

머신러닝/딥러닝 개요 및 용어정리

임 경 태

00. 용어 정리

- 인공지능

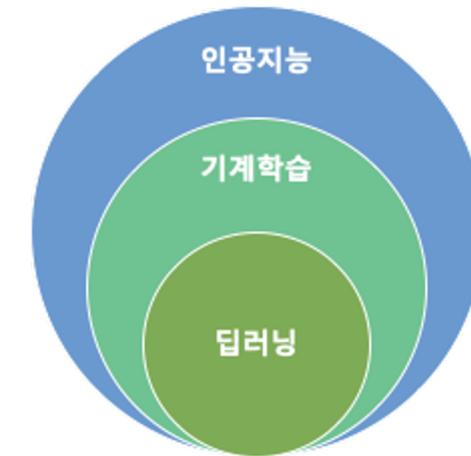
- 인간의 **지능**으로 할 수 있는 **사고**, **학습**, 자기개발 등을 컴퓨터가 할 수 있도록 하는 방법

- 기계학습 (머신러닝)

- 인간이 자연적으로 수행하는 **학습 능력과** 같은 기능을 컴퓨터에서 실현하려는 기술이나 방법

- 빅데이터

- 인공지능/기계학습 을 실현 할 수 있는 데이터



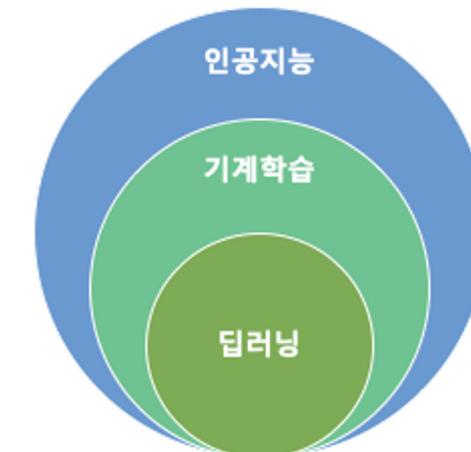
00. 용어 정리

- 인공지능은 어떻게 구현하는가?

- 주로 기계학습 알고리즘을 활용함
- 특정한 과제에 대해 경험을 통해 성능을 향상시키는 것

- **기계학습:**

- 경험을 통해
- **(빅)데이터**를 모아서
- 패턴을 **분석**하여
- 특정 문제를 해결한다



01. 머신러닝이란?

- 머신러닝은 데이터에서부터 학습하도록 컴퓨터를 프로그래밍하는 과학(또는 예술)
 - “머신러닝은 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야”
-아서 새뮤얼(Arthur Samuel), 1959
 - “어떤 작업 T 에 대한 컴퓨터 프로그램의 성능을 P 로 측정했을 때 경험 E 로 인해 성능이 향상됐다면, 이 컴퓨터 프로그램은 작업 T 와 성능 측정 P 에 대해 경험 E 로 학습한 것”
-톰 미첼(Tom Mitchell), 1997
- 스팸 필터는 (사용자가 스팸이라고 지정한) 스팸 메일과 일반 메일의 샘플을 이용해 스팸 메일 구분법을 배울 수 있는 머신러닝 프로그램의 하나
- 기본 용어
 - 훈련 세트(training set): 시스템이 학습하는 데 사용하는 샘플
 - 훈련 사례(training instance)(혹은 샘플): 각 훈련 데이터,
 - 이 경우 작업 T 는 새로운 메일이 스팸인지 구분하는 것
 - 경험 E 는 훈련 데이터(training data)
 - 성능 측정 P 는 직접 정의해야 하며, 이 성능 측정을 정확도accuracy라고 부르며 분류 작업에 자주 사용됨

02. 머신러닝의 탄생 배경

- 전통적 프로그래밍 기법으로는 규칙이 점점 길고 복잡해지므로 유지 보수하기 매우 힘듦
 - 머신러닝 기법에 기반을 둔 스팸 필터는 일반 메일에 비해 스팸에 자주 나타나는 패턴을 감지하여 어떤 단어와 구절이 스팸 메일을 판단하는 데 좋은 기준인지 자동으로 학습합니다
- 전통적인 방식으로는 너무 복잡하거나 알려진 알고리즘이 없는 분야(예:음성인식)

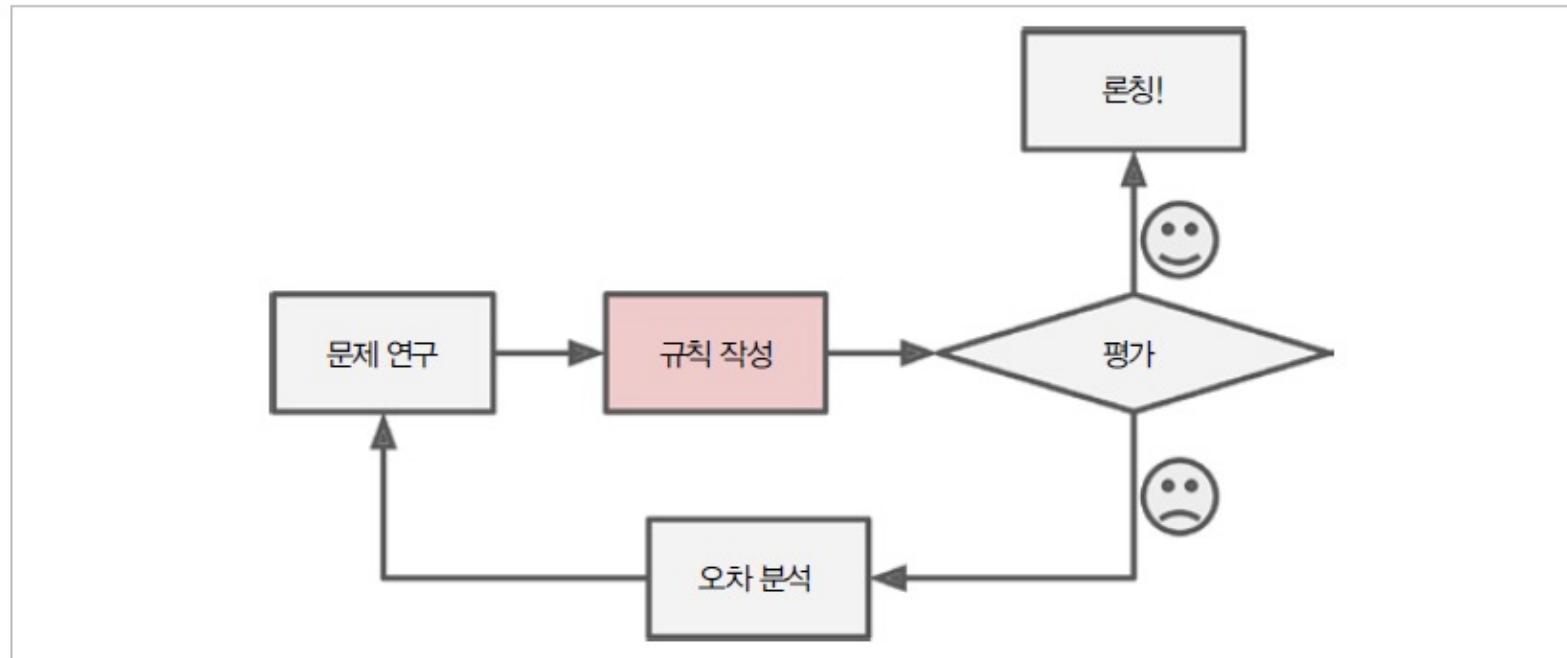


그림 1-1 전통적인 접근 방법

03. 머신러닝을 왜 사용하는가?

