2020 D&A

MACHINE LEARNING SESSION

1. Introduction to Machine Learning

D&A

머신러닝의 여러 가지 정의들

머신러닝 (Machine Learning; 기계학습)

기계(컴퓨터)가 스스로 데이터를 학습해서 각 변수들 간의 관계를 찾아나가는 과정

데이터에서 지식을 추출해내는 작업

이미 존재하는 데이터를 가지고 기계 스스로 규칙을 만들고 성능을 향상하는 것

통계학, 인공지능, 컴퓨터 과학이 얽혀 있는 연구 분야

* example

영화추천, 음식 주문, 쇼핑, 맞춤형 온라인 라디오 방송, 사진에서 친구 얼굴을 찾아주는 일, 페이스북, 아마존, 넷플릭스

기존 프로그래밍과 머신러닝의 차이

기존 프로그래밍?

인간이 만든 규칙을 기계(컴퓨터)에 입력해 정답을 도출

if와 else의 무한 반복인 엄청난 하드코딩으로 이뤄짐

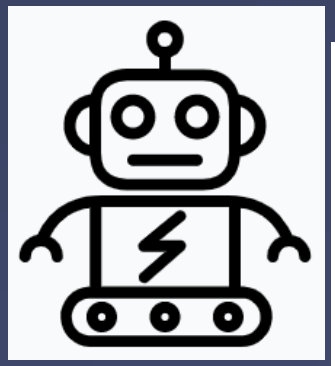
세분화된 한 분야에만 국한되어 사용할 수 있음

작업이 조금만 변경되어도 시스템 전체를 갈아엎어야 함

사용 분야에 대한 전문가급의 높은 이해도와 프로그래밍 실력이 결합되어야 함

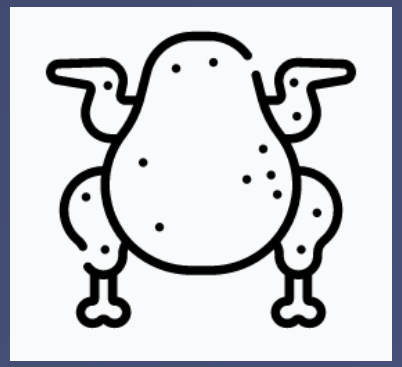
-> 개발도 힘들고 사용도 까다로움!

기존 프로그래밍과 머신러닝의 차이



```
if 목 in 닭:
자른다
if 양날개 in 닭:
두동강 낸다
if 용도 == 치킨
if 발 in 닭:
자른다
else:
남긴다
if 내장 in 닭:
제거한다
if 닭껍질 in 닭:
KFC에 판다
```

로봇



생닭

인간

기존 프로그래밍과 머신러닝의 차이

머신러닝은!

인간이 만든 규칙을 가지고 기계(컴퓨터)에 입력해 정답을 도출
-> 이미 존재하는 정답을 가지고 기계(컴퓨터)가 스스로 학습

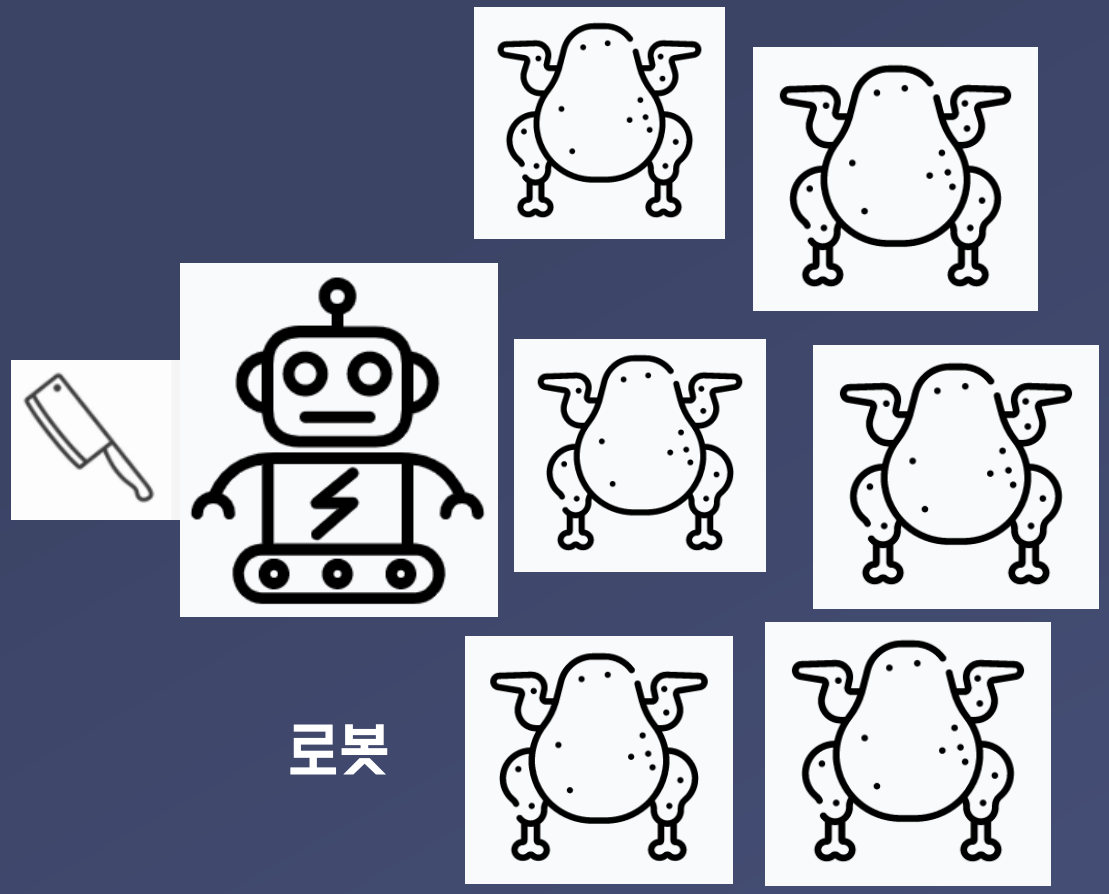
if와 else의 무한 반복인 엄청난 하드코딩으로 이뤄짐
-> 여러 라이브러리로 이미 존재하는 모델을 활용

