머신러닝 이론 및 실습

1주차. 한눈에 보는 머신러닝

유선호

0 CONTEXT

1. 머신러닝이란?
2. 왜 머신러닝을 사용하는가?
3. 애플리케이션 사례
4. 머신러닝 시스템의 종류
5. 머신러닝의 주요 도전 과제
6. 테스트와 검증

1 머신러닝이란?

머신러닝은 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야다. – 아서 새뮤얼, 1959

어떤 작업 T에 대한 컴퓨터 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상됐다면,

이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로 학습한 것이다. – 톰 미첼, 1997

1 머신러닝이란?

머신러닝 프로그램 예제

스팸 필터 : 스팸 메일과 일반 메일의 샘플을 이용해 스팸 메일 구분법을 배울 수 있는 머신러닝 프로그램

작업 T : 새로운 메일이 스팸인지를 구분하는 것

경험 E : 훈련 데이터

성능 측정 P : 사용자 정의 (ex. 정확히 분류된 메일의 비율(정확도, accuracy))

*훈련세트(training set) : 시스템이 학습하는 데 사용하는 샘플

*훈련샘플(training instance) : 하나의 훈련 데이터

2 왜 머신러닝을 사용하는가?

전통적인 프로그래밍 기법으로 스팸 필터 만들기

1. 스팸에 어떤 단어들이 주로 나타나는지 살펴본다. 그 결과, '4U', '신용카드', '무료', '굉장한' 같은 단어나 구절이 제목에 많이 나타나는 경향을 알아냈다. 보낸이의 이름이나 메일 주소, 본문이나 이메일의 다른 요소에서도 다른 패턴을 감지한다.

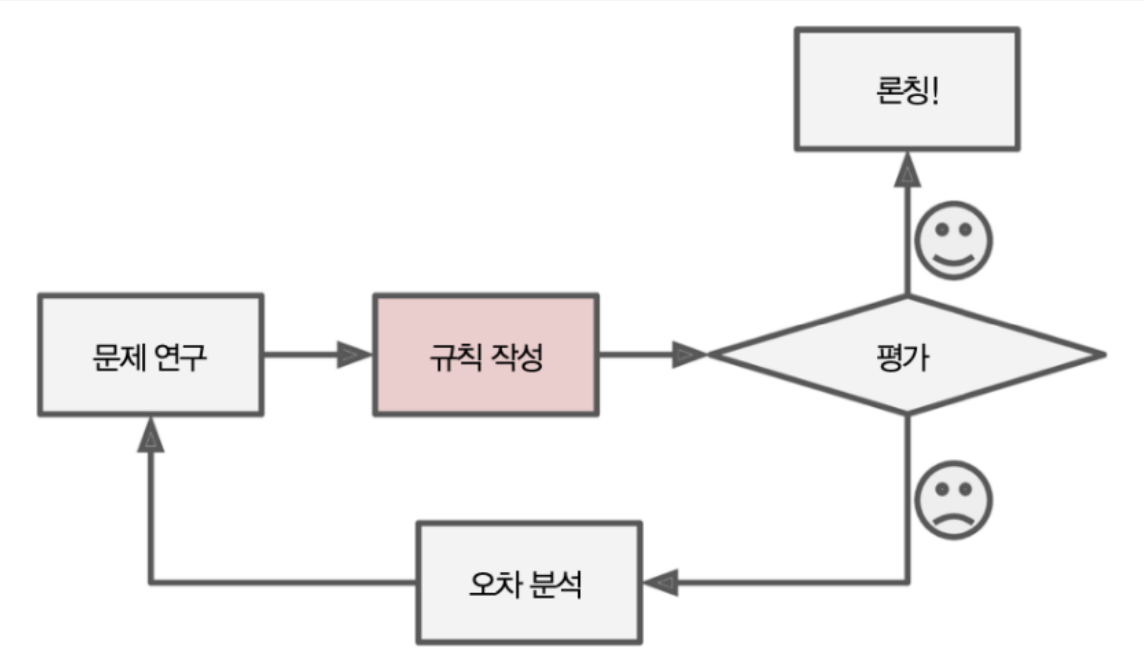
2. 발견한 각 패턴을 감지하는 알고리즘을 작성하여 프로그램이 이런 패턴을 발견했을 때 그 메일을 스팸으로 분류하게 한다.

3. 프로그램을 테스트하고 론칭할 만큼 충분한 성능이 나올 때까지 1단계와 2단계를 반복한다.

2 왜 머신러닝을 사용하는가?

전통적인 접근 방법 vs 머신러닝 접근 방법

전통적인 프로그래밍



문제 연구 : 누군가가 문제를 해결하기 위해 해결책을 찾음

규칙 작성 : 결정된 규칙을 개발자가 프로그램을 작성

평가 : 만들어진 프로그램을 테스트

문제가 없다면 **론칭**, 문제가 있다면 **오차를 분석**한 후 처음 과정부터 다시 실시

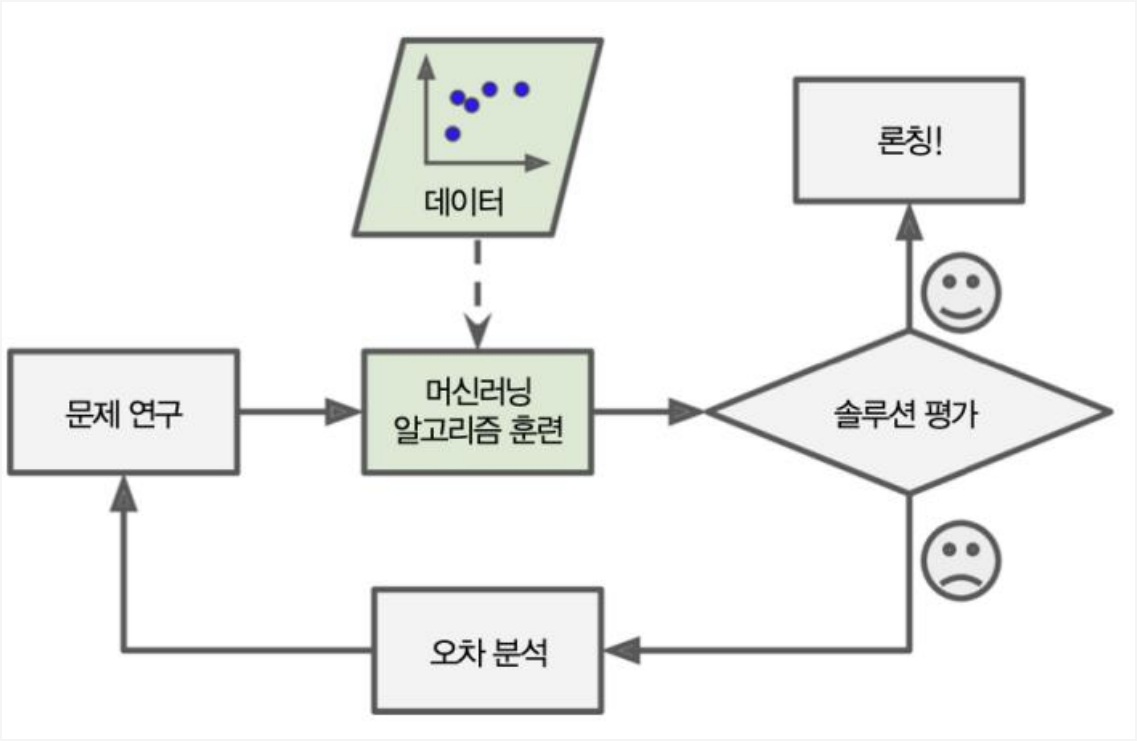
새로운 규칙이 생겼을 때 사용자가 매번 업데이트를 시켜야 하기 때문에 유지 보수가 어려움

WeTIE

2 왜 머신러닝을 사용하는가?

