



『국방 인공지능(AI) 중급과정』
오리엔테이션

운영사무국

2020. 11. 23



목 차

공개SW생태계가 모두 함께
대한민국 IT 미래를 향해

- I. 교육 개요
- II. 교육 소개
- III. 강사 소개
- IV. 교육 일정
- V. 협조 사항
- VI. 교육 자료



교육 개요

군(軍)현장, 대상별 기술역량 강화 및
문제해결 서비스 개발·활성화를 위한 프로젝트 중심의 교육과정 개발·운영

주최/주관

과학기술정보통신부·국방부 / 정보통신산업진흥원

교육 일정

2020년 11월 23일(월) ~ 11월 27일(금) (35H)

교육 장소

프론트원 7층 NIPA교육장(서울)

교육 대상

국직부대 소속 총 20명

교육 일정

선수교육

01

[활용] 실전 데이터
분석 및 인공지능
활용

교육수강

02

[초급] 인공지능과
데이터분석의 이해

교육수료

03

정보통신산업진흥원
온라인 수료증 발급

후속지원

04

온라인 교육 콘텐츠
Help Desk 운영





▶ [중급] AI 초급 개발자

교육소개

AI에 대한 전반적 지식을 이해하고, 머신 러닝의 구조에 대해 이해하고, 이를 코딩 및 테스트 실습을 통해 이해하고 구현한다.

교육기간

2020년 11월 23일(월) ~ 11월 27일(금) (35H)

학습목표

- AI의 개념과 기계학습에 대해 이해한다.
- 다양한 기계학습 모델을 살펴보고 활용한다.

프로그램

구 분	내 용
대상요건	python, numpy, pandas, matplotlib에 대한 경험과 이해
교육시간	<ul style="list-style-type: none">- 선수과정(온라인) : 실전 데이터 분석 및 인공지능 활용 (15h) ※ nipa.elice.io- 본 과정(집체) : 인공지능과 데이터 분석의 이해 3h + 이론 및 실습 32h
실습환경	Python 3.8, Anaconda 3.8, Jupyter notebook
강사구성	주강사 1명
평 가	퀴즈, 실습 예제 풀이
성과측정	만족도조사, 사전/사후 테스트

강사 소개

▶ 강사소개



윤재성

(주)소프트캠퍼스 대표

- 학력사항**
- 방송통신대학교 대학원 재학중
 - 방송통신대학교 학사

- 경력사항**
- (주) 소프트캠퍼스 대표이사
 - (주) 나르샤소프트 팀장
 - (주) 트리시스 팀장

- 강의활동**
- 전주정보문화산업진흥원 (머신러닝 기술을 활용한 웹 어플리케이션 전문 개발자 과정)
 - 대전 ETRI 파이썬 프로그래밍
 - 멀티캠퍼스 4차산업 딥러닝 (파이썬, 하둡, 스파크)
 - 멀티캠퍼스 청년취업 아카데미(파이썬을 활용한 데이터 분석 사이트 개발)
 - 부산대학교 혁신성장 빅데이터 (빅데이터를 활용한 데이터 분석 과정-- 파이썬, 하둡, R프로그래밍, 스파크)
 - 전주정보문화산업진흥원 (빅데이터를 활용한 데이터 분석 과정--파이썬, 하둡, R프로그래밍, 스파크)

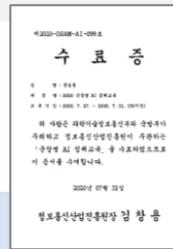
▶ 교육 일정표

일정		교육 내용
1일 11.23(월)	오전	인공지능 개념 및 동작원리의 이해
	오후	지도학습과 비 지도학습 및 분류와 회귀
2일 11.24(화)	오전	선형회귀, 로지스트 회귀, 결정트리, 랜덤포레스트
	오후	Naive Bayes, SVM
3일 11.25(수)	오전	인공신경망, KNN
	오후	K-means, 군집화
4일 11.26(목)	오전	PCA, NMF
	오후	t-SNE, 연관성 분석
5일 11.27(금)	오전	지도학습 응용 실습
	오후	비지도학습 응용 실습

협조사향

▶ 참가자 준수사항

- ✓ 입실 전 발열 체크, 교육장 내 마스크 착용
- ✓ 교육 기간 중 명찰 패용
- ✓ 교육시작 10분 전 입실
- ✓ 1회 무단 조퇴 시 결석처리, 2회 이상 무단 조퇴 시 제적 처리
- ✓ 핸드폰은 매너모드(진동)로 설정



잠깐만!

교육 전에는 꼭! **사전테스트**를 실시 해주세요!

교육 후에는 꼭! **사후테스트**

※ 사전역량진단 및 사후역량진단 미참여시 수료증 발급 불가

[사전역량진단 및 사후역량진단 참여방법]

- 교육 첫날과 마지막날 링크 공유하여 실시

평가 방법

교육 수료 조건

- 별도로 성적을 산출하지 않으나 교육시간 총계 이수시간(35시간)의 **80% 이상** 수료증 발급
※ 출결정보는 매일 국방부에 전달 예정
- 지각 / 조퇴 / 외출 3회 이하
- 온라인 선수과목 이수 권고

설문 및 후기

만족도 조사

현장에서
공지된 링크를 통해
접속하여 참여

교육 후기

<https://aiam.kr>

사이트 접속 후 강의 선택
↓
교육후기

06 교육 자료

중급5회

교육자료

AI 초급 개발자

윤재성

※ 본 강의자료는 저작권보호법에 의거 외부유출 및 무단배포를 엄격히 금지합니다.
위반시 법적 책임을 받을 수 있음을 알려드립니다.

