INDUSTRIAL AI RESEARCH CENTER

# 어프렌티스 프로젝트

## 1주차 - 오리엔테이션

충북대학교 산업인공지능연구센터 김 재영

Dept. of Industrial AI

# 1. 소개

INDUSTRIAL AI RESEARCH CENTER

## *강사*

- 김 재영 산업인공지능연구센터 초빙교수
  - jaykim@cbnu.ac.kr
  - ☎ 043-249-1464
  - 충북대학교 오창캠퍼스 융합기술원 C655호

## *수강 학생*

- 충북대학교 산업인공지능학과
  - 석사과정 25명

# 2. 교과목 소개

INDUSTRIAL AI RESEARCH CENTER

## *교과 개요*

| | | | | | | |
|---|---|---|---|---|---|---|
| **강의개요** | 본 교과목에서는 산업현장에 직접 적용 가능한 지능화 기술을 습득하기 위한 기초 지식을 학습하고, 머신러닝 개요, 절차 및 알고리즘을 학습하며, 프로젝트 수행을 통해 실제로 적용해 본다. | | | | | |
| **학습목표** | - 머신러닝을 위한 기초 수학 및 통계에 대한 이해<br>- 머신러닝 개요 및 절차에 대한 이해<br>- 데이터 분석 및 머신러닝 알고리즘의 이해<br>- 실제 머신러닝 프로젝트 수행 능력 함양 | | | | | |
| **문제해결방법** | 강의와 실험/실습 병행 | | | | | |
| **수업진행방법** | **강의** | **토의/토론** | **실험/실습** | **현장학습** | **개별/팀별 발표** | **기타** |
| | 60% | 0% | 20% | 0% | 20% | 0% |
| | **상세정보** | 대면 수업과 비대면 수업(ZOOM을 통한 실시간 온라인 강의 또는 동영상 강의) 병행 | | | | |
| **평가방법** | **중간고사** | **기말고사** | **출석** | **퀴즈** | **과제** | **기타** |
| | 0% | 0% | 10% | 0% | 90% | 0% |
| | **상세정보** | 중간 프로젝트 30%, 기말 프로젝트 30%, Homework 30%, 출석 10% | | | | |
| **프로그램 학습성과 평가** | 실습을 통한 과제 제출과 중간/기말 프로젝트를 통해 머신러닝에 대한 이해도와 실용적 구현 능력을 평가 | | | | | |
| **교재 및 참고문헌** | 1. 주교재 : Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 오렐리앙 제롬 지음, 박해선 옮김 한빛미디어, 2018<br>2. 부교재 : 비즈니스 애널리틱스를 위한 데이터마이닝 in 파이선 Galit Shmueli 저, 조성준 등 옮김 한빛아카데미, 2023<br>3. 부교재 : 데이터 과학을 위한 기초수학 with 파이썬, 이병준, 한빛아카데미, 2021 | | | | | |

# 2. 교과목 소개

INDUSTRIAL AI RESEARCH CENTER

## *강의 일정*

| 어프렌티스 프로젝트 | | | | | |
|---|---|---|---|---|---|
| **담당교수** | **김재영** | **수강대상** | **2023학번(00명)** | | |
| **주차** | **날짜** | **강의실** | **강의내용** | **과제물** | **평가** |
| 1 | 09/04(월) | | 오리엔테이션 | | |
| 2 | 09/11(월) | A704 | 데이터 처리 기초 – Array, Dataframe | | |
| 3 | 09/18(월) | | 기초 수학 - 함수, 통계 확률분포, 가설 검증 | | |
| 4 | 09/25(월) | | 머신러닝(1) - 환경구성, 데이터 구조 조사, 테스트 세트 | 과제(1) | |
| 5 | 10/02(월) | 보강(?) | 특강(1) | | |
| 6 | 10/09(월) | 보강(?10/13) | 과제(1) 발표 | | 과제 평가(10%) |
| 7 | 10/16(월) | | 특강(2) | | |
| 8 | 10/23(월) | A704 | 머신러닝(2) - 데이터 탐색(시각화) | | |
| 9 | 10/30(월) | A704 | 중간발표 | | 평가(30%) |
| 10 | 11/06(월) | | 머신러닝 실습(3) - 예측성능 평가 | 과제(2) | |
| 11 | 11/13(월) | | 머신러닝 실습(4) - 회귀분석 | | 과제 평가(10%) |

# 2. 교과목 소개

INDUSTRIAL AI RESEARCH CENTER

## *출석*

- 5회 이상 결석 – D(평점 0)
- 출석 체크 : 수업시작시 호명

## *어프렌티스 프로젝트 과목 평가*

- 출석 10%, 과제 30%, 중간 평가 30%, 기말 평가 30%

## *프로젝트 수행*

- 중간 프로젝트 – 개별 수행
- 기말 프로젝트 – 팀 수행

# 3. 인공지능의 개요 및 발전 과정

INDUSTRIAL AI RESEARCH CENTER

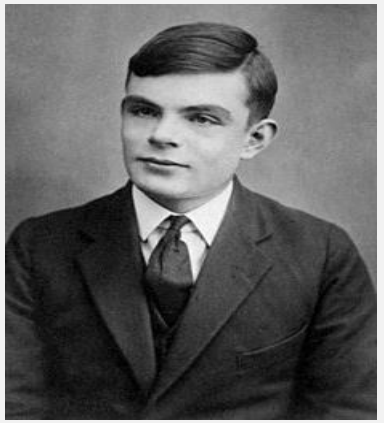
## *인공지능이란*

- 기계 또는 컴퓨터 시스템이 인간과 유사한 지능적인 작업을 수행하도록 만들기 위하여 개발된 기술 및 능력을 의미

ChatGPT

AlphaGo

## *앨런 튜링(Alan Mathison Turing)*



- 튜링 테스트(Turing Test)
- 계산이론(Computability Theory)
- 튜링 완전성(Turing Completeness)
- 암호학(Cryptography)

# 3. 인공지능의 개요 및 발전 과정

INDUSTRIAL AI RESEARCH CENTER

**1950년대**

앨런 튜링에 의한 컴퓨터 과학의 발전과 인공지능 개념의 태동.

다트머스 컨퍼런스에서 인공지능의 개념이 처음으로 정의됨.

**1960년대**

인공지능 연구에 대한 긍정적인 기대와 함께 기호주의 인공지능이 발전.

