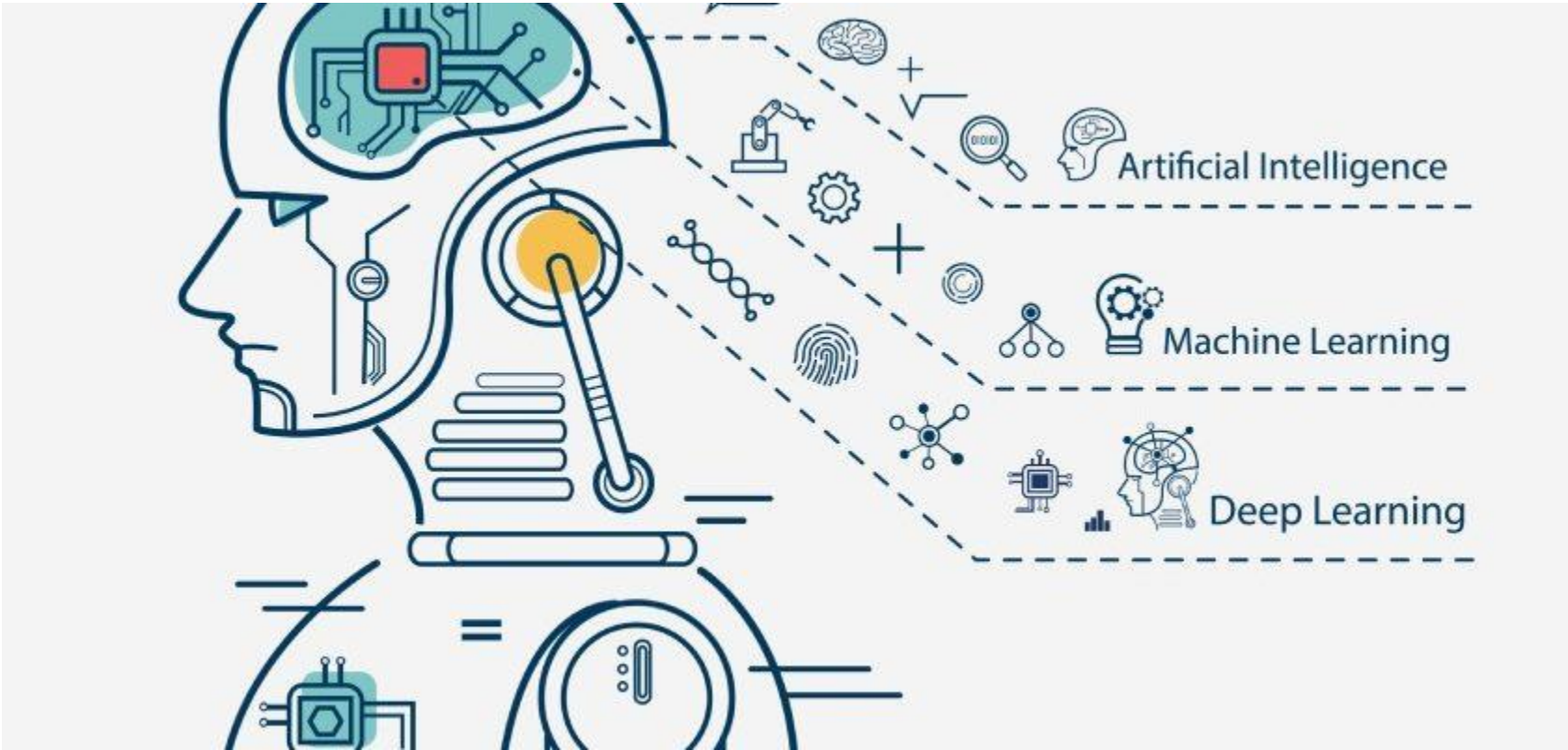
K-NET ML/DL STUDY GROUP MEETING #2

Hyungyu Kim
K-NET
2023-04-14

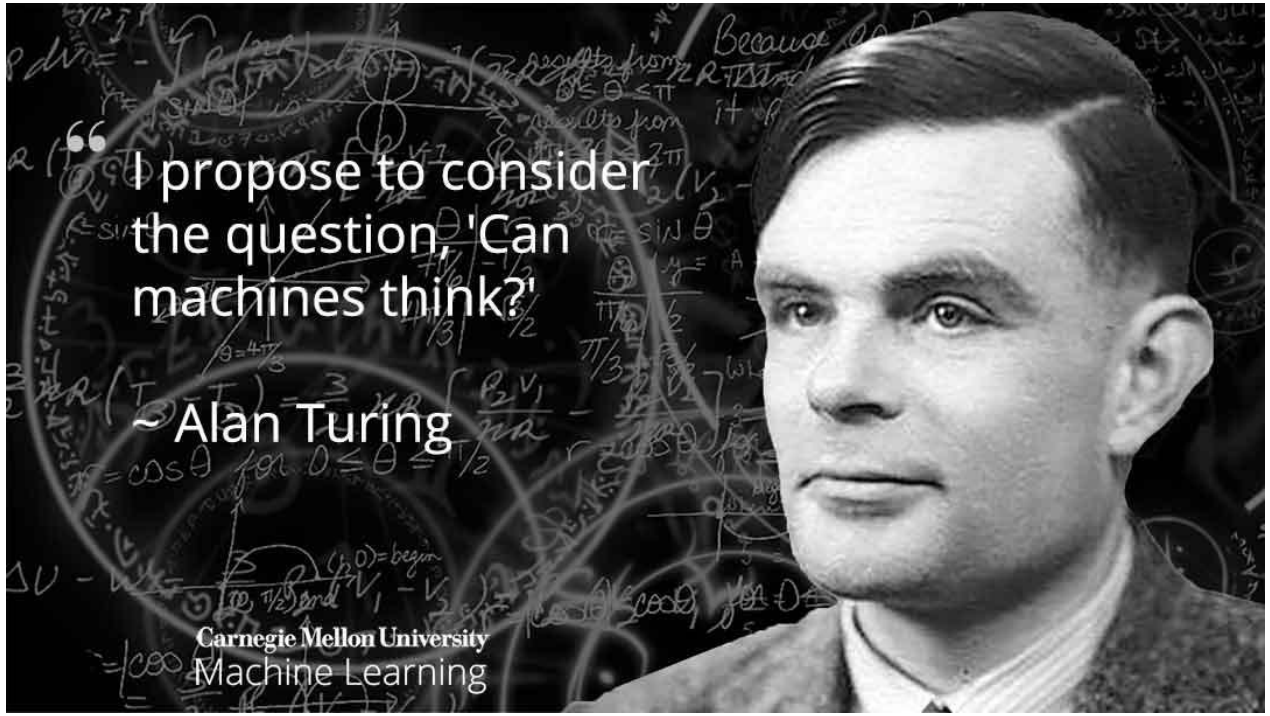
이번 시간에 다룰 주제

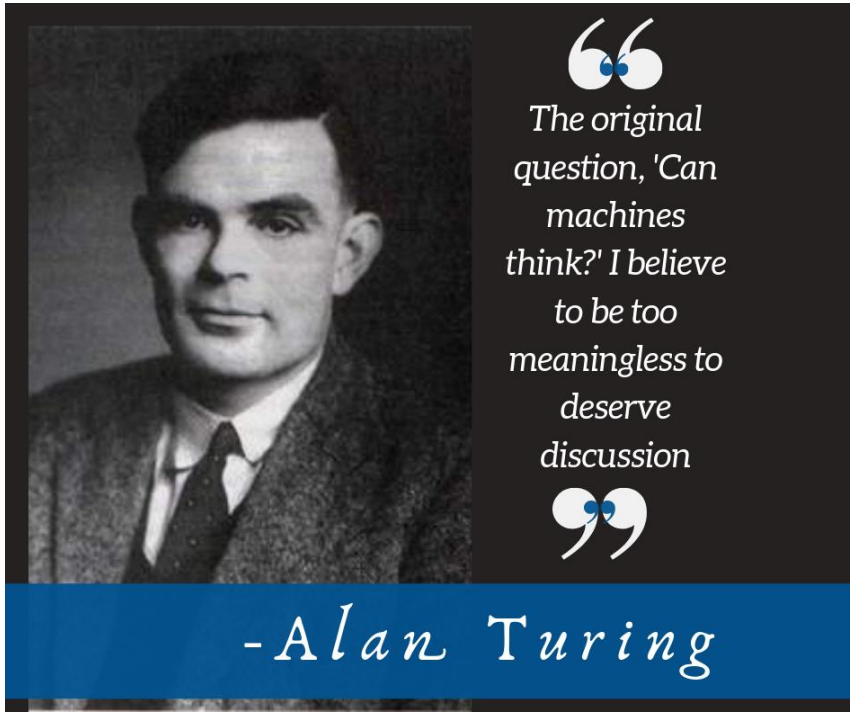
- ML/DL이란 무엇인가?



기계도 생각, 학습을 할 수 있을까?

- 기계는 어떤 방식으로 생각하고 학습할까? 사람과 같은 방식일까?



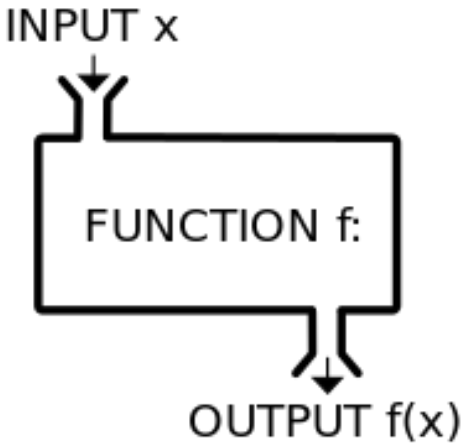


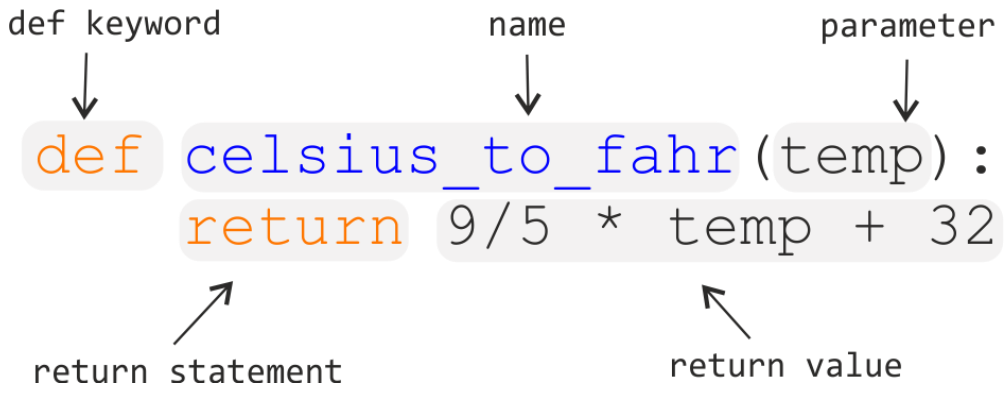
머신러닝이란?