전통적인 접근 방법 vs 머신러닝 접근 방법

머신러닝



문제 연구 : 알고리즘이 데이터로부터 스스로 스팸의 특징을 찾음

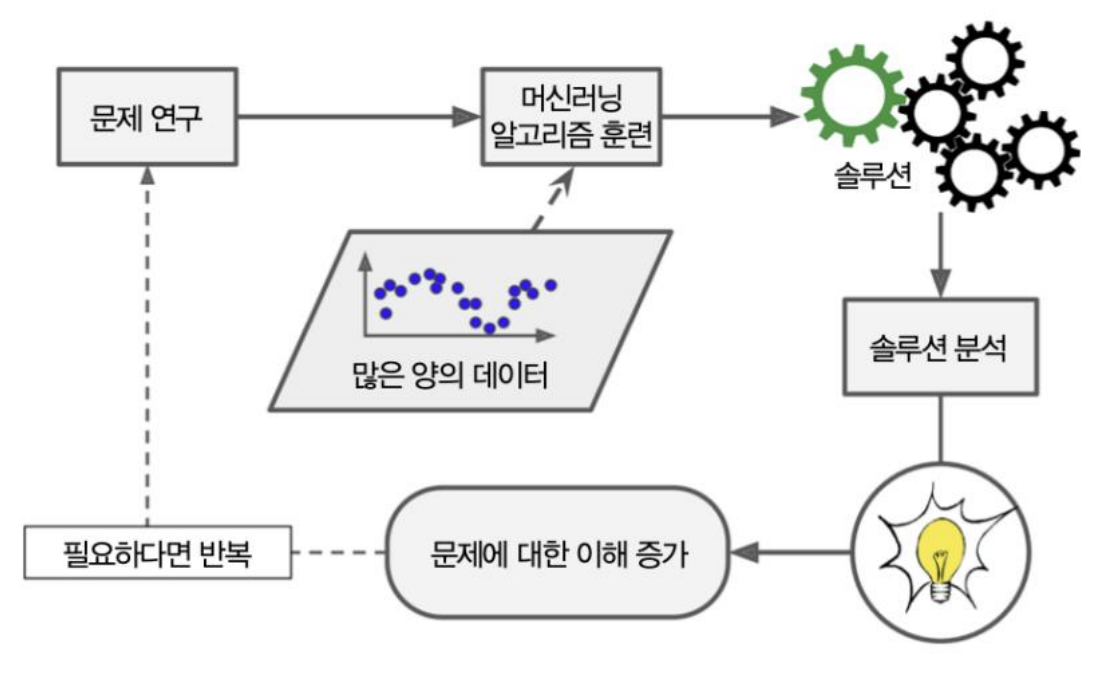
머신러닝 알고리즘 훈련 : 주어진 데이터를 바탕으로 훈련

솔루션 평가 : 문제가 없다면 론칭, 문제가 있다면 오차를 분석한 후 처음 과정부터 다시 실시

갑자기 A라는 단어가 자주 나타는 것을 **자동으로 인식하고 별도의 작업을 하지 않아도** 자동으로 이 단어를 스팸으로 분류할 수 있다.

2 왜 머신러닝을 사용하는가?

머신러닝을 통해 배울 수 있다!



복잡한 문제와 대량의 데이터에서 보이지 않던 패턴, 통찰 얻기 (데이터 마이닝, Data Mining)

ex. 스팸 필터가 충분한 스팸 메일로 훈련되었다면, 스팸을 예측하는데 가장 좋은 단어 및 단어의 조합이 무엇인지 확인할 수 있다

가끔 예상치 못한 연관 관계나 새로운 추세가 발견되어 해당 문제를 더 잘 이해하도록 도와준다.

WeTIE

2 왜 머신러닝을 사용하는가?

머신러닝은 다음 분야에 뛰어나다!

기존 솔루션으로 많은 수동 조정과 규칙이 필요한 문제 :
(하나의 머신러닝 모델이 코드를 간단하게 만들고 전통적인 방법보다 더 잘 수행되도록 할 수 있다.)

전통적인 방식으로 해결 방법이 없는 복잡한 문제

유동적인 환경 : 머신러닝 시스템은 새로운 데이터에 적응할 수 있다.

복잡한 문제와 대량의 데이터에서 통찰 얻기

3 애플리케이션 사례

1. 생산 라인에서 제품 이미지를 분석해 자동을 분류하기

2. 뇌를 스캔하여 종양 진단하기

3. 자동으로 뉴스 기사를 분류하기

4. 토론 포럼에서 부정적인 코멘트를 자동으로 구분하기

5. 챗봇 또는 개인 비서 만들기

6. 다양한 성능 지표를 기반으로 회사의 내년도 수익을 예측하기

7. 음성 명령에 반응하는 앱을 만들기

8. 신용 카드 부정 거래 감지하기

9. 구매 이력을 기반으로 고객을 나누고 각 집합마다 다른 마케팅 전략을 계획하기

10. 고차원의 복잡한 데이터셋을 명확하고 의미 있는 그래프로 표현하기

11. 과거 구매 이력을 기반으로 고객이 관심을 가질 수 있는 상품 추천하기

12. 지능형 게임 봇 만들기

4 머신러닝 시스템의 종류

1. 지도, 비지도, 준지도, 강화학습 : 사람의 감독하에 훈련하는 것인지 그렇지 않은 것인지

2. 온라인 학습과 배치 학습 : 실시간으로 점진적인 학습을 하는지 아닌지

3. 사례 기반 학습과 모델 기반 학습 : 단순하게 알고 있는 데이터 포인트와 새 데이터 포인트를 비교하는 것인지

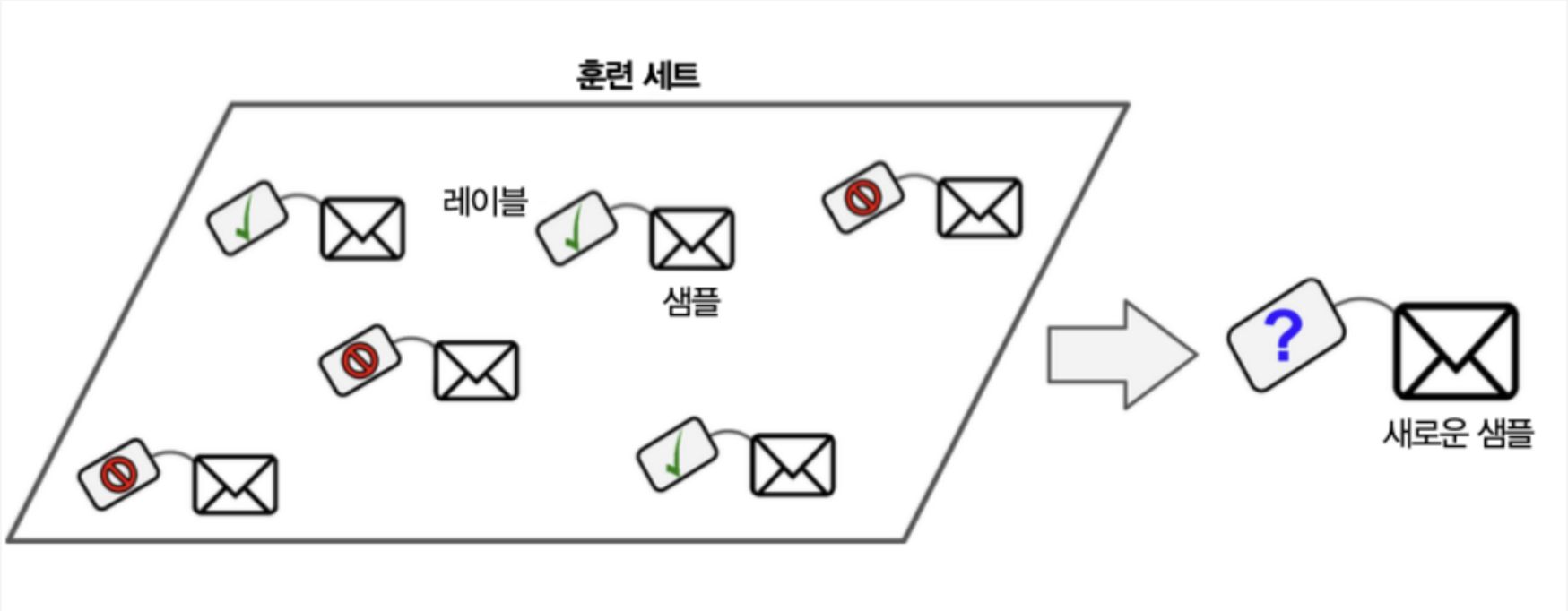
   훈련 데이터셋에서 패턴을 발견하여 예측 모델을 만드는 것인지

분류 기준은 상호 배타적이지 않다!

