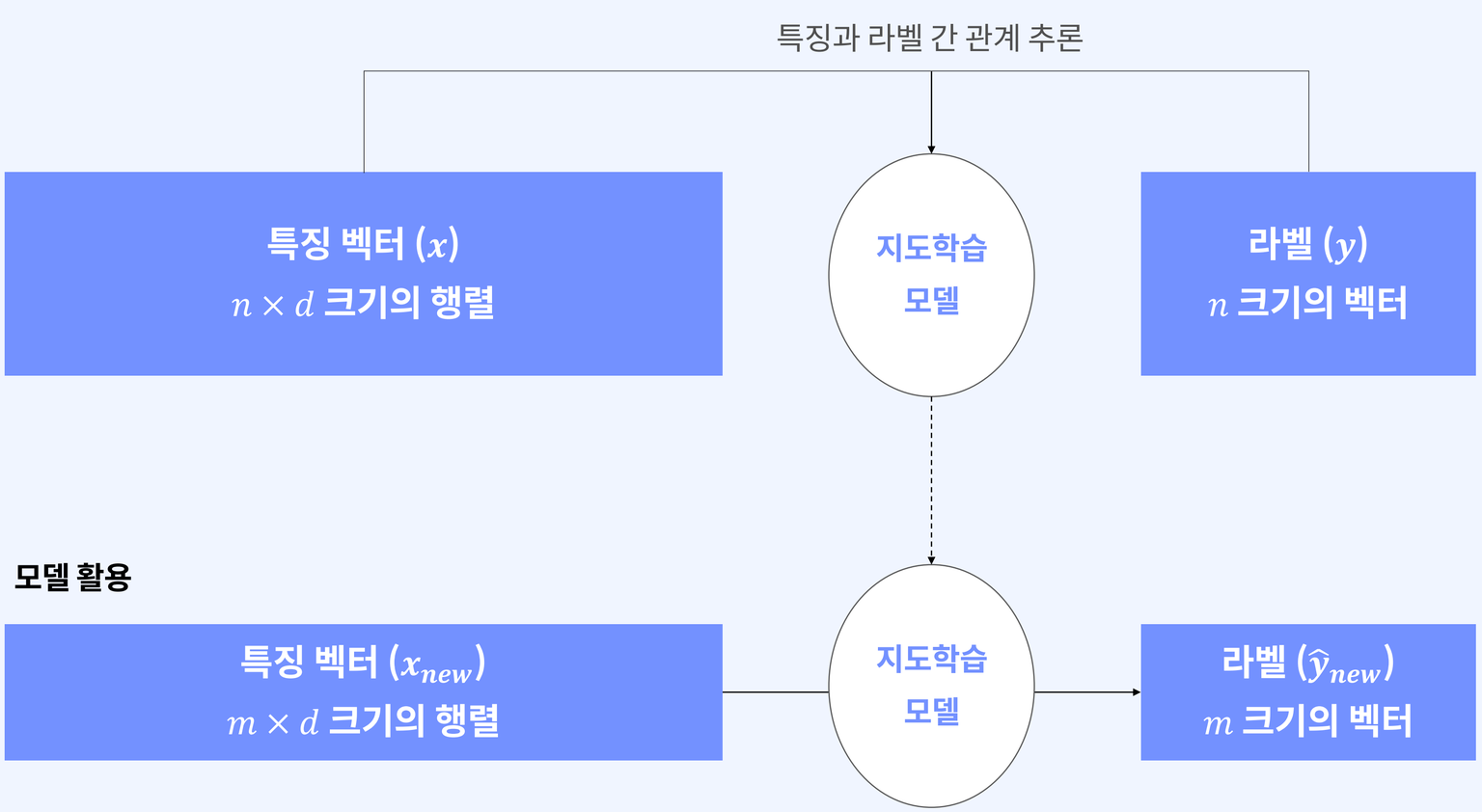
1. 머신러닝 과제의 분류

1 지도 학습 개요

지도 학습이란?

지도 학습은 데이터로 주어진 입력과 출력 간 관계를 학습하여 새로운 입력에 대해 적절한 출력을 내는 머신러닝의 대표적인 과제입니다.