엘리자와 같은 인공지능 프로그램이 개발됨.

**1970년대**

전문가 시스템의 등장으로 전문적인 지식을 활용한 문제 해결을 시도.

MYCIN, DENDRAL 등의 전문가 시스템이 개발되었음.

**1980년대**

연결주의 인공지능과 지식기반 시스템의 발전.

인공신경망, 전문가 시스템, 지식 표현 방법론 등의 연구 활발.

**1990년대**

인공지능 기술이 상용화되기 시작.

머신 러닝, 데이터 마이닝, 컴퓨터 비전 등의 분야에서 발전.

**2000년대 이후**

딥러닝 알고리즘과 대용량 데이터를 통한 머신 러닝의 발전.

음성 인식, 이미지 인식, 자율 주행 등 다양한 분야에서 인공지능의 혁신적인 성과.

**2011년**

IBM의 인공지능 시스템 왓슨이 퀴즈 프로그램 '제퍼디'에서 인간 퀴즈 챔피언을 이김.

**2016년**

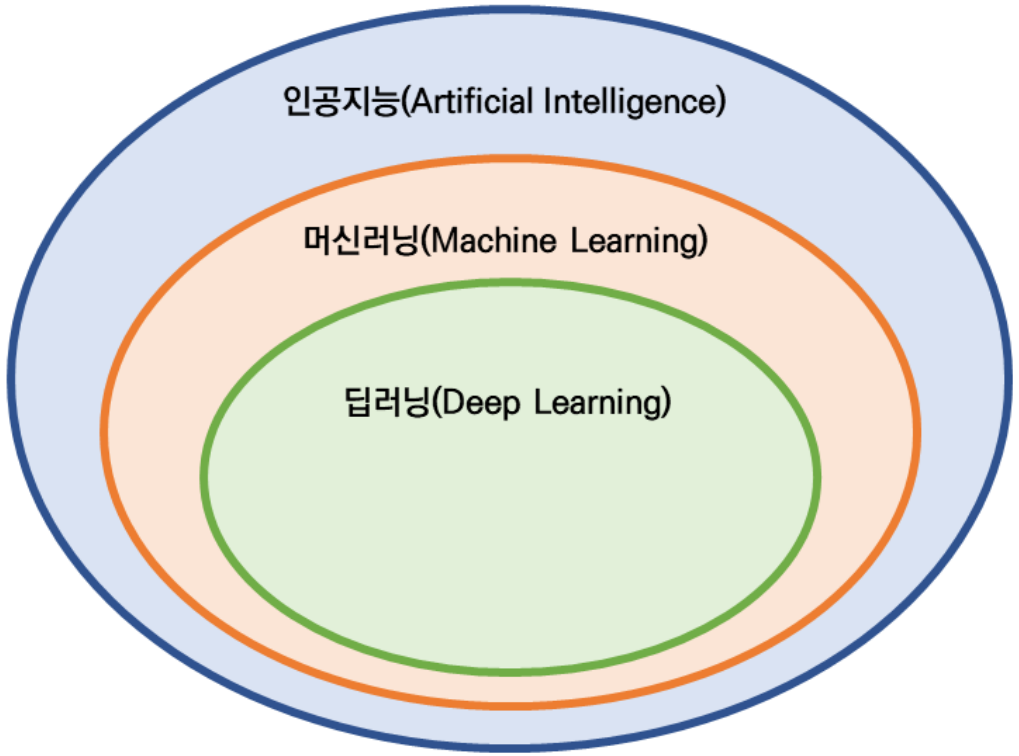
알파고, 구글 딥마인드의 인공지능 프로그램이 바둑 세계 챔피언을 이김.

**2023년 3월**

OpenAI의 최신 언어모델인 GPT-4가 출시

# 4. 머신러닝




인공지능(Artificial Intelligence)
머신러닝(Machine Learning)
딥러닝(Deep Learning)

# 4. 머신러닝

INDUSTRIAL AI RESEARCH CENTER

## *머신러닝이란*

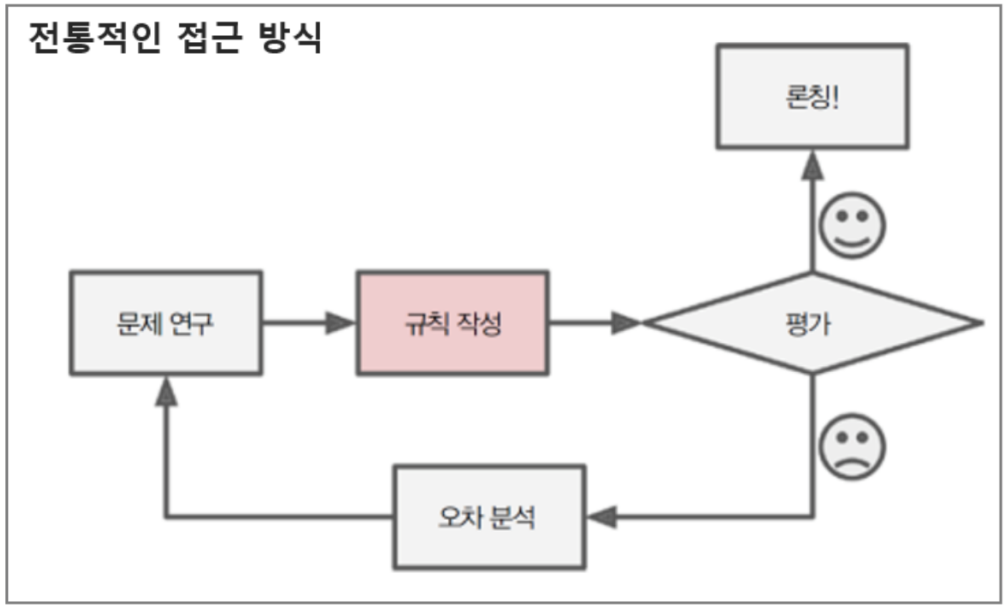
- 머신러닝은 데이터에서부터 학습하도록 컴퓨터를 프로그래밍하는 과학(또는 예술)
  - "머신러닝은 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야"
    -아서 새뮤얼(Arthur Samuel), 1959
  - "어떤 작업 T에 대한 컴퓨터 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상됐다면, 이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로 학습한 것"
    -톰 미첼(Tom Mitchell), 1997
- 스팸 필터는 (사용자가 스팸이라고 지정한) 스팸 메일과 일반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

