강원지역혁신플랫폼

기계학습

Machine Learning

기계학습의 주요 도전과제

학습목표

- 기계학습의 주요 도전과제를 설명할 수 있습니다.

01 | 기계학습의 주요 도전과제

기계학습의 주요 도전과제

- 우리의 주요 작업은 학습 알고리즘을 선택해서 어떤 데이터에 훈련 시키는 것임
  - 여기서 문제가 될 수 있는 두 가지는 나쁜 데이터와 나쁜 알고리즘임

1 나쁜 데이터

- 충분하지 않은 양의 훈련 데이터
- 대표성 없는 훈련 데이터
- 낮은 품질의 데이터
- 관련 없는 특성

2 나쁜 알고리즘

- 훈련 데이터에 과대적합
- 훈련 데이터에 과소적합

02 | 나쁜 데이터: 충분하지 않은 양의 훈련 데이터

충분하지 않은 양의 훈련 데이터

- 어린아이에게 사과에 대해 알려주려고 "빨간 사과"를 가르침
  - 그러면 아이는 색상과 모양이 달라도 모든 종류의 사과를 구분할 수 있음
    - 즉, 초록 사과도 사고로 인식할 수 있음
      - 이러한 인간의 능력을 분별 능력이라고 부름

이것은 인간에게 아주 쉬운 문제

02 | 나쁜 데이터: 충분하지 않은 양의 훈련 데이터

충분하지 않은 양의 훈련 데이터

- 분별 능력을 기계에게 어떻게 가르칠 수 있을지 생각해 보자.
  - 대부분의 기계학습 알고리즘이 잘 작동하려면 데이터가 많아야 가능함
    - 아주 간단한 문제에서 조차도 수천 개의 데이터가 필요함
      - 이미지나 음성 인식 같은 복잡한 문제라면 수백만 개가 필요할지도 모름
      - 예를 들어 아래 그림과 같이 '더하기' 문제를 풀이하기 위해 학습을 하는 경우를 생각해 보자.

얼마나 많은 데이터가 필요할까요?