4차 산업 혁명

4차 산업혁명

- ▶ 기업들이 제조업과 정보통신기술을 융합해 작업 경쟁력을 제고하는 차세대 산업혁명
 - 클라우드 슈밥

“모든 것이 연결되고 보다 지능적인 사회로의 진화”

- 다보스 포럼, 2016 -



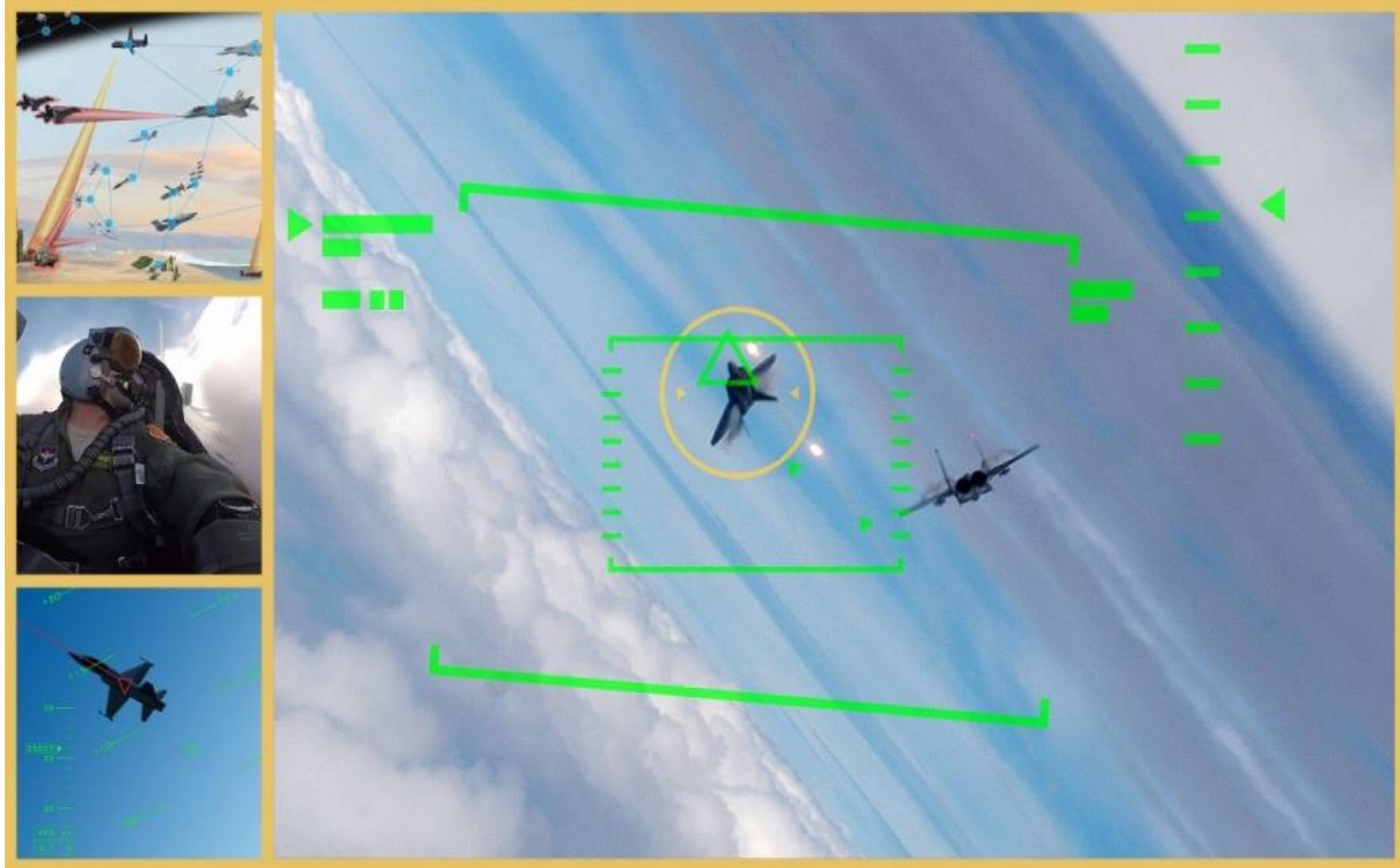
DNA (Data – Network – AI)



알파고 쇼크



인공지능 조종사 (5:0) 인간 조종사[전투기]



인공지능의 역사

	일어난 일	주목을 모았던 것	등장한 과제
제1차 붐 (1956년~ 1970년쯤)	<ul style="list-style-type: none"> • 다트머스 회의가 개최되다 • 그 회의에서 'Artificial Intelligence(=인공지능)'이라는 표현이 처음으로 사용되다 	<ul style="list-style-type: none"> • 기호 처리 분야 	<ul style="list-style-type: none"> • 인공지능은 사회에 도움이 될 수 있을 것인가? • 프레임 문제(사고범위 문제)에 어떻게 대처할 것인가?
제2차 붐 (1980년~ 1995년쯤)	<ul style="list-style-type: none"> • 인공지능을 활용한 시스템이 의사보다 높은 성과를 내다 • 일본에서 '제5세대 컴퓨터 프로젝트'가 시작된다 	<ul style="list-style-type: none"> • 지식 표현 분야 • 신경망 • 전문가 시스템 	<ul style="list-style-type: none"> • 인공지능에 지식을 가르치기 위한 비용은 어떻게 할 것인가? • 명확하게 적어낼 수 없는 지식은 어떻게 할 것인가? (지식 습득의 병목)
제3차 붐 (2010년쯤 ~현재)	<ul style="list-style-type: none"> • 인공지능의 산업 응용이 진행되다 • 인공지능이 '고양이'의 특징을 자동으로 습득하다 ('Google의 고양이') • 인공지능이 바둑/장기로 인간에게 승리하다 	<ul style="list-style-type: none"> • 기계학습(딥러닝) • AlphaGo • 휴먼 컴퓨터이션 	<ul style="list-style-type: none"> • 인공지능 학습에 필요한 데이터를 어떻게 모을 것인가?