- “명시적으로 프로그램을 작성하지 않고 컴퓨터에 학습할 수 있는 능력을 부여하기 위한 연구 분야”
- 우리가 백준 문제를 풀 때는 모든 테스트 케이스를 만족하도록 규칙을 명시적으로 프로그래밍해주어야 한다.
- 그런데 우리가 자율 주행 자동차나 축구를 하는 로봇을 만들 때, “이 상황에서는 이렇게, 저 상황에서는 저렇게” 식으로 프로그래밍을 하는 것은 거의 불가능하다.

머신러닝이란?

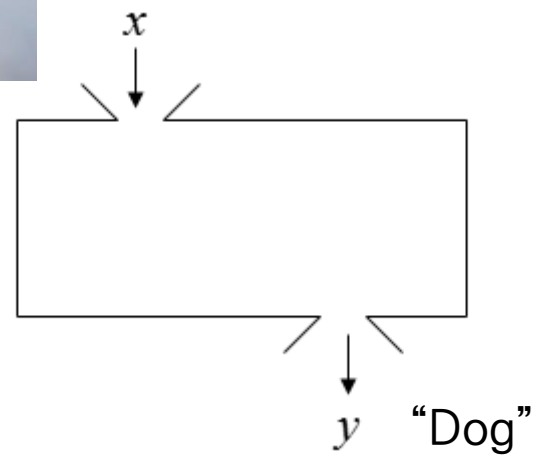
- ChatGPT: “기계 학습은 데이터와 알고리즘을 이용하여 모델을 구성하고, 이를 사용하여 새로운 데이터에 대한 예측이나 분류를 수행하는 것”
- 쉽게 말해 컴퓨터에게 스스로 데이터를 가장 잘 설명하는 패턴, 규칙(내지는 함수)를 찾도록 시키는 것이다.
- 잠깐, 머신러닝이 함수라고?