지도학습 모델은
특징 벡터와 라벨 간
관계를 표현하는 함수

모델 학습

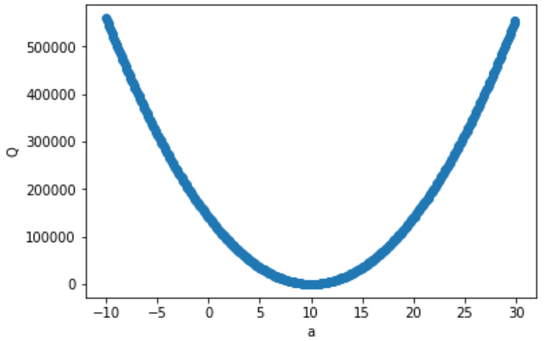
모델 학습은 손실 함수가 최소화되도록 모델의 파라미터를 추정하는 작업을 의미합니다.

손실 함수란 파라미터에 따른 예측 오차를 나타내는 함수를 의미함

(예시)

- **모델:** $y = ax$
- **손실 함수:** $\mathcal{L}(a) = \sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$
- **모델 학습:** $\hat{a} = \underset{a}{\operatorname{argmin}}\,\mathcal{L}(a)$

i	x	y
1	10	100
2	20	200
3	30	300



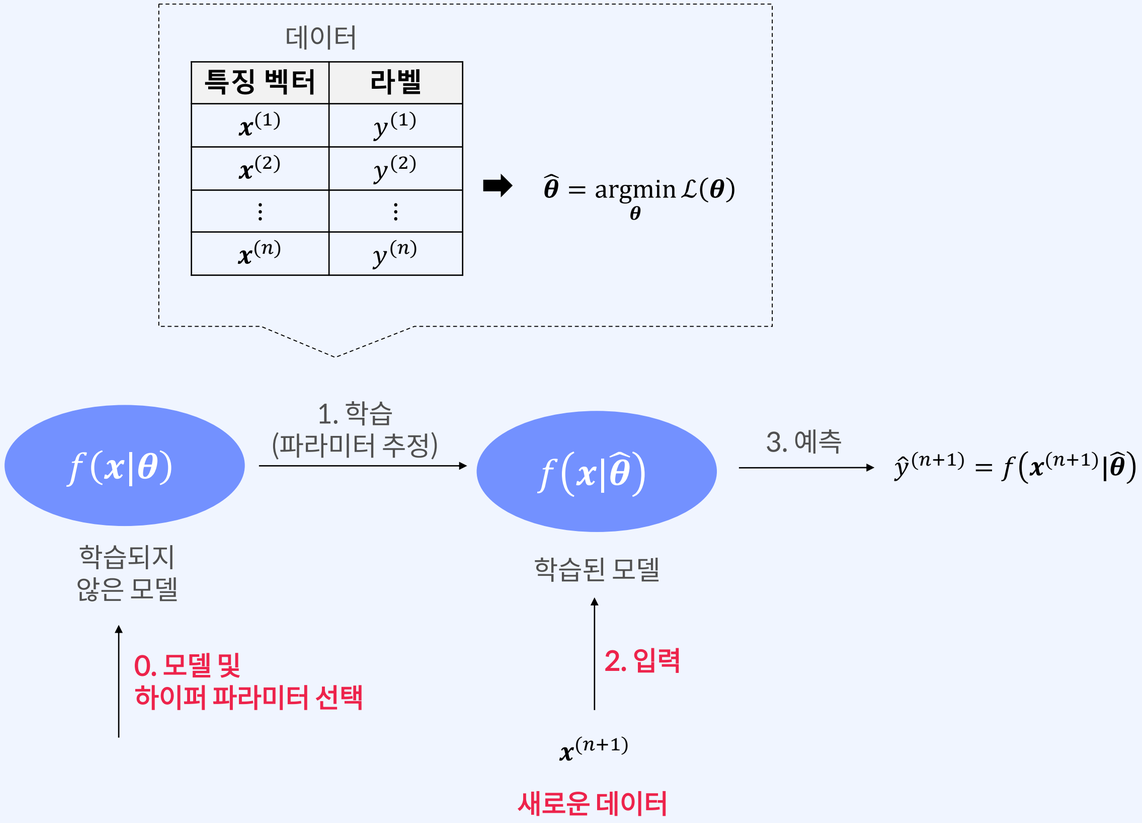
$$\hat{a} = \underset{a}{\operatorname{argmin}}\,\mathcal{L}(a) = 10$$

모델 학습은 머신러닝에서 가장 중요한 내용임

BUT 모델 학습은 컴퓨터가 하고, 라이브러리 등을 사용하지 않고 학습 코드를 작성할 일은 극히 드묾

실무자의 관점에서 바라본 지도 학습

실무자의 관점에서 중요한 것은 머신러닝 모델을 선택하는 것과 학습된 모델을 활용하는 것입니다.