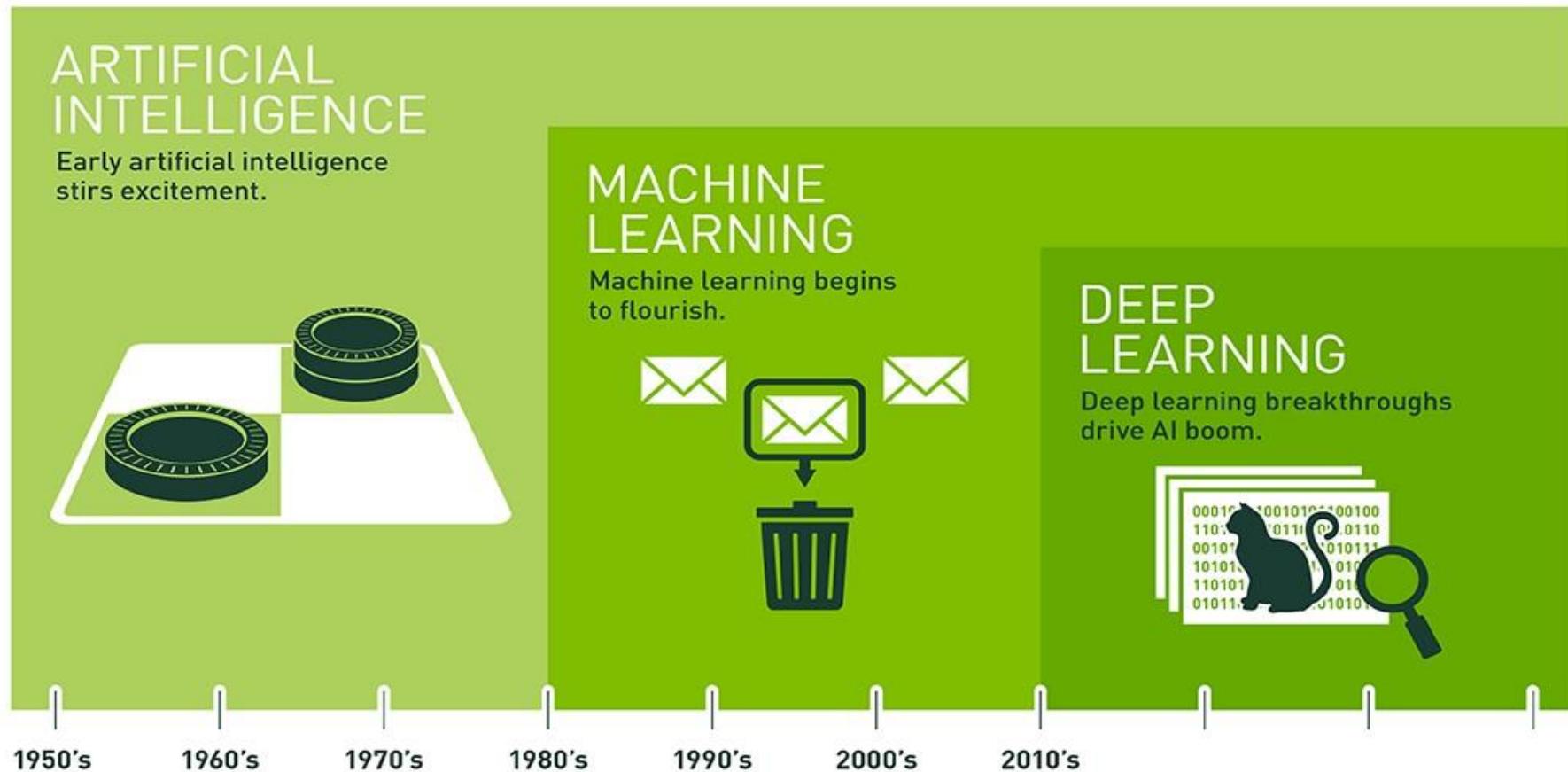
▲ 그림 2-5 인공지능의 역사

인공지능 사례 및 발전사

- 2011년 미국의 유명한 TV 퀴즈 쇼 제퍼디에서 IBM 왓슨 우승
- 왓슨(Watson) 의료 : 가천대 병원에서 활용 중, 의사의 판단과 왓슨의 판단이 다를 때 환자들이 왓슨을 더욱 신뢰한다고 함.
- 로스(ROSS, 왓슨(Watson)과 연계된 인공지능(AI) 변호사) : 미국 로펌에서 변호사 업무 중
- 2016년 알파고 : 이세돌 9단에게 4대 1로 승리
- 2017년 알파고 제로(제로에서 학습) : 학습 시작 36시간 만에 알파고를 능가, 72시간(490만판)을 학습한 뒤 100전 100승
- 2020년 인공지능 전투기 조종사 5대 0으로 인간 조종사에 승리

Artificial Intelligence Machine Learning and Deep Learning

인공지능 vs 머신러닝 vs 딥러닝



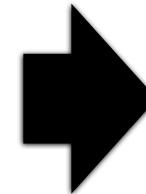
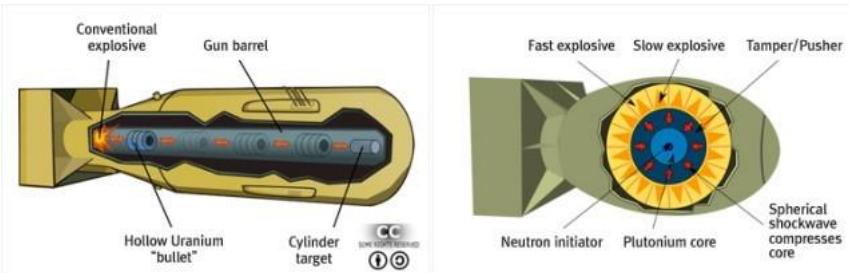
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Manhattan Project

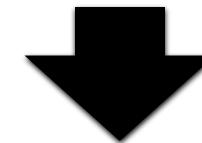
▣ 맨해튼 프로젝트 (Manhattan Project)

- 히틀러 나치 정권에 앞서 원자폭탄을 만들기 위한 미국의 비밀 프로젝트
- 미국 내 최고 과학자들이 모두 참여하여, 원자폭탄을 만드는 연구에 참여

<최초의 원자폭탄 구성>



- 연구비 20억 달러
- 고용인원 최대 13만명



태평양 전쟁 승리
냉전 체제에서 미국의 우위



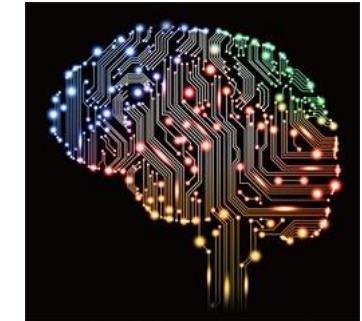
AI Manhattan Project of Google

□ 인공지능 맨해튼 프로젝트 (AI Manhattan Project)

- 인공지능 기술 개발을 위한 구글의 세기적 프로젝트
- 관련 IT 기업과 전문가들을 확보하기 위한 거대 인수합병을 진행

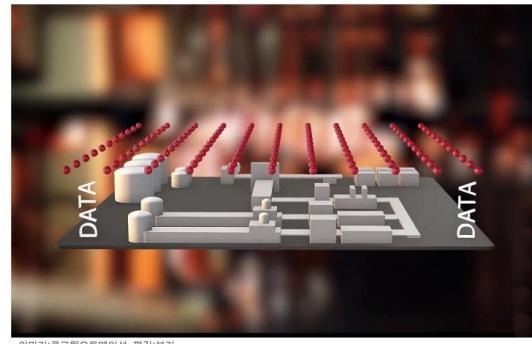


- 1000명 이상의 인력이 구글의 인공지능/머신러닝 연구 진행 중 (아주경제)
- 2014년 1월 Deep Mind 인수 (6억 5000만 달러)
- 2014년 1월 Nest Labs 인수 (32억 달러)
- 기타 50여 개의 인공지능 기업들이 맨해튼 프로젝트에 참여, 연구 진행