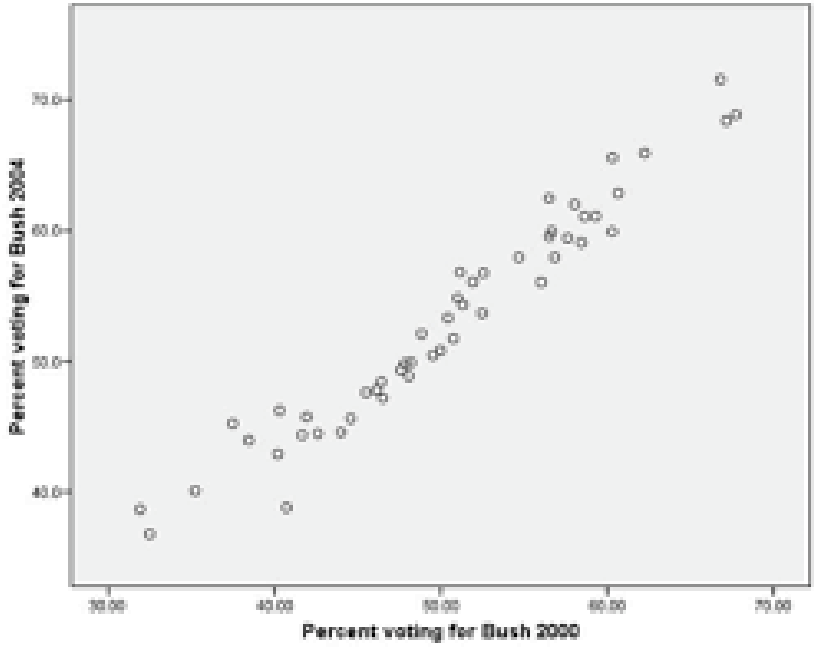
머신러닝이란?

- 개 사진을 입력 받으면 “개”라는 문자열을 출력하고 고양이 사진을 입력 받으면 “고양이” 라는 문자열을 출력하는 인공지능 모델을 생각해보자.
- 이는 함수일까? 만약 함수라면 어떻게 생겼을까?



머신러닝이란?

- 순서쌍(튜플) (x, y)를 생각해보자. 아래 그래프는 (x, y)들을 나타낸 것이다.
- 우리의 목적은 피처 x와 타깃 y의 관계를 최대한 잘 설명하는 함수를 찾는 것이다.
- 왠지 그래프에 직선(일차함수)을 하나 긋고 싶은 욕망이 생길 것이다.



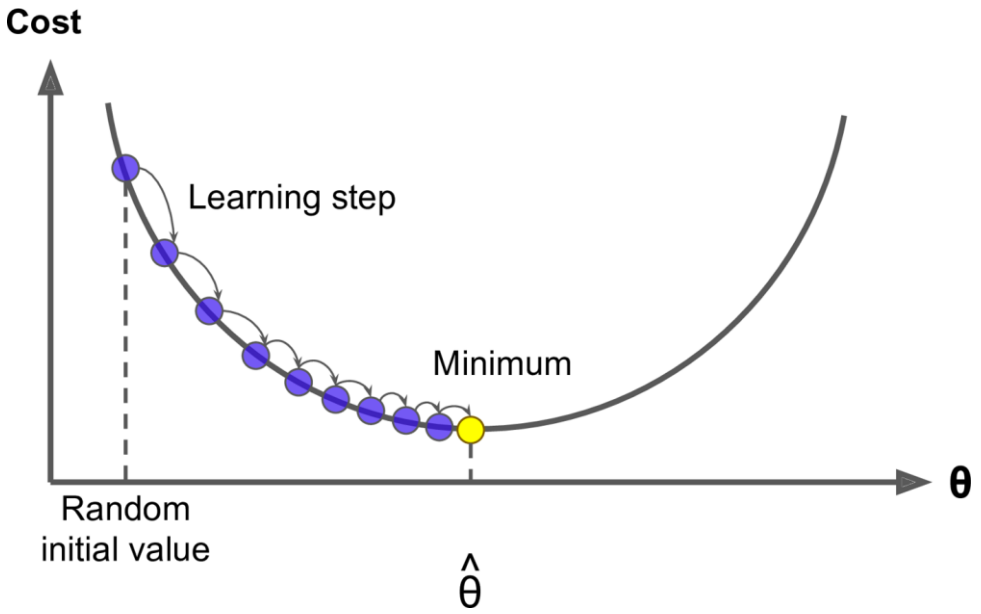
머신러닝이란?

- 생각해보니 일차함수는 y = ax + b의 형태이다.
- 이 최적의 a, b(파라미터)를 찾는 것을 “학습”이라고 한다.
- 이 함수의 형태는 결국 사람이 정해주어야 한다.
- 참고: No Free Lunch(NFL) Theorem에 의해 모든 데이터에 다 잘 맞는 최적의 머신 러닝 알고리즘은 존재하지 않는다는 것이 증명되었다.

⇒ 각각의 문제와 데이터에 따라 적절한 알고리즘과 방법론을 선택해야 한다.

학습은 어떻게 할까?

- 아까 예시에서 최적의 파라미터 a, b는 어떻게 찾을까?
- 일반적으로 머신러닝 모델들은 "오차"를 최소화하는 방향으로 학습한다. 이 오차를 "손실 함수" 또는 "비용 함수"라고 한다. loss function, cost function이라고도 한다.
- 즉 최적 파라미터는 손실 함수(비용 함수)를 최소화하여 찾는다.