- 머신러닝의 장점

- 기존 솔루션으로는 많은 수동 조정과 규칙이 필요한 문제: 하나의 머신러닝 모델이 코드를 간단하게 만들고 전통적인 방법보다 더 잘 수행되도록 할 수 있음
- 전통적인 방식으로는 해결 방법이 없는 복잡한 문제: 가장 뛰어난 머신러닝 기법으로 해결 방법을 찾을 수 있음
- 유동적인 환경: 머신러닝 시스템은 새로운 데이터에 적응 가능
- 복잡한 문제와 대량의 데이터에서 통찰 얻기

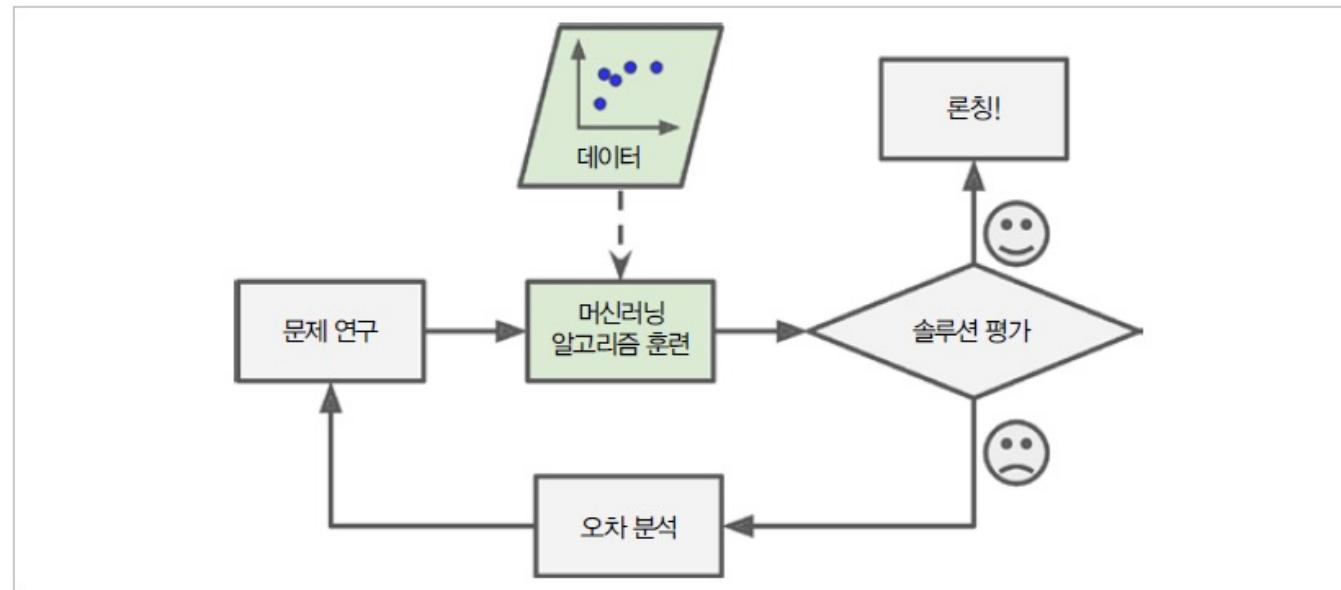
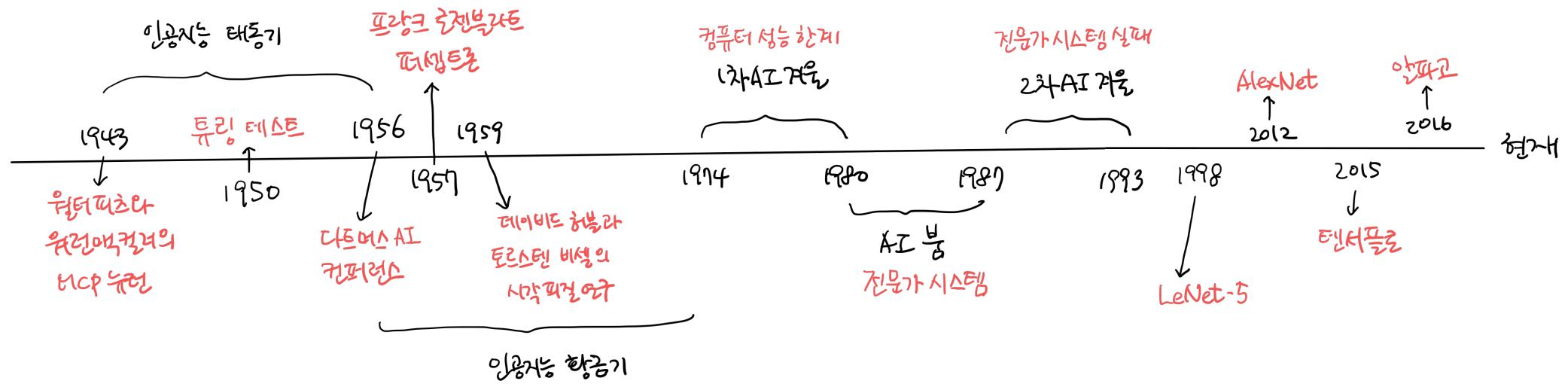


그림 1-2 머신러닝 접근 방법

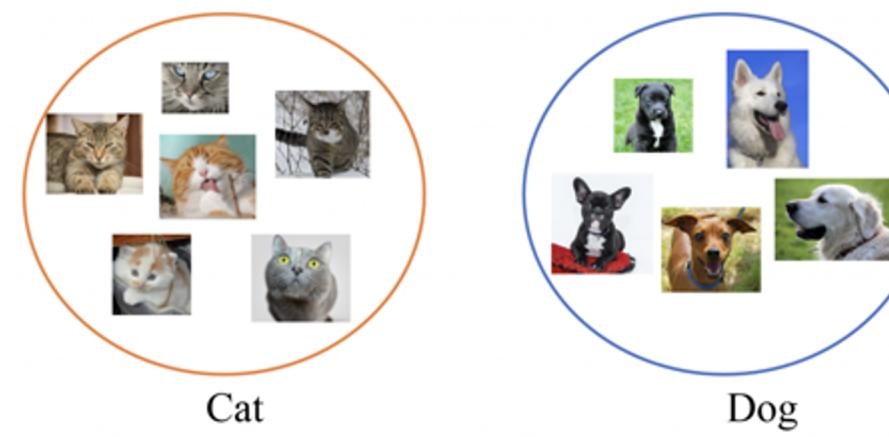
03. 머신러닝을 왜 사용하는가?