실제로 고민해야 할 질문

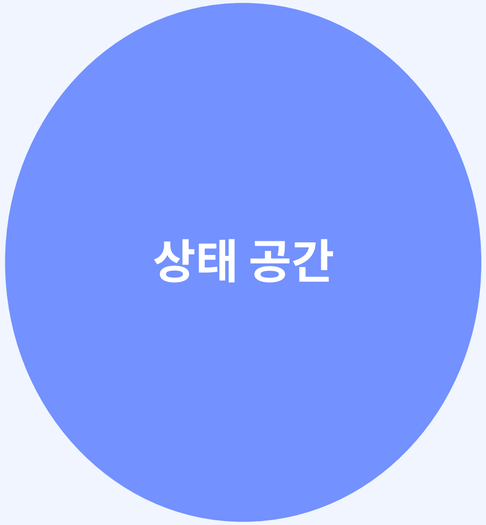
- 어떤 모델을 선택해야 하지?
- 어떤 하이퍼 파라미터를 선택하지?
- 이 모델을 실제 환경에서 사용할 수 있을까?

고민할 필요가 없는 질문

- 학습 알고리즘이 어떻게 작동하지?
- 학습 알고리즘을 내가 이해하고 구현할 수 있을까?

상태 공간
(state space)

지도 학습 과제는 상태 공간의 크기에 따라 분류(classification)와 회귀(regression)으로 구분할 수 있습니다.



- 한 변수가 취할 수 있는 값의 집합
- 상태 공간의 크기가 유한한 변수를 **범주형 변수**라 하며, 무한한 변수를 **연속형 변수**라 함
- 라벨이 범주형 변수라면 **분류**라고 하며, 연속형 변수면 **회귀**라고 함
- 실제로는 **도메인 지식을 활용**하거나 데이터에서는 **한 변수가 취하는 모든 값**을 바탕으로 변수의 유형을 판단함
  (주의: 실수 자료형이라고 해서 반드시 연속형 변수는 아님)

상태 공간에 따른 범주형 변수와 연속형 변수의 구분은 데이터를 탐색하거나 전처리할 때도 중요하게 사용됨

1. 머신러닝 과제의 분류

2 일반화와 과적합

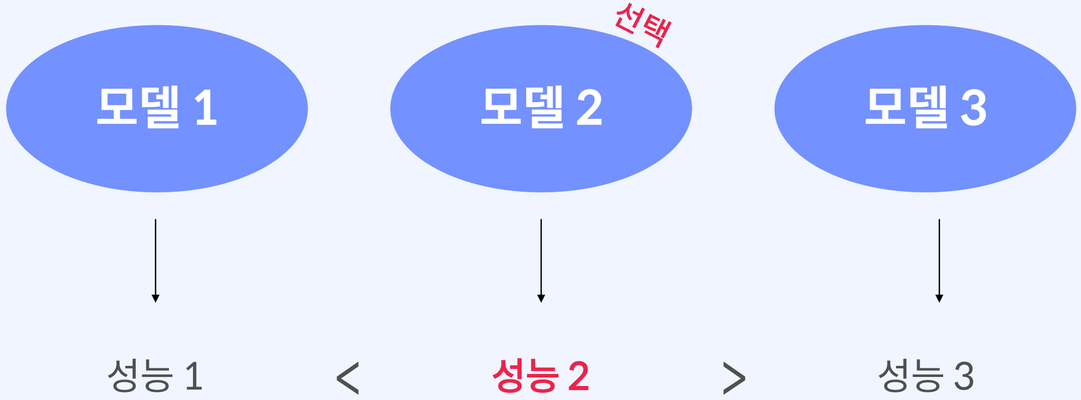
객관적인 평가

지도 학습은 라벨이 있기 때문에 모델의 성능을 객관적으로 평가할 수 있습니다.