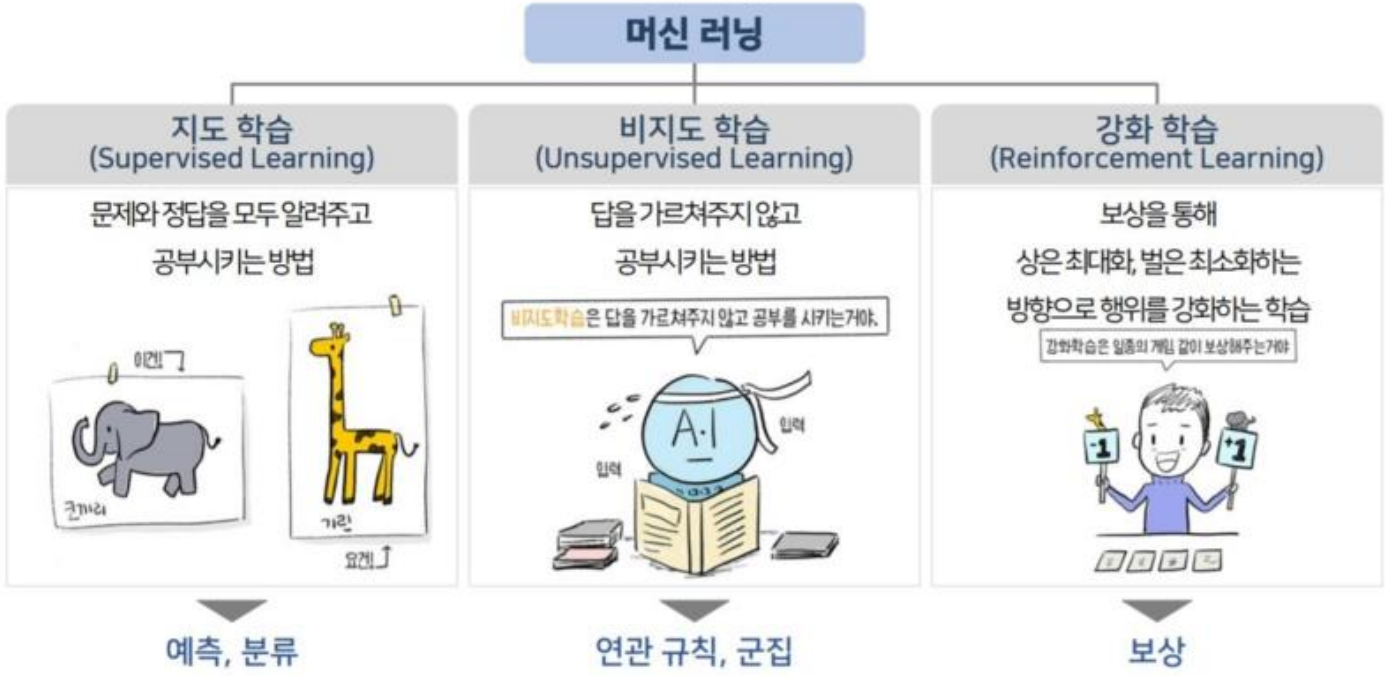


학습은 어떻게 할까?

- 사람이 직접 정해줘야 하는 파라미터도 있다. 이를 하이퍼파라미터라고 한다.
- kNN 알고리즘을 적용하려고 하면 두 가지 사항에 대해 사람이 직접 결정해야 한다. (k는 얼마인지, 어떤 거리를 쓸 것인지)
- 하이퍼파라미터는 학습을 통해 얻지 못하지만, 모델의 성능에 크게 영향을 준다.
- 따라서 실험을 통해 모델의 성능이 가장 좋은 하이퍼파라미터를 찾아야 한다. 이를 하이퍼파라미터 튜닝이라고 한다.
- 하이퍼파라미터 튜닝에는 그리드 서치, 랜덤 서치, 베이지안 최적화 등의 방법이 있다.

머신러닝의 종류

- 지도 학습: 입력 X와 정답 y가 주어지고, X, y 사이의 패턴(함수)를 학습하는 방법
- 비지도 학습: 입력 X만 주어지고, X 내에서의 패턴, 유사성을 학습하는 방법
- 강화 학습: 환경이 주어지고, 보상을 최대화하는 방향으로 행동을 학습하는 방법

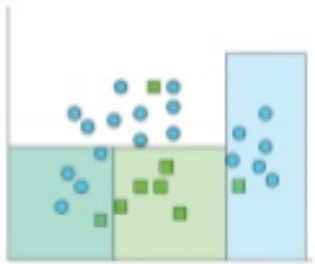


기계학습
(Machine Learnig)

지도학습
(Supervised Learning)

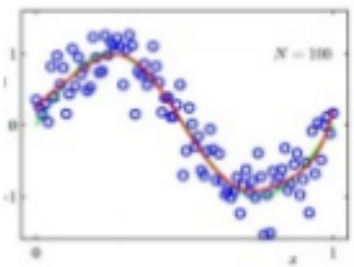
- 레이블된 데이터로 학습
- 미래 데이터 예측
- y = f(x)

분류
(classification)



- MNIST 필기체 인식
- 스팸 메일 분류

회귀
(regression)



- 주가 예측

비지도학습
(Unsupervised Learning)

- 레이블 없이 학습
- 데이터의 숨겨진 구조/특징 발견
- x ~ p(x), x = f(x)

클러스터링
(clustering)

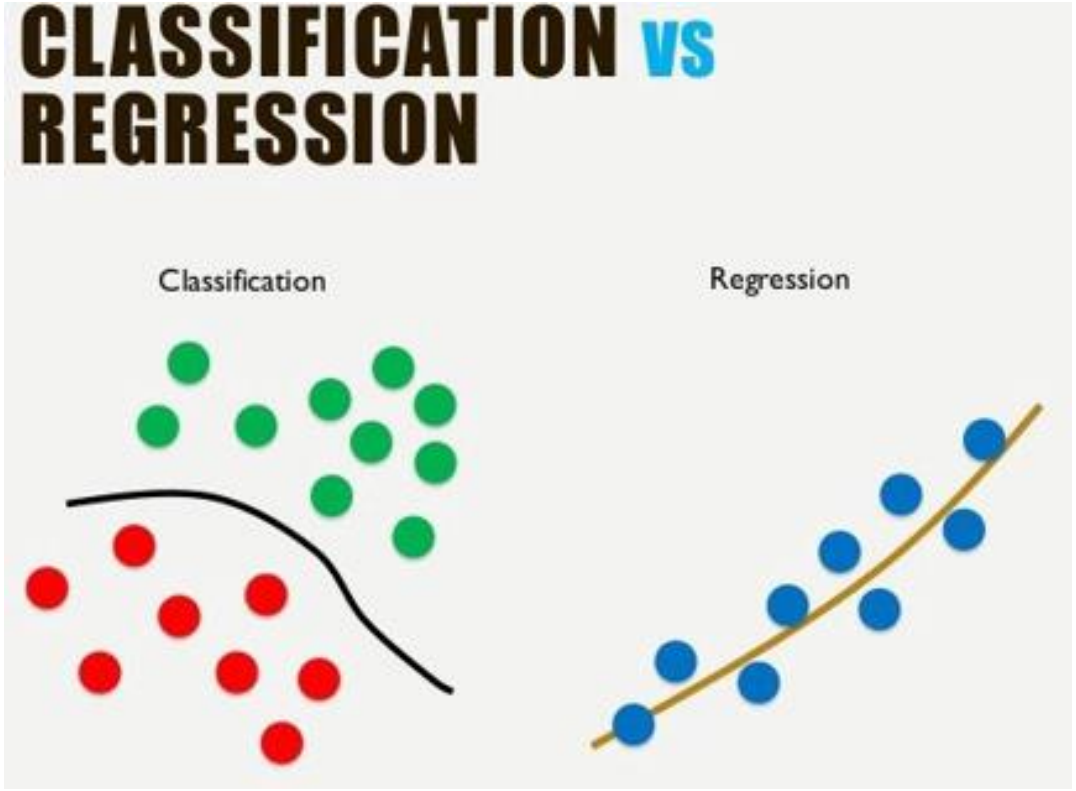
- 마케팅 고객 그룹화

강화학습
(Reinforcement Learning)

- 보상 시스템으로 학습
- 의사결정을 위한 최적의 액션 선택

지도 학습의 분류

- 회귀: 연속적인 값을 예측하는 것 (예시: 부동산 가격, 주식 가격 예측)
- 분류: 이산적인 값을 예측하는 것 (예시: 환자가 정상인지 암인지 분류, 스팸 메일 필터)