4 머신러닝 시스템의 종류

4.1. 지도 학습과 비지도 학습

지도학습 : 훈련 데이터에 레이블이라는 답을 포함하고 있다.



4 머신러닝 시스템의 종류

4.1. 지도 학습과 비지도 학습

대표적인 지도 학습

분류 : 특성을 사용하여 데이터를 분류하는 문제

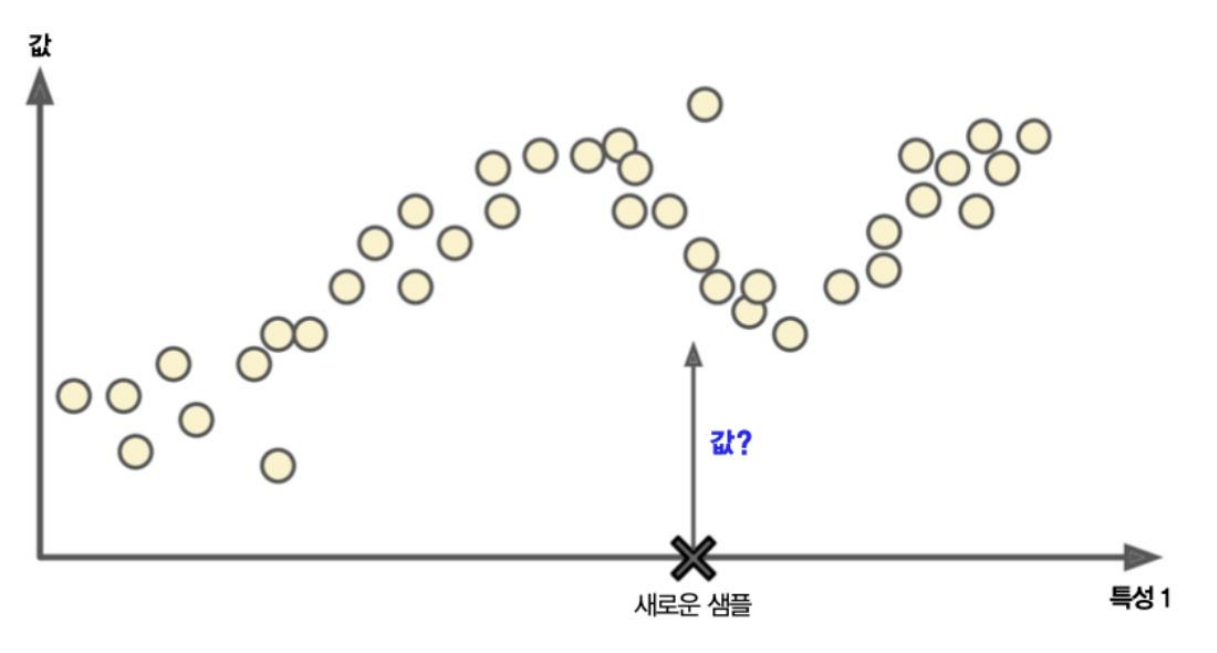
특성(feature) : 소속 정보, 특정 단어 포함 여부 등

타깃(target) : 스팸여부

회귀 : 특성을 사용해 타깃 수치를 예측하는 문제

특성(feature) : 주행거리, 연식, 브랜드 등

타깃(target) : 중고차 가격



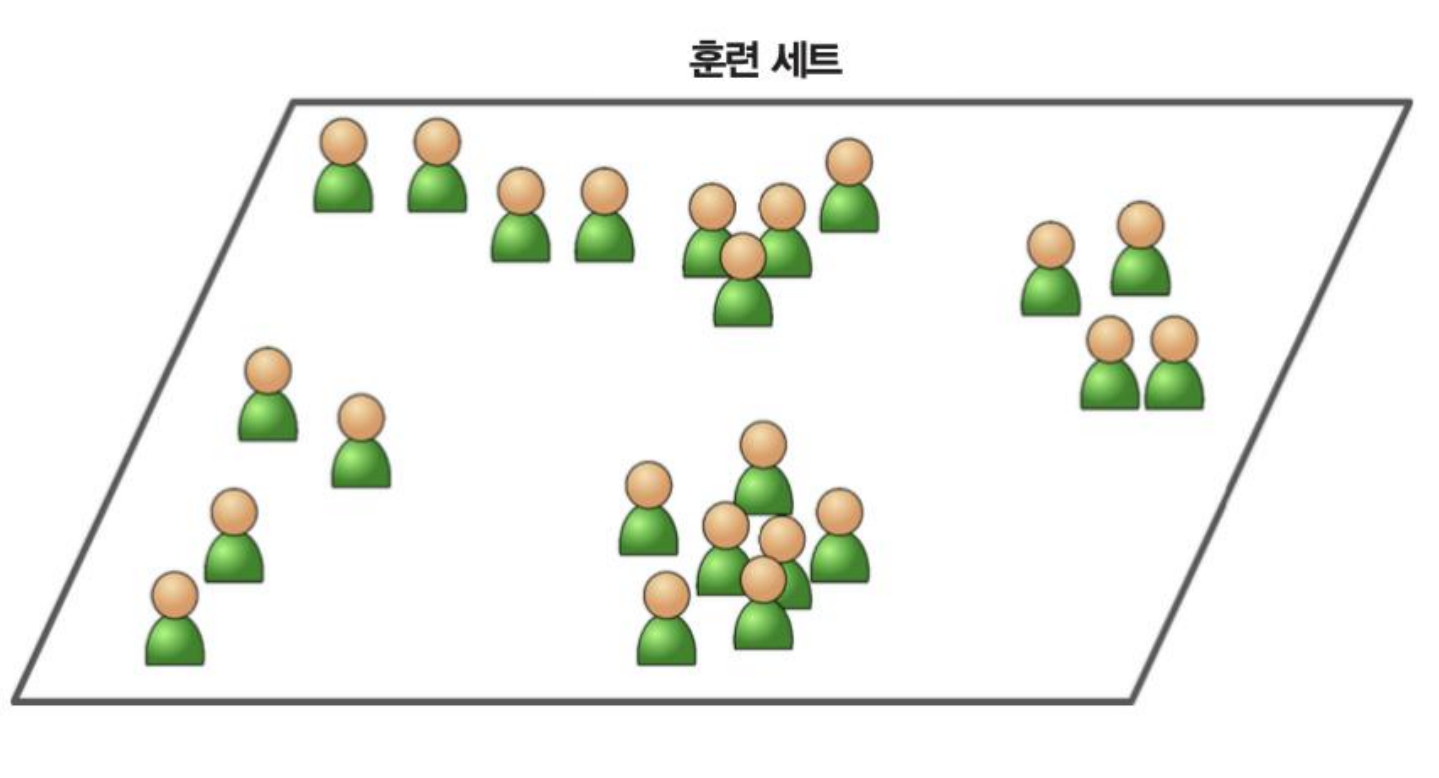
일부 회귀 알고리즘을 분류에 사용할 수 있고, 일부 분류 알고리즘을 회귀에 사용할 수 있다.

지도 학습 알고리즘 종류 : k-최근접 이웃(k-nearest neighbors), 선형 회귀(linear regression), 로지스틱 회귀(logistic regression), 서포트 벡터 머신(Support vector machine), 결정 트리(Decision Tree), 랜덤 포레스트(Random Forest), 신경망(Neural networks)

4 머신러닝 시스템의 종류

4.1. 지도 학습과 비지도 학습

비지도학습 : 레이블이 없는 훈련 데이터를 이용하여 시스템 스스로 학습한다.