입력						
5	7	+	5			
6	2	8	+	5	2	1
2	2	0	+	8		

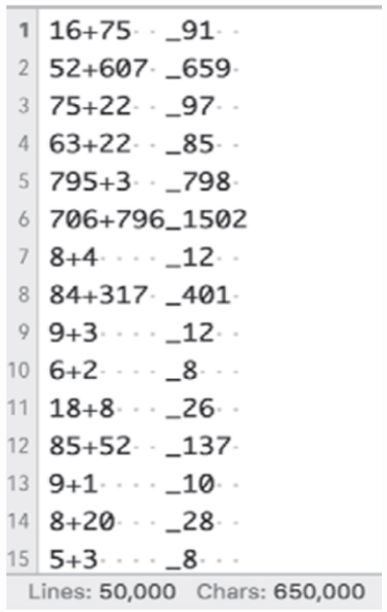
출력				
_	6	2		
_	1	1	4	9
_	2	2	8	

더하기 문제 풀기

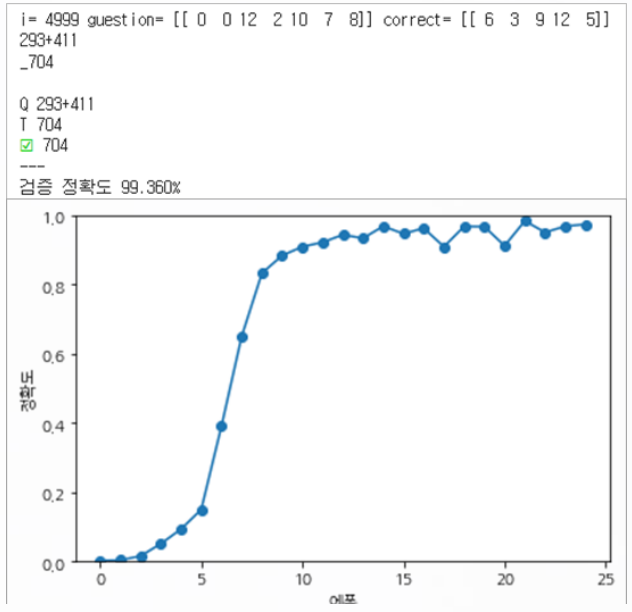
02 | 나쁜 데이터: 충분하지 않은 양의 훈련 데이터

충분하지 않은 양의 훈련 데이터

- '더하기' 문제를 풀이하기 위해 학습에 활용되는 데이터셋은 아래와 같음
  - 훈련용 데이터 집합: 50,000 개
    - 기계학습의 모델 성능 평가 결과: 정확도 약 99.36%



'덧셈' 학습용 데이터 집합



PeekSeq2seq

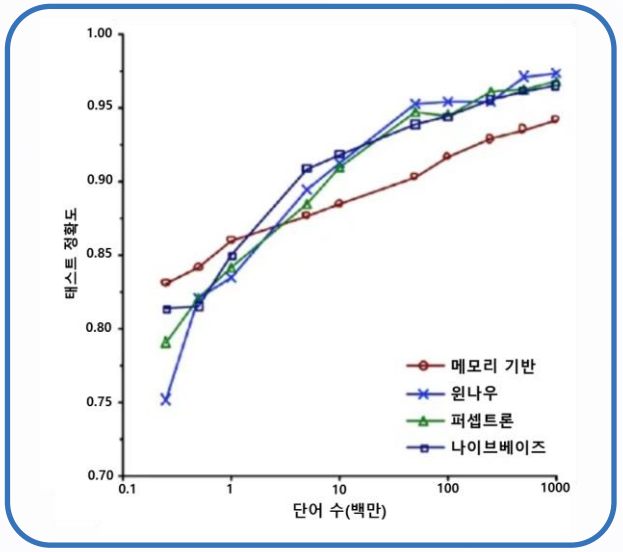
출처: https://github.com/keras-team/keras/blob/2.0.0/examples/addition_rnn.py

02 | 나쁜 데이터: 충분하지 않은 양의 훈련 데이터

충분하지 않은 양의 훈련 데이터

믿을 수 없는 데이터의 효과

- 2001년에 발표한 유명한 논문에서 MS 연구자인 미셸 반코(Michele Banko)와 에릭 브릴(Eric Brill)은 아주 간단한 모델을 포함하여 여러 다른 머신러닝 실험을 진행함
  - 이 실험에서 충분한 데이터가 주어지면 복잡한 자연어 중의성 해소 문제를 거의 비슷하게 잘 처리하는 것을 보여주었음
    - 중의성 해소 문제: 문맥에 따라 'to', 'two', 'too' 중 어떤 것을 써야 할지 아는 것임