모델 학습

반복적(iterative)으로 오차를 계산하고
오차를 줄이는 방향으로 학습할 수 있음

모델 비교를 통한 선택



성능을 기준으로 모델을 비교하고
가장 나은 모델을 선택할 수 있음

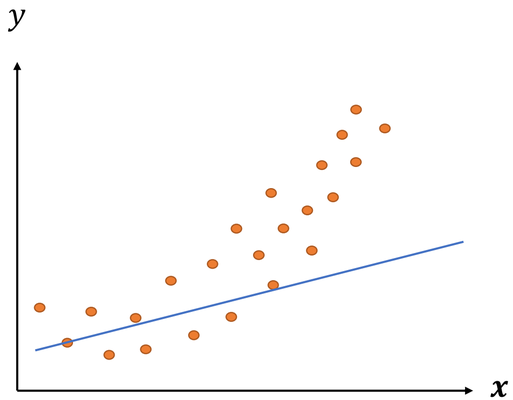
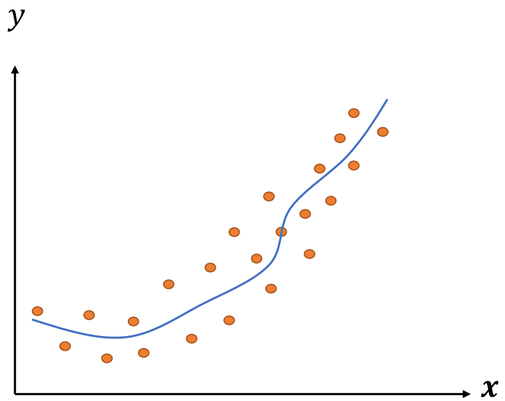
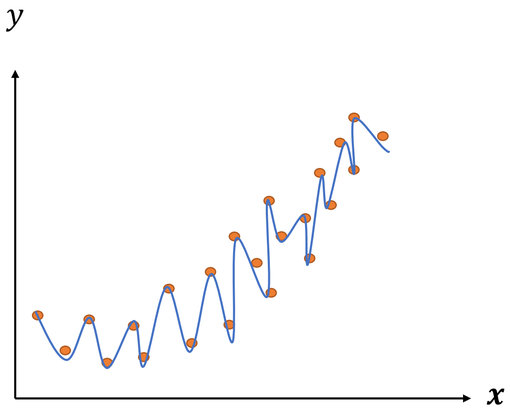
일반화

모든 지도 학습의 목표는 라벨을 알지 못하는 새로운 데이터를 잘 분류하거나 예측하는 모델을 개발하는 것입니다.

과적합과 과소적합

라벨을 알고 있는 **학습 데이터만 잘 맞추는 복잡한 모델을 과적합**됐다고 하며, **학습 데이터도 잘 맞추지 못하는 단순한 모델을 과소 적합**됐다고 합니다.

데이터 $\mathrm{D} = \{(\boldsymbol{x}^{(i)}, y^{(i)}) | i = 1, 2, \cdots, n\}$로 학습한 세 종류의 회귀 모델 f_1, f_2, f_3과 데이터

모델	 f_1 (과소 적합)	 f_2 (적정 적합)	 f_3 (과적합)
학습 데이터에 대한 오차	크다	작다	매우 작다
새로운 데이터에 대한 오차	크다	작다	크다

과적합과 과소적합의 원인

학습 데이터 대비 모델이 복잡할수록 과적합 가능성이, 단순할수록 과소 적합 가능성이 커집니다.

요인	효과	
데이터 크기	• 샘플 수가 많을수록 과적합 가능성 감소 • 특징 수가 많을수록 과적합 가능성 증가	데이터 관련: 제어 불가
특징 유형	• 상태 공간의 크기가 큰 특징일수록 과적합 가능성 증가	
모델 유형 및 하이퍼 파라미터	• 복잡한 모델 유형과 하이퍼 파라미터일수록 과적합 가능성 증가	모델 관련: 제어 가능
이터레이션 수	• 이터레이션 수가 적을수록 과소 적합 가능성 증가	