세분화된 한 분야에만 국한되어 사용할 수 있음
-> 하나의 모델은 굉장히 다양한 분야에 적용할 수 있음

작업이 조금만 변경되어도 시스템 전체를 갈아엎어야 함
-> 데이터가 변경되어도 모델의 변경은 불필요

사용 분야에 대한 전문가급의 높은 이해도와 프로그래밍 실력이 결합되어야 함
-> 전문가급 이해, 프로그래밍 실력 둘다 뛰어나지 않아도 됨

기존 프로그래밍과 머신러닝의 차이



로봇

생닭

인간

머신러닝의 종류

지도학습 (Supervised Learning)

- 분류 (Classification)
- 회귀 (Regression)

비지도학습 (Unsupervised Learning)

- 군집 분석 (Clustering)
- 차원 축소 (Dimensional Reduction)

지도학습

주어진 데이터(피쳐, x)와 결과(라벨, y)가 존재할 때 학습하는 방법
기존의 데이터를 학습한 모델을 이용해 새로운 데이터의 결과를 예측할 수 있음

-> 일반화된 모델을 만드는 것

- 회귀 : 결과가 연속형인 변수
- 분류 : 결과가 이산형, 범주형인 변수

* example

편지 봉투에 손으로 쓴 우편번호 숫자 판별
의료 영상 이미지에 기반한 종양 판단
의심되는 신용카드 거래 감지

비지도학습

주어진 데이터만 있고 결과가 없어서 모델 스스로 규칙을 도출하는 방법
기계(컴퓨터)가 만드는 것이기 때문에 이해하거나 평가하기는 쉽지 않음

- 군집 : 데이터를 비슷한 것끼리 묶음
- 차원 축소 : 변수의 차원을 줄임

* example

  블로그 글의 주제 구분
  고객들을 취향이 비슷한 그룹으로 묶기
  비정상적인 웹사이트 접근 탐지

머신러닝 과정

데이터 전처리

↓

데이터 분리

↓

모델 학습

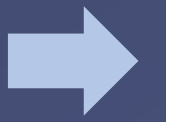
↓

모델 평가

데이터의 특징을 나타낼 수 있는 다양한 변수를 생성

사용할 모델에 맞게 데이터를 알맞은 형태로 전처리

	cust_id	tran_date	store_nm	goods_id	gds_grp_nm	gds_grp_mclas_nm	amount
0	0	2007-01-19 00:00:00	강남점	127105	기초 화장품	화장품	850000
1	0	2007-03-30 00:00:00	강남점	342220	니 트	시티웨어	480000
2	0	2007-03-30 00:00:00	강남점	127105	기초 화장품	화장품	3000000
3	0	2007-03-30 00:00:00	강남점	342205	니 트	시티웨어	840000
4	0	2007-03-30 00:00:00	강남점	342220	상품군미지정	기타	20000
...	...	...	...	...	...	...	...
231999	3499	2007-12-17 00:00:00	본 점	127129	상품군미지정	기타	-135000
232000	3499	2007-12-23 00:00:00	노원점	285136	시티웨어	시티웨어	6380000
232001	3499	2007-12-23 00:00:00	노원점	39107	야채	농산물	40800
232002	3499	2007-12-27 00:00:00	본 점	740120	어덜트	명품	4880000
232003	3499	2007-12-27 00:00:00	본 점	740120	수입의류	명품	610000



