

MACHINE LEARNING WITH PYTHON

PART III

**MODEL & DATA &
LEARNING**

ML 모델과 데이터와 학습

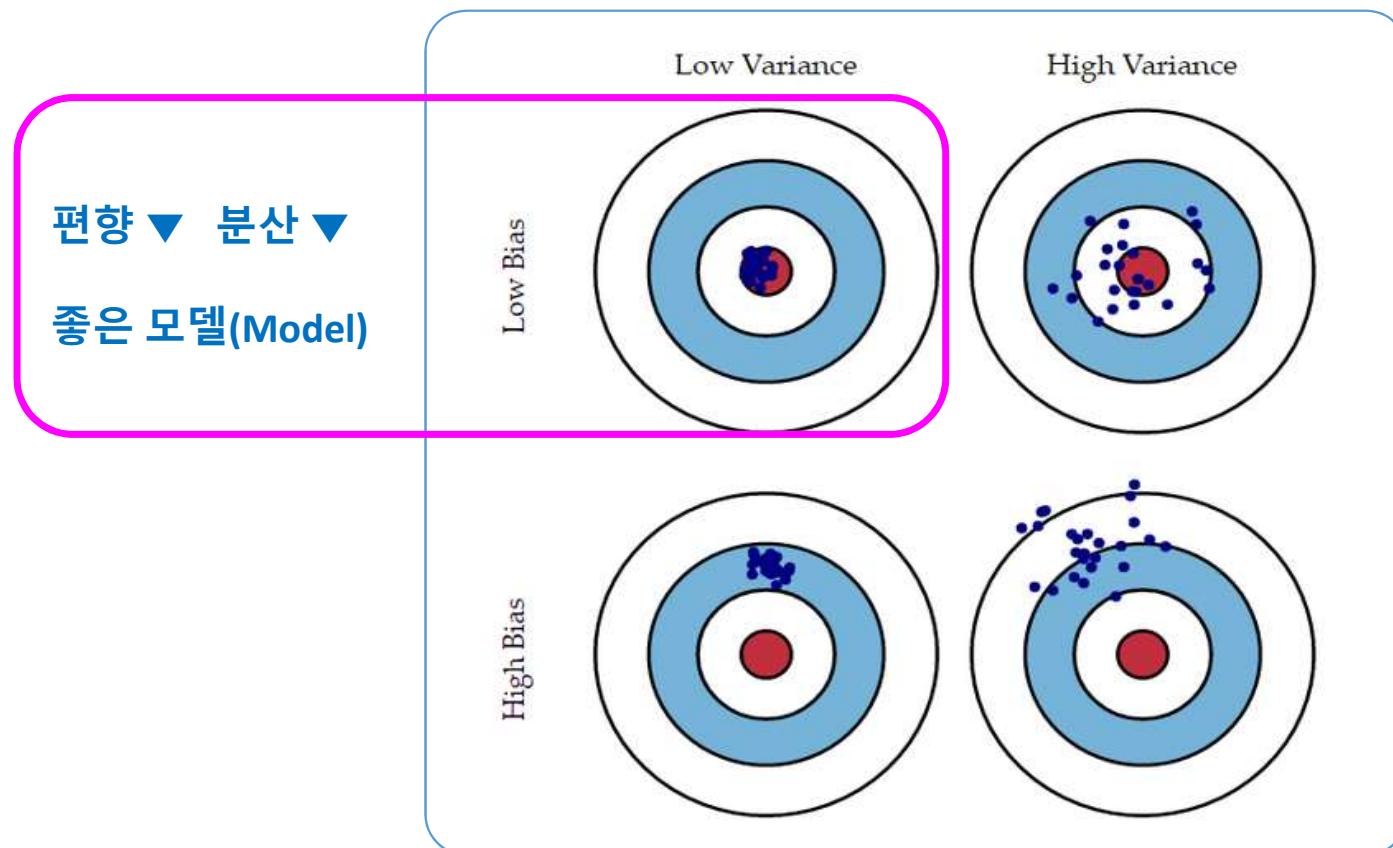
◆ 편향과 분산

❖ 모델 예측값에 대한 표현

- 편향 (Bias) → 예측값과 정답 간의 관계
- 분산(Variance) → 예측값 끼리의 관계

ML 모델과 데이터와 학습

◆ 편향과 분산



ML 모델과 데이터와 학습

◆ 과대적합(Overfitting)