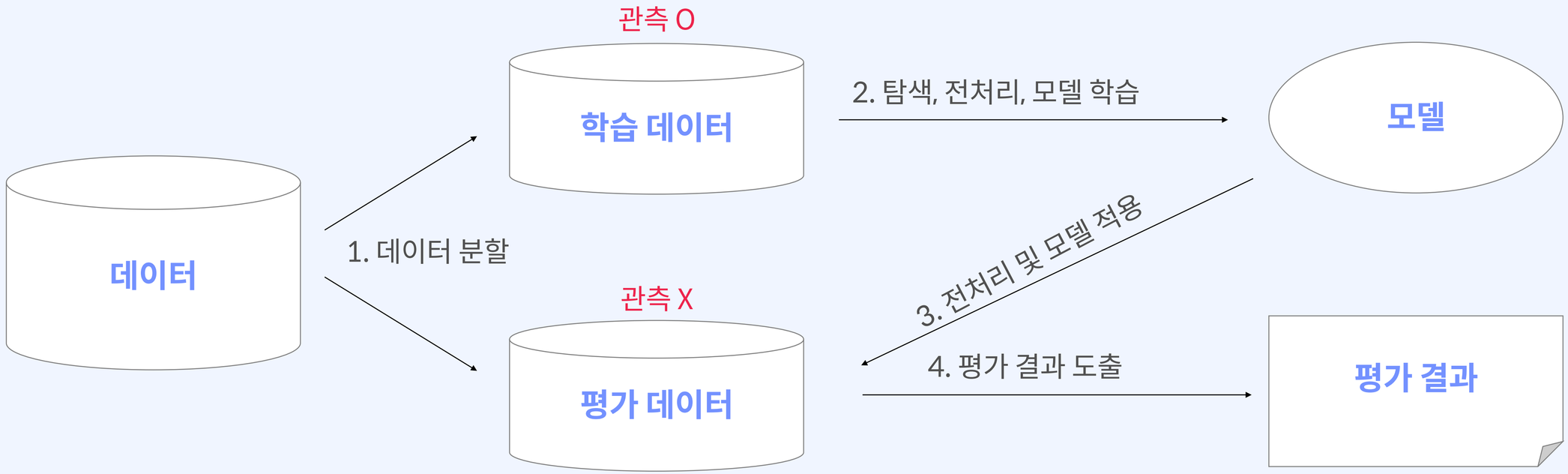
넓은 공간(많은 특징 수, 상태 공간의 크기가 큰 특징)에 데이터가 적을 때 복잡한 모델을 과하게 학습하면 과적합이 일어난다고 할 수 있음

1. 머신러닝 과제의 분류

3 데이터 분할

학습 데이터와 평가 데이터
: 개요

학습에 사용한 데이터를 그대로 모델을 평가하는데 사용하면 적절하게 적합된 모델보다 과적합된 모델을 좋게 평가하는 문제가 발생합니다.
따라서 모델 학습에 사용할 데이터(학습 데이터)와 학습된 모델을 평가하는데 사용할 데이터(평가 데이터)로 임의로 분할해야 합니다.



객관적인 모델 평가를 위해 평가 데이터는 탐색, 전처리, 파라미터 추정 등 모델을 학습하는 전 과정에서 활용해서는 안 됨

학습 데이터와 평가 데이터
: 문제점

단순히 학습 데이터와 평가 데이터로 분할하면 객관적인 평가가 안 될 위험이 있습니다.

