

Inclass 01: Introduction to Machine Learning

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.ac.kr)

School of AI Convergence & Department of Artificial Intelligence, Dongguk University

Course Introduction

담당 교수 및 조교

- 담당 교수: 박성식 *주말 x, 시간 x*
 - e-mail: s.park@dgu.ac.kr
 - tel: 02) 2260 - 3784
 - web: spark-lab.com
- 조교: 강민지 *지남학기 수강생*
 - e-mail: 01minjik@naver.com
 - 직접 연락 x, 저한테기 연락.*
- 교재 *강의자료 ok,*
 - ➔ C. Bishop, Pattern Recognition and Machine Learning *[PRML]* 
 - A. Geron, Hands-On Machine Learning with Scikit-Learn and Tensorflow *4*

- 주차별 강의계획: 구글 시트 ↗
 - 플립 러닝 (flipped learning)
 - 사전강의(preclass): 이론 수업
 - 본강의(inclass) / 수목: 사전 강의 커버
 - 본강의(inclass) / 월화: 파이썬 실습
- 오픈 카톡방
 - 월수 분반
 - 화목 분반 ↗

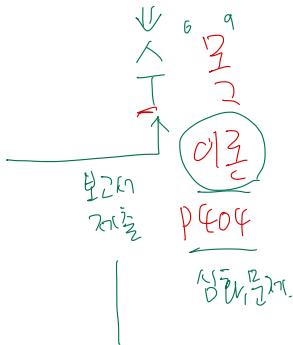
☆ 수업 의견: 구글 설문 (아래) ← ★ brake, 다영 메시지

- 유튜브: <https://youtube.com/@spark-lab> ↗

이

사전강의
pre class

기본이론



호수

시수
4

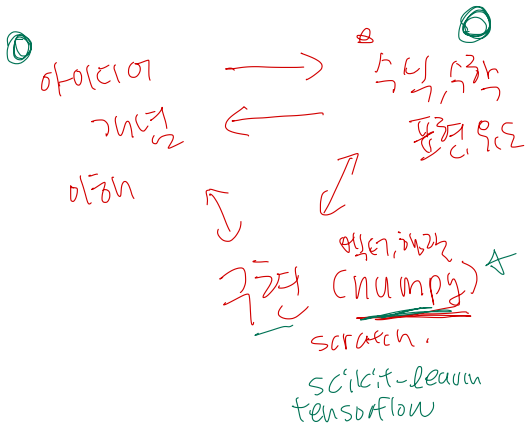
Q

구편

① 필기, 손으로 쓴 거

② on/off

강의 목표



노력, 시간투자 ∞ 가해가 될 것

① ML 모델을 설계, 구현. + 평가.

② $\frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{y}_i)$ 최소화.

③ test

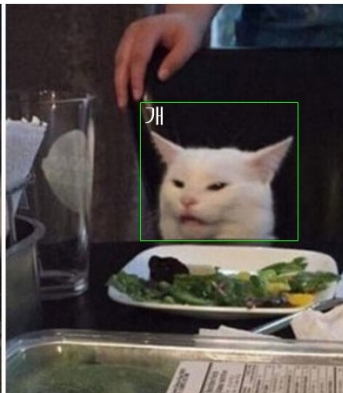
Introduction to Machine Learning

What is Machine Learning?

AI가 세계를 지배할 거라는
AI 잘못들:



내가 만든 AI:



What is Machine Learning?

기재 학습

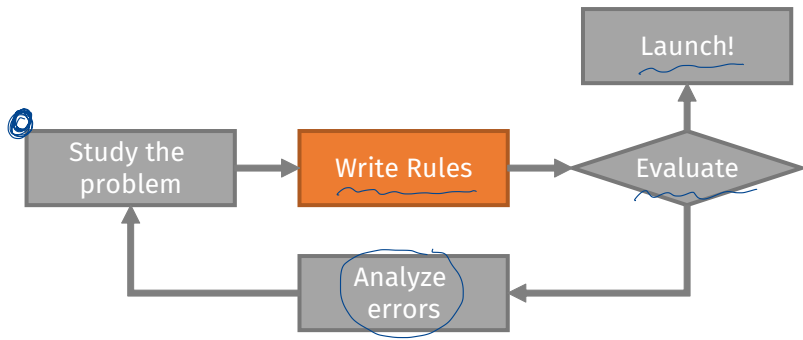
“[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed” - Arthur Samuel, 1959