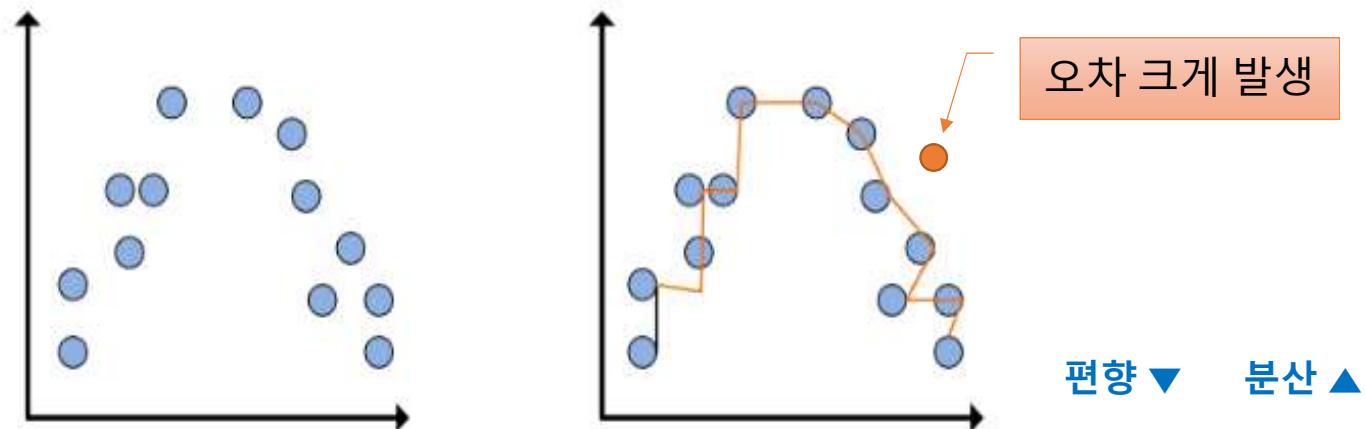
- 훈련 데이터 셋 특화된 즉 최적화된 모델
- 새로운 데이터에 대한 **오차가 매우 커짐**

- 원인
 - 데이터(특성)이 많아 **모델** 지나치게 **복잡**
 - 너무 **많은 학습**

ML 모델과 데이터와 학습

◆ 과대적합(Overfitting)

- 훈련 데이터 : 오차 없음
- 새로운 데이터 : 오차 크게 발생 확률 높음!



ML 모델과 데이터와 학습

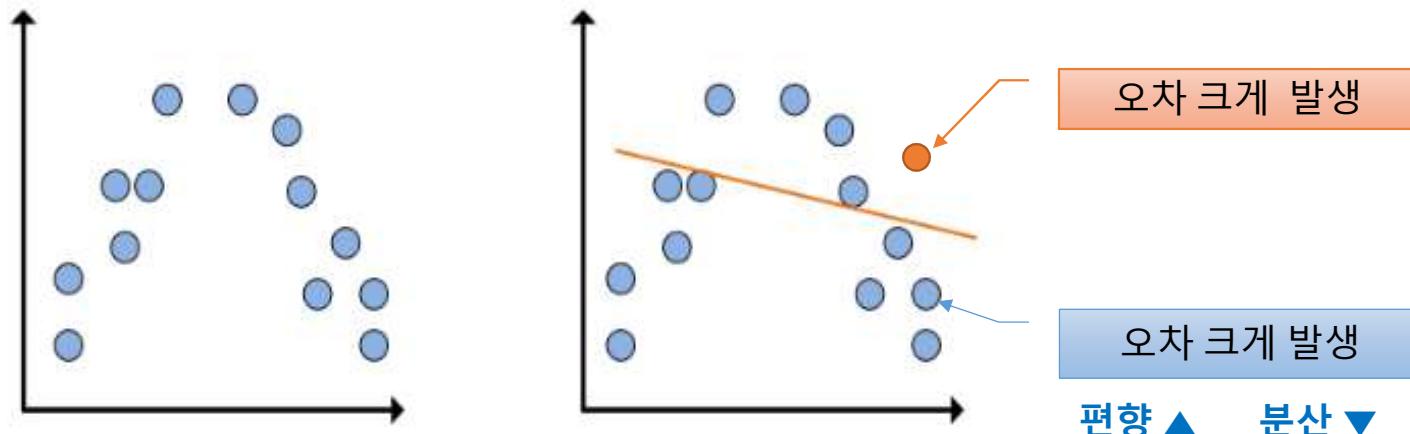
◆ 과소적합(Underfitting)