# 4. 머신러닝

INDUSTRIAL AI RESEARCH CENTER

*왜 머신러닝을 사용하는가?*

- 전통적 프로그래밍 기법으로는 규칙이 점점 길고 복잡해지므로 유지 보수하기 매우 힘듦
  1. 스팸에 어떤 단어들이 주로 나타나는지 조사
  2. 발견한 각 패턴을 감지하는 알고리즘(규칙)을 작성
  3. 프로그램을 테스트하고 충분한 성능이 나올 때까지 "1", "2"를 반본

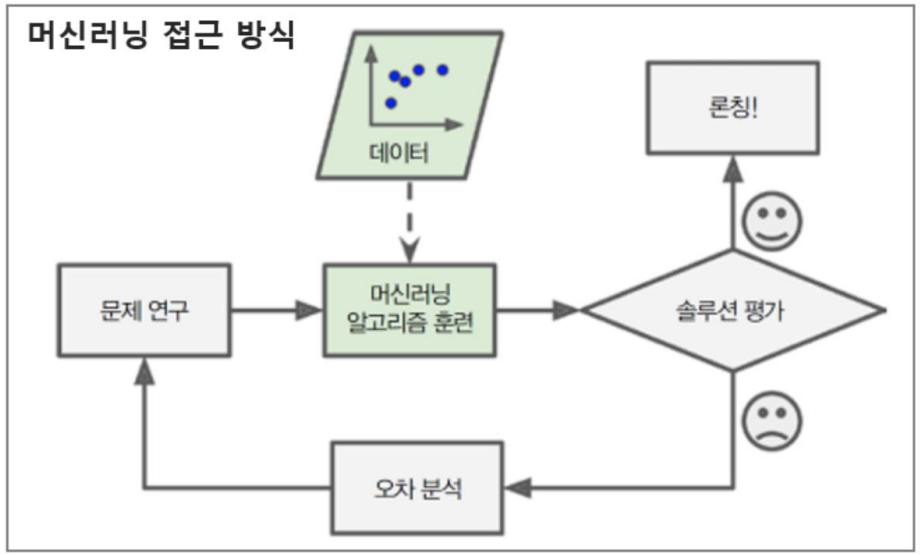


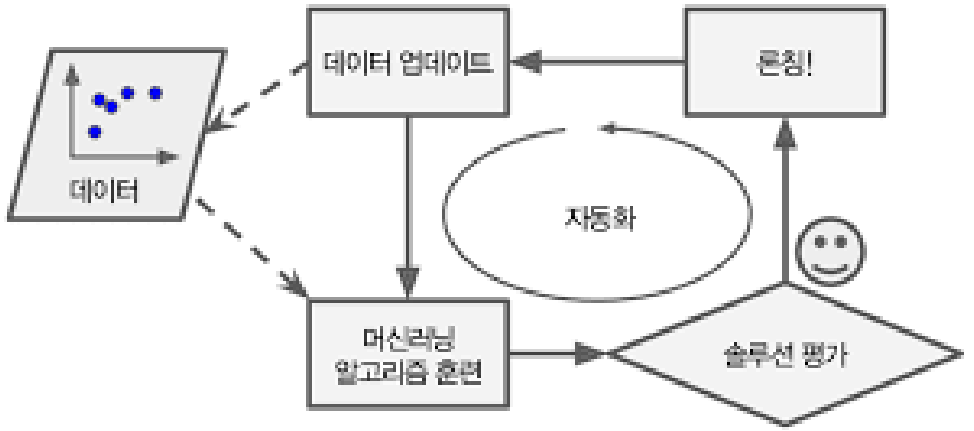
# 4. 머신러닝

INDUSTRIAL AI RESEARCH CENTER

## *왜 머신러닝을 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 없거나 복잡한 문제(음성인식): 가장 뛰어난 머신러닝 기법으로 해결 방법을 찾을 수 있음
  - 유동적인 환경: 머신러닝 시스템은 새로운 데이터에 적응 가능
  - 복잡한 문제와 대량의 데이터에서 통찰 얻기



# 4. 머신러닝

INDUSTRIAL AI RESEARCH CENTER

## *머신러닝 애플리케이션 사례*

- 생산 라인에서 제품 이미지를 분석해 자동으로 분류하기 : CNN
- 뇌를 스캔하여 종양 진단하기 : CNN
- 자동으로 뉴스 기사를 분석 : NLP, RNN, CNN, 트랜스포머
- 토론 포럼에서 부정적인 코멘트 자동으로 구분하기 : NLP
- 긴 문서를 자동으로 요약하기 : NLP
- 챗봇 또는 개인 비서 만들기 : NLU, NLP
- 다양한 선능 지표를 기반으로 회사의 내년도 수익을 예측하기 : 회귀, 인공 신경망, RNN, CNN
- 음성 명령에 반응하는 앱 만들기 : RNN, CNN, 트랜스포머
- 신용 카드 부정 거래 감지하기 : 이상치 탐지 작업
- 구매 이력을 기반으로 고객을 나누고 다른 마케팅 전략을 계획 : 군집 작업
- 고차원의 복잡한 데이터셋을 명확하고 의미 있는 그래프로 표현하기 : 차원 축소

# 5. 머신러닝 시스템의 종류

INDUSTRIAL AI RESEARCH CENTER

## *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

# 5. 머신러닝 시스템의 종류

INDUSTRIAL AI RESEARCH CENTER

*지도 학습과 비지도 학습*

## 지도 학습

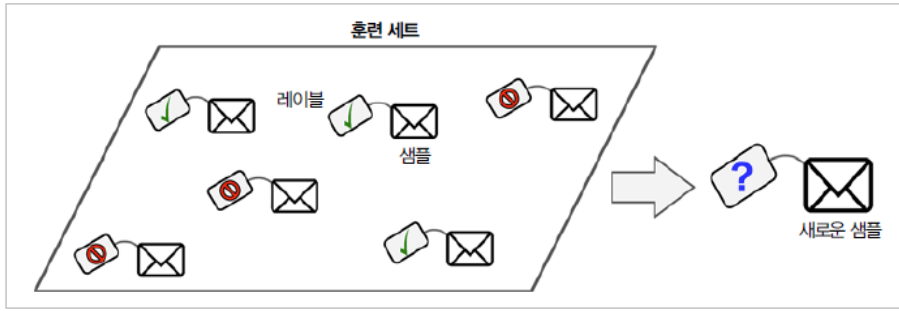
알고리즘에 주입하는 훈련 데이터에 레이블(label) 이라는 원하는 답이 포함된다.