• 머신러닝 발전 역사



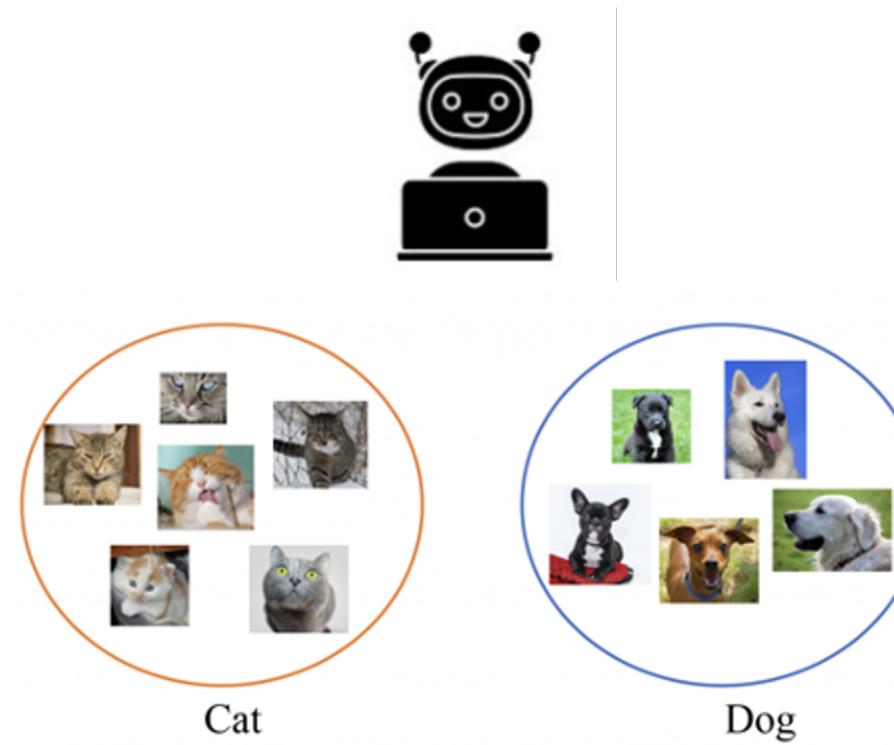
04. 머신러닝은 어떻게 작동할까? Overview

- 사람은 개와 고양이를 구분하는 방법을 어떻게 배울까?
 - 경험을 통해 특징점을 **분석**하거나 혹은 **데이터**(e.g., 책)를 통해 지식으로 습득