ML 모델이 처리하는 데이터의 종류

- 정형 데이터: 행과 열에 의해 데이터의 속성이 구별되는 데이터(엑셀 생각하면 됨)
- 비정형 데이터: 정형 데이터가 아닌 것을 비정형 데이터라고 한다. 텍스트, 오디오, 이미지, 비디오 등

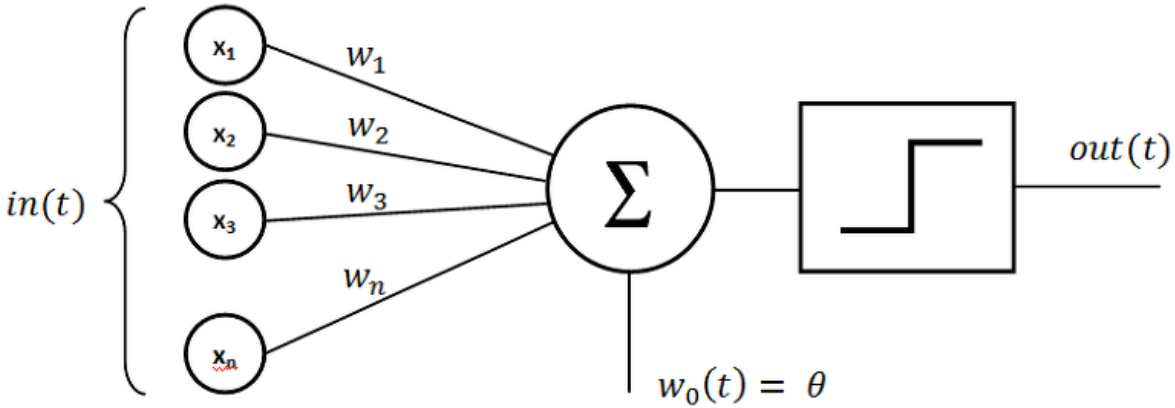
	A	B	C	D	E	F	G	H
1	일자	요일	시간대	업종	시도	시군구	읍면동	통화건수
2	20180601	금	0	음식점-족	서울특별시	강남구	논현동	5
3	20180601	금	0	음식점-족	서울특별시	강동구	길동	5
4	20180601	금	0	음식점-족	서울특별시	강서구	내발산동	5
5	20180601	금	0	음식점-족	서울특별시	동대문구	제기동	5
6	20180601	금	0	음식점-족	서울특별시	서대문구	창천동	7
7	20180601	금	0	음식점-족	서울특별시	서초구	양재동	5
8	20180601	금	0	음식점-족	서울특별시	성동구	성수동2가	5
9	20180601	금	0	음식점-족	서울특별시	성북구	동선동2가	5
10	20180601	금	0	음식점-족	서울특별시	송파구	송파동	5
11	20180601	금	0	음식점-족	서울특별시	영등포구	문래동3가	5

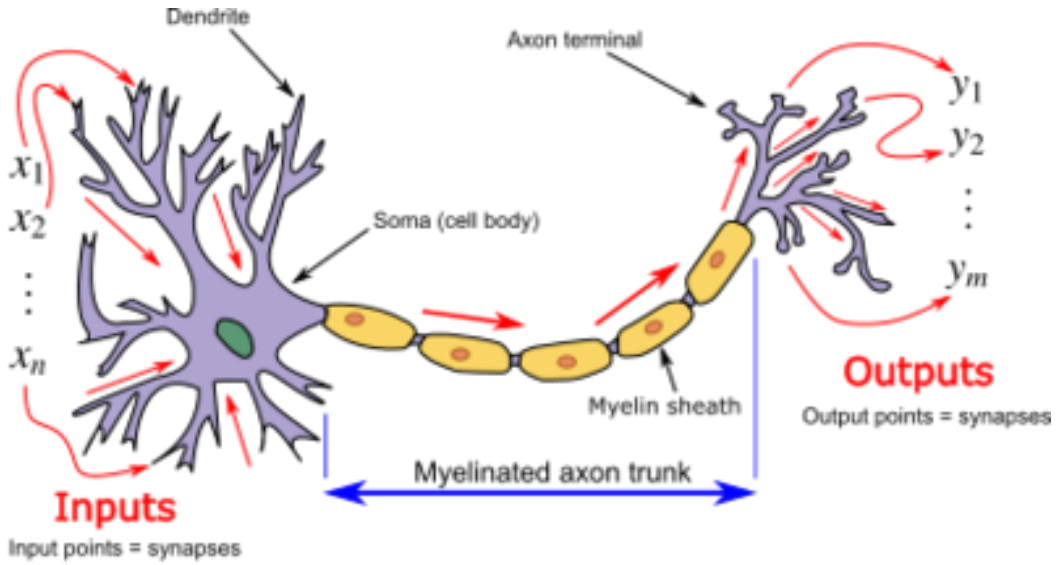
ML 모델이 처리하는 데이터의 종류

- 정형 데이터
  - 이산형 변수: 셀 수 있는 형태의 값을 가지고 있는 변수
    - 명목형(범주형): 데이터가 속하는 종류(class)들 간에 순서, 서열이 없는 경우. 자동차 부품의 종류가 파이프 A, 파이프 B, 파이프 C인지 구분할 때 이 셋 간에는 순서나 서열이 없으므로 명목형이다.
    - 순서형: 데이터가 속하는 종류(class)들 간에 순서나 서열이 있는 경우. 백준 티어, 롤 티어에는 순서와 서열이 있으므로 순서형에 포함된다.
  - 연속형 변수: 실수 형태의 연속적인 값을 가지고 있는 변수

딥러닝이란?

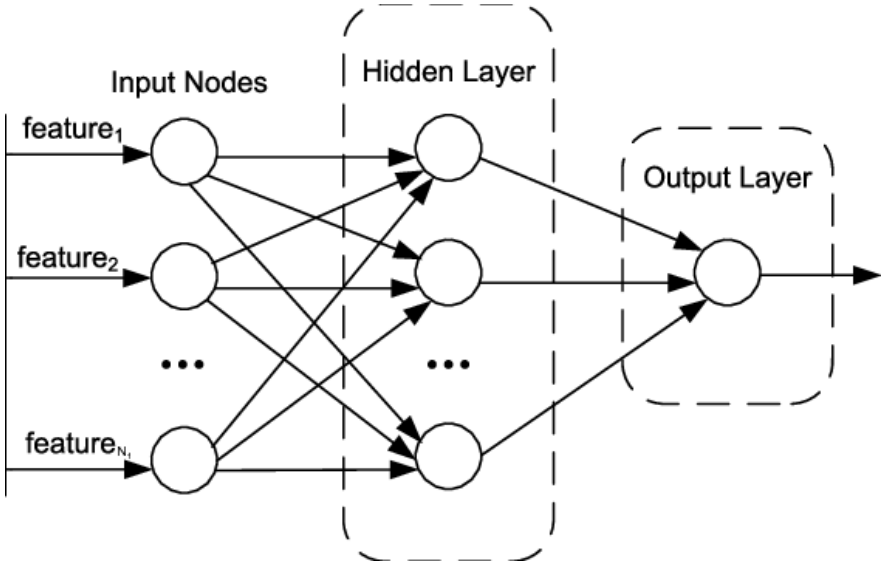
- 퍼셉트론: 인간의 뉴런을 본떠 만든 인공 구조.
- 입력층과 출력층이 있다. 비선형 함수(직선이 아닌 함수)를 포함한다.