- 분류
- 회귀

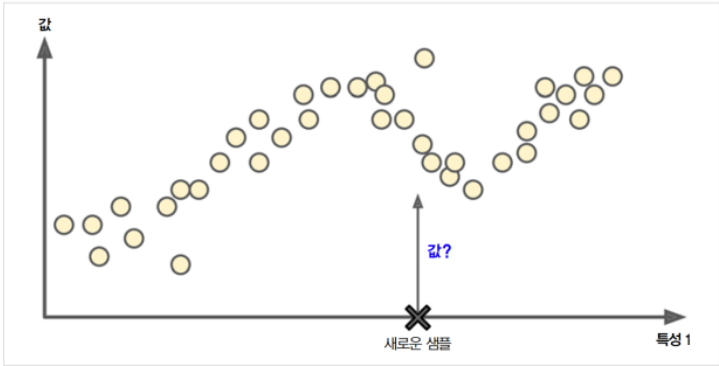
## 중요한 지도학습 알고리즘들

알고리즘에 주입하는 훈련 데이터에 레이블(label) 이라는 원하는 답이 포함된다.

- k-최근접 이웃
- 선형 회귀
- 로지스틱 회귀
- 서포트 벡터 머신
- 결정 트리와 랜덤 포레스트
- 신경망




스팸 분류를 위한 레이블된 훈련 세트




회귀 문제: 주어진 입력 특성으로 값을 예측

# 5. 머신러닝 시스템의 종류

INDUSTRIAL AI RESEARCH CENTER

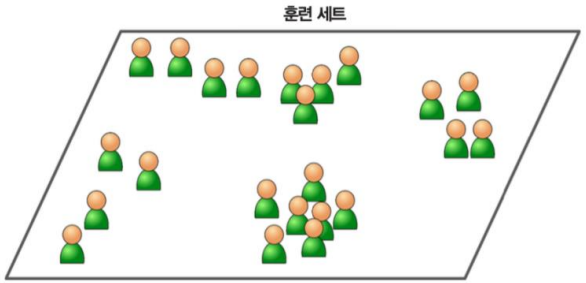
*지도 학습과 비지도 학습*

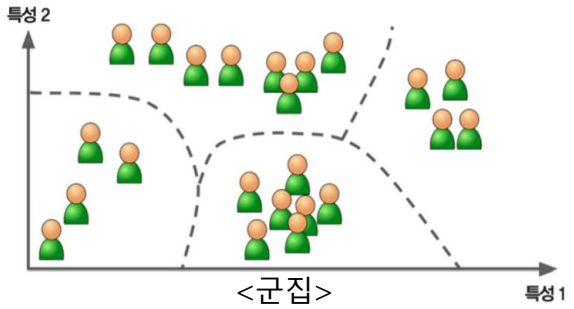
## 비지도 학습

훈련 데이터에 레이블이 없어서, 시스템이 아무런 도움 없이 학습해야 한다.

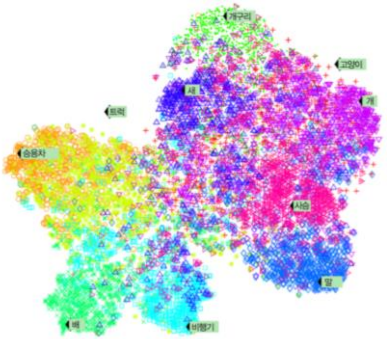
## 중요한 비지도학습 알고리즘들

- 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
- 이상치 탐지(outlier detection)
- 특이치 탐지(novelty detection)
- 연관 규칙 학습
  - 어프라이어리(Apriori)
  - 이클렛(Eclat)

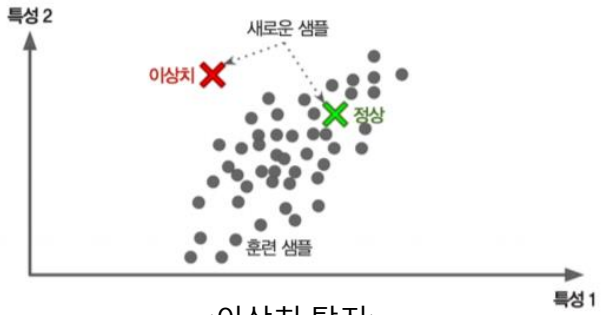



<군집>



<t-SNE 시각화>




<이상치 탐지>

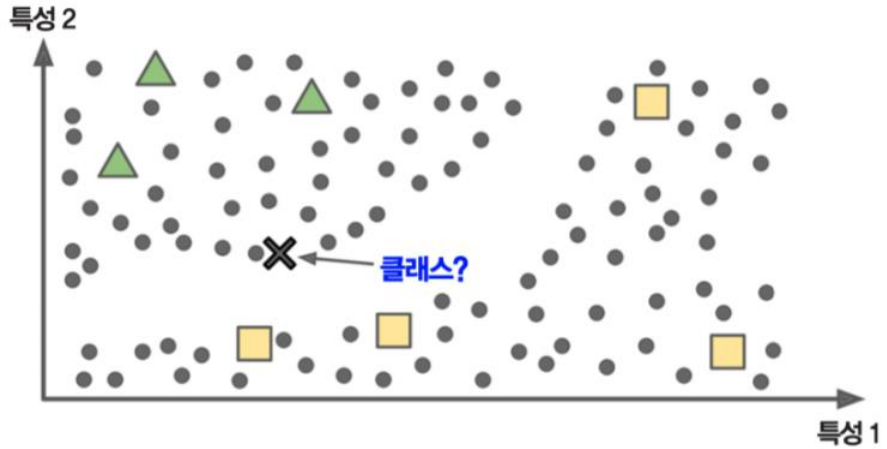
# 5. 머신러닝 시스템의 종류



*지도 학습과 비지도 학습*

## 준지도 학습

- 알고리즘에 주입하는 훈련 데이터 중 일부만 레이블이 있는 경우
- 지도 학습과 비지도 학습의 중간 형태로, 레이블이 부분적으로만 제공되는 학습 방법
  - 레이블 전파(label propagation)