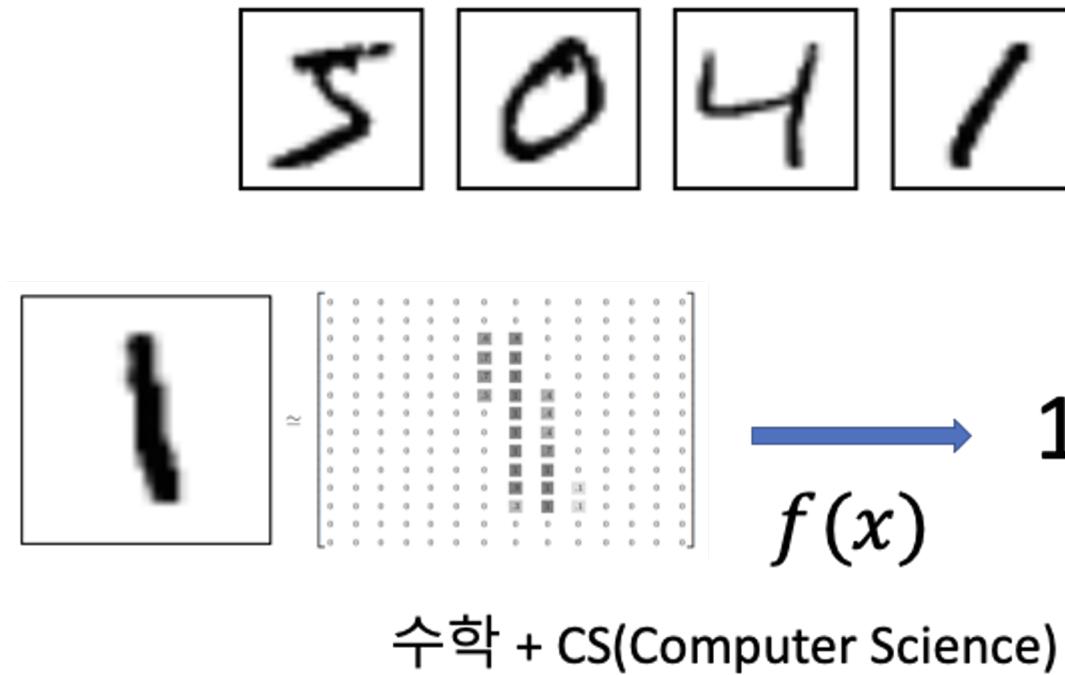
04. 머신러닝은 어떻게 작동할까? Overview

- 기계학습으로 개와 고양이를 어떻게 구분할 수 있을까?
 - 데이터(e.g., 사진)를 통해 특징점을 **분석**하고 이를 지식으로 습득



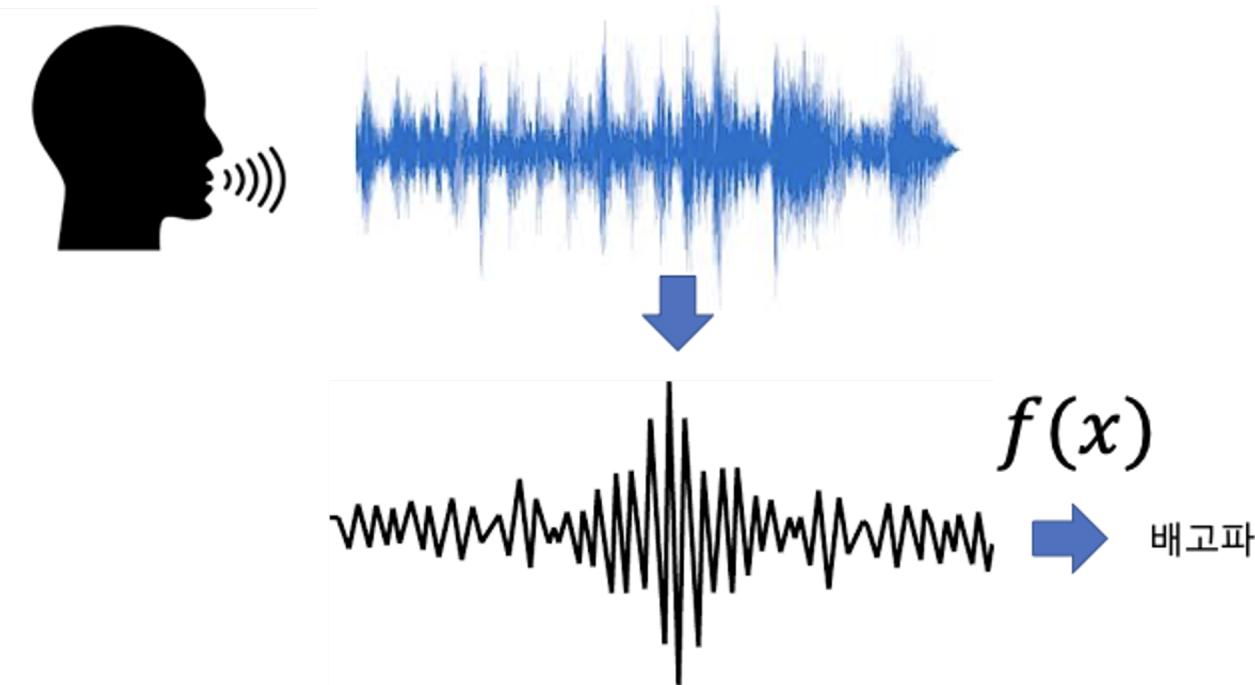
04. 머신러닝은 어떻게 작동할까? Overview

- 기계학습 기반의 이미지 인식
 - 이미지의 패턴을 찾는 함수를 찾는다!



04. 머신러닝은 어떻게 작동할까? Overview

- 기계학습 기반의 음성 인식
 - 음성데이터의 패턴을 찾는 함수를 찾는다!



04. 머신러닝은 어떻게 작동할까? Overview

- 기계학습 기반의 언어 번역
 - 한국어와 영어의 문장 패턴을 찾는 함수를 찾는다!



04. 머신러닝은 어떻게 작동할까? Overview

- 인공지능의 기술 수준 (실패 사례)
 - Object Detection



04. 머신러닝은 어떻게 작동할까? Overview

- 인공지능의 기술 수준 (실패 사례)
 - Object Detection

Damn google, these getting hard



04. 머신러닝은 어떻게 작동할까? Overview

- 인공지능의 기술 수준 (실패 사례)
 - Object Detection



04. 머신러닝은 어떻게 만드나?

- 코딩으로 구현함

- 프로그래밍 언어: Python
- 머신러닝 라이브러리: sk-learn, keras, pytorch, tensorflow

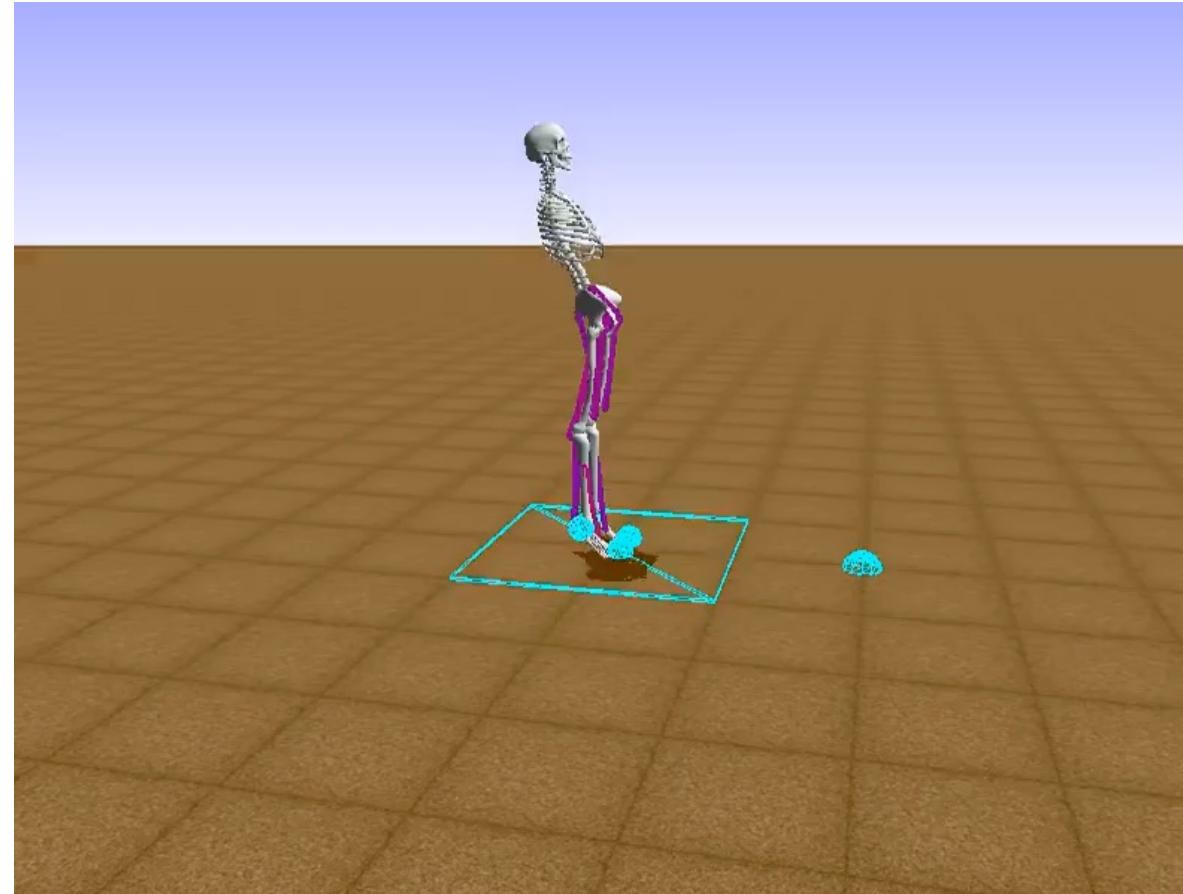


05. 머신러닝이 어디에 활용될까?

- 이미지 분류 작업: 생산 라인에서 제품 이미지를 분석해 자동으로 분류
- 시맨틱 분할 작업: 뇌를 스캔하여 종양 진단
- 텍스트 분류(자연어 처리): 자동으로 뉴스 기사 분류
- 텍스트 분류: 토론 포럼에서 부정적인 코멘트를 자동으로 구분
- 텍스트 요약: 긴 문서를 자동으로 요약
- 자연어 이해 : 챗봇(chatbot) 또는 개인 비서 만들기
- 회귀 분석: 회사의 내년도 수익을 예측하기
- 음성 인식: 음성 명령에 반응하는 앱
- 이상치 탐지: 신용 카드 부정 거래 감지
- 군집 작업: 구매 이력을 기반으로 고객을 나누고 각 집합마다 다른 마케팅 전략을 계획
- 데이터 시각화: 고차원의 복잡한 데이터셋을 명확하고 의미 있는 그래프로 표현하기
- 추천 시스템: 과거 구매 이력을 기반으로 고객이 관심을 가질 수 있는 상품 추천하기
- 강화 학습: 지능형 게임 봇(bot) 만들기