WeTIE

4 머신러닝 시스템의 종류

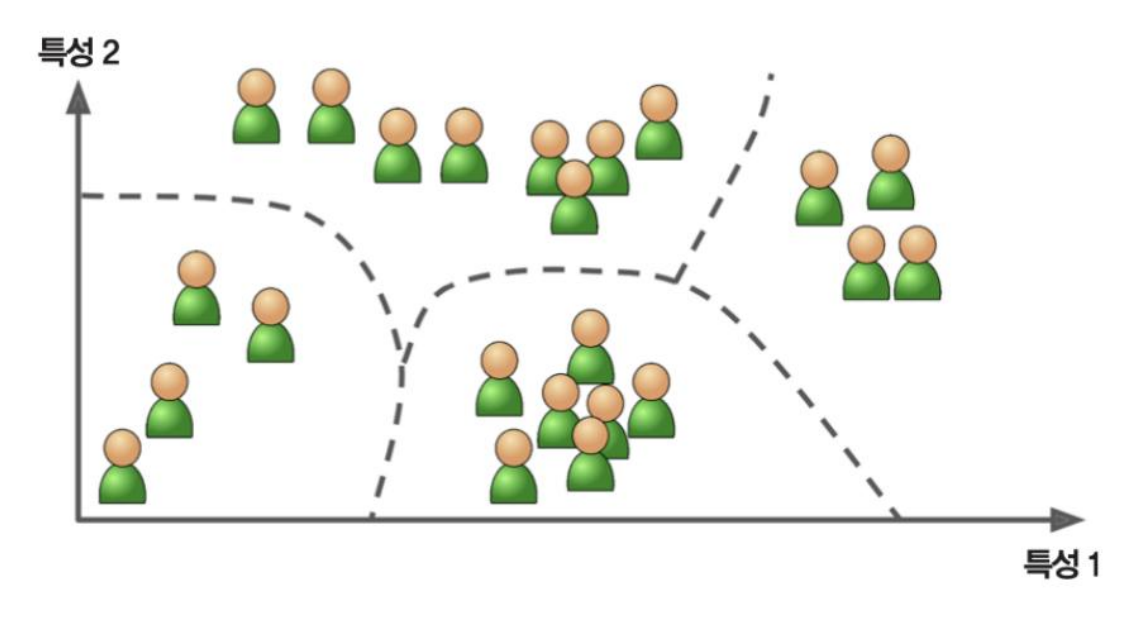
4.1. 지도 학습과 비지도 학습

대표적인 비지도 학습

군집(clustering) : 데이터를 비슷한 특징을 가진 몇 개의 그룹으로 나누는 것

ex. 블로그 방문자들을 그룹으로 묶기 : 남성, 여성, 주말, 주중, 만화책, SF 등등

종류 : k-means, DBSCAN, 계층 군집 분석(HCA)



WeTIE

4 머신러닝 시스템의 종류

4.1. 지도 학습과 비지도 학습

대표적인 비지도 학습

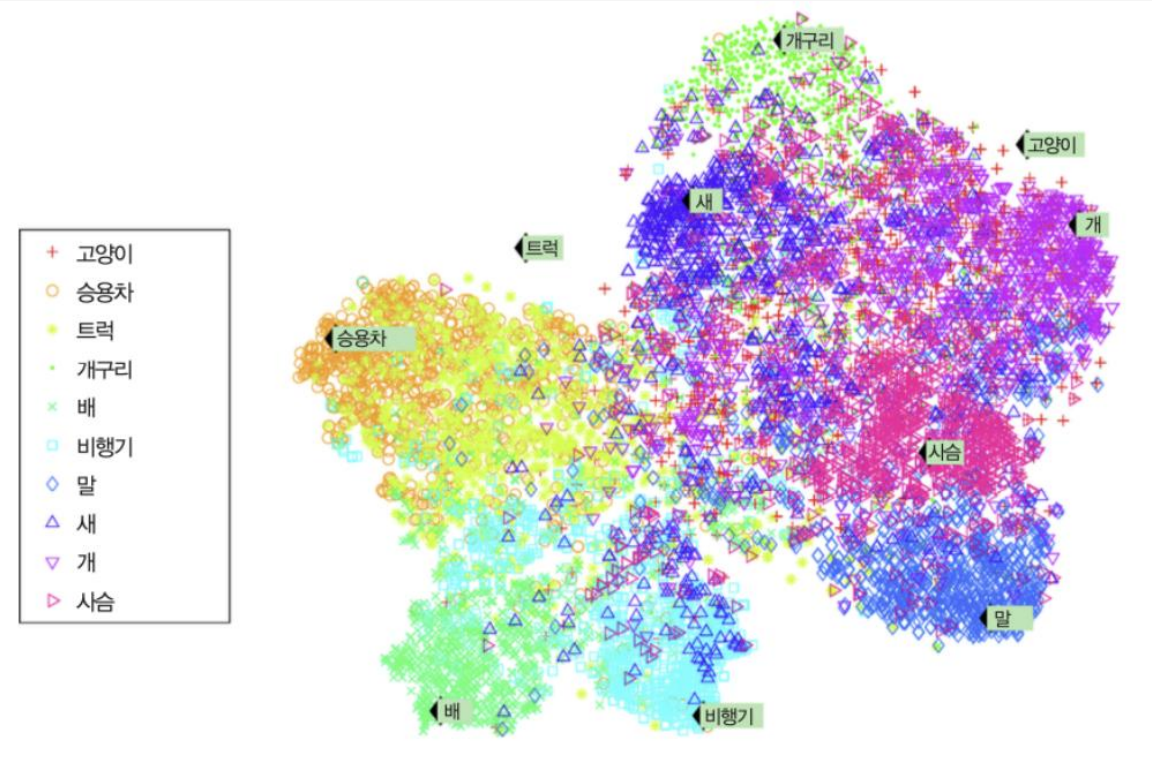
시각화(visualization)와 차원 축소(dimensionality reduction):

1. 다차원 특성을 가진 데이터셋을 2D 또는 3D로 표현하기
2. 상관관계가 있는 여러 특성을 하나로 합쳐 데이터의 특성 수 줄이기
3. 시각화를 하기 위해서는 데이터의 특성을 2가지로 줄여야 함

ex. 자동차의 주행거리와 연식은 상관관계가 높음.

-> 차의 '마모정도'라는 하나의 특성으로 합칠 수 있다.

장점 : 머신러닝 알고리즘의 성능 향상, 훈련 실행 속도 빠름, 메모리 사용 공간이 줄어듦



4 머신러닝 시스템의 종류

4.1. 지도 학습과 비지도 학습

대표적인 비지도 학습

이상치 탐지와 특이치 탐지

이상치 탐지(anomaly detection) : 정상 샘플을 이용하여 훈련 후 입력 샘플의 정상여부 판단

특이치 탐지(novelty detection) : 전혀 오염되지 않은(clean) 훈련 세트 활용, 훈련 세트에 포함된 데이터와 다른 데이터 감지

ex. 부정거래 사용 감지, 제조 결함 잡기, 이상치 자동 제거

이상치 탐지 vs. 특이치 탐지

수 천장의 강아지 사진에 치와와 사진이 1% 정도 포함되어 있는 경우

특이치 탐지 알고리즘은 새로운 치와와 사진을 특이한 것으로 간주하지 않음

이상치 탐지 알고리즘은 새로운 치와와 사진을 다른 강아지들과 다른 종으로 간주할 수 있음

종류 : 원-클래스 SVM, Isolation Forest

4 머신러닝 시스템의 종류

4.1. 지도 학습과 비지도 학습

대표적인 비지도 학습

연관 규칙 학습(association rule learning)

데이터 간의 흥미로운 관계 찾기

ex. 마트 판매 기록 : 바비큐 소스와 감자 구매와 스테이크 구매 사이의 연관성이 밝혀지면 상품을 서로 가까이 진열해야 함.