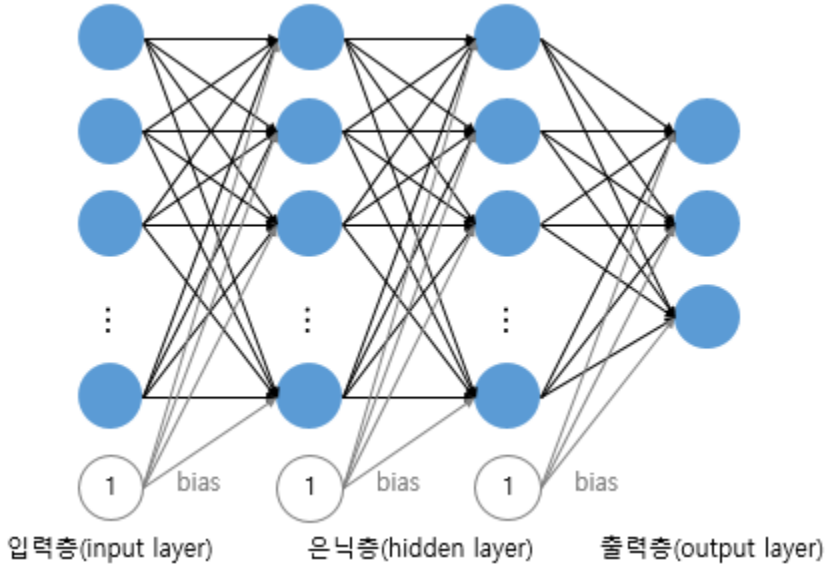
딥러닝이란?

- 다층 퍼셉트론(MLP): 퍼셉트론(뉴런)을 여러 개 쌓아 인간의 두뇌를 본떠 데이터를 처리하도록 만든 구조.
- MLP는 여러 개의 층이 있으며, 입력층과 출력층 사이의 층을 은닉층이라 부른다. “인공 신경망” 혹은 간단히 “신경망”이라고도 한다.



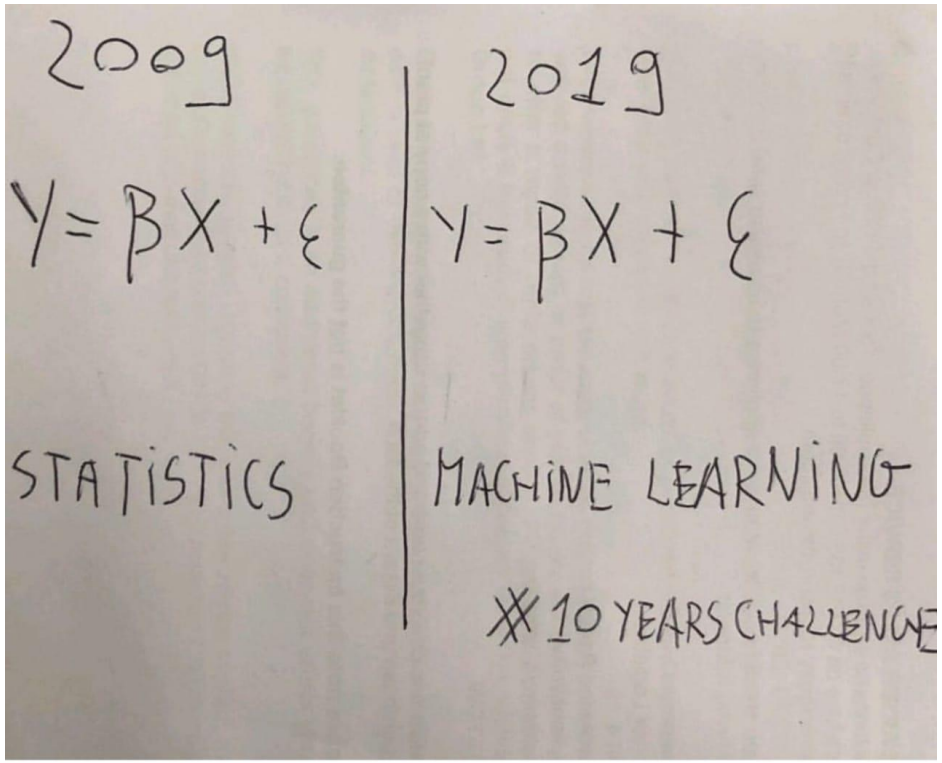
딥러닝이란?

- 심층 신경망: 은닉층이 2개 이상인 인공 신경망
- 딥러닝: 심층 신경망을 학습시키는 기계 학습의 한 유형
- 쉽게 말해 비선형 함수를 여러 번 써서 복잡한 함수를 만드는 것이다.
- 복잡한 함수를 만들면 복잡한 패턴을 잘 설명할 수 있게 된다.(아까 개 사진 예시)