<두 개의 클래스(삼각형과 사각형)을 사용한 준지도 학습>

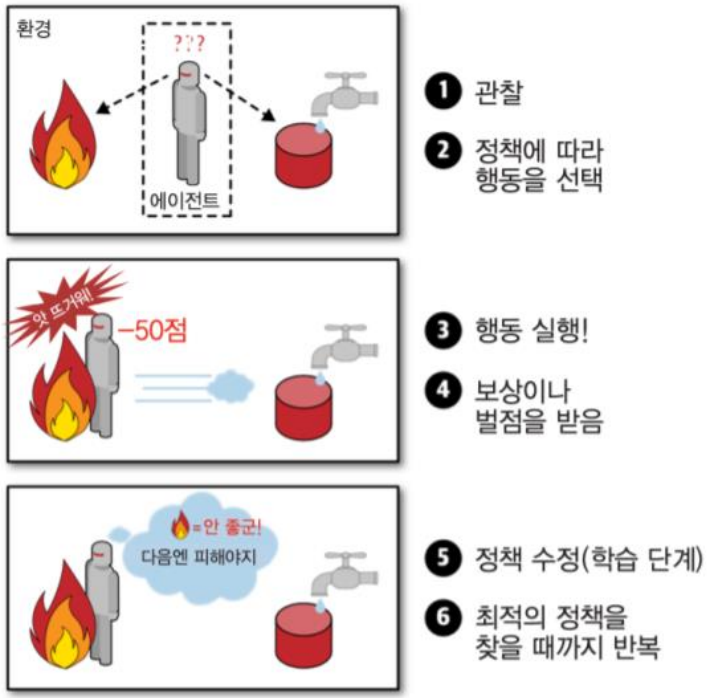
# 5. 머신러닝 시스템의 종류



*지도 학습과 비지도 학습*

## 강화 학습

- 주어진 환경에서 보상이 최대가 되는 행동을 선택 (= 정책) 하도록 에이전트를 훈련하는 것



# 5. 머신러닝 시스템의 종류

INDUSTRIAL AI RESEARCH CENTER

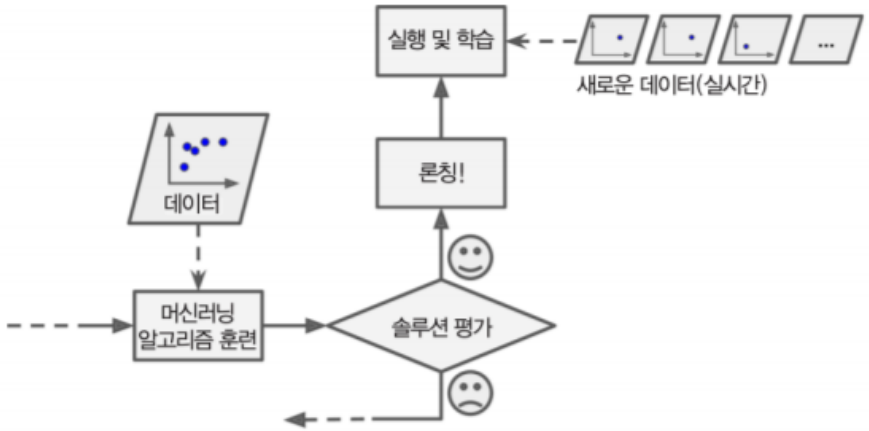
*배치 학습과 온라인 학습*

## 배치 학습

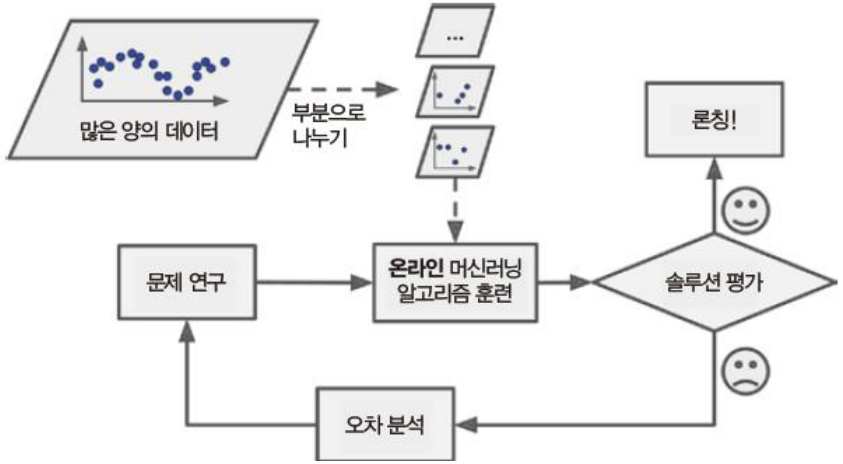
- 주어진 학습데이터를 모두 사용해서 학습
- 시스템이 점진적으로 학습할 수 없음
- 가용한 데이터를 모두 사용하여 훈련시켜야 함
- 학습한 것을 단지 적용만 하는 것
- 오프라인 학습이라고도 칭함

## 온라인 학습

- 데이터를 순차적으로 한 개 또는 미니배치(mini-batch)를 이용해서 학습하는 방법
- 연속적으로 데이터를 받고, 빠른 변화에 스스로 적응해야 하는 시스템에 적합 (예 - 주식 가격)
- 온라인 학습에 가장 중요한 파라미터 하나는 learning rate
- learning rate는 변화하는 데이터에 얼마나 빠르게 적응할 것인지를 결정하는 파라미터