□ 현재 인공지능 기술은 분야를 가리지 않고 다양하게 이용되고 있음

<스마트 팩토리 적용 사례>



이미지: 로크월드미션, 편집: 본지

<구글의 자율주행자동차>



<금융 지표 예측>



<영유아 발달 관리>



인공지능으로 영유아 발달 관리 하는 '써모케어 AI 헤칭(사진: 엔트리케어)

<핀테크 업체 챗봇 예시>



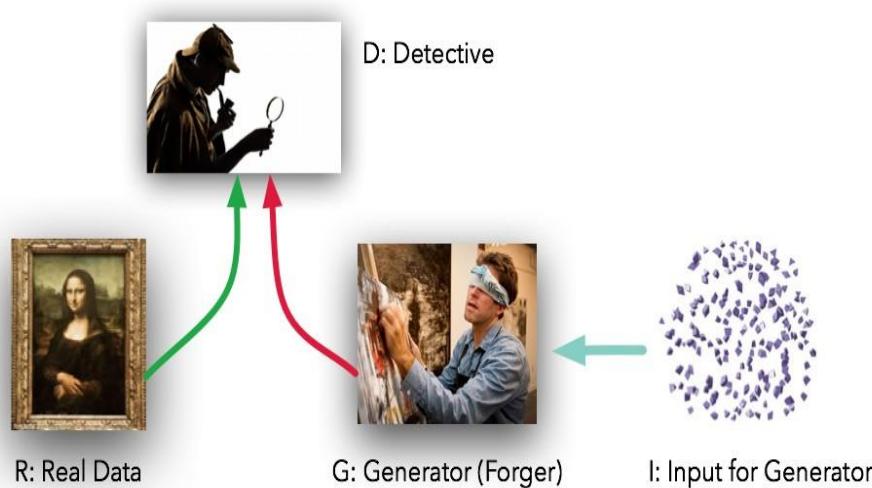
<음악 작곡 등 예술 창작>



flowmachines

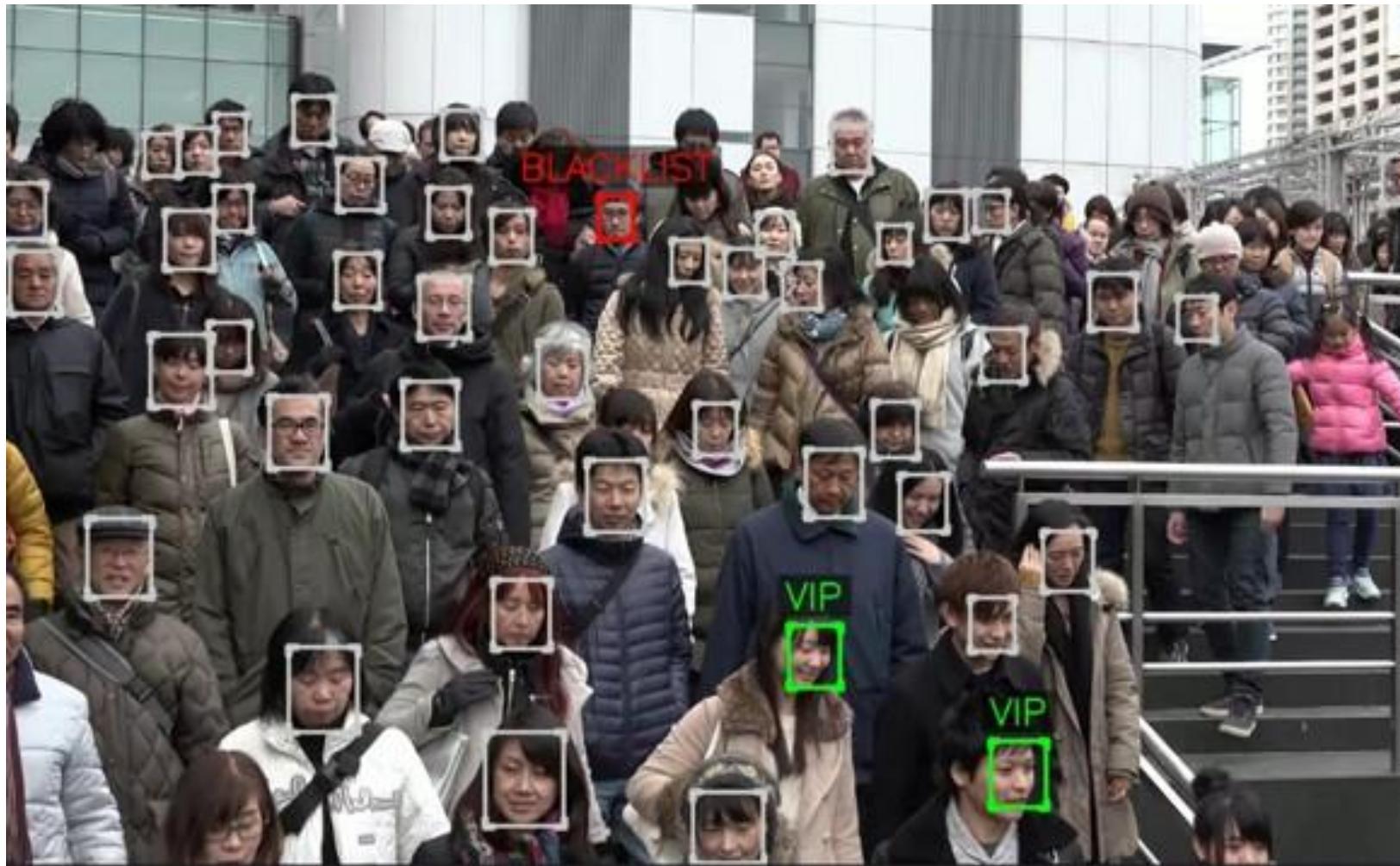
Generative Model Examples

□ GAN (Generative Adversarial Networks)