출처: https://homl.info/6

02 | 나쁜 데이터: 대표성 없는 훈련 데이터

대표성 없는 훈련 데이터

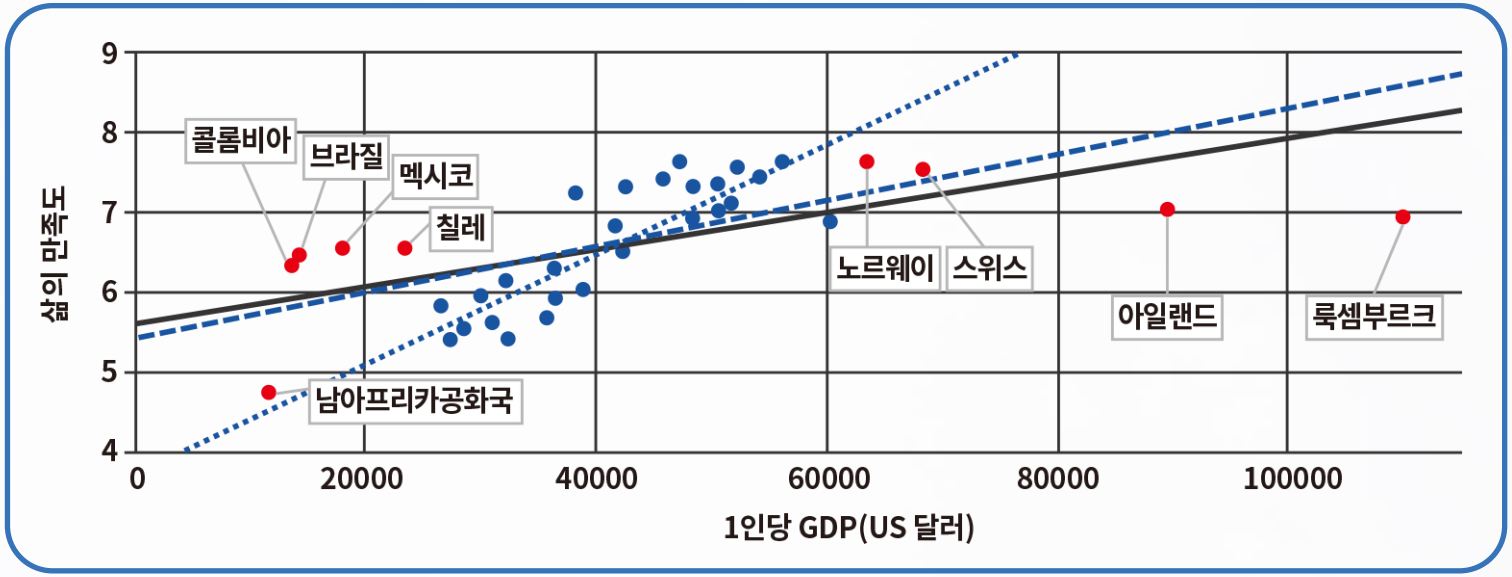
- 일반화가 잘되기 위해서는 우리가 일반화하기 원하는 새로운 사례를 훈련 데이터가 잘 대표하는 것이 중요함
  - 이는 사례 기반 학습이나 모델 기반 학습 모두 동일함
  - 예를 들어 돈이 사람을 행복하게 만드는지 알아본다고 가정해보자.
    - OECD 웹사이트(https://homl.info/4)에서 더 나은 삶의 지표 (Better Life Index)와 IMF 웹사이트(https://homl.info/5)에서 1인당 GDP 통계 데이터를 내려 받음
      - 두 데이터를 합치면 1인당 GDP로 정렬하면 다음과 같음

국가	1인당 GDP (US달러)	삶의 만족도
헝가리	12,240	4.9
대한민국	27,195	5.8
프랑스	37,675	6.5
호주	50,962	7.3
미국	55,805	7.2

02 | 나쁜 데이터: 대표성 없는 훈련 데이터

대표성 없는 훈련 데이터

- 아래의 그래프는 앞의 1인당 GDP와 삶의 만족도 데이터로 선형 모델을 훈련시킨 결과임
  - 아래의 그래프에서 점선은 일부 나라가 빠져 있어 대표성이 완벽하지 못함
    - 실선은 누락된 나라를 추가해서 얻은 학습된 모델
      - 매우 부유한 나라가 중간 정도의 나라보다 행복하지 않고,
        반대로 일부 가난한 나라가 부유한 나라보다 행복한 것 같음



출처: https://m.hanbit.co.kr/channel/category/category_view.html?cms_code=CMS3238544209