기존의 방법 (rule-base)

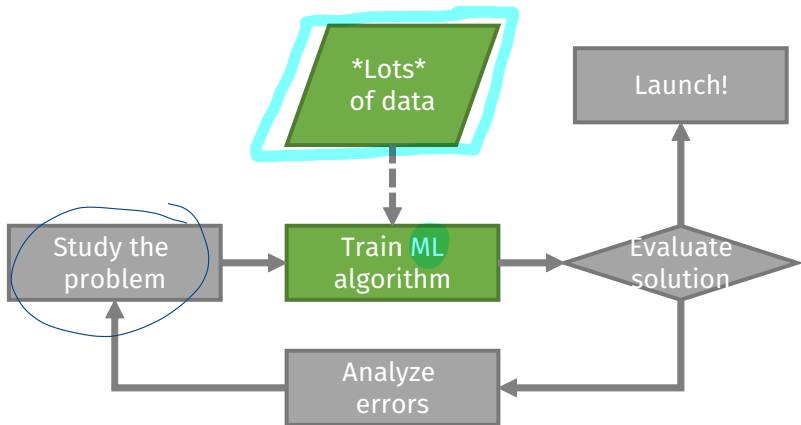
SPAM FILTER is a Machine Learning Program

- not by an explicit algorithm, e.g., filtering by special words
- given examples of spam emails (e.g., flagged by users) } data
- given examples of regular (nospam) emails
- it can learn to flag spam

Traditional approach



Machine learning approach



학습
데이터

데이터

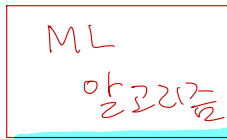


sample
instance

label
from
target

input

입력
데이터



output

출력
데이터

예측
prediction

학습
데이터

테스트
데이터



training dataset

instance, sample

label, truth, target

input - output

prediction

인공지능 Artificial Intelligence

- 가장 넓은 의미에서의 인공지능으로서 인공적으로 구현한 모든 수준의 지능을 포괄
- 컴퓨터 과학 측면에서 보면, 환경을 인지하여 어떤 목표를 성취할 수 있는 가능성이 최대화되도록 행동을 취하는 지능적 객체에 대한 연구

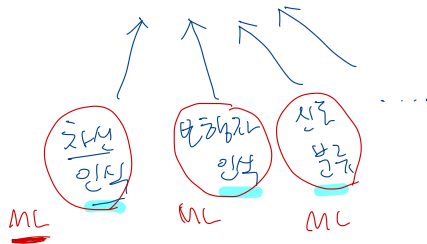
머신러닝 Machine Learning

- 구체적으로 프로그래밍하지 않아도 스스로 학습하여 특정 임무를 수행할 수 있는 능력을 컴퓨터가 갖도록 구현하는 분야
- 개체에 대한 속성을 잘 학습할 수 있는 feature에 대한 고민 필요

딥러닝 Deep Learning

- 데이터에 대한 다층적 표현과 추상화를 통해 학습하는 머신 러닝의 기법
- 특징 공학 없이도 개체 속성을 스스로 이해하고 학습하여 결과를 내도록 할 수 있다는 것이 강점