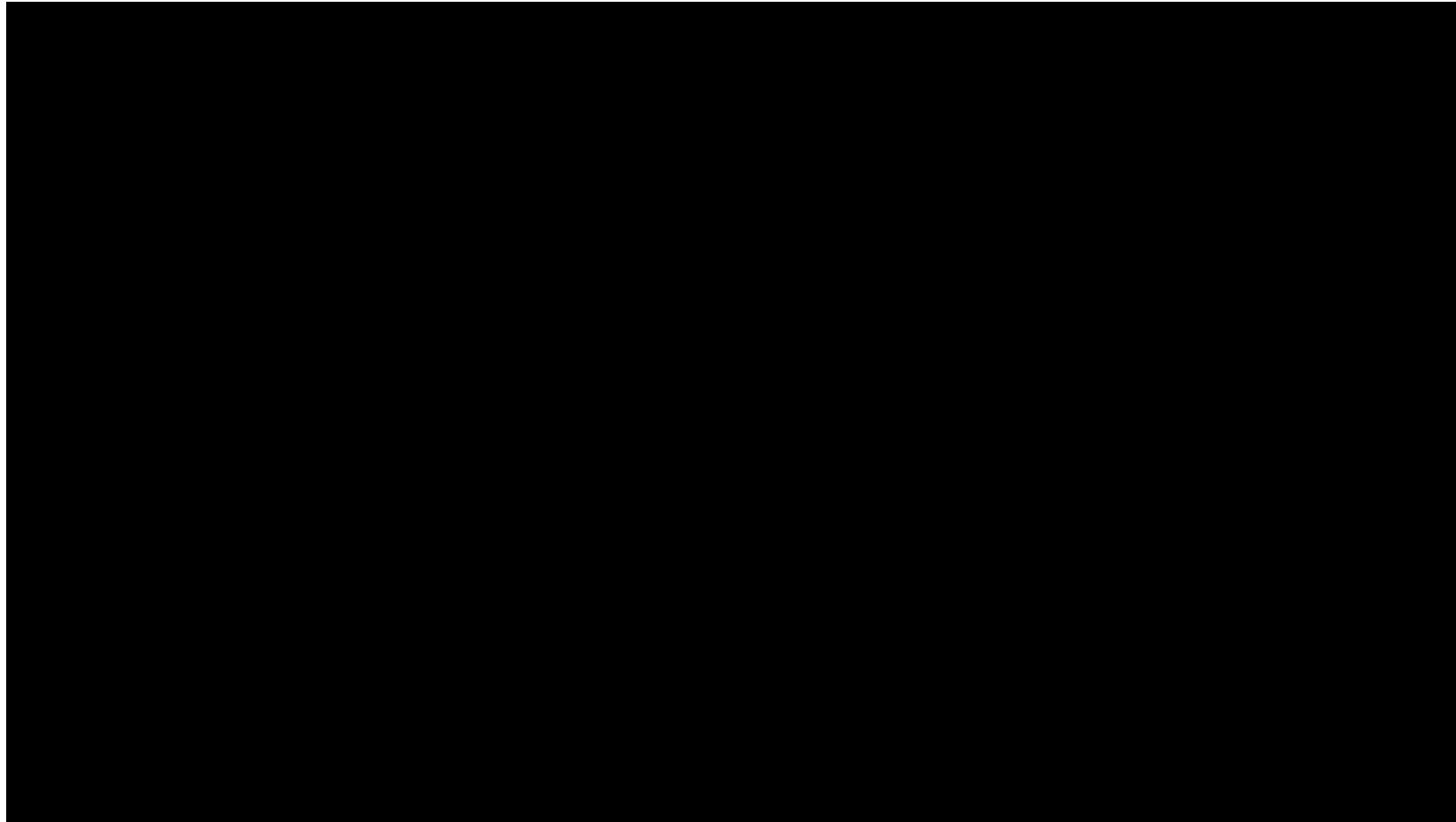
05. 머신러닝이 어디에 활용될까?

- 인공지능의 기술 수준 (성공 사례)
 - 최적제어



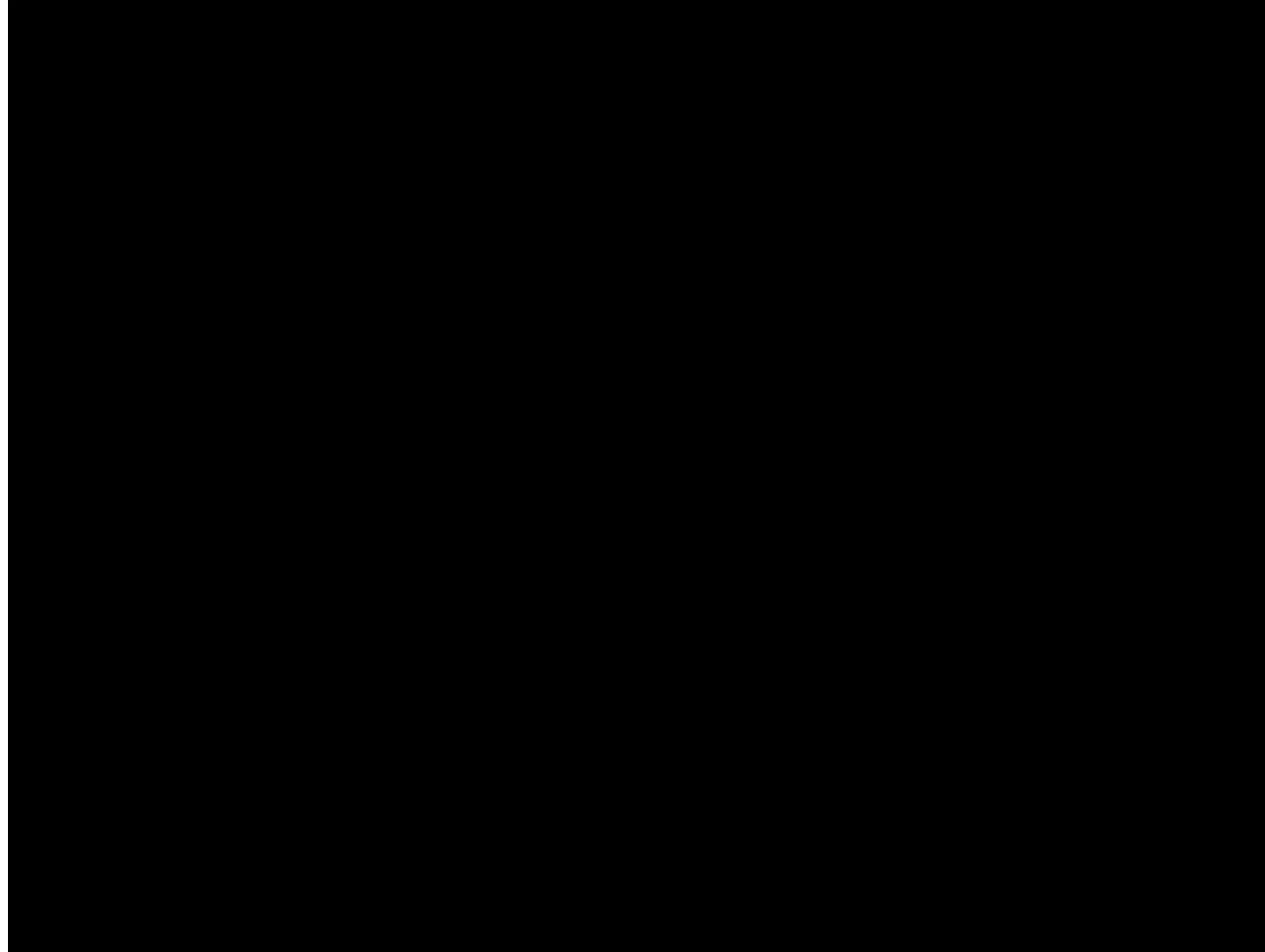
05. 머신러닝이 어디에 활용될까?