02 | 나쁜 데이터: 대표성 없는 훈련 데이터

대표성 없는 훈련 데이터

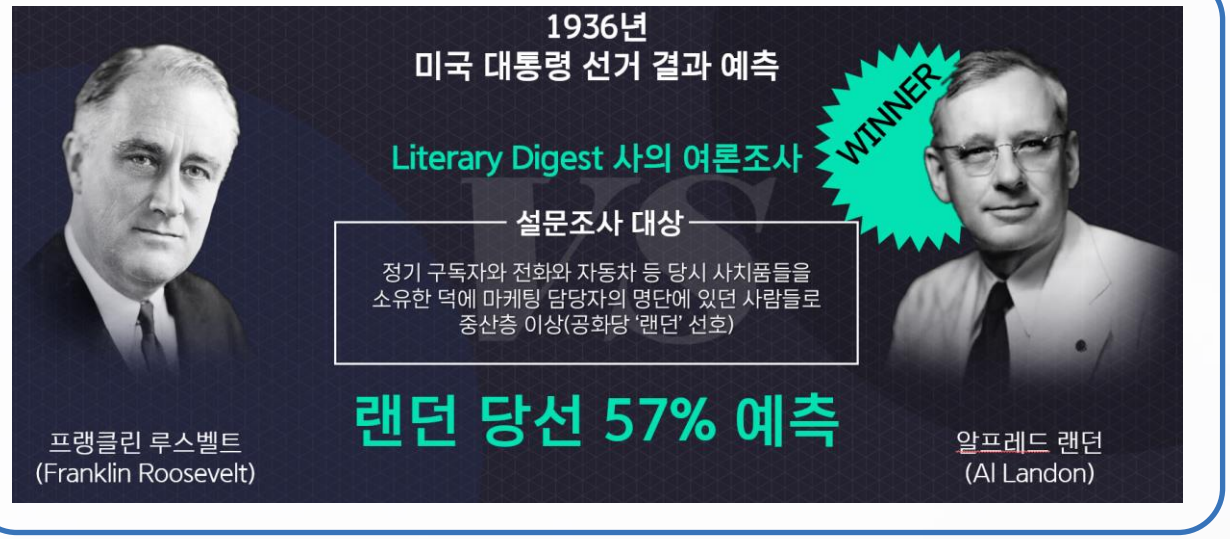
- 일반화하려는 사례들을 대표하는 훈련 세트를 사용하는 것이 매우 중요함
  - 하지만, 이것이 생각보다 어려울 때가 많음
    - 샘플이 작으면 샘플링 편향(sampling noise)이 생김
    - 매우 큰 샘플도 표본 추출 방법이 잘못되면 대표성을 띠지 못할 수 있음
      - 이것을 샘플링 편향(sampling bias)이라고 부름

02 | 나쁜 데이터: 대표성 없는 훈련 데이터

대표성 없는 훈련 데이터

[유명한 샘플링 편향 사례]

- 1936년 랜던과 루즈벨트의 대통령 선거에서 『The Literary Digest』 잡지사가 천만 명에게 우편물을 보내 수행한 여론조사
  - ✦ 여론 조사에서 240만 명의 응답을 받았고 랜던이 선거에서 57% 득표를 예측함
  - ✦ 실제 투표에서는 루즈벨트가 60.8% 득표로 당선됨



출처: https://www.britannica.com/biography/Franklin-D-Roosevelt | https://www.britannica.com/biography/Alf-Landon

02 | 나쁜 데이터: 대표성 없는 훈련 데이터

대표성 없는 훈련 데이터

[유명한 샘플링 편향 사례]

- 이 여론조사에서 **문제**는 **샘플링 방법**에 있음
  - **첫째**, 여론 조사를 얻기 위해 전화번호부, 자사의 구독자 명부, 클럽 회원 명부 등을 사용했음
    - 이런 명부는 모두 **공화당**에 **투표할 가능성**이 **높은 부유한 계층**에 **편중**된 경향
  - **둘째**, 우편물 수신자 중 **25% 미만의 사람**이 **응답**했음
    - 이는 정치에 관심 없는 사람, 잡지사를 싫어하는 사람, 대표성이 있는 중요한 그룹을 제외시킴
      - 특히, 이러한 종류의 샘플링 편향을 **비응답 편향**(nonresponse bias)라고 부름