인공지능 : 자율주행, Siri



더 많은 것이
가능하다

ML $\supset$ DL

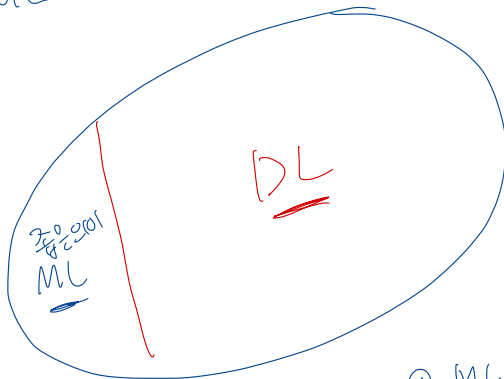
ML



DL

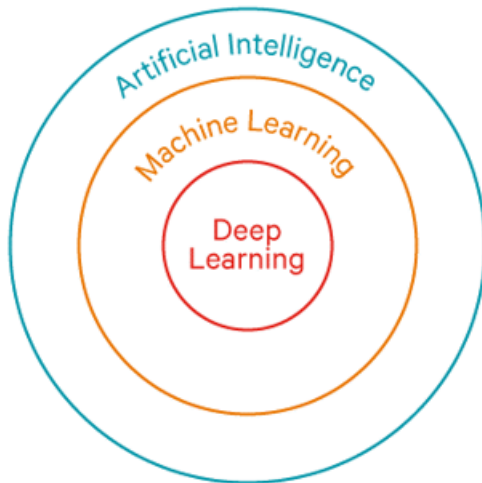
neural net,
layer ↑

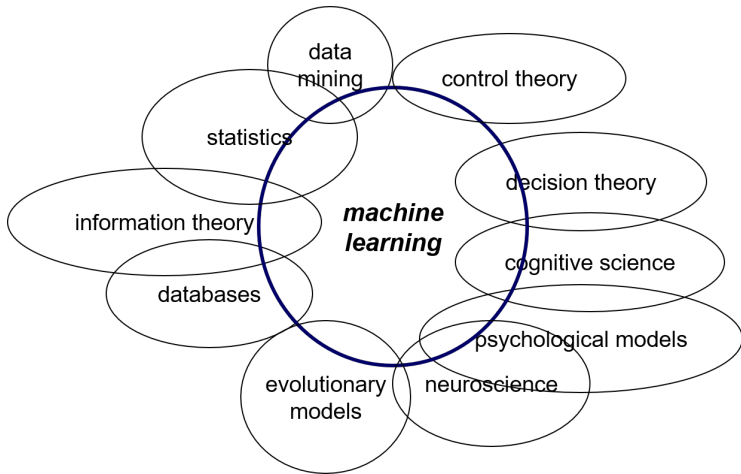
주요인 ML



① ML 기본

② 다 같은 것

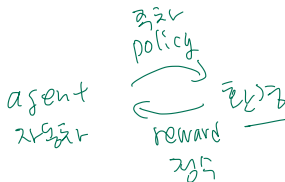
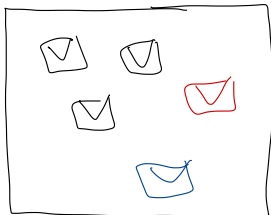




Types of Machine Learning

- 지도(supervision) 유무

- 지도 학습 supervised learning : 모든 data에 label이 있을 때.
- 비지도 학습 unsupervised learning : " label이 없을 때
 - 준지도 학습 semi-supervised learning
- 강화 학습 reinforcement learning



Types of Machine Learning

- output
- 목표가 되는 출력값의 종류에 따라
 - 분류 classification: 분절된, 범주화된 출력값 discrete, categorical 정수
 - 회귀 regression: 연속된 출력값 continuous 실수
 - 탐지 detection: 참/거짓 binary T/F

Types of Machine Learning

- 데이터에 대한 점진적 학습 여부
 - batch learning
 - mini-batch learning
 - online learning, stochastic learning

→ 다음시간

Supervised vs. unsupervised

- Supervised learning
 - 레이블(label, target value, ground truth)이 붙은 데이터셋으로 학습
 - k-Nearest Neighbors (NN), regression, support vector machine (SVM), random forests, neural network
- Unsupervised learning
 - 레이블이 없는 데이터에서 패턴이나 구조를 찾는 방법
 - clustering, hierarchical clustering, expectation-maximization (EM), principal component analysis (PCA), anomaly detection