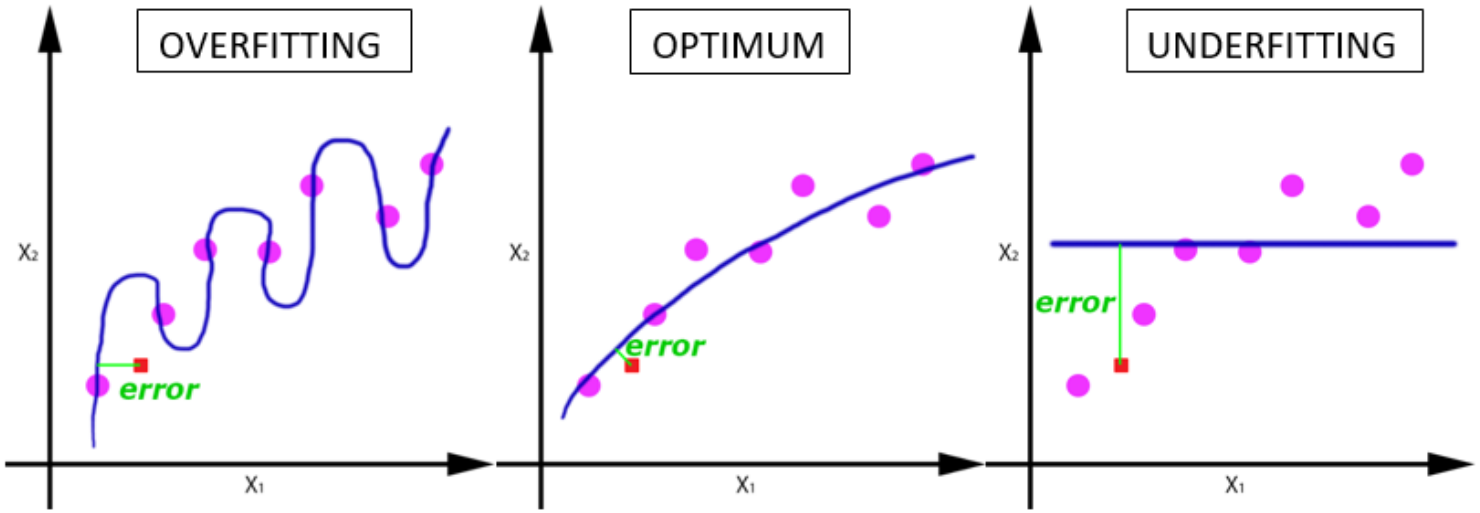
머신러닝/딥러닝의 목적

- 현재 가지고 있는 데이터를 통해 미지의 데이터를 예측하는 것 (통계와의 차이점)
- 따라서 과대 적합(과적합, 오버피팅), 과소 적합(언더피팅) 등을 조심해야 한다.



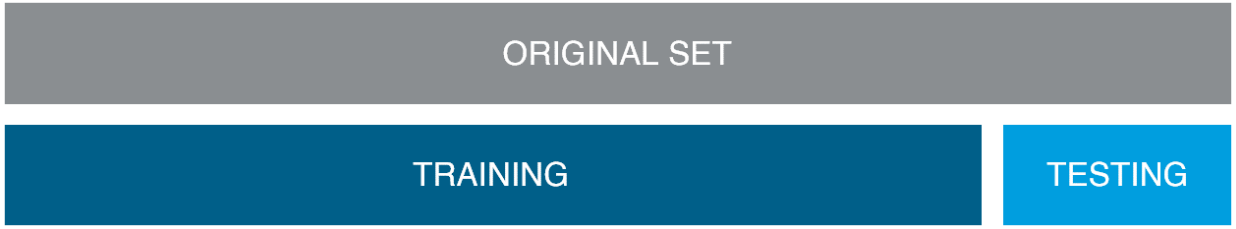
머신러닝/딥러닝의 목적

- 과대 적합(과적합, 오버피팅): 훈련 시에 사용한 데이터에만 높은 성능을 보이고, 미지의 데이터는 예측하지 못하는 상황
- 과소 적합(언더피팅): 모델이 너무 단순하여 예측력이 떨어지는 상황



머신러닝/딥러닝의 목적

- 과대적합을 피하기 위해, 데이터셋을 훈련 세트(train set), 테스트 세트(test set)로 나눈다.
- 훈련 시에는 훈련 세트를 사용하고, 테스트 세트로 모델의 성능을 평가한다.
- 그러나 테스트 세트의 분포가 훈련 세트의 분포와 다를 경우 일반화 성능을 신뢰하기 힘들며, 모든 데이터를 적극적으로 활용하지 못하게 된다.

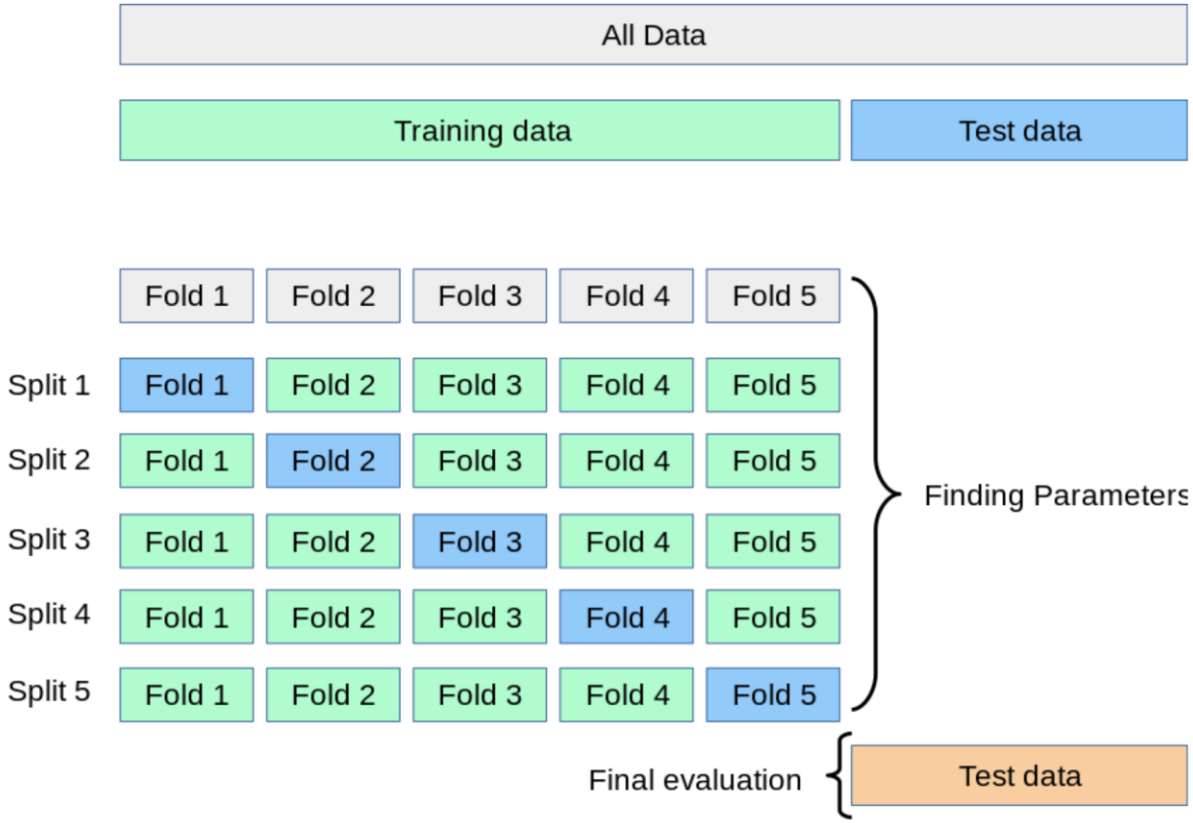


머신러닝/딥러닝의 목적

- 이를 막기 위한 아이디어가 검증 세트(Validation set)을 사용해 성능을 검증하는 교차 검증(Cross Validation)이다.
- 교차 검증을 하면 훈련 도중에 모델의 성능과 과대적합 여부를 간접적으로 확인할 수 있다는 점이고, 모든 데이터를 학습과 성능 평가에 활용하므로 효율적이다.