02 | 나쁜 데이터: 낮은 품질의 데이터

낮은 품질의 데이터

- 훈련 데이터가 에러, 이상치, 잡음으로 가득하다면 기계학습 시스템이 내재되어 있는 패턴을 찾기 어려울 수 있음
  - 이러한 이유로 훈련 데이터 정제에 시간을 투자할 만한 가치는 충분함
    - 실제 대부분의 데이터 과학자가 데이터 정제에 많은 시간을 쓰고 있음

Garbage in, Garbage out (GIGO)

쓰레기가 들어가면 쓰레기가 나온다.

02 | 나쁜 데이터: 낮은 품질의 데이터

낮은 품질의 데이터

- 훈련 데이터에 데이터 정제가 필요한 경우는 다음과 같음
  - 일부 샘플의 특성 값이 이상치가 명확한 경우 무시하거나 수동으로 잘못된 것을 고치는 것이 좋음
  - 일부 샘플에 특성 값 몇 개가 빠져있는 경우 다음을 결정해야 할 것임
    - 이 특성을 모두 무시할지
    - 이 샘플을 무시할지
    - 빠진 값을 채울지
    - 이 특성을 넣은 모델과 제외한 모델을 따로 훈련시킬 것인지

02 | 나쁜 데이터: 관련 없는 특성

관련 없는 특성

- 훈련 데이터에 관련 없는 특성이 적고 관련 있는 특성이 충분해야 기계학습 시스템이 학습을 잘 진행할 수 있음
  - 성공적인 기계학습 프로젝트의 핵심 요소는 훈련에 사용할 좋은 특성들을 찾는 것임
    - 이 과정을 특성 공학(feature engineering)이라 부름

