01

Machine Learning의 개념과 종류

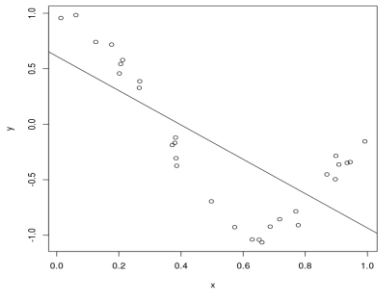
| 과적합(Overfitting)이란

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

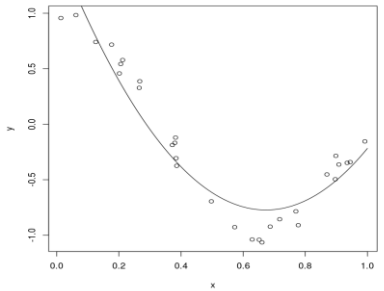
강사. 이경택

I 과적합(Overfitting)

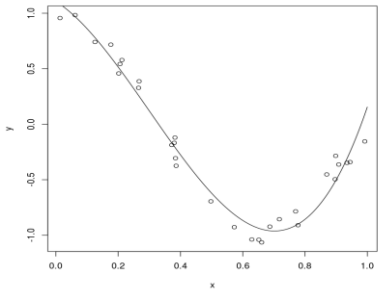
- 과적합이란
  - 복잡한 모형일수록, 데이터가 적을수록 과적합이 일어나기 쉬움
  - 아래 그림은 회귀분석에서 고차항을 넣었을때 만들어지는 직선
  - 과적합은 data science 뿐만 아니라 AI전반적으로 매우 큰 이슈



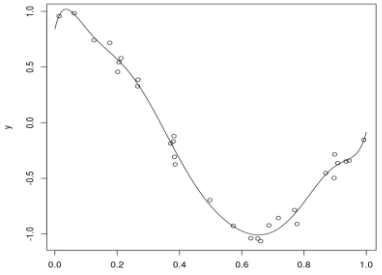
1차항 고려



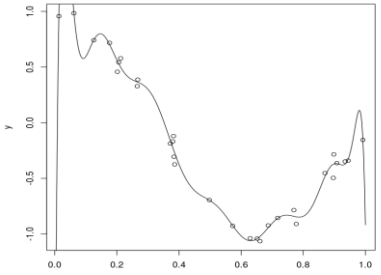
2차항 고려



3차항 고려



4차항 고려 (과적합 발생)



5차항 고려 (과적합 발생)

I 과적합(Overfitting)

- 분산(Variance)과 편파성(Bias)의 트레이드오프(Tradeoff) (Dilemma)
  - 모형 $\hat{f}(X)$ 로 모집단의 전체 데이터를 예측할 때 발생하는 총 error를 계산하면 reducible error와 irreducible error 로 표현되며, reducible error는 다시 분산과 편파성으로 구성

$$E\left(Y-\hat{Y}\right)^2 = E\left[f(X)+\epsilon-\hat{f}(X)\right]^2$$

$$= Var\left(\hat{f}(X)\right) + \left[Bias\left(\hat{f}(x)\right)\right]^2 + Var(\epsilon)$$

오류자승의 평균 (기대 값)
mean squared error (MSE)

irreducible error

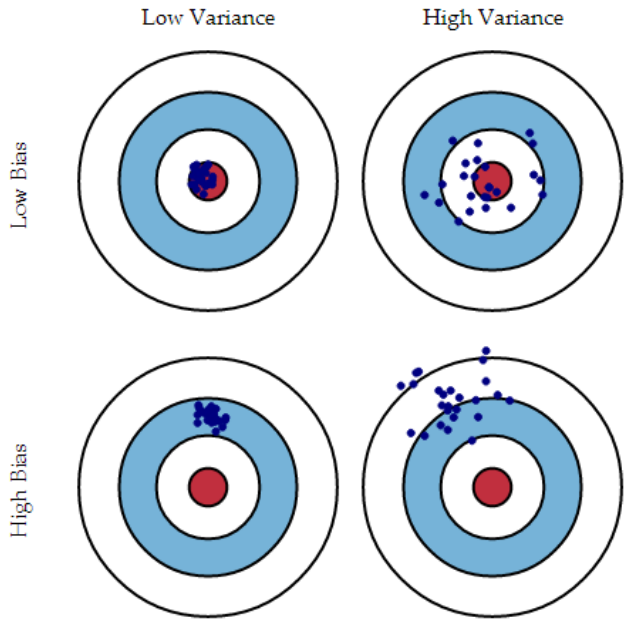
분산

편파성

  - 분산: 전체 데이터 집합 중 다른 학습 데이터를 이용했을 때, $\hat{f}$이 변하는 정도(복잡한 모형일수록 분산이 높음)
  - 편파성: 학습 알고리즘에서 잘못된 가정을 했을 때 발생하는 오차(간단한 모형일수록 편파성이 높음)
  - 복잡한 모형 $\hat{f}(X)$을 사용하여 편파성을 줄이면, 반대로 분산이 커짐 (간단한 모형일 경우엔 반대의 현상이 발생)
  - 따라서 분산과 편파성이 작은 모형을 찾아야 함

I 과적합(Overfitting)

- 분산(Variance)과 편파성(Bias)의 트레이드오프(Tradeoff) (Dilemma)



적절한 모형 선택과 실험 설계를 통한 과적합 방지