- 인공지능의 기술 수준 (성공 사례)
 - 최적제어



05. 머신러닝이 어디에 활용될까?

- 인공지능의 기술 수준 (성공 사례)
 - 게임



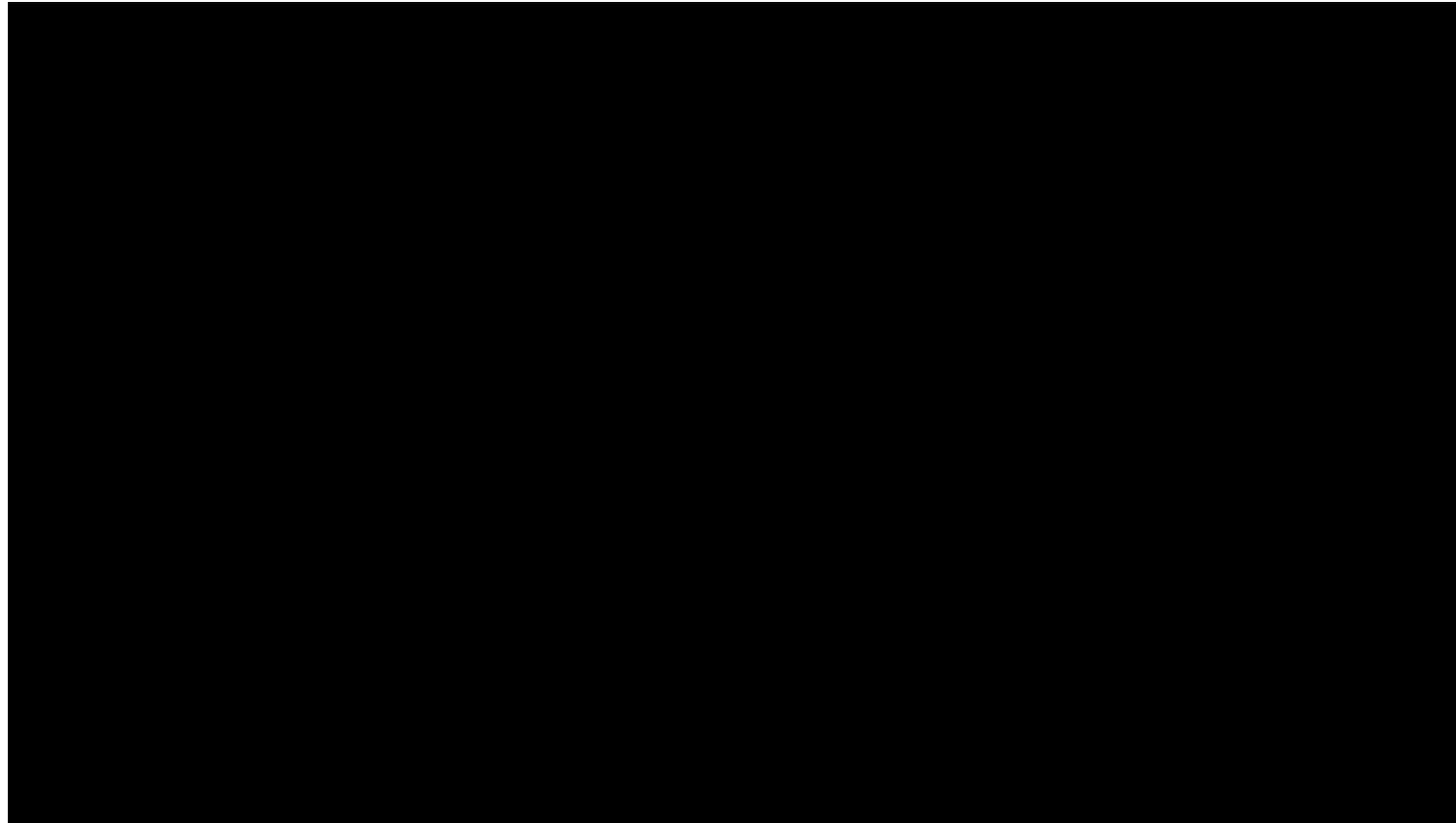
05. 머신러닝이 어디에 활용될까?

- 인공지능의 기술 수준 (성공 사례)
 - 이미지 합성



05. 머신러닝이 어디에 활용될까?

- 인공지능의 기술 수준 (성공 사례)
 - 예술



06. 머신러닝 시스템의 분류

- 넓은 범주의 분류

- 사람의 감독하에 훈련하는 것인지 그렇지 않은 것인지: 지도, 비지도, 준지도, 강화 학습
- 실시간으로 점진적인 학습을 하는지 아닌지: 온라인 학습과 배치 학습
- 단순하게 알고 있는 데이터 포인트와 새 데이터 포인트를 비교하는 것인지 아니면 과학자처럼 훈련 데이터셋에서 패턴을 발견하여 예측 모델을 만드는지: 사례 기반 학습과 모델 기반 학습