	cust_id	총구매액	구매건수	평균구매액	최대구매액
0	0	68282840	74	922741	11264000
1	1	2136000	3	712000	2136000
2	2	3197000	4	799250	1639000
3	3	16077620	44	365400	4935000
4	4	29050000	3	9683333	24000000
...	...	...	...	...	...
3495	3495	3175200	2	1587600	3042900
3496	3496	29628600	13	2279123	7200000
3497	3497	75000	1	75000	75000
3498	3498	1875000	2	937500	1000000
3499	3499	263101550	92	2859799	34632000

D&A

머신러닝 과정

데이터 전처리

↓

데이터 분리

↓

모델 학습

↓

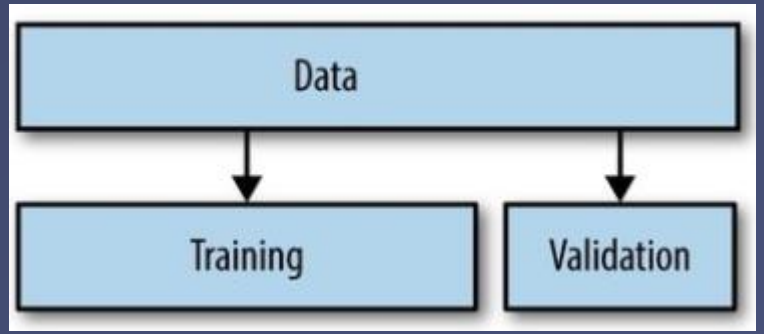
모델 평가

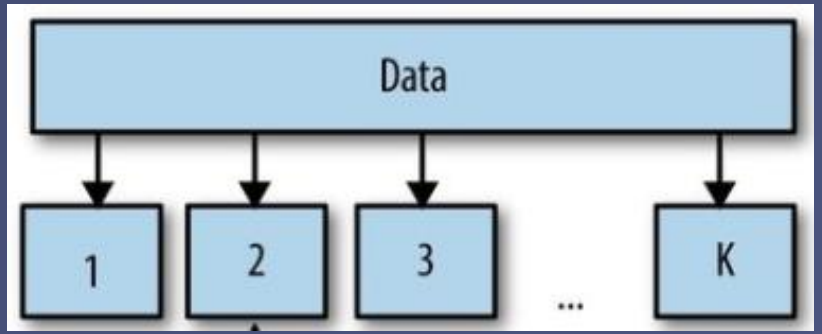
데이터를 학습시키기 위해서는 보통 훈련(train)데이터와 평가(validation)데이터로 분리해서 사용하게 됨

보통 훈련 데이터를 70~80% 정도의 비율로 두고, 평가데이터를 20~30%의 비율로 두고 데이터를 사용함

데이터 분리에는 여러가지 방법이 존재함

ex) Hold-Out, K-fold





머신러닝 과정

데이터 전처리

↓

데이터 분리

↓

모델 학습

↓

모델 평가

데이터 학습 과정은 무조건 train(훈련)데이터만 사용해서 학습을 진행함

오차를 줄이는 방향으로 모델의 파라미터가 갱신됨 (Gradient Descent)

모델의 성능을 높이기 위해서 피쳐를 새로 만들거나 하이퍼 파라미터를 조정함.

* 학습 모델의 종류

Logistic Regression, Decision Tree, Neural Network 등등..

머신러닝 과정

데이터 전처리

↓

데이터 분리

↓

모델 학습

↓

모델 평가

학습이 된 모델을 사용해 학습에 사용되지 않은 남은
validation데이터를 모델에 넣어 검증

훈련데이터의 스코어와 비슷하거나 조금 더 좋은 게 잘 학습된 것

최고 성능을 위해 앞의 과정을 반복하며 학습

but!
일반화된 모델을 만들어야 함
데이터별로 과적합의 기준이 명확하지 않음
(과적합된 모델이란? 훈련데이터에만 잘 맞아 일반화되지 않은 모델)
- 과대적합 : 훈련데이터를 너무 외워버린 모델
- 과소적합 : 훈련데이터를 잘 학습하지 못한 모델

ex) train score : 0.5인데 test score가 0.6인 경우에 과적합이라
고 판단할 수도 있고, 잘 학습되었다고 판단할 수도 있음

과제

조이름, 조장 결정하기

(개인 과제)
오늘 배운 내용 워드 한 페이지로 요약한 보고서 써오기

https://drive.google.com/drive/folders/1ScQRSHsMQ8t4OCOcQ2VHvRZXA1ON2LRD

D&A