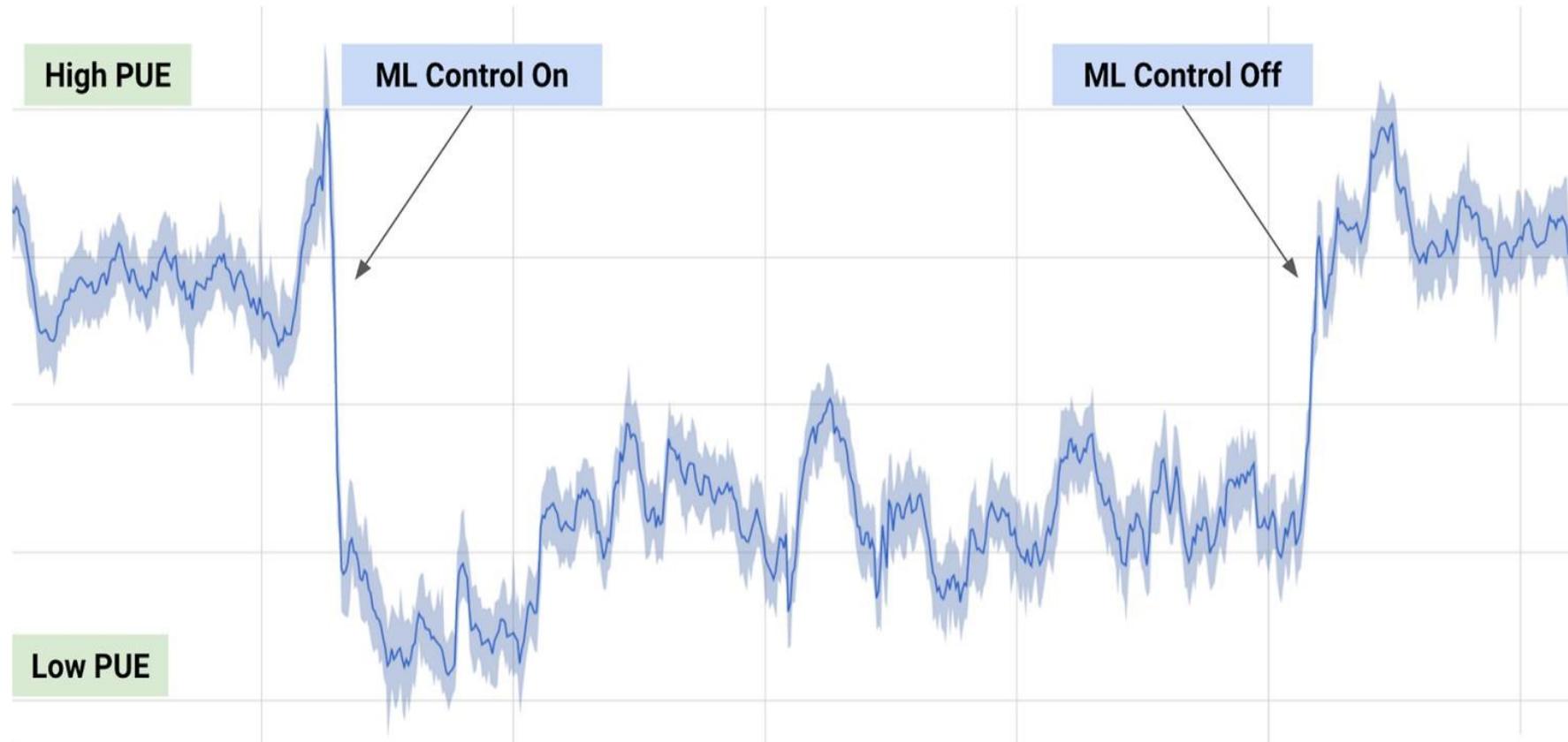
Vision Application of Deep Learning

□ Image Detection and Recognition



Reinforcement Learning for Data Center Control

- 데이터 센터의 발열량을 조절하기 위해서 인공지능 기술이 적용되기도 함

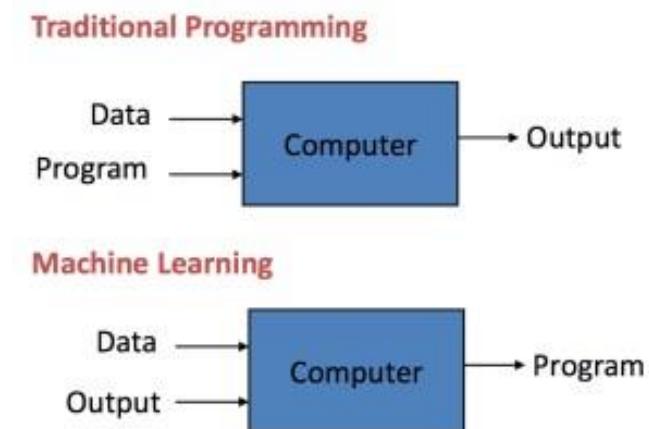
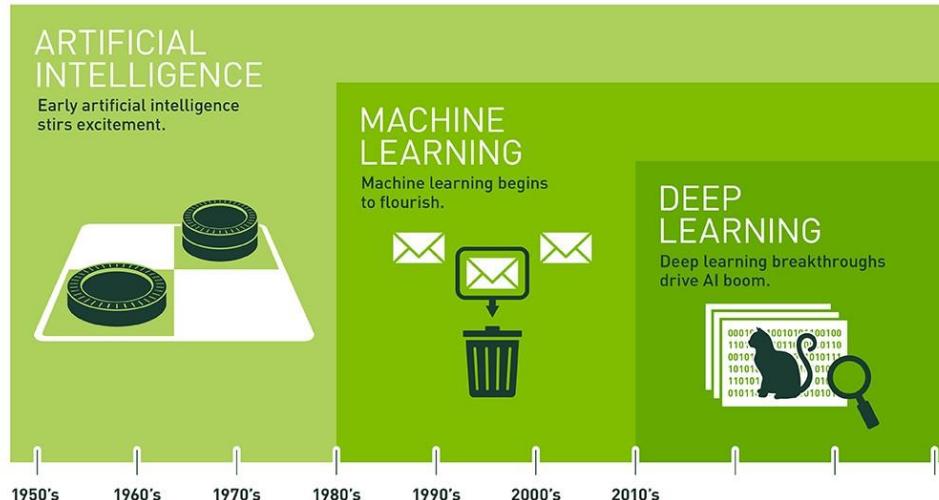