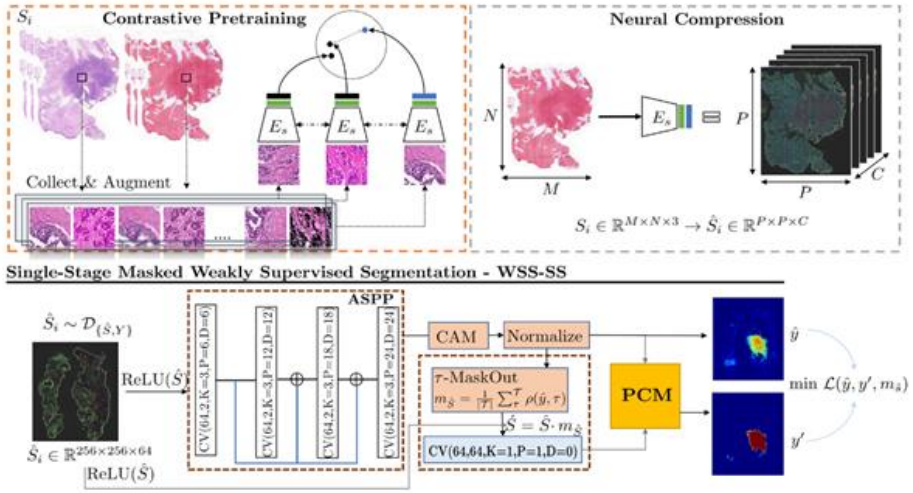
머신러닝/딥러닝의 목적

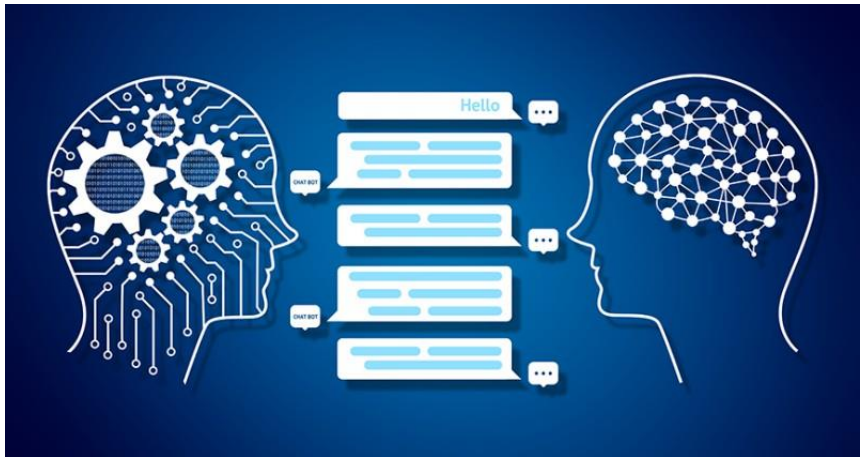
- 교차 검증 기법으로는 K-Fold Cross Validation(K-폴드 교차 검증), Stratified K-Fold Cross Validation(계층화 K-폴드 교차 검증) 등이 있다.



ML/DL로 해결 가능한 문제는?

- 이미지 분류: 의료 이미지에서 암세포를 탐지하는 데에 딥러닝(컴퓨터 비전) 기술이 사용
- 음성 인식: 음성을 자동으로 인식하여 텍스트로 변환 (음성 비서, 유튜브 자동 자막 등)
- 자연어 처리: ML/DL 기술을 사용해 문장에서 단어의 의미를 추론하거나 언어 간 번역, 감성 분석 등에 사용됨





ML/DL로 해결 가능한 문제는?

- 추천 시스템: 사용자의 행동 데이터를 분석해 제품이나 서비스를 추천 가능
- 자율 주행: 자동차, 드론 등을 ML/DL 기술을 이용해 자율적으로 운전
- 게임 AI: 예를 들어 AlphaGo에는 딥러닝 기술(강화 학습)이 사용되었음

예시

- 성조가 생략된 베트남어 문장에 성조를 달아주는 인공지능 모델
- 베트남 사람들은 채팅할 때 귀찮아서 성조를 달지 않는 경우가 있는데 간혹 오해가 생김

"Co ve an com khong con cho?"

1) Có về ăn cơm không **còn chờ**?
(밥먹으러 오니? 기다릴까?)
2) Có về ăn cơm không **con chó**?
(밥먹으러 오니, 멍멍아?)

The danger of typing Vietnamese without diacritics

The text:

Anh oi em dang coi quan. Den ngay di anh, muon lam roi. A, tien the mua bao moi nhe, o nha toan bao cu thoi. Ma thoi khong can mua bao nua dau, em vua mat kinh roi, khong nhin duoc nua anh oi, den ngay di, muon lam roi.

One way to fill in the diacritics:

Anh ơi, em đang coi quán. Đến ngay đi anh, muộn lắm rồi. À, tiện thể mua báo mới nhé, ở nhà toàn báo cũ thôi. Mà thôi không cần mua báo nữa đâu, em vừa mất kính rồi, không nhìn được nữa anh ơi, đến ngay đi, muộn lắm rồi.
(Honey, I'm looking after the shop. Please get here quick, it's late. Oh, please get the newspaper by the way. There's only old newspapers at home. Oh wait but no need. I just lost my glasses, can't see anything. Just please get here quick, it's really late.)

Another way:

Anh ơi, em đang cởi quần. Đến ngay đi anh, muốn lắm rồi. À, tiện thể mua bao mới nhé, ở nhà toàn bao cũ thôi. Mà thôi không cần mua bao nữa đâu, em vừa mất kinh rồi, không nhịn được nữa anh ơi, đến ngay đi, muốn lắm rồi.
(Honey, I'm taking off my pants. Please get here quick, I want it real bad. Oh please get new condoms by the way. There's only old condoms at home. Oh wait but no need. I just missed my period. Can't stand it anymore. Just please get here quick, I want it real bad.)

머신러닝 주요 라이브러리

- numpy, scipy: 과학 계산을 위한 도구
- pandas: 데이터를 표현하기 위한 도구
- matplotlib, seaborn: 데이터를 시각화하기 위한 도구
- sklearn: 머신러닝 모델을 학습하기 위한 도구
- pytorch, tensorflow, huggingface: 딥러닝 모델을 학습하기 위한 도구

요약

- 머신러닝과 딥러닝이란 무엇인지 살펴봤습니다.
- ‘학습’이란 무엇이며, 어떻게 이루어지는 것인지 살펴봤습니다.
- 머신러닝의 종류를 살펴봤고, 그 중 지도학습의 종류에 대해서도 살펴봤습니다.
- 데이터의 종류에 대해 살펴봤습니다.
- 머신러닝과 딥러닝의 목적이 무엇인지 살펴봤습니다.
- 머신러닝과 딥러닝으로 해결 가능한 문제에 대해 살펴봤습니다.

다음 시간에 다룰 주제

- ML/DL과 수학의 관계