평가 데이터는 모델 학습에 전혀 기여하지 못하는데, 데이터가 작을수록 모델의 성능 감소로 이어질 가능성이 큼

아무리 데이터를 임의로 분할하더라도 평가 결과가 객관적이지 않을 수 있음 → 평가 데이터에만 잘 맞는 모델을 좋게 평가 할 위험이 있기 때문

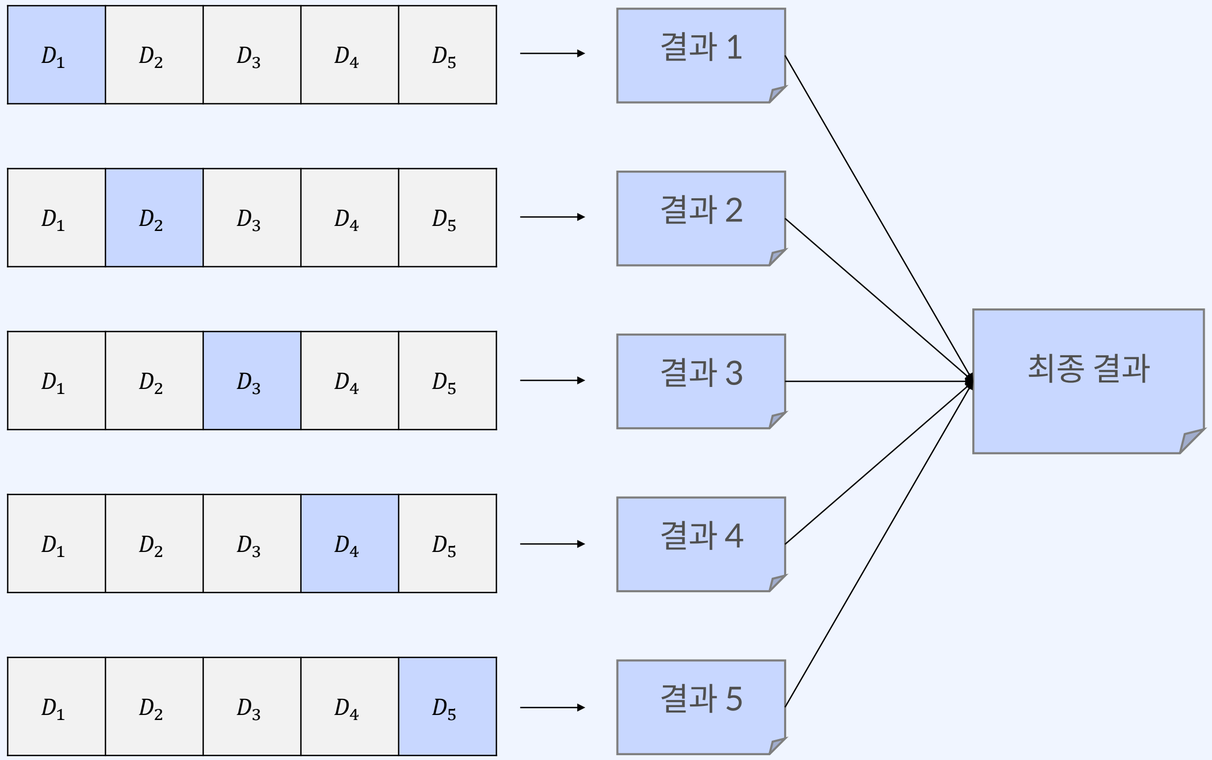
학습 데이터와 평가 데이터를 분리하면 과적합된 모델을 좋게 평가할 위험이 크게 줄어들지만, 여전히 객관적인 평가가 어려움

→ k-겹 교차 검증의 필요성

k-겹 교차 검증

k-겹 교차 검증은 데이터를 서로 겹치지 않는 k개의 작은 데이터인 폴드(fold)로 분할한 뒤, 각각의 폴드를 평가 데이터로 사용하고 나머지 폴드의 합집합을 학습 데이터로 사용하는 방식입니다.

5-겹 교차 검증 예시



활용 팁

k가 클수록 실제 모델을 사용했을 때의 예상 성능과 유사한 검증 결과를 얻을 수 있습니다.

k가 클수록 검증 시간이 오래 걸립니다.

결과 요약은 보통 평균을 사용하나, 보수적으로 검증할 때는 최솟값을 사용합니다.

최종 결과가 만족스럽다면 전체 데이터를 가지고 모델을 재학습합니다.

1. 머신러닝 과제의 분류

4 비지도 학습

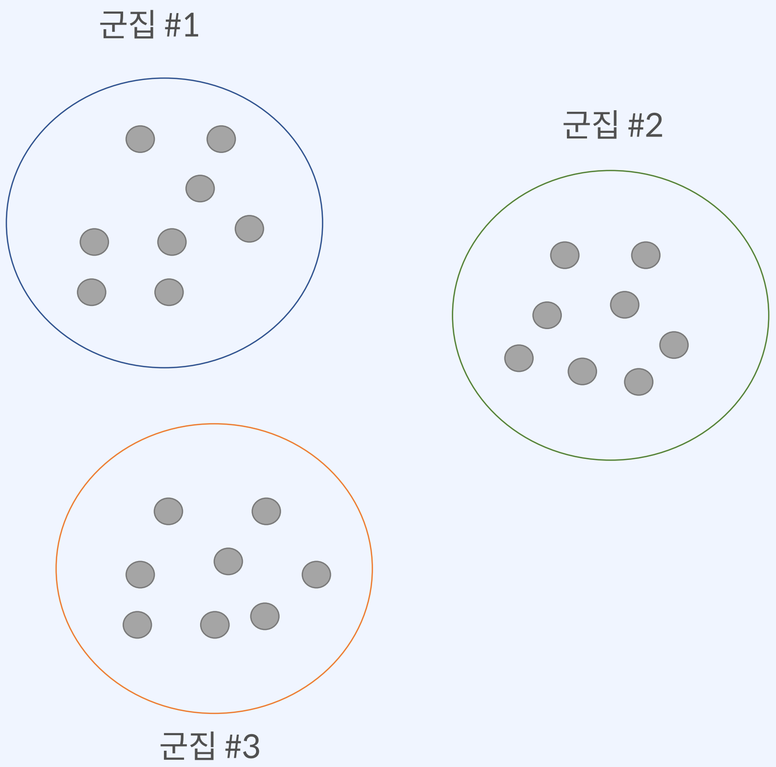
개요

비지도 학습은 라벨이 없는 데이터로 수행하는 머신러닝 과제입니다.