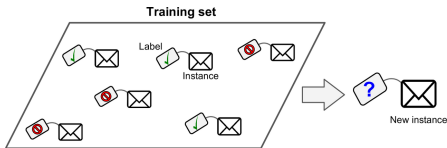


Figure 1: Supervised training dataset

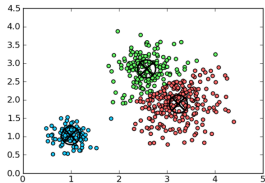


Figure 2: Unsupervised learning (k-Means clustering)

Batch and online learning

배치 러닝 batch learning

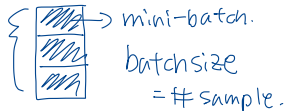
- 가용한 데이터를 한 번에 모두 사용해서 학습
- 시간 및 컴퓨팅 자원이 많이 들어감
- 새로운 데이터셋이 주어지면 처음부터 다시 학습시켜야함
- 부드럽게 해에 접근하지만 지역 최적값(local optima)에서 빠져나오지 못 할 수도 있음

$$\nabla J(\theta) = 2X^T(X\theta - y)$$

↑
[]

미니 배치 러닝 mini-batches learning

- 데이터셋을 미니 배치라고 하는 작은 그룹으로 나누어 학습
- 배치 방식과 온라인 방식의 절충
- 기울기(gradient)가 온라인 학습보다 정확하고 지역 최적값 문제에서 배치 학습보다 우수
- 실제 대부분 프로그램에서 미니 배치 방식을 사용



온라인 러닝 online learning

- 데이터를 순차적으로 처리하면서 점증적으로 학습
- 각 학습 단계가 빠름
- 배치 러닝보다 불규칙하게 접근하지만 지역 최적값 문제에서 유리
- 데이터가 순차적으로 들어오는 경우에 적합
- 컴퓨팅 자원이 매우 제한적일 때 적합

$$\nabla J(\theta) = 2(\theta^T x_n - y_n)x_n$$

online
Stochastic

→ 증강학습

- 테스트의 목적: 일반화 오차(Generalization error)가 관건

① 테스트 (Test)

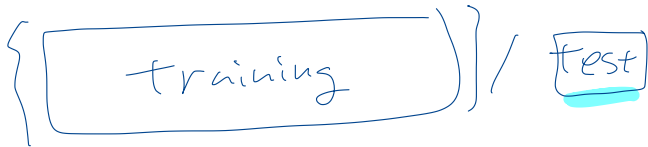
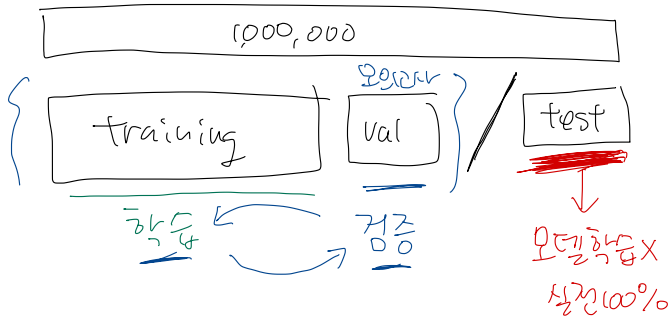
- 실무 적용 직전에 모델 성능을 평가하는 것
- 테스트 데이터셋: 전체 데이터셋에서 학습 전에 별도로 떼어 놓은 데이터셋 (전체 데이터의 약 20%)

☆ 훈련(학습) 도중에 절대 절대로 답(Test dataset)을 들여다 보면 안됨

② 검증 (Validation)

- 모델 선택 및 미세 조정을 위해 실시
- 검증 데이터셋 (hold-one-out set) : 나머지 학습 데이터셋의 약 20%
- 한번 검증으로는 보통 불충분할 수도 있음 ↵
- 그리고, 가까운 데이터가 낭비로 볼 수 있음
- Cross validation (교차 검증) 필요

학습 데이터



Reference and further reading

- “Chap 1” of A. Geron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
- “Chap 1” of C. Bishop, Pattern Recognition and Machine Learning