Machine Learning

□ Machine Learning (기계학습)

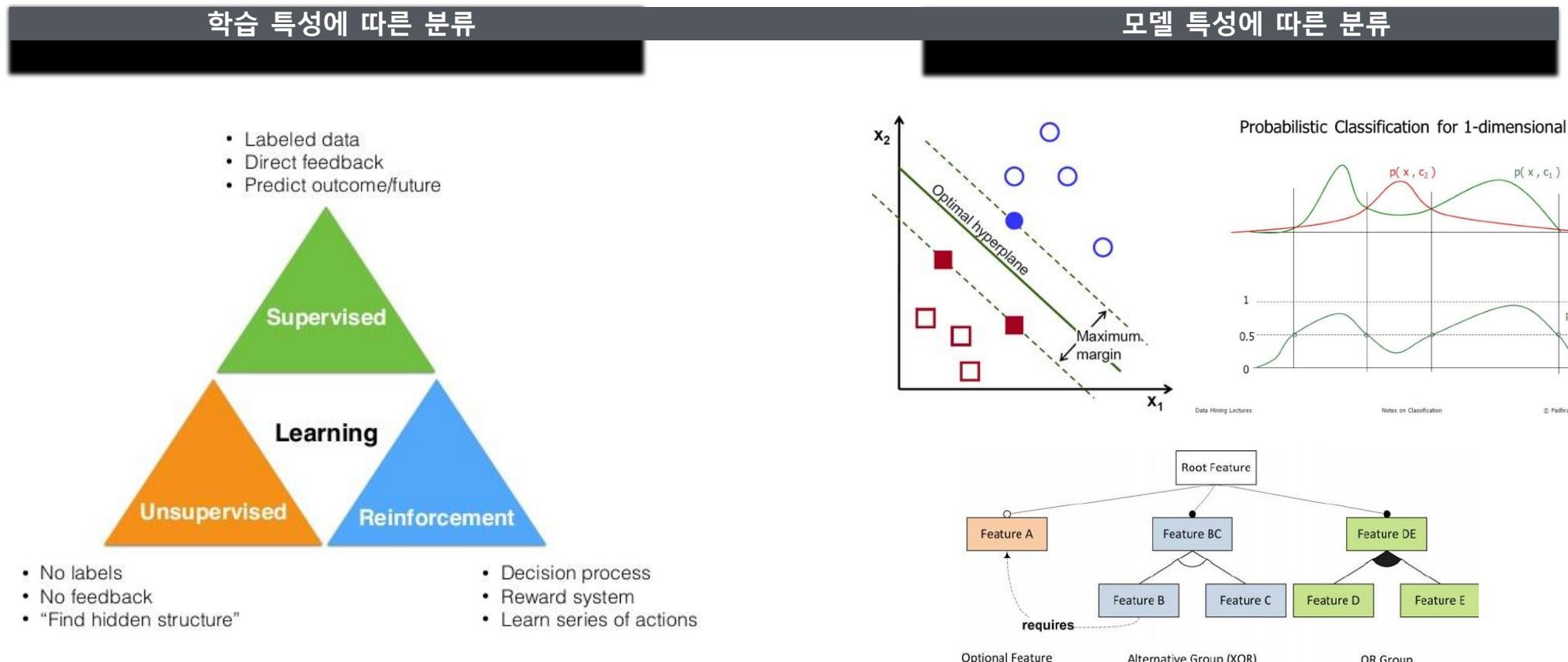
- 인공 지능의 한 분야. 컴퓨터가 학습할 수 있게 하는 알고리즘 분야 (Wikipedia)
- Field of study that gives computers the ability to learn **without being explicitly programmed**. (Arthur Samuel, 1959)
- A computer program is said to **learn from experience E** with respect to **some task T** and some **performance measure P**, if its performance on T, as measured by P improves with experience E. (Tom Mitchell, 1998)

<AI & Machine Learning 차이>



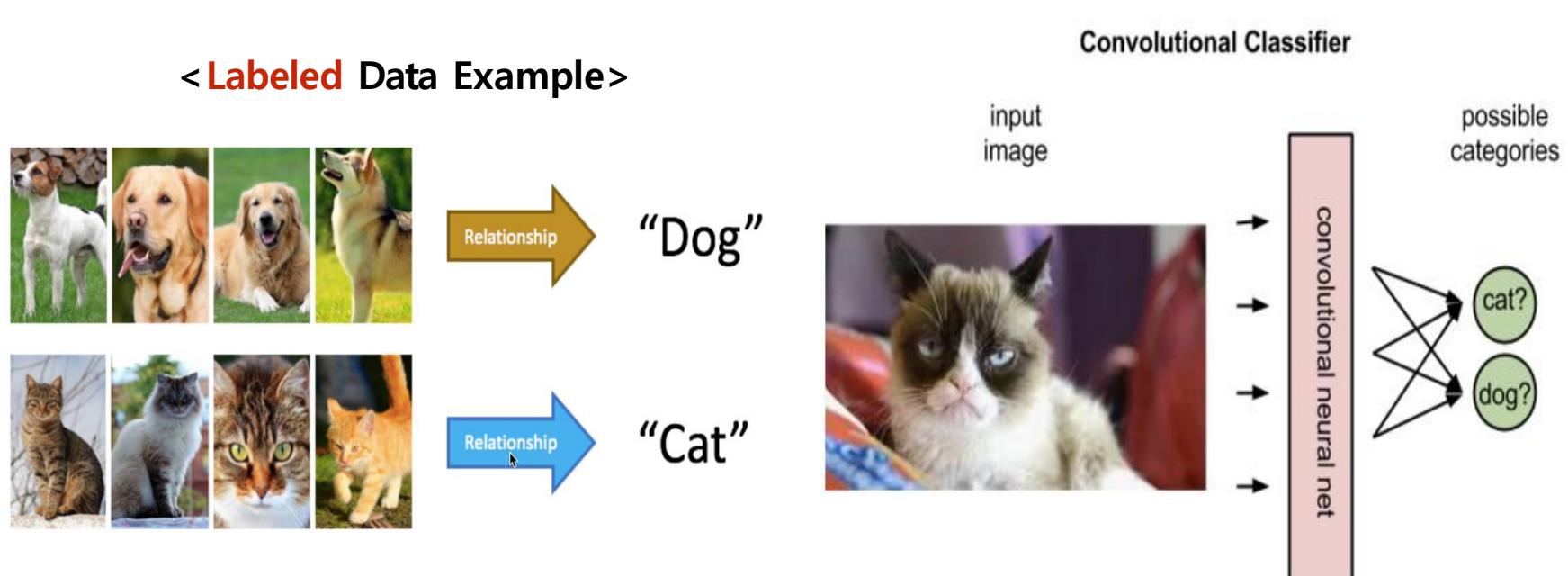
Types of Machine Learning

- 머신러닝은 다양한 기준으로 세부 분류가 가능함
- 학습 특성에 따른 분류: 1) Supervised, 2) Unsupervised, 3) Reinforcement Learnings
- 모델 특성에 따른 분류: 1) Geometric, 2) Probabilistic, 3) Logical Models