<온라인 학습>




<외부 메모리 학습>

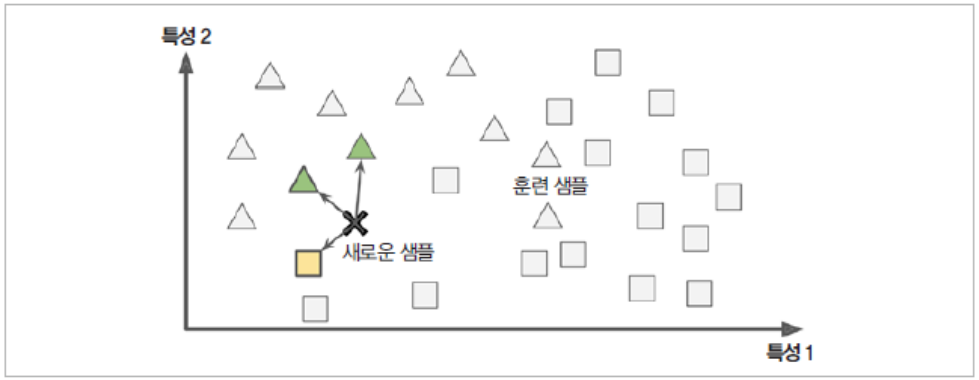
# 5. 머신러닝 시스템의 종류



*사례 기반 학습과 모델 기반 학습*

## 사례 기반 학습 (Case-Based Learning)

- 시스템이 훈련 샘플을 기억함으로써 학습, 즉, 학습은 샘플을 기억하는 과정이라 할 수 있음
- 유사도 측정을 사용하여 새로운 데이터와 학습한 데이터를 비교하는 방식으로 일반화
- 최근접 이웃(K-Nearest Neighbors, KNN) 알고리즘
  - 새로운 데이터 포인트를 기존 데이터 포인트와 비교하여 가장 유사한 데이터를 찾고, 해당 데이터의 레이블을 사용하여 예측을 수행



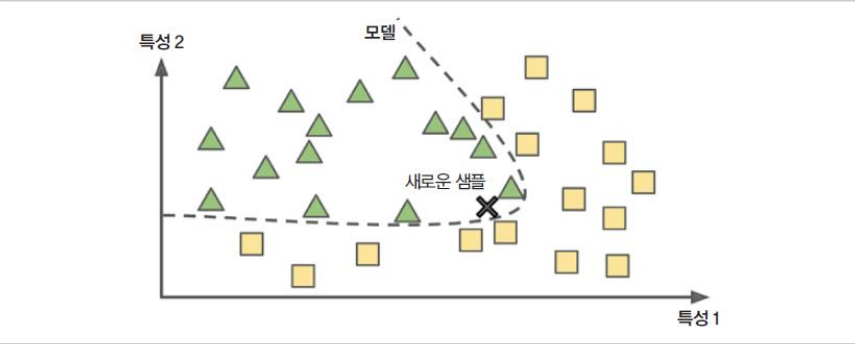
# 5. 머신러닝 시스템의 종류

INDUSTRIAL AI RESEARCH CENTER

## *사례 기반 학습과 모델 기반 학습*

### 모델 기반 학습 (Model-Based Learning)

- 주어진 데이터로부터 일반화된 모델을 생성하여 새로운 데이터에 대한 예측을 수행하는 접근 방식으로 "규칙을 배우는" 것과 유사
- 데이터의 패턴을 파악하여 모델을 구축하고, 이 모델을 사용하여 예측을 수행
- 대표 알고리즘
  - 선형 회귀(Linear Regression)
  - 의사결정 트리(Decision Tree)



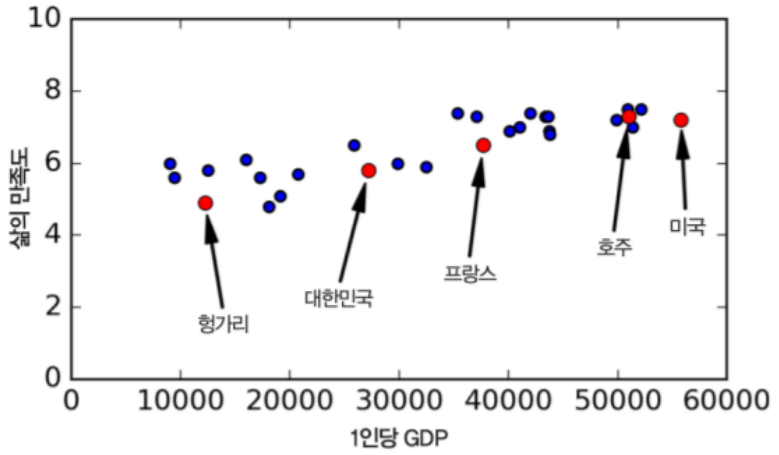
# 5. 머신러닝 시스템의 종류

INDUSTRIAL AI RESEARCH CENTER

## *모델 기반 학습 사례*

◦ 더 나은 삶의 지표 데이터 + 1인당 GDP

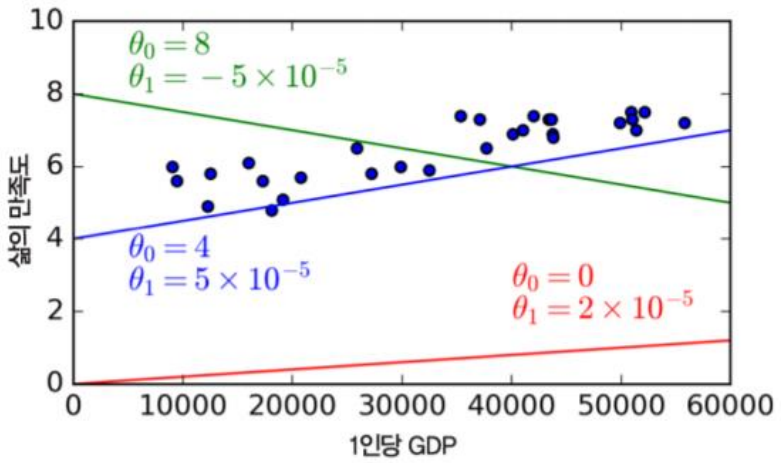
| 국가 | 1인당 GDP(미국 달러) | 삶의 만족도 |
|---|---|---|
| 헝가리 | 12,240 | 4.9 |
| 한국 | 27,195 | 5.8 |
| 프랑스 | 37,675 | 6.5 |
| 호주 | 50,9672 | 7.3 |
| 미국 | 55,805 | 7.2 |




◦ 간단한 선형모델

'삶의 만족도' = $\theta_0 + \theta_1 \times$ 1인당 GDP'