- 지도 학습과 비지도 학습

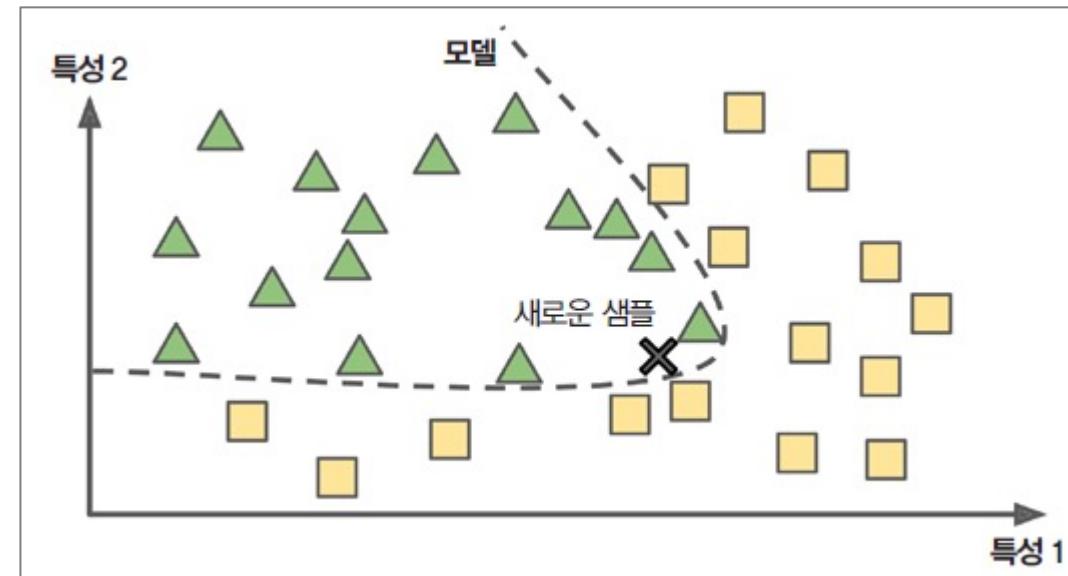
- 지도 학습
- 비지도 학습
- 준지도 학습
- 강화 학습

- 배치 학습과 온라인 학습

- 배치 학습
- 온라인 학습

- 사례 기반 학습과 모델 기반 학습

- 사례 기반 학습
- 모델 기반 학습



▲ 모델 기반 학습

06. 머신러닝 시스템의 분류

지도 학습과 비지도 학습

- **지도 학습:** 알고리즘에 주입하는 훈련 데이터에 레이블(label)이라는 원하는 답이 포함된다.
 - 분류
 - 특성(예측 변수)
 - 회귀
- 중요한 지도학습 알고리즘들
 - k-최근접 이웃
 - 선형 회귀
 - 로지스틱 회귀
 - 서포트 벡터 머신
 - 결정 트리와 랜덤 포레스트
 - 신경망



그림 1-5 스팸 분류를 위한 레이블된 훈련 세트(지도 학습의 예)

06. 머신러닝 시스템의 분류

지도 학습과 비지도 학습

- **비지도 학습:** 훈련 데이터에 레이블이 없어서, 시스템이 아무런 도움 없이 학습해야 한다.
- 중요한 비지도학습 알고리즘들

- 군집
 - k-평균
 - DBSCAN
 - 계층 군집 분석
 - 이상치 탐지와 특이치 탐지
 - 원-클래스 SVM
 - 아이솔레이션 포레스트
- 시각화와 차원 축소
 - 주성분 분석(PCA)
 - 커널 PCA
 - 지역적 선형 임베딩
 - t-SNE
- 연관 규칙 학습
 - 어프라이어리(Apriori)
 - 이클랫(Eclat)

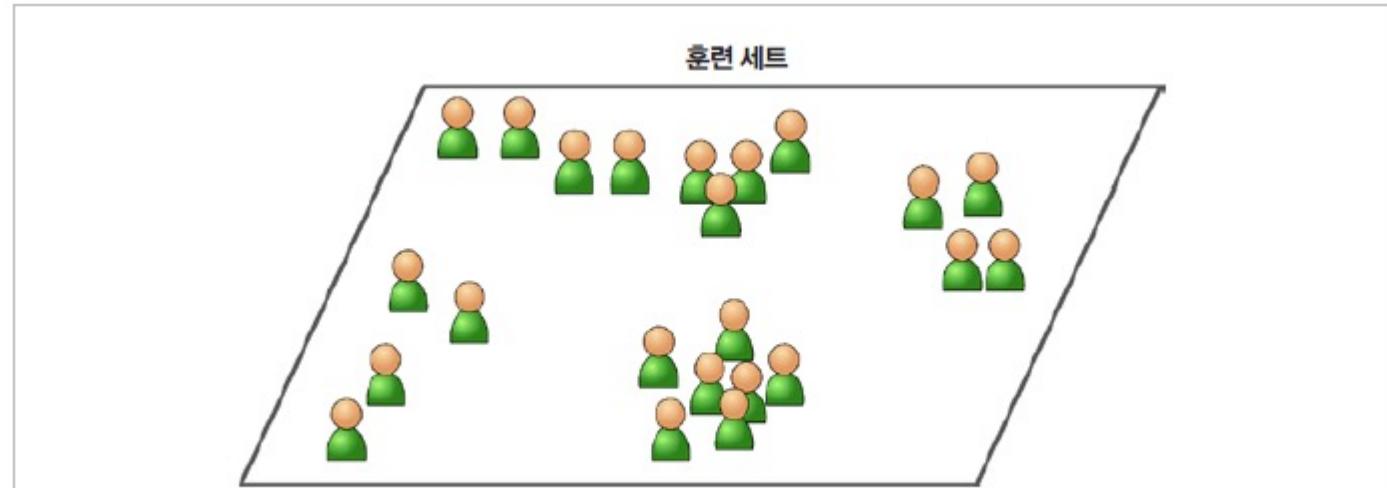


그림 1-7 비지도 학습에서 레이블 없는 훈련 세트

07. 머신러닝 시스템의 학습 방법

배치 학습과 온라인 학습

- **배치 학습:** 시스템이 점진적으로 학습할 수 없다.
- **온라인 학습:** 데이터를 순차적으로 한 개씩 또는 미니배치(mini-batch)라 부르는 작은 묶음 단위로 주입하여 시스템을 훈련한다.