- 훈련 데이터 셋의 규칙/패턴 반영되지 않는 모델
- 훈련 데이터와 새로운 데이터 대한 오차가 매우 커짐
- 훈련 데이터 오류가 줄어들지 않음
- 원인
 - 데이터(특성)이 부족하여 모델 지나치게 단순
 - 학습 횟수 부족

ML 모델과 데이터와 학습

◆ 과소적합(Underfitting)

- 훈련 데이터 → 오차 크게 발생
- 새로운 데이터 → 오차 크게 발생 확률 높음!



ML 모델과 데이터와 학습

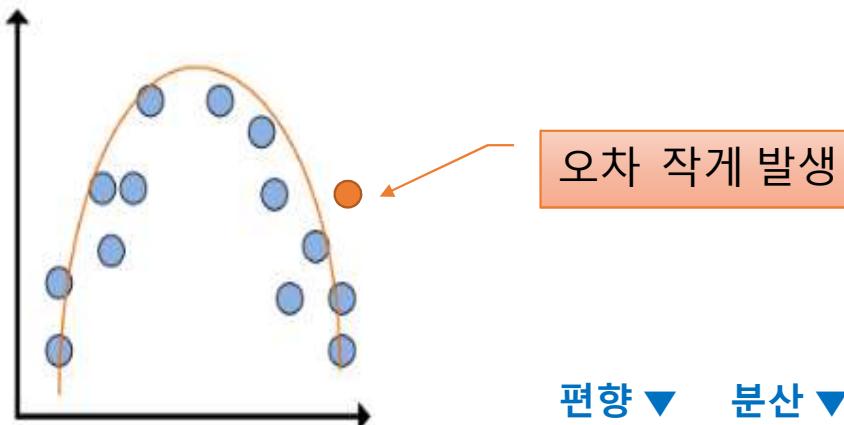
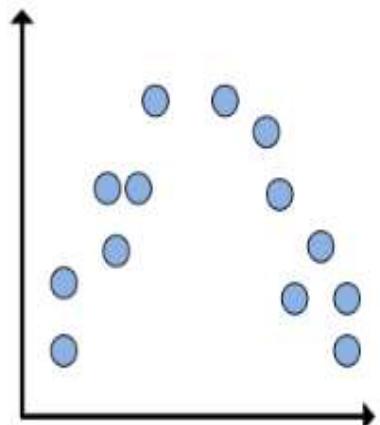
◆ 최적적합(Optimalfitting)

- 훈련 데이터 셋의 규칙/패턴이 일반화(Generalization) 된 모델
- 훈련 데이터셋과 새로운 데이터 대한 오차 및 정확도 비슷
- 새로운 데이터에 대한 정확도 높음

ML 모델과 데이터와 학습

◆ 최적적합(Optimal fitting)

- ❖ 훈련 데이터 : 오차 약간 발생
- ❖ 새로운 데이터 : 오차 약간 발생 확률 높음!



오차 작게 발생

편향 ▼ 분산 ▼

ML 모델과 데이터와 학습

◆ 최적적합(Optimalfitting)

❖ 조건

- 편중되지 않은 **다양성 갖춘 데이터**로 학습 진행
- **양질의 많은 데이터**
- **규제(Regularization)** 통한 **모델 복잡도 적정수준** 설정

ML 모델과 데이터와 학습

◆ 모델 복잡도(Model Complexity)

- ❖ 내부 구조가 이해하기 어려운 모델
- ❖ 모델마다 복잡성 기준 및 결정 다름
- ❖ 복잡도 고려 사항
 - 특성 파라미터 개수
 - 하이퍼파라미터 개수
 - 입력 데이터셋 개수
 - 학습 반복 횟수
 - 양상별 모델 개수
 - 모델 차수