종류 어프라이어리(Apriori), 이클렛(Eclat)

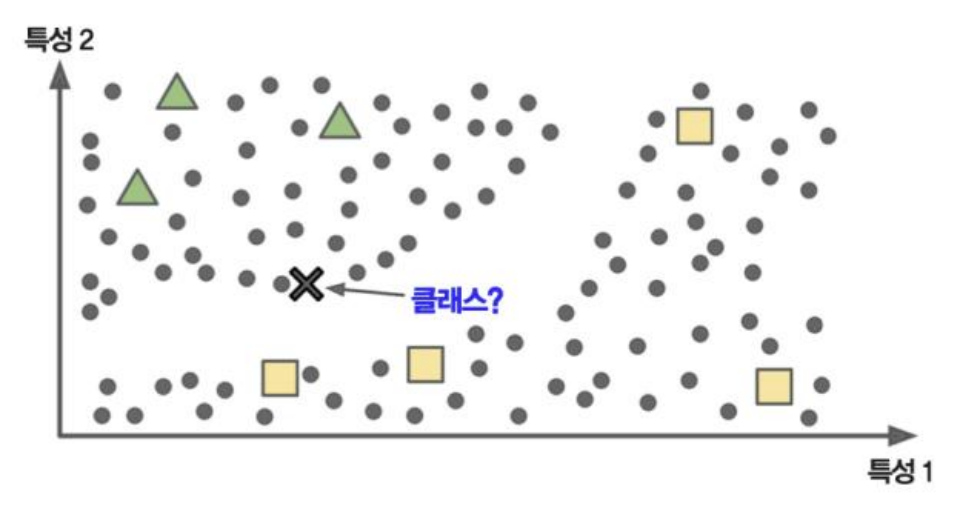
4 머신러닝 시스템의 종류

4.1. 준지도 학습

데이터에 레이블을 다는 것은 시간과 비용이 많이 들기에 일반적으로 레이블이 없는 샘플이 많다.

적은 수의 샘플에 레이블을 적용함.

비지도 학습을 통해 군집을 분류한 후, 샘플들을 활용해 지도 학습에 활용함.



새로운 사례 x를 세모에 더 가깝다고 판단함.

구글 포토 호스팅 : 가족 사진 몇 장에만 레이블 적용. 이후 모든 사진에서 가족사진 확인 가능.

종류 : 심층 신뢰 신경망(Deep belief network, DBN)

4 머신러닝 시스템의 종류

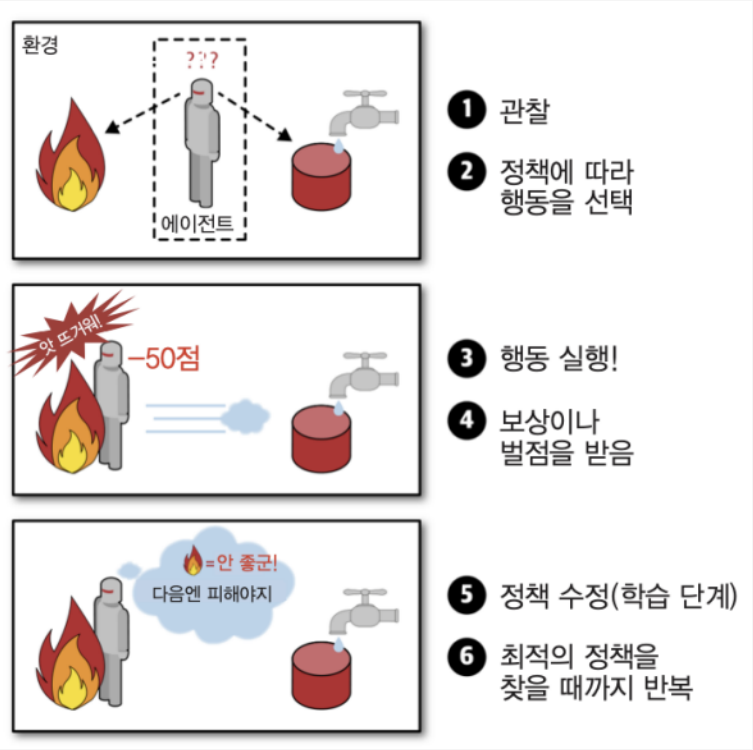
4.1. 강화 학습

에이전트가 환경을 관찰하여 행동을 실행하고 그 결과로 보상 혹은 벌점을 받는다.

시간이 지나면서 가장 큰 보상을 얻기 위해 정책이라고 부르는 최상의 전략을 학습한다.

정책 : 주어진 상황에서 에이전트가 어떤 행동을 선택해야 할지 정의한다.

ex. 보행 로봇, 딥마인드의 알파고



WeTIE

4.2. 배치 학습과 온라인 학습

배치 학습(batch learning)

주어진 훈련 세트 전체를 사용하여 오프라인에서 훈련

먼저 시스템을 훈련시킨 후 더 이상의 학습 없이 제품 시스템에 적용

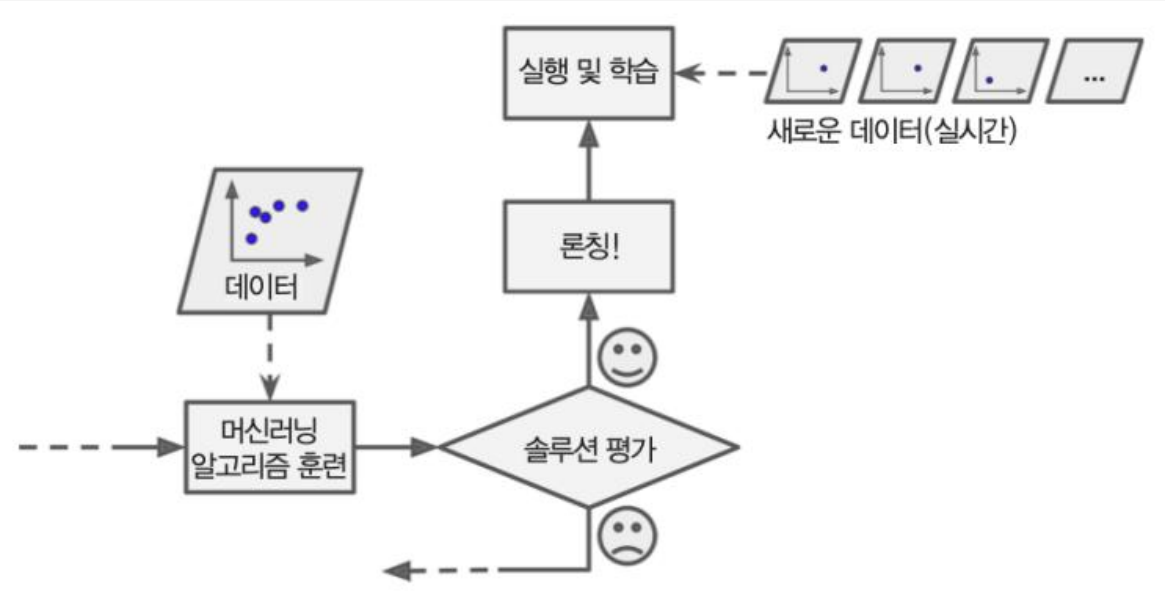
단점 : 컴퓨팅 자원이 충분한 경우에만 사용 가능, 새로운 데이터가 들어오면 처음부터 새롭게 학습해야 함

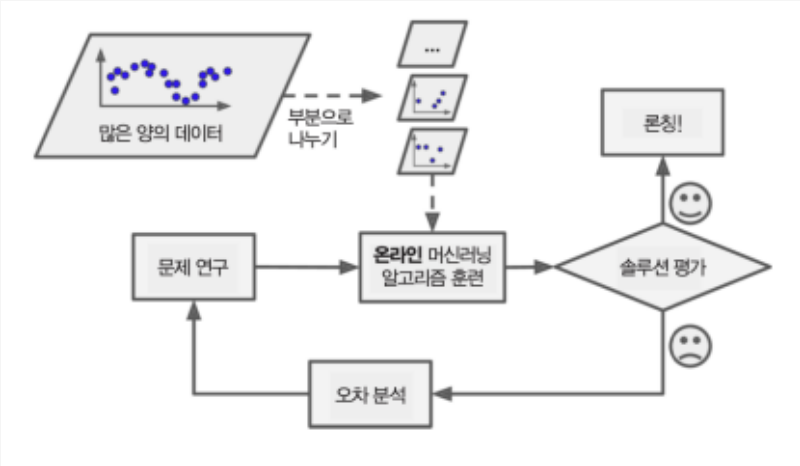
4.2. 배치 학습과 온라인 학습

온라인 학습(online learning)

적은 양의 데이터(미니배치, mini-batch)를 사용해 점진적으로 훈련

단점 : 나쁜 데이터가 주입되는 경우 시스템 성능이 점진적으로 떨어질 수 있음, 지속적인 시스템 모니터링이 필요함.