02 | 나쁜 데이터: 관련 없는 특성

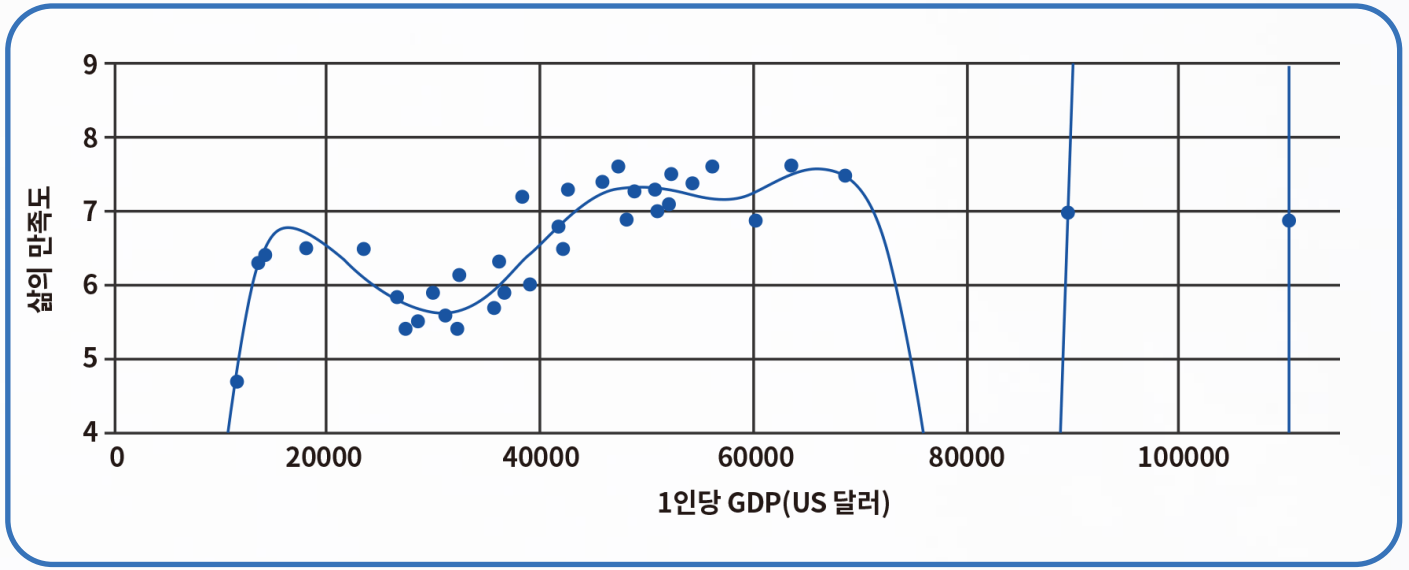
관련 없는 특성

- 특성 공학은 다음과 같은 작업을 의미함
  - 특성 선택(feature selection)
    - 가지고 있는 특성 중에서 훈련에 가장 유용한 특성을 선택함
  - 특성 추출(feature extraction)
    - 특성을 결합하여 더 유용한 특성을 만듦
    - 차원 축소 알고리즘이 도움이 될 수 있음
  - 새로운 데이터를 수집해 새 특성을 만듦

03 | 나쁜 알고리즘: 훈련 데이터 과대 적합

훈련 데이터 과대 적합

- 과대 적합(overfitting)은 학습 데이터를 과하게 잘 학습한 것을 의미함
  - 아래의 그래프는 고차원의 다항 회귀 모델이 1인당 GDP와 삶의 만족도 훈련 데이터에 크게 과대 적합된 사례를 보여줌
    - 이 모델이 훈련 데이터에 잘 적응되었지만, 새로운 데이터에도 예측이 잘 될지는 모름

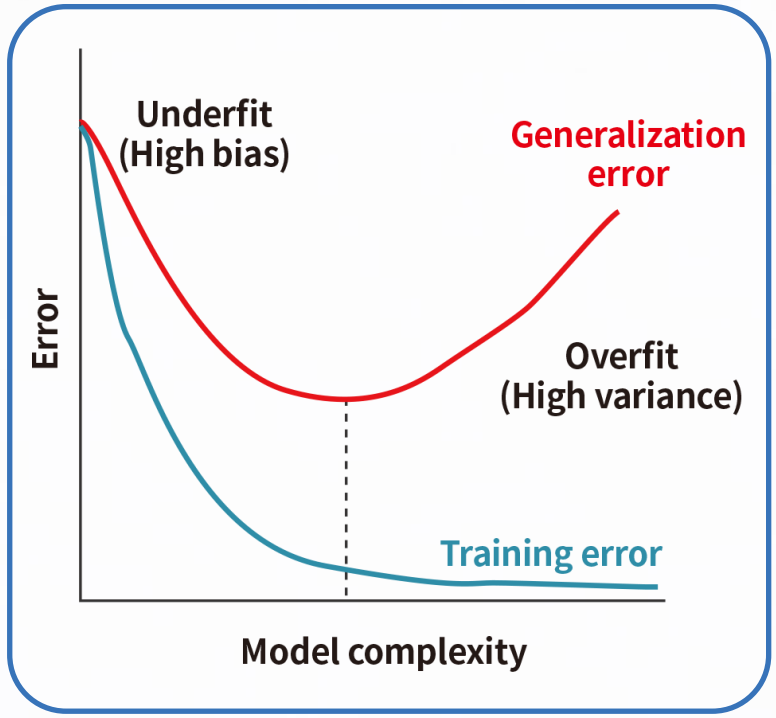


출처: https://m.hanbit.co.kr/channel/category/category_view.html?cms_code=CMS3238544209