- 라벨이 없으므로 객관적인 평가와 오차에 따른 피드백이 불가능함
- 데이터에 숨겨진 특징이나 구조를 발견하는 데 주로 활용함
- 도메인 지식만을 활용하여 비지도 학습 기반의 분류 및 예측을 수행하기도 함
- 이상 탐지, 군집화 등이 비지도 학습에 속함
- 도메인 지식 등 주관이 많이 개입해서 머신러닝 자동화의 대상이 아니지만, 다양한 결과를 리포트 형태로 출력하는 시스템을 개발할 수도 있음

군집화

군집화(clustering)란 유사한 샘플을 하나의 그룹으로 묶고 유사하지 않은 샘플을 다른 그룹으로 묶는 비지도 학습 과제입니다.



두 샘플이 유사하다

= 두 샘플 간 거리가 가깝다

군집화를 위해 필요한 개념

거리 및 유사도를 측정하는 방법과 군집화 알고리즘에 따라 다른 군집화 결과가 나올 수 있습니다.

거리 및 유사도 측정 방법

- 유클리디안 거리/유사도
- 맨하탄 거리 / 유사도
- 코사인 유사도
- 매칭 유사도
- 자카드 유사도

군집화 알고리즘

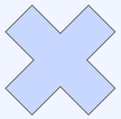
- k-평균 군집화 알고리즘
- 계층적 군집화 알고리즘
- DBSCAN 알고리즘

군집화와 머신러닝 자동화

군집화 결과는 다른 비지도 학습과 마찬가지로 어느 결과가 좋은지를 객관적으로 평가할 수 없으며 도메인 지식에 의존해서 평가해야 합니다.
따라서 군집화는 여러 모델을 학습하여 비교한 뒤 가장 좋은 모델을 반환하는 머신러닝 자동화 시스템에 사용하기 부적합합니다.

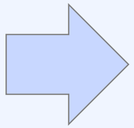
거리 및 유사도 측정 방법

- 유클리디안 거리/유사도
- 맨하탄 거리 / 유사도
- 코사인 유사도
- 매칭 유사도
- 자카드 유사도



군집화 알고리즘

- k-평균 군집화 알고리즘
- 계층적 군집화 알고리즘
- DBSCAN 알고리즘



리포트

분석 결과 # 1

- 알고리즘: k-평균 군집화 알고리즘
- 거리 척도: 유클리디안 거리
- 군집 수: 5

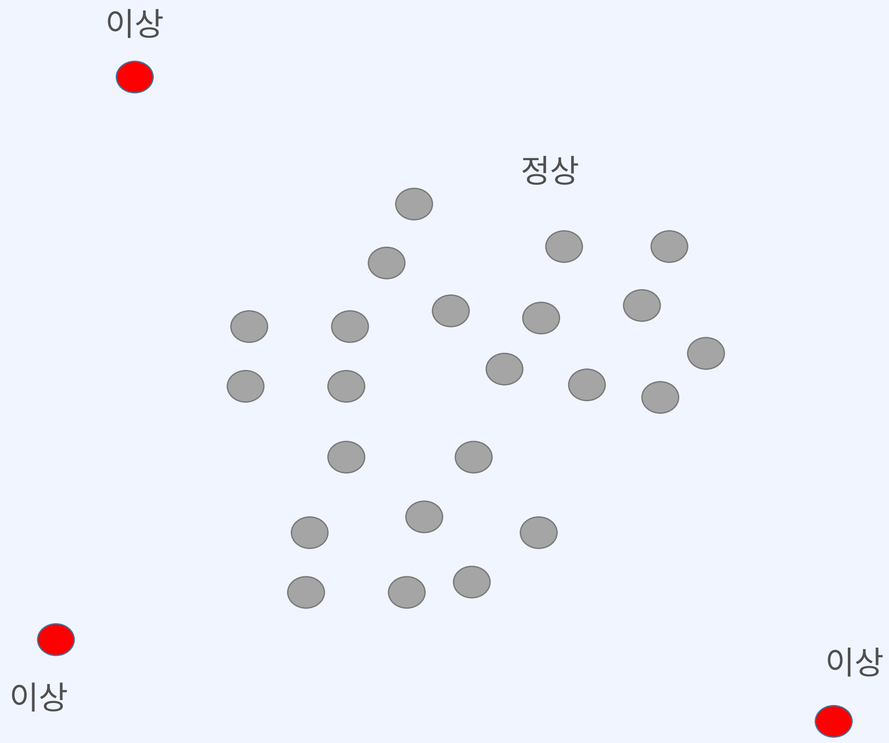
분석 결과 # 2

- 알고리즘: 계층적 군집화 알고리즘
- 거리 척도: 코사인 거리
- 군집 수: 5

...