주식 가격 시스템 등 실시간 반영이 중요한 시스템

스마트폰 등 제한된 자원의 시스템

4.3. 사례 기반 학습과 모델 기반 학습

머신러닝 시스템의 일반화 방식에 따른 분류

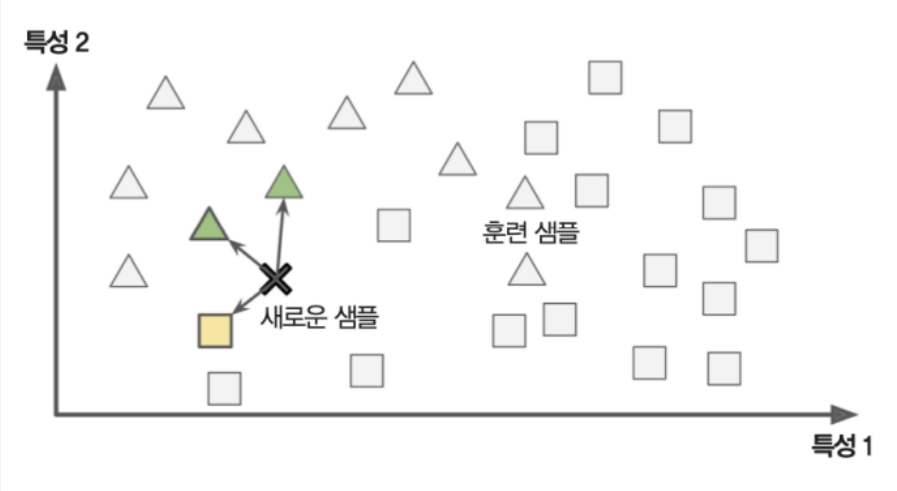
일반화 : 새로운 데이터에 대한 예측을 잘한다는 의미

사례 기반 학습

샘플을 기억하는 것이 훈련의 전부

예측을 위해 기존 샘플과의 유사도 측정

ex. K-NN 알고리즘, 새로운 샘플 X가 기존에 세모인 샘플과의 유사도가 높기 때문에 세모로 분류



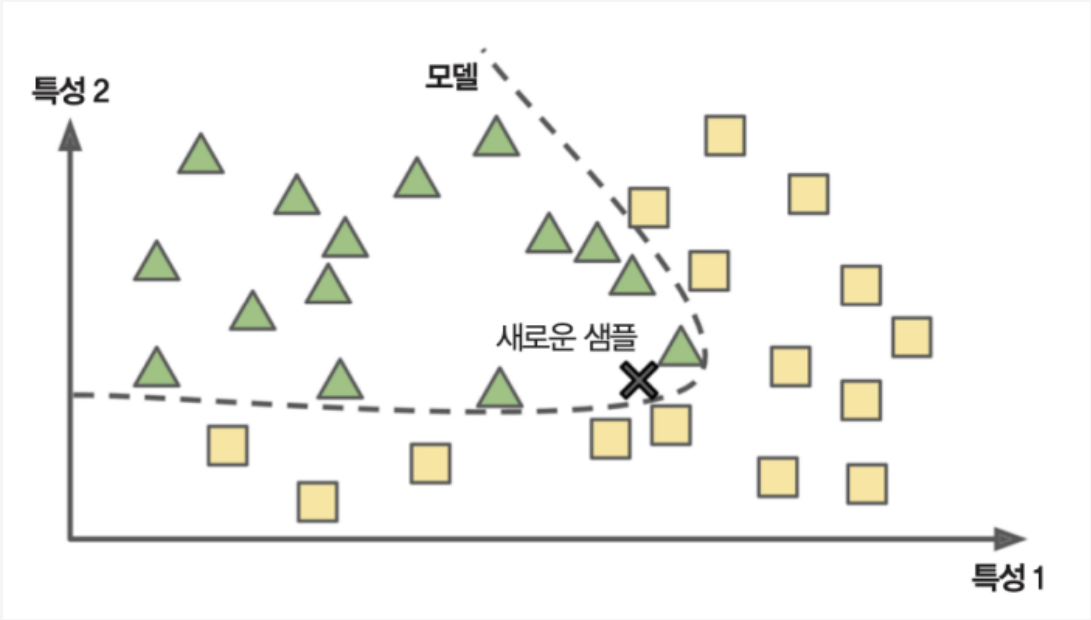
4.3. 사례 기반 학습과 모델 기반 학습

모델 기반 학습

모델을 미리 지정한 후 훈련 세트를 사용하여 모델을 훈련시킴

훈련된 모델을 사용해 새로운 데이터에 대한 예측 실행

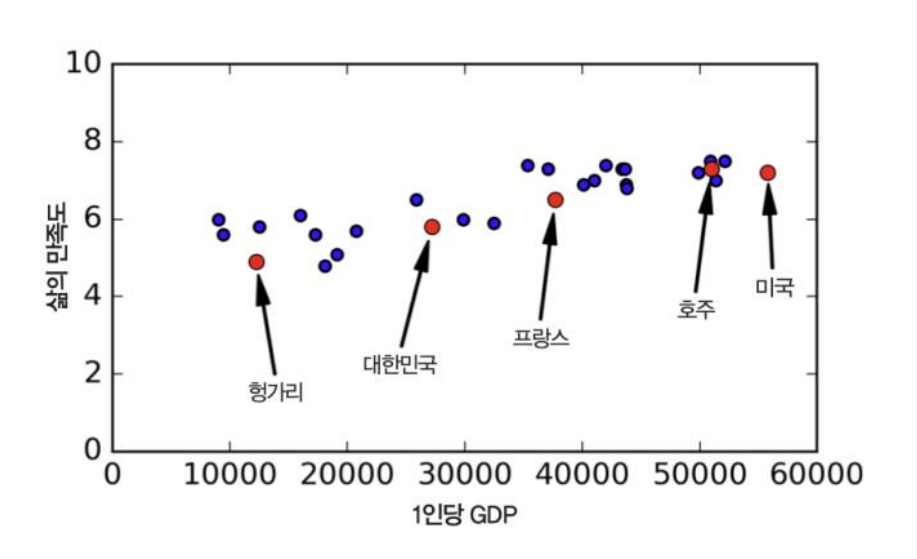
ex. 세션에서 다루는 대부분의 알고리즘, 학습된 모델을 이용하여 새로운 데이터 x를 세모 클래스로 분류



4.3. 사례 기반 학습과 모델 기반 학습

선형 모델 학습 예제

목표 : OECD 국가의 1인당 GDP(1인당 국가총생산)와 삶의 만족도 사이의 관계 파악



1인당 GDP가 증가할수록 삶의 만족도가 선형으로 증가하는 것처럼 보임.