03 | 나쁜 알고리즘: 훈련 데이터 과대 적합

훈련 데이터 과대 적합

- 아래 그래프와 같이 학습 데이터에 대해서는 오차가 감소하지만, 실제 데이터에 대해서는 오차가 증가하는 지점이 존재함
  - 아래 그림에서는 테스트 에러가 감소하다 갑자기 치솟는 부분에서 과대 적합이 발생했다고 볼 수 있음



03 | 나쁜 알고리즘: 훈련 데이터 과대 적합

훈련 데이터 과대 적합

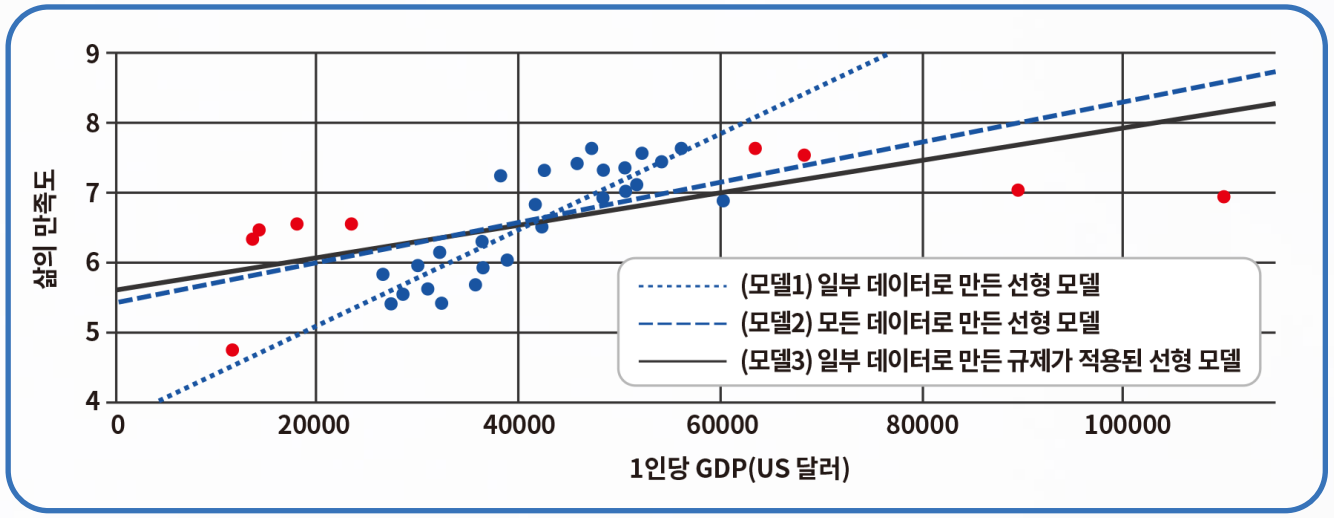
과대 적합의 해결 방법은 다음과 같음

- 모델 파라미터 수가 적은 모델을 선택함
- 특성 수를 줄이거나, 모델에 제약(규제)을 가하여 단순화시킴
  (규제는 모델 파라미터의 값을 작게하거나 0으로 만듦)
- 훈련 데이터를 더 많이 확보함
- 훈련 데이터의 잡음을 줄임(Outlier, Error 제거)

03 | 나쁜 알고리즘: 훈련 데이터 과대 적합

훈련 데이터 과대 적합

- 아래 그림은 1인당 GDP와 삶의 만족도 데이터로 만든 세 가지 선형 모델임
  - 점선(··)은 (사각형 제외) 동그라미로 표시된 나라로 훈련된 원래 선형 모델(모델1)
  - 실선(-)은 (동그라미와 사각형) 모든 데이터로 만든 선형 모델(모델2)
  - 대시선(--)은 일부 데이터로 만든 규제가 적용된 선형 모델임(모델3)
    - 모델3의 경우 훈련 데이터(동그라미)에는 덜 맞지만 새로운 샘플(사각형)에는 더 잘 일반화됨



출처: https://m.hanbit.co.kr/channel/category/category_view.html?cms_code=CMS3238544209