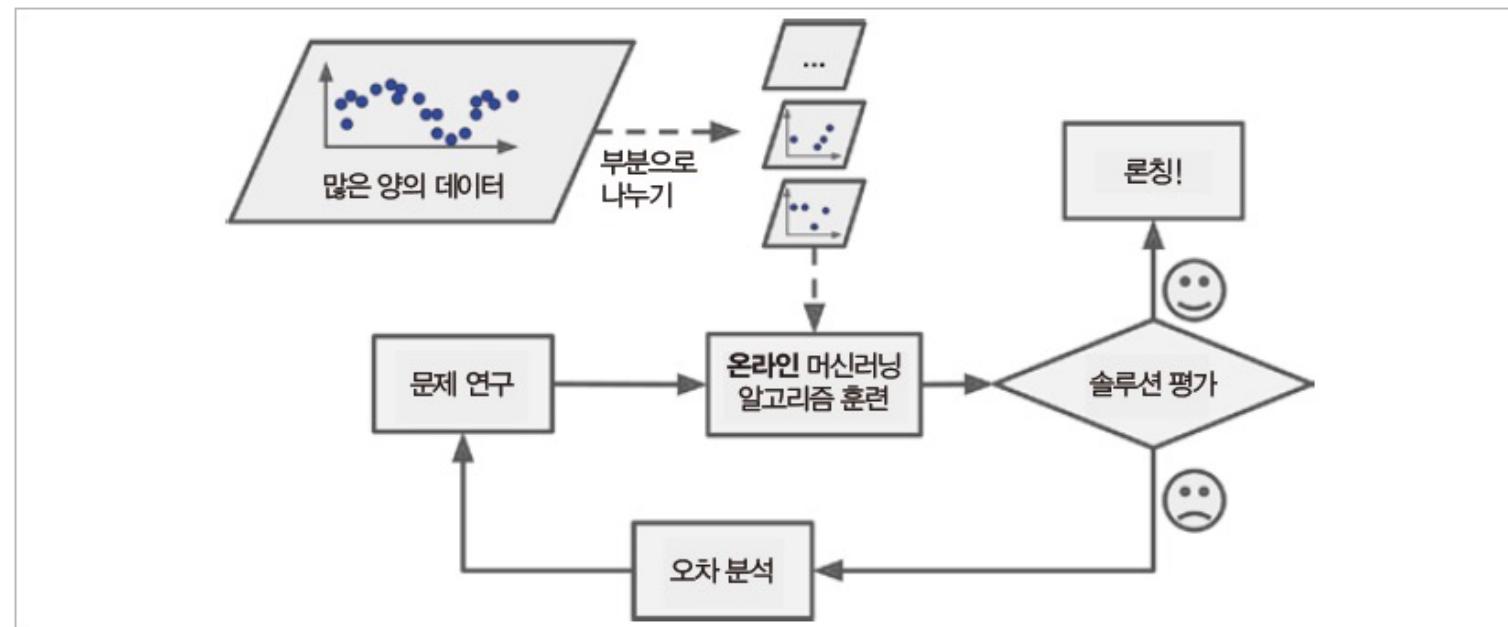


그림 1-14 온라인 학습을 사용한 대량의 데이터 처리

07. 머신러닝 시스템의 학습 방법

사례 기반 학습과 모델 기반 학습



그림 1-15 사례 기반 학습

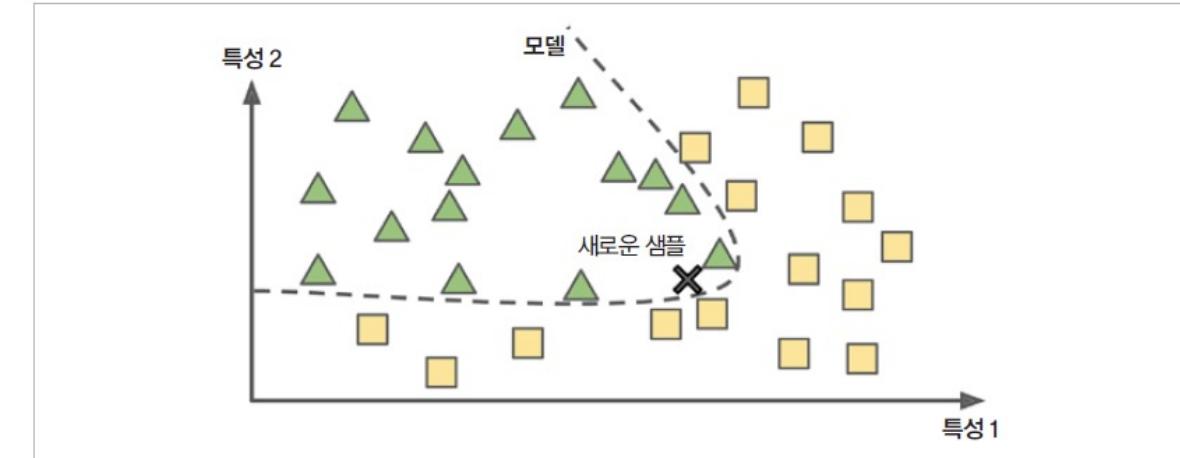
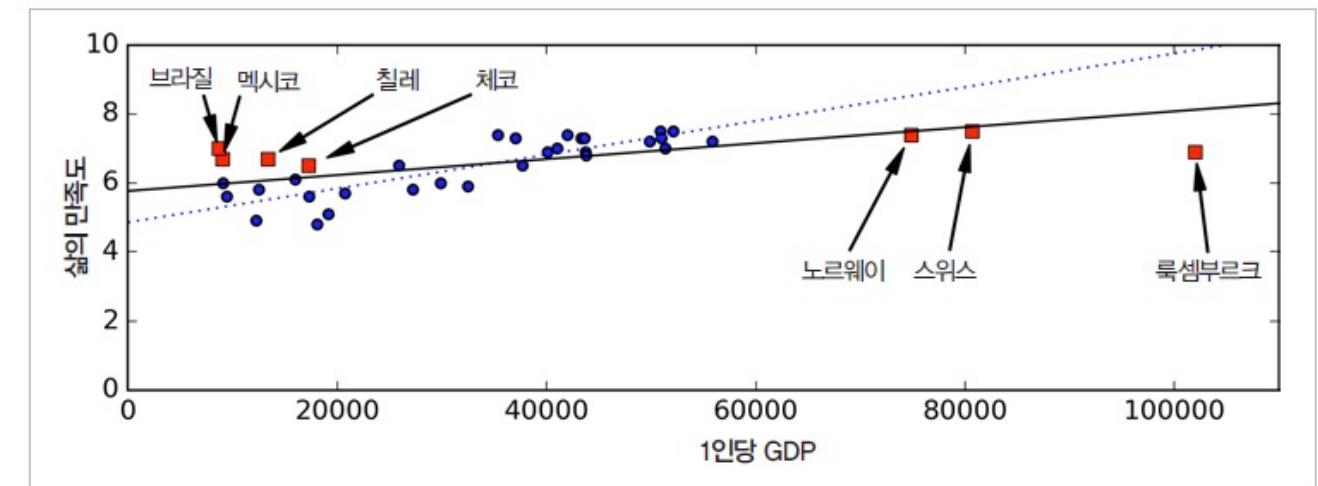


그림 1-16 모델 기반 학습

08. 머신러닝의 주요 도전 과제

- 머신러닝의 주요 작업은 학습 알고리즘을 선택해서 어떤 데이터에 훈련시키는 것
 - '나쁜 데이터'
 - 충분하지 않은 양의 훈련 데이터
 - 대표성 없는 훈련 데이터
 - 낮은 품질의 데이터
 - 관련 없는 특성
 - '나쁜 알고리즘'
 - 훈련 데이터 과대적합
 - 훈련 데이터 과소적합



▲ 대표성이 더 큰 훈련 샘플

09. 머신러닝 시스템의 테스트와 검증

- 훈련 세트와 테스트 세트 두 개로 나누어 검증
 - 데이터의 80%를 훈련에 20%는 테스트용으로 분리하며, 데이터셋 크기에 따라 비율이 다름
 - 훈련 세트를 사용해 모델을 훈련하고 테스트 세트를 사용해 모델을 테스트
 - 새로운 샘플에 대한 오류 비율: 일반화 오차 또는 외부 샘플 오차
 - 테스트 세트에서 모델을 평가함으로써 이 오차에 대한 추정값으로, 이전에 본 적이 없는 새로운 샘플에 모델이 얼마나 잘 작동할지 예측
 - 훈련 오차가 낮지만(즉, 훈련 세트에서 모델의 오차가 적음) 일반화 오차가 높다면 이는 모델이 훈련 데이터에 과대적합되었다는 뜻
 - 하이퍼파라미터 튜닝과 모델 선택
 - 홀드아웃 검증(holdout validation): 간단하게 훈련 세트의 일부를 떼어내어 여러 후보 모델을 평가하고 가장 좋은 하나를 선택
 - 검증 세트가 작을 경우, 반복적으로 교차 검증 수행
 - 데이터 불일치
 - NFL(No Free Lunch): 데이터에 관해 완벽하게 어떤 가정도 하지 않으면 한 모델을 다른 모델보다 선호할 근거가 없음 – 데이비드 월퍼트, 1996

감사합니다.