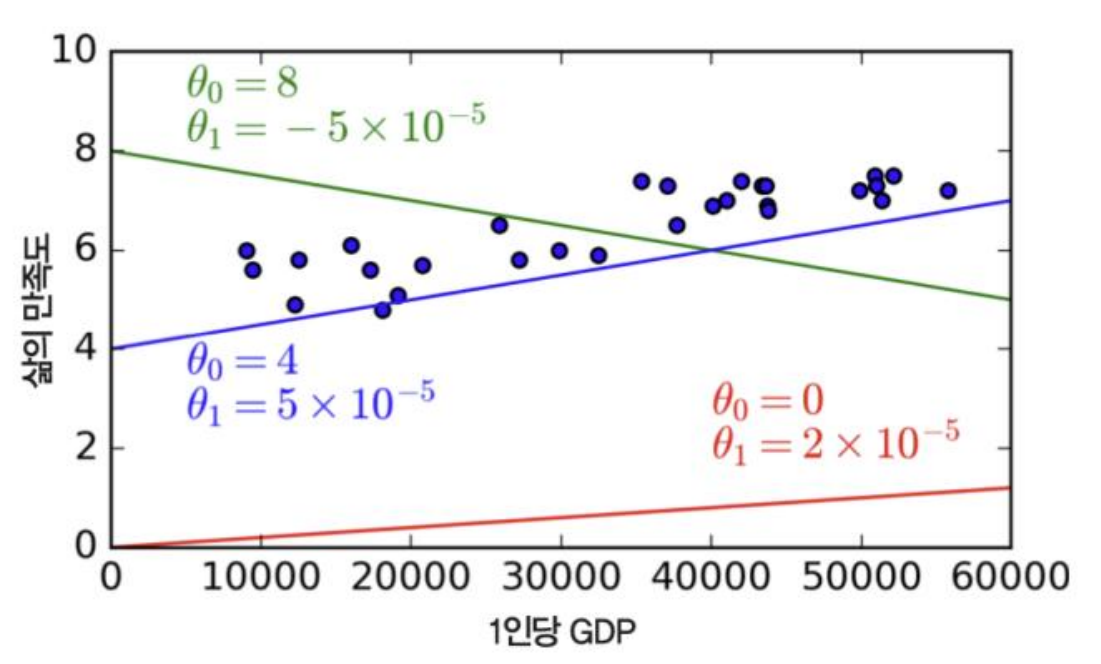
데이터를 대표하는 하나의 직선(선형 모델)을 찾기

4.3. 사례 기반 학습과 모델 기반 학습

선형 모델 학습 예제

선형 모델 : 삶의 만족도 = $\theta_0 + \theta_1 \times$ 1인당 GDP

데이터를 대표할 수 있는 선형 방정식을 찾아야 함



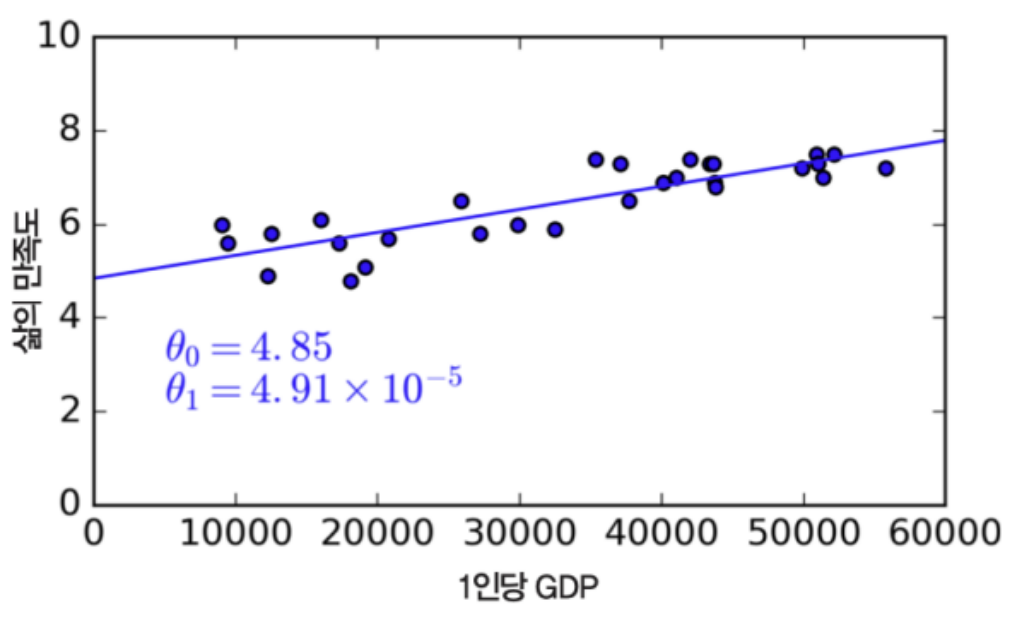
4.3. 사례 기반 학습과 모델 기반 학습

선형 모델 학습 예제

학습되는 모델의 성능평가기준을 측정하여 가장 적합한 모델 학습

효용 함수 : 모델이 얼마나 좋은지 측정

비용 함수 : 모델이 얼마나 나쁜지 측정



WeTIE

4.3. 사례 기반 학습과 모델 기반 학습

실습) 사이킷런을 이용한 선형 모델의 훈련과 실행

4.3. 사례 기반 학습과 모델 기반 학습

실습 요약

1. 데이터를 분석한다.
2. 모델을 선택한다.
3. 훈련 데이터로 모델을 훈련시킨다. (즉, 학습 알고리즘이 비용 함수를 최소화하는 모델 파라미터를 찾는다)
4. 마지막으로 새로운 데이터에 모델을 적용해 예측(추론)을 한다. 모델이 잘 일반화되기를 기도한다.