Types of Machine Learning

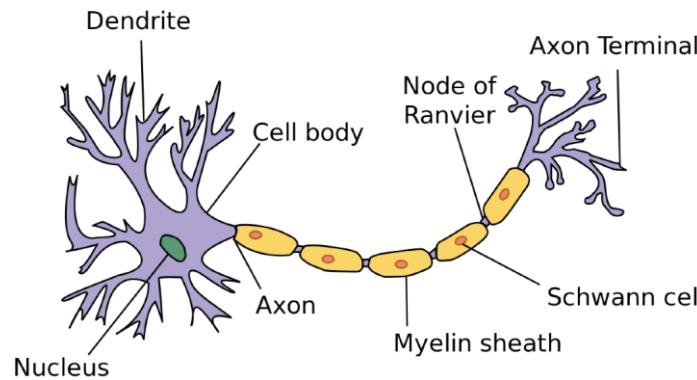
- 머신러닝은 다양한 기준으로 세부 분류가 가능함
- 학습 특성에 따른 분류: 1) Supervised, 2) Unsupervised, 3) Reinforcement Learnings
- 모델 특성에 따른 분류: 1) Geometric, 2) Probabilistic, 3) Logical Models



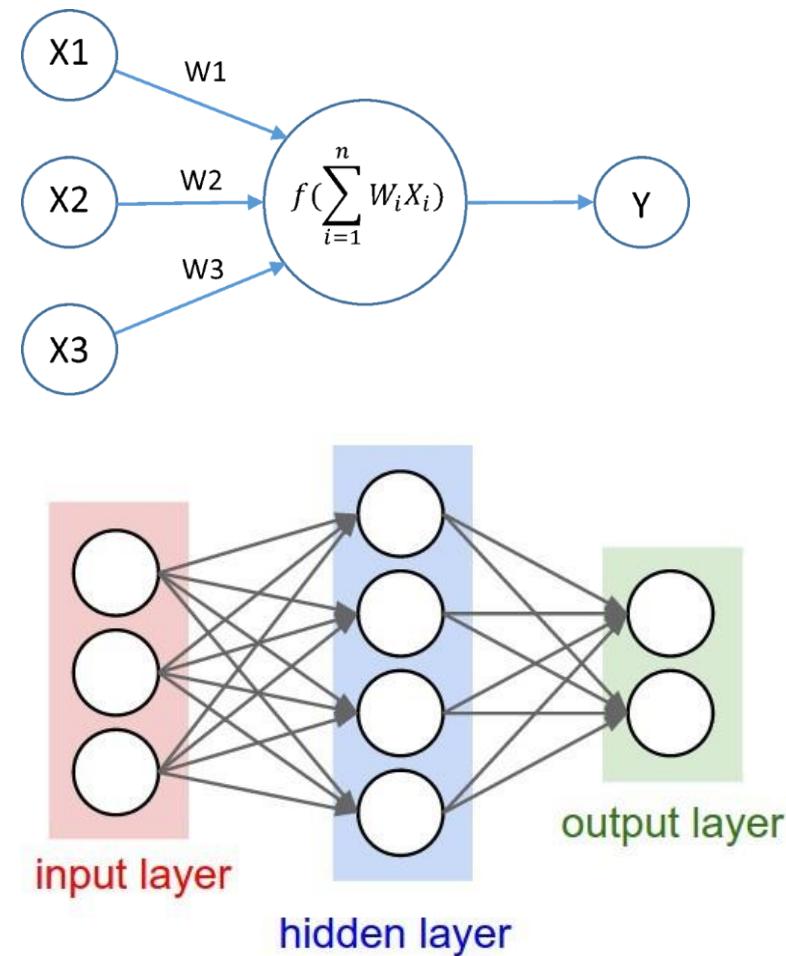
Neuron and Neural Network

- 인간의 뇌는 뉴런을 단위로 뉴런의 연결을 통해 복잡한 계산을 효과적으로 수행함

<Typical Cortical Neuron >



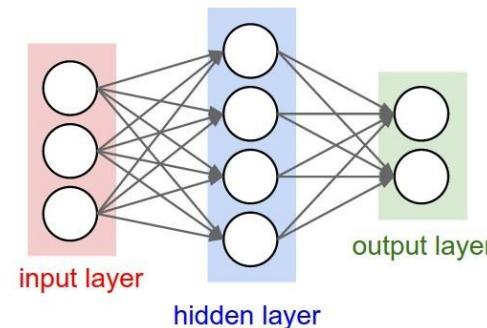
<Overview of Neural Network>



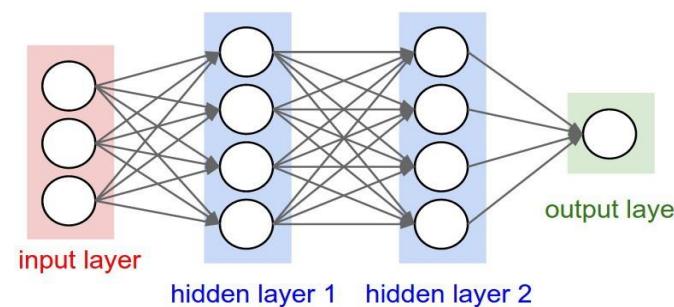
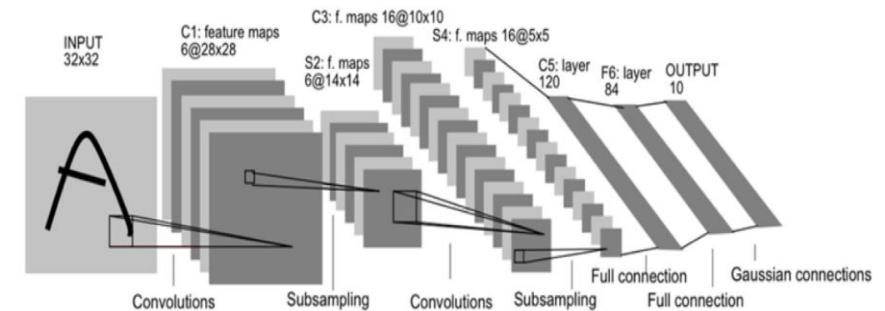
□ Deep Learning