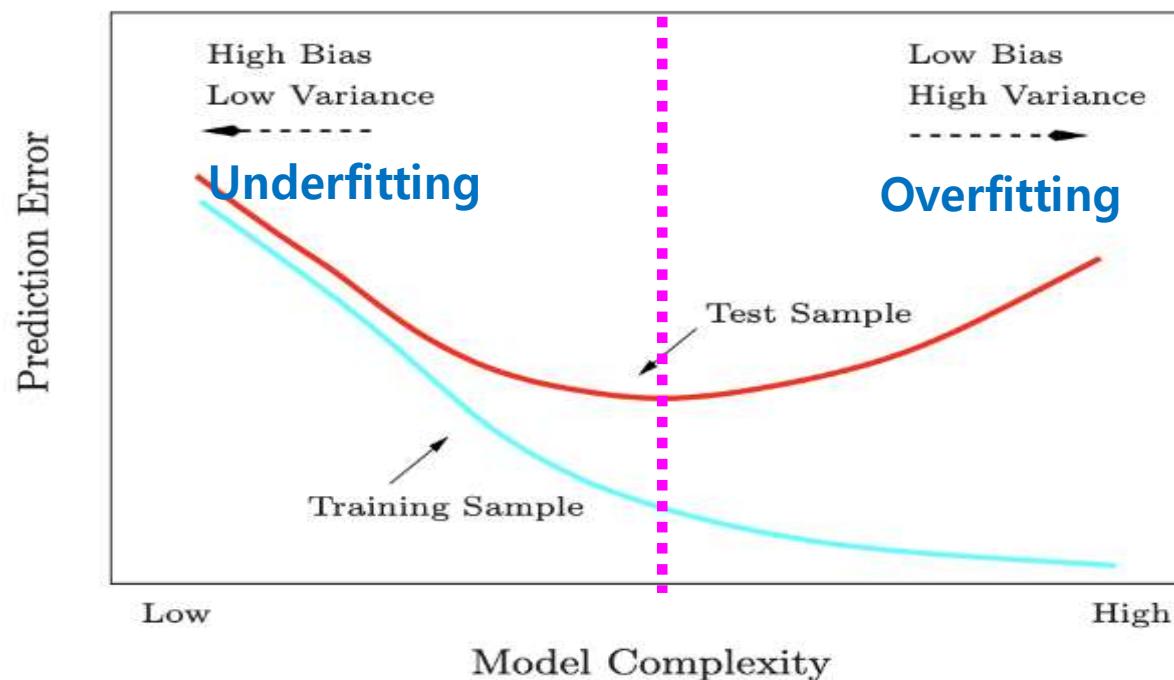


관점에 따라 다양하게 정의

ML 모델과 데이터와 학습

◆ 모델 복잡도(Model Complexity)

❖ 과대/과소



ML 모델과 데이터와 학습

◆ 모델 복잡도(Model Complexity)

Occam's Razor (오캄의 면도날)

같은 현상을 설명하는 두 개의 주장이 있다면, 간단한 쪽을 선택하라.

Given two equally accurate theories, choose the one that is less complex



모델 자체가 복잡

→ 덜 복잡한 모델 사용

데이터 부족

→ 다양성 보장 데이터 제공

변수가 너무 많아서 복잡

→ 변수 개수 줄임

학습이 너무 과해서 복잡

→ 적당 선에서 중단하

학습 데이터 불량

→ 데이터 깨끗이 정제

ML 모델과 데이터와 학습

◆ 모델 복잡도(Model Complexity)

❖ 복잡도 제어 방법

- 입력 데이터 셋 늘리기 ▲ ↪ 시간, 비용, 어려움 | 교차검증
- 학습 반복 횟수 조절 ↪ Early Stopping
- 모델 차수 ▼
- 특성 파라미터 개수 줄이기 ▼
- 특성 파라미터마다 패널티(규제) 부여 즉 정칙화(Regularization)
- 양상블 모델 개수

ML 모델과 데이터와 학습

◆ 모델 복잡도(Model Complexity)

❖ 가중치 감소(Weight Decay)

- 모델 학습 과정 **과도한 가중치 부여 제어**하는 방법
- 가중치가 클 수록 더 큰 패널티 부여
 - ➔ L(Loss Function) 1 : cost function에 **가중치 절대값** 더함
 - ➔ L(Loss Function) 2 : cost function에 **가중치 제곱값** 더함
 - ➔ ElasticNet : L1+L2 합친것

ML 모델과 데이터와 학습

◆ 데이터와 학습

- 목적에 맞는 **다양성이 보장된 데이터** 많아야 함
- 데이터 양이 많을 수록 일반화 좋음 **단, 다양한 데이터!!**
- **편증된 데이터는 오히려 역효과**

- 학습 데이터
- **검증 데이터** → 학습 진행 시 사용되어 학습 중단 시점 체크
- 테스트 데이터