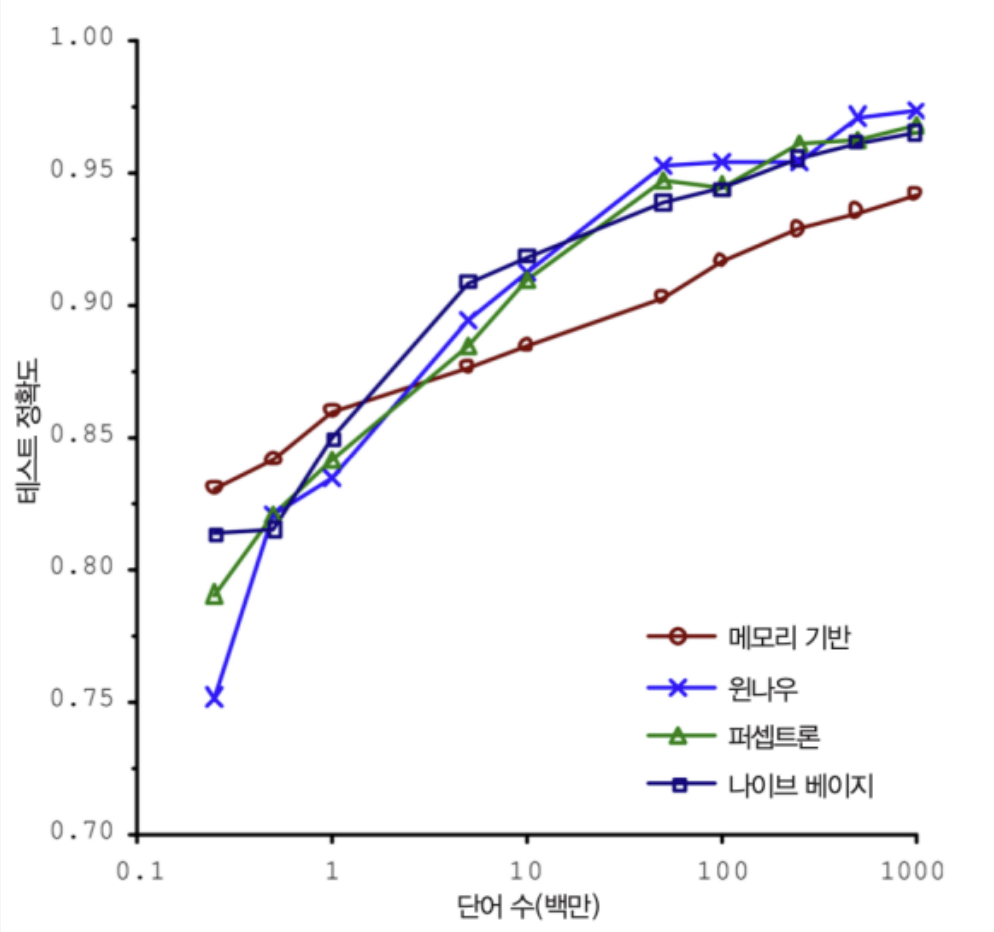
5 머신러닝의 주요 도전 과제

5.1. 충분하지 않은 양의 훈련 데이터

간단한 문제라도 수천 개의 데이터가 필요

이미지나 음성 인식 같은 문제는 수백만 개가 필요할 수도 있음

데이터가 부족하면 알고리즘 성능 향상 어려움

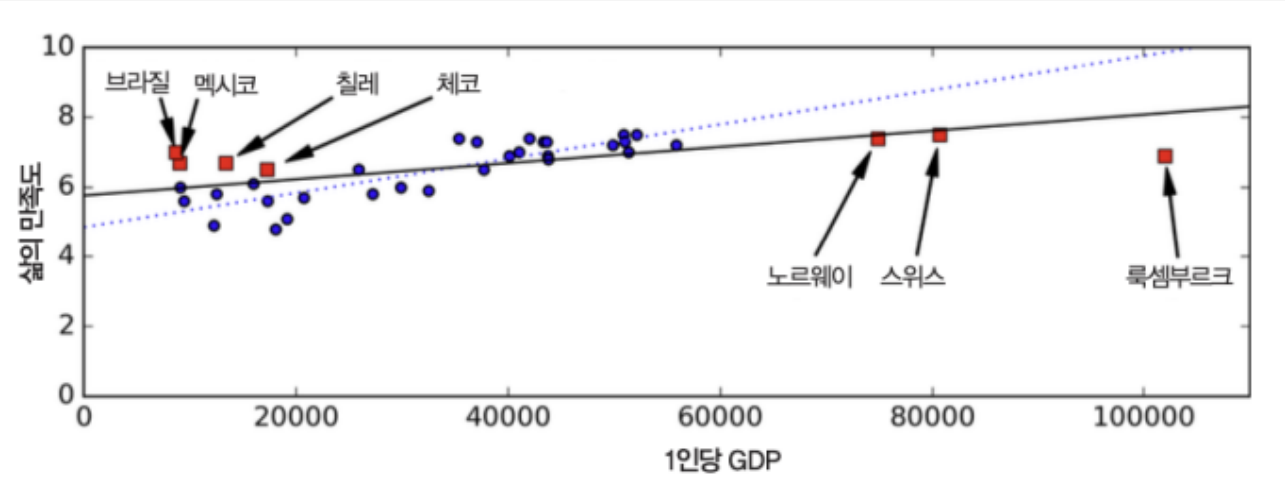


WeTIE

5.2. 대표성 없는 훈련 데이터

샘플링 잡음 : 우연에 의해 대표성이 없는 데이터

샘플링 편향 : 표본 추출 방법이 잘못된 대표성이 없는 데이터



5.3. 낮은 품질의 데이터

이상치 샘플이라면 고치거나 무시

특성이 누락되었다면

- 해당 특성을 제외
- 해당 샘플을 제외
- 누락된 값을 채움
- 해당 특성을 넣은 경우와 뺀 경우 각기 모델을 훈련

5.4. 관련이 없는 특성

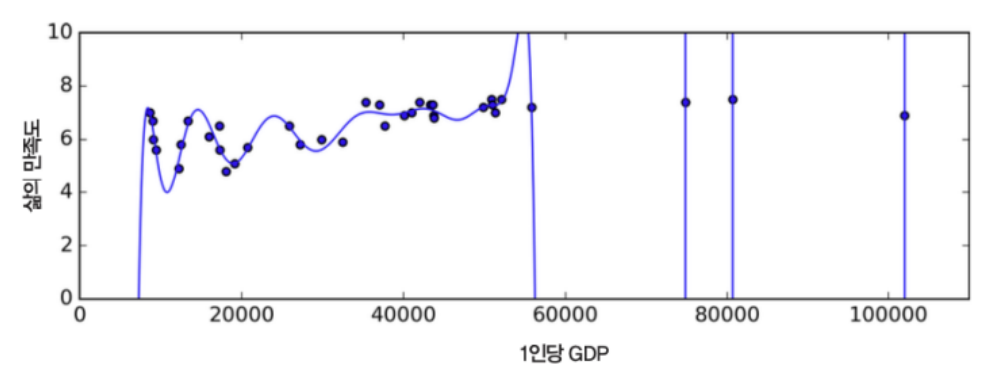
특성 공학 : 풀려는 문제에 관련이 높은 특성 찾기

특성 선택: 준비되어 있는 특성 중 가장 유용한 특성을 찾음

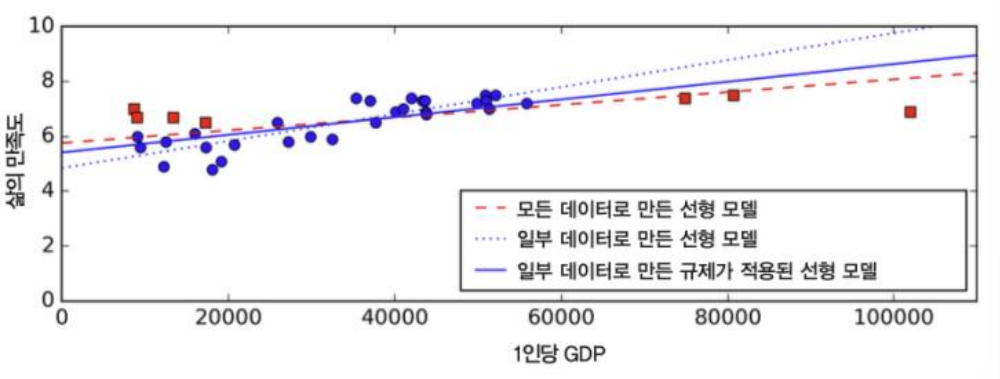
특성 추출: 특성을 조합하여 새로운 특성을 만듦

5.5. 훈련 데이터 과대적합(Overfitting)

훈련 세트에 특화되어 일반화 성능이 떨어지는 현상



여러 규제를 적용해 과대적합을 감소시킬 수 있음



WeTIE

5.6. 훈련 데이터 과소적합(Underfitting)

모델이 너무 단순해서 훈련 세트를 잘 학습하지 못함

해결 방법

- 모델 파라미터가 더 많은 복잡한 모델을 사용
- 특성 공학으로 더 좋은 특성을 찾음
- 규제의 강도를 줄임

6 테스트와 검증

6.1. 검증

훈련된 모델의 성능 평가 : 테스트 세트 활용

전체 데이터셋을 훈련 세트(80%)와 테스트 세트(20%)로 구분

- 훈련 세트 : 모델 훈련용
- 테스트 세트 : 모델 테스트용
- 데이터셋이 매우 크면 테스트 세트 비율을 낮출 수 있음

검증 기준 : 일반화 오차(테스트 데이터 정확도)

- 새로운 샘플에 대한 오류 비율
- 학습된 모델의 일반화 성능의 기준

과대 적합 : 훈련 오차에 비해 일반화 오차가 높은 경우

6.2. 하이퍼파라미터 튜닝과 모델 선택

하이퍼파라미터

- 알고리즘 학습 모델을 정의하는데 사용되는 파라미터
- 훈련 과정에 변하는 파라미터가 아님
- 하이퍼파라미터를 조절하면서 가장 좋은 성능의 모델 선정

모델 선택 : 선형 모델과 다항 모델 중 어떤 것을 선택해야 하는가?

6.3. 교차검증

검증 세트(홀드아웃 검증)

- 검증 세트 : 훈련 세트의 일부로 만들어진 데이터셋
- 다양한 하이퍼파라미터 값을 후보 모델 평가용으로 예비표본을 검증세트로 활용하는 기법

교차 검증

- 여러 개의 검증세트를 사용한 반복적인 예비표본 검증 적용 기법
- 장점 : 교차 검증 후 모든 모델의 평가를 평균하면 훨씬 정확한 성능 측정 가능
- 단점 : 훈련 시간이 검증 세트의 개수에 비례해 늘어남

E.O.D

1주차. 한눈에 보는 머신러닝