03 | 나쁜 알고리즘: 훈련 데이터 과소 적합

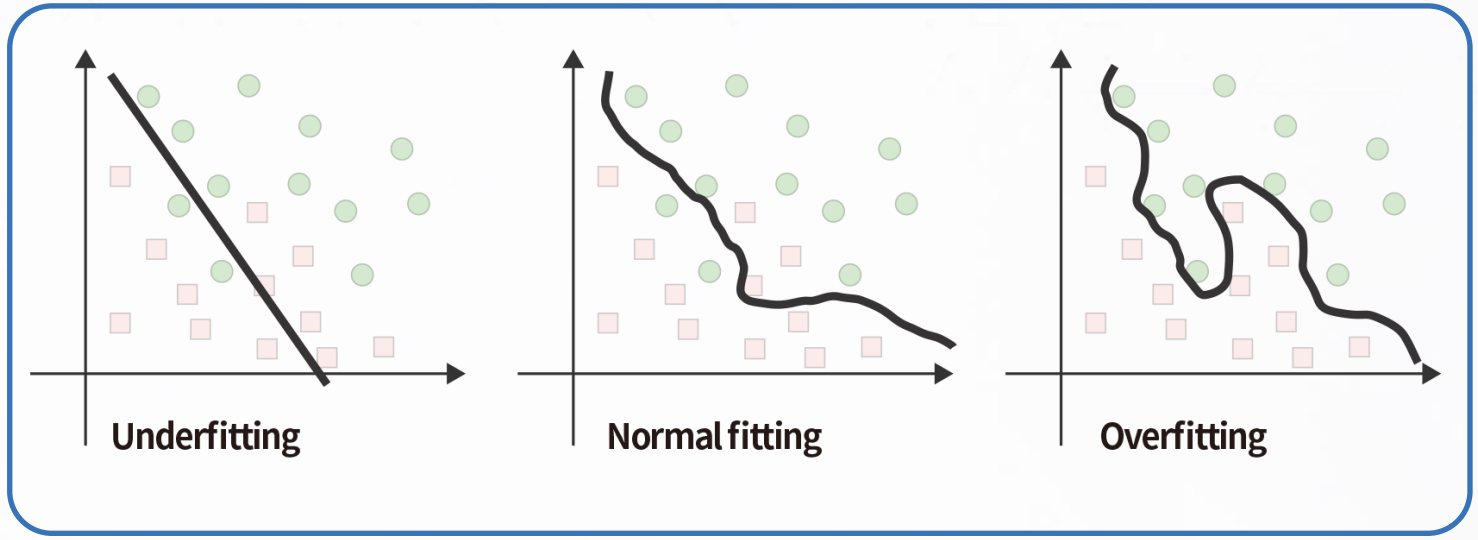
훈련 데이터 과소 적합

- 과소 적합(underfitting)은 과대 적합의 반대임
  - ✦ 모델이 너무 단순해서 데이터의 내재된 구조를 학습하지 못할 때 발생함
  - ✦ 과소 적합 문제를 해결하는 주요 기법은 다음과 같음
    - › 모델 파라미터가 더 많은 강력한 모델을 선택함
    - › 학습 알고리즘에 더 좋은 특성을 제공함(특성 공학)
    - › 모델의 제약을 줄임
      - 예를 들면 규제 하이퍼 파라미터를 감소시킴

03 | 나쁜 알고리즘: 훈련 데이터 과소 적합

훈련 데이터 과소 적합

- 훈련 데이터를 올바르게 학습시키기 위해서는 과대 적합과 과소 적합의 중간점을 찾는 것이 바람직함
  - ✦ 아래 그림에서 보이는 바와 같이 너무 잘 분류해도, 분류하지 못해도 올바른 모델이라고 할 수 없음



출처:https://tkleen.cafe24.com/1-머신러닝-학습-알고리즘의-구현/