이상 탐지

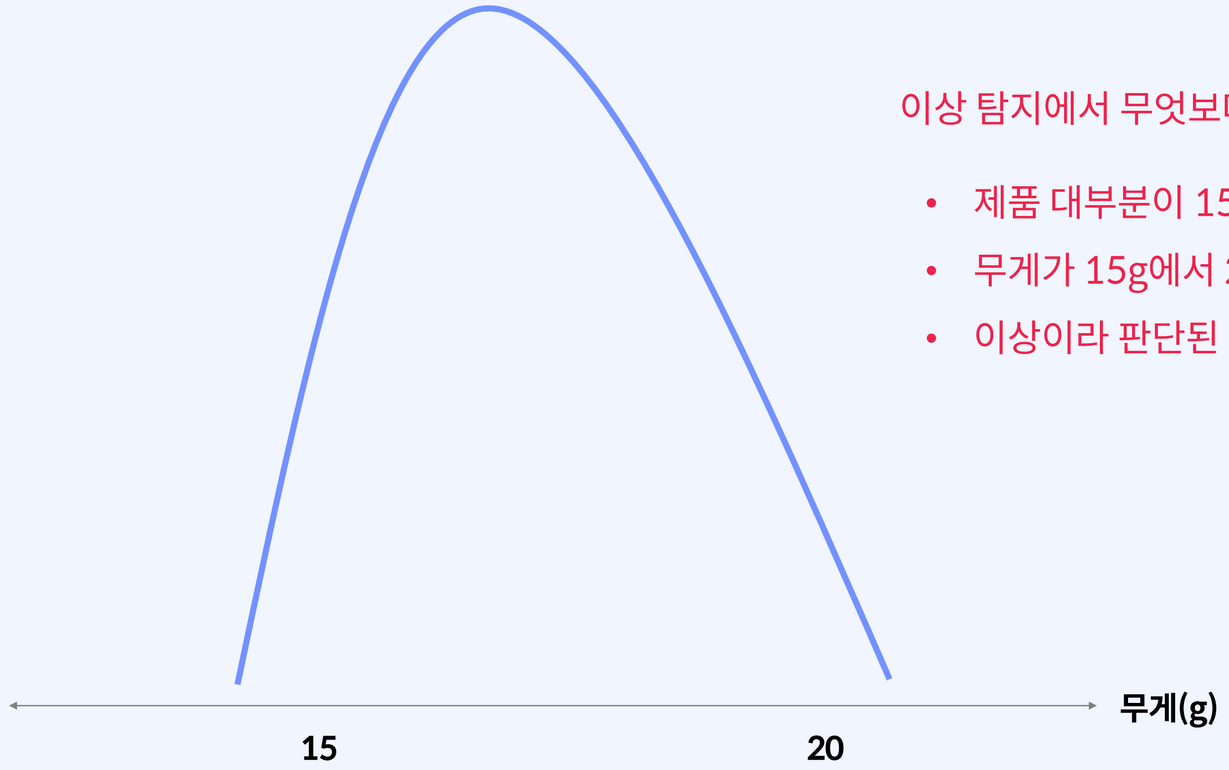
이상 탐지는 데이터에서 이상한 샘플을 찾아내는 비지도 학습 과제입니다.



- 이상: 문제가 있는 샘플이 아니라 다른 샘플과 많이 다른 샘플
- 정상: 문제가 없는 샘플이 아니라 다른 샘플과 비슷한 샘플

이상 탐지를 이용한 불량 탐지 예시

제품 무게 기반의 불량 탐지: 무게의 분포



이상 탐지에서 무엇보다 중요한 것은 도메인에 기반한 특징

- 제품 대부분이 15g에서 20g 사이임
- 무게가 15g에서 20g 사이면 정상, 그렇지 않으면 이상이라 판단
- 이상이라 판단된 샘플은 불량인가? → 그럴 수도 있고 그렇지 않을 수도 있음