

## <1. Introduction to ML and ANN>

### ● What is machine learning?

#### - 기계 학습의 타입

- 1. Supervised learning: label 예시가 주어지기 때문에, 옳은 행동으로 가도록 유도
- 2. Unsupervised learning: label이 없는 대신, 데이터에서의 interesting한 패턴을 찾는다.
- 3. Reinforcement learning: 학습 시스템이 reward signal을 받으면서, reward signal이 최대화되는 방향으로 학습함.

### ● Supervised learning examples

#### - Handwritten digit classification with MNIST dataset

- Input: 이미지 입력
- Target: 숫자 클래스

#### - Object recognition

#### - ImageNet dataset

#### - data augmentation: Lots of variability in viewpoint, lighting

#### - Error rate dropped from 25.7% ~ 5.7%

#### - Neural Machine Translation

- A 언어 -> 실체(raw data) -> 벡터(if 같은 도메인, 다른 카테고리, Encoder, Decoder 활용) -> 다른 카테고리 raw data -> B 언어

#### - Caption generation

- 이미지가 주어지면 텍스트로 출력
- 이미지 raw data -> vision Deep CNN -> Language Generating RNN -> 텍스트 출력

### ● Unsupervised learning examples

#### - generative modeling에서 natural images으로 구성된 dataset을 distribution을 하도록 학습하길 원한다.

#### - data의 distribution에 집중

#### - generative model을 샘플링함으로써 평가할 수 있다.

- 샘플링(Sampling): 전체 데이터 집합에서 일부 데이터를 선택하여 분석하거나 모델 학습에 사용하는 과정을 말한다. 데이터가 너무 많거나 다루기 어려울 때, 전체 데이터를 다루는 대신 대표성을 지닌 일부 샘플을 선택함으로써 시간과 계산 자원을 절약하면서 유의미한 결과를 얻을 수 있다.

#### - New state-of-the-art

#### - The progress of generative models

#### - CycleGAN: 어떤 카테고리의 많은 이미지를 다른 카테고리 이미지로 바꿔서 생성할 수 있다.

- Reinforcement learning

- Agent는 환경과 상호작용한다.
- each time step of Reinforcement learning
  - 1. 관찰자는 Agent에게 상태 정보 준다.
  - 2. Agent는 받은 상태 정보에 맞게 행동을 선택함.
  - 3. Agent는 주기적으로 보상을 받는다.
  - 4. Agent는 policy를 배우길, 관찰자에 맞게 행동하길, 시간이 지나면서 평균 reward가 최대화되기를 원한다.

- What is Statistical Learning?

- Statistical Learning: 데이터에서 패턴 구조를 발견하고, 이를 바탕으로 예측하거나 결론을 도출하기 위해 통계학과 수학적 모델을 사용하는 방법론임.
- 주어진 데이터를 가장 잘 나타낼 수 있도록 학습함.

- Statistical Learning이 어디에 쓰이나?

- 전립선암 위험을 식별할 수 있다.
- log-peropdogram(로그 주기도)에 근거해서 기록된 현상을 기록할 수 있다.
- 스팸 이메일인지 아닌지 감지할 수 있다.
- 유전자 발현 프로필을 통해서 여러 개의 암 종의 조직 샘플을 분류할 수 있다.
- image segmentation(픽셀 단위로 영역화하여 나타내는 것)이 가능하다.
  - LANDSAT 이미지의 픽셀을 분류할 수 있다.

- The Supervised Learning Problem

- 시작점
  - 1. Y: 결과 측정값(= 종속변수, 반응 변수, target)
  - 2. X: Input, regressors, features, 독립변수
- In regression problem, Y = 수치나 양을 측정할 수 있는 데이터 (금액, 혈압 등)
- In classification problem, Y = 정렬이 되지 않은 집합의 유한한 값을 가짐. (class 0, class 1, class 2, class 3 등)
- training data =  $(x_1, y_1), \dots, (x_N, y_N)$

- Objectives

- 훈련 데이터를 바탕으로 Objective를 아래와 같이 하고자 한다.
  - 1. 보이지 않는 test case를 정확히 예측하고자 할 때
  - 2. inputs이 outcome에 어떤 영향을 미치는지 알고자 할 때
  - 3. prediction이나 inferences를 quality를 평가할 때

- Philosophy

- Statistical learning은 현대 data scientist의 training에 기초적인 성분이다.

- Statistical Learning vs Machine Learning

- Machine Learning: subfield of Artificial Intelligence
- Statistical Learning: subfield of Statistics
- 공통점: 두 분야, 둘 다 supervised 와 unsupervised에 집중하고 있다.
- 차이점
  - ML은 크기가 큰 applications와 prediction 정확도를 강조함.
  - Statistical Learning은 models와 이해도, 정확성, 불확실성을 강조함.
- 그러나 이러한 경계서는 모호해졌고, 두 영역에서 많은 상호 교류가 이뤄지고 있다.

- What are neural networks?

- 생물학적 디테일의 대부분은 필수적이지 않다. 그래서 우리는 광범위하게 뉴런의 간단한 모델 구조를 사용한다.
- neural nets을 선형대수적, 통계학적으로 나타냄.
- 즉, Neural networks는 유용한 계산을 수행하기 위해 간단한 processing units의 수 백만개의 집합을 말한다.
- why neural nets?
  - 뇌로부터 영감을 얻는다.
    - neural architecture는 보고 들을 수 있는 개념이 증명됨.
  - 효과적인 applications의 범위: vision, text, speech, medicine, robotics ... etc
  - Academia와 tech industry에서 넓게 사용된다.
  - powerful한 software frameworks(PyTorch, Tensorflow)이 생겨남.

- Deep Neural Networks

- DL: processing의 많은 layers(stages) 존재함.

- Software frameworks

- Array processing: Numpy
- Neural net frameworks: Torch, PyTorch, TensorFlow, Theano
  - 1. 미분 자동 수행
  - 2. 다양한 라이브러리
  - 3. GPU를 기반한 병렬 연산이 가능