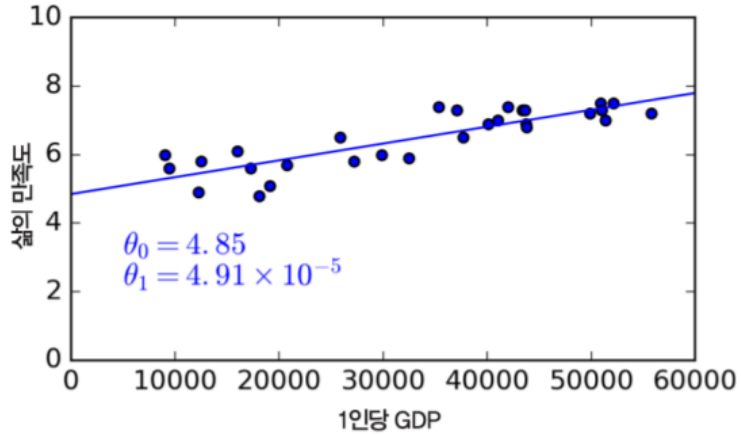
- 1인당 GDP 특성의 선형 모델
- 두 개의 모델 파라미터 $\theta_0$와 $\theta_1$




◦ 모델을 사용하기 위해 두 개의 모델 파라미터 $\theta_0$와 $\theta_1$ 정의 필요

◦ 측정지표

- 효용 함수(utility function)/적합도 함수(fitness function)
- 비용 함수(cost function)



# 5. 머신러닝 시스템의 종류



## *사이킷런을 이용한 선형 모델의 훈련과 실행*

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

# 데이터 적재
data_root = "https://github.com/ageron/data/raw/main/"
lifesat = pd.read_csv(data_root + "lifesat/lifesat.csv")
X = lifesat[["GDP per capita (USD)"]].values
y = lifesat[["Life satisfaction"]].values

# 데이터 시각화
lifesat.plot(kind='scatter', grid=True,
x="GDP per capita (USD)", y="Life satisfaction")
plt.axis([23_500, 62_500, 4, 9])
plt.show()

# 선형 모델 선택
model = LinearRegression()

# 모델 훈련
model.fit(X, y)

# 예측
X_new = [[37_655.2]] # 키프로스 1인당 GDP
print(model.predict(X_new)) # 결과: [[6.30165767]]
```

## 작업 요약

- 데이터 분석
- 모델을 선택
- 훈련데이터로 모델을 훈련
- 새로운 데이터에 모델을 적용해 예측

## 머신러닝 프로젝트 진행 단계

1. 목적 결정
2. 데이터 수집
3. 데이터 탐색 및 시각화
4. 머신러닝 알고리즘을 위해 데이터 준비
5. 모델을 선택하고 훈련
6. 모델의 상세 조정
7. 솔루션 제시
8. 시스템 론칭 및 모니터링

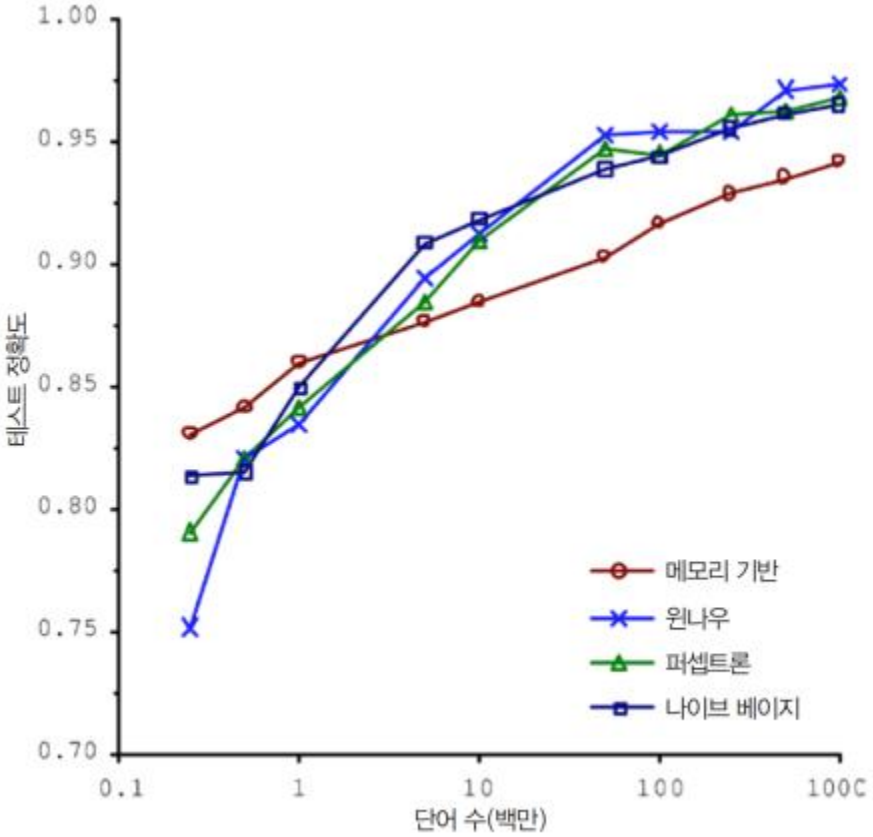
# 6. 머신러닝의 주요 도전 과제

INDUSTRIAL AI RESEARCH CENTER

*나쁜 데이터*

**충분하지 않은 양의 훈련 데이터**

- 데이터의 추가 수집과 알고리즘 개발 사이에서의 트레이드 오프.
  - 대부분의 머신러닝 알고리즘은 잘 작동하기 위해 충분히 많은 양의 데이터를 필요로 한다.
  - 특히, 복잡한 자연어 중의성 해소 논문의 저자들이 밝히듯, 데이터가 충분히 많아지면 여러 머신러닝 알고리즘은 비슷한 성과를 낸다..




알고리즘 대비 데이터의 중요성

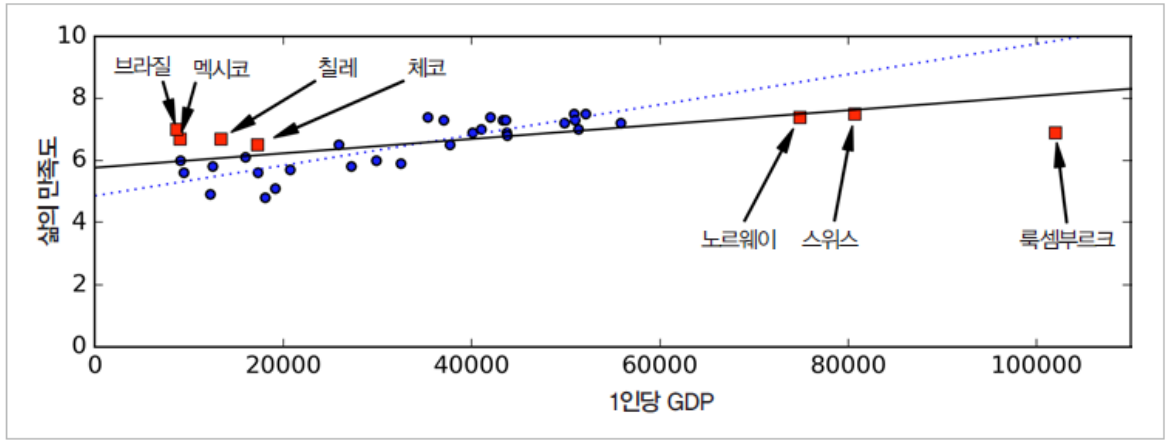
# 6. 머신러닝의 주요 도전 과제

INDUSTRIAL AI RESEARCH CENTER

*나쁜 데이터*

**대표성 없는 훈련 데이터**

- 훈련 데이터가 일반화하기를 원하는 새로운 사례를 잘 대표하지 못한다면, 일반화가 되지 않는다.
- 고려해야 할 주요 문제점
  - 대표성 없는 데이터의 문제로 인해 샘플을 늘릴 때, 새로운 샘플이 이전의 샘플이나, 혹은 다른 새로운 데이터들을 대표하지 못한다면?
  - 샘플이 작아도, 커도 대표성 문제가 생길 수 있다. 샘플이 너무 작으면 샘플링 잡음(우연에 의한 데이터 편향)이, 샘플이 너무 크면 표본 추출 방법에 따라 샘플링 편향이 생길 수 있다.




대표성이 더 큰 훈련 샘플

Dept. of Industrial AI

# 6. 머신러닝의 주요 도전 과제

INDUSTRIAL AI RESEARCH CENTER

## *나쁜 데이터*

### 낮은 품질의 데이터

- 이상치인 경우엔 해당 데이터를 수정하거나 무시힌다.
- 나이, 성별 등 데이터의 특성 중 일부가 누락된 경우 아래 방식에서 선택한다.
  - 해당 특성 제외
  - 해당 데이터 제외
  - 누락된 특성값을 평균값 등으로 채우기

### 관련 없는 특성

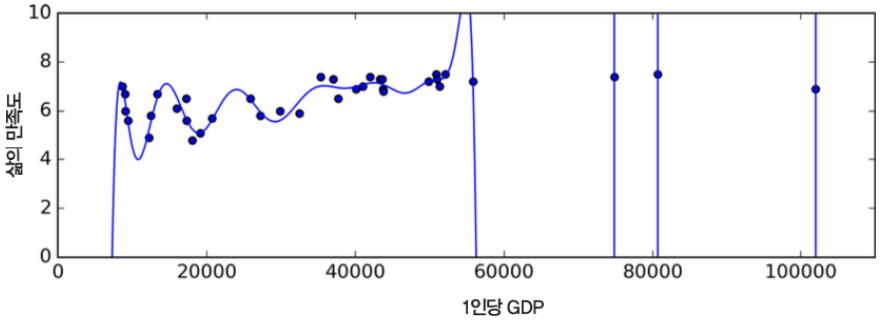
- 특성 선택feature selection: 예측해야 할 값과 가장 연관성이 높은 특성을 선택한다.
- 특성 추출feature extraction: 특성을 조합하여 보다 유용한 새로운 특성을 생성한다.
- 모델 학습에 유용하다고 판단되는 새로운 특성을 추가한다.

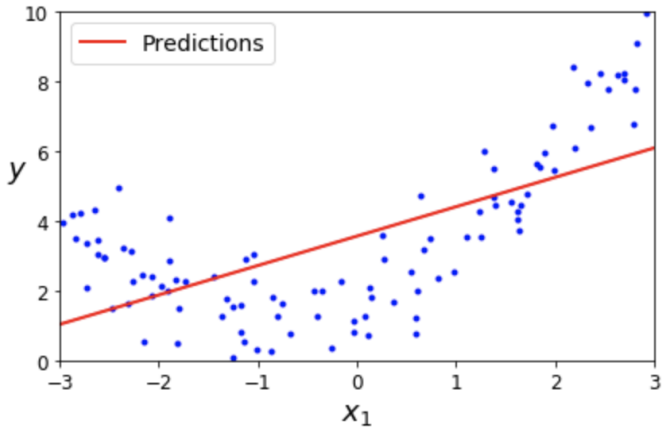
# 6. 머신러닝의 주요 도전 과제

*머신러닝의 주요 작업은 학습 알고리즘을 선택해서 어떤 데이터에 훈련시키는*

## 나쁜 알고리즘

- 과대적합
  - 모델이 훈련 과정에서 훈련셋에 특화되어 일반화 성능이 떨어지는 현상
- 과소적합
  - 모델이 너무 단순해서 훈련셋을 제대로 대변하지 못하는 경우



# 7. 테스트와 검증

INDUSTRIAL AI RESEARCH CENTER

*모델이 새로운 샘플에 얼마나 잘 일반화되는지를 평가하기 위해 테스트을 진행*

- 훈련 데이터 분할 : 일반적으로 데이터의 80%를 훈련에 20%는 테스트용으로 분리하며, 데이터셋 크기에 따라 비율이 다름
  - 훈련 세트를 사용해 모델 훈련
  - 테스트 세트를 사용해 모델 테스트
- 일반화 오차 : 새로운 샘플에 대한 오류 비율

*검증 세트: 모델과 하이퍼파라미터가 테스트 세트에 최적화된 모델을 만드는 것을 방지.*

- 훈련 데이터 분할 : 훈련 세트, 테스트 세트, 검증 세트.
  - 훈련 세트 : 다양한 하이퍼파라미터로 여러 모델 훈련.
  - 검증 세트 : 최상의 성능을 내는 모델 및 하이퍼파라미터 선택.
  - 테스트 세트 : 단 한 번의 최종 테스트.
- 교차 검증 : 훈련 데이터에서 검증 세트로 너무 많은 양을 사용하지 않기 위해, 훈련 세트를 여러 subset으로 나누고, 각 모델을 subset의 조합으로 훈련시킨 뒤 나머지 부분으로 검증.




- 다음주(9/11) 수업은 **Off-Line (노트북 지참)**
- 강의실 **A704(오창 융합기술원)**

김 재영

jaykim@cbnu.ac.kr