- "Deep" Neural Network를 활용하여 학습하는 머신러닝 알고리즘을 의미
- Deep : Hidden Layer의 수가 **2개 이상**일 경우

<DNN 예시(Below)>



<Convolutional Neural Network LeNet5>

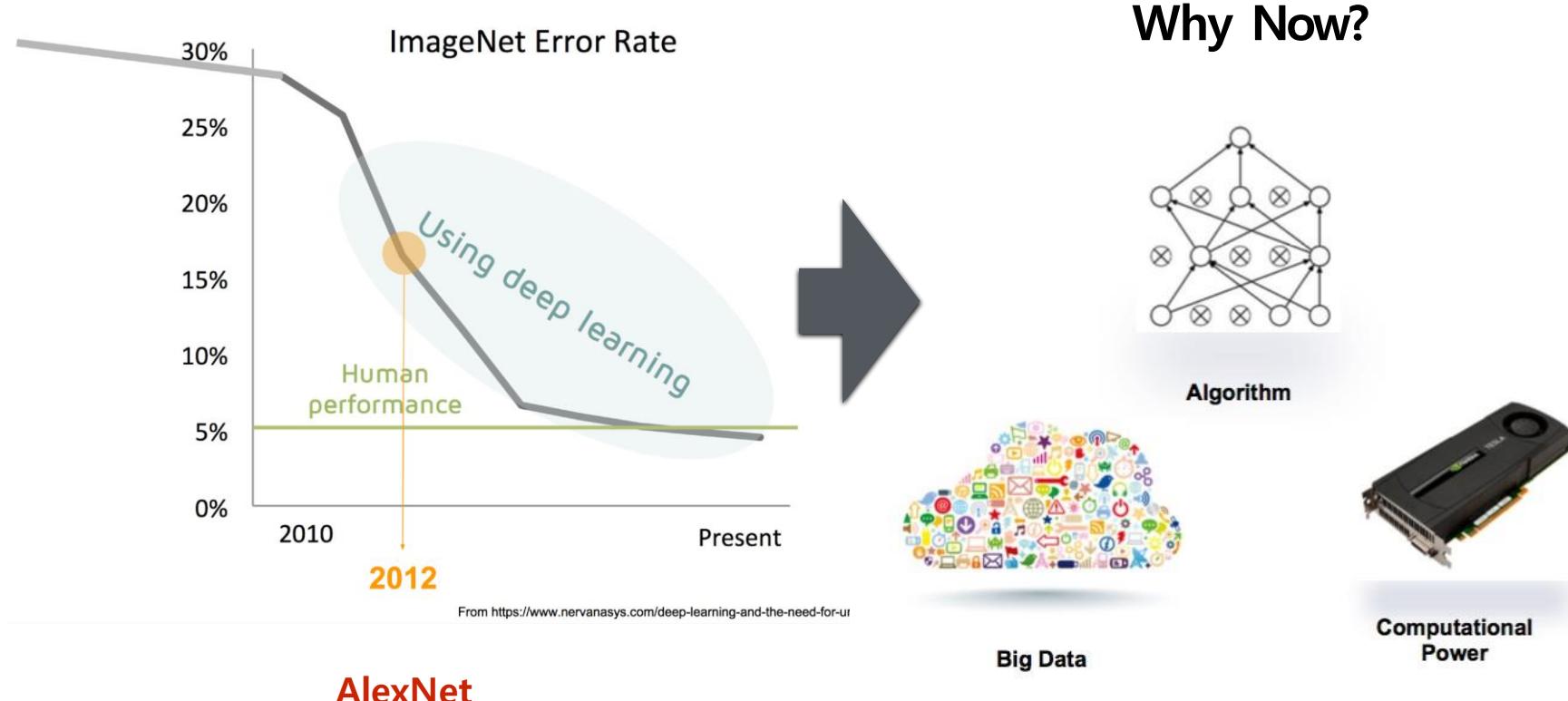


<Object Detection>



□ Deep Learning 알고리즘 자체는 최신이 아님

- Alexnet 2012 based on CNN (LeCunn, 1989)
- Alpha Go based on RL and MCTS (Sutton, 1998)

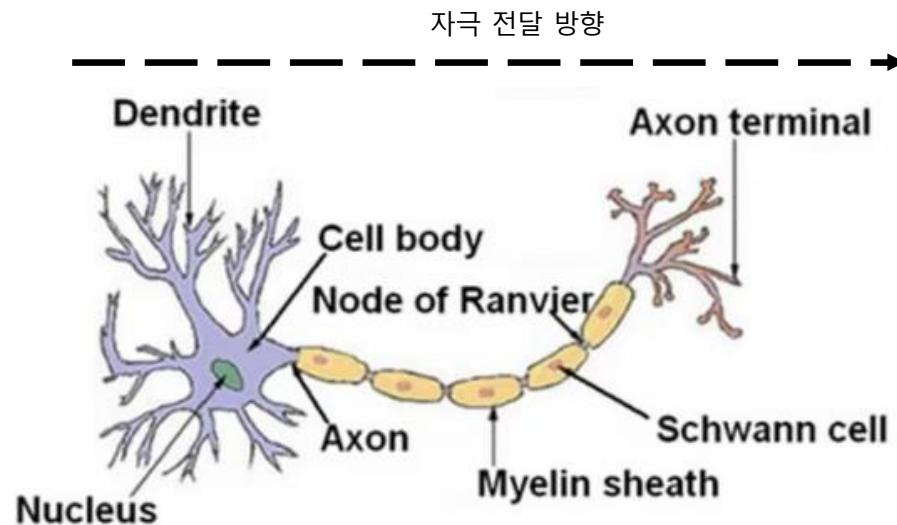


Principle of Cortical Neuron

□ Neuron (신경세포, 뉴런)

- 신경계를 구성하는 구조적 및 기능적 단위이며, 전기적인 방법으로 신호를 전달함
- 수상돌기(Dendrite), 핵 (Nucleus), 축색돌기(Axon) 등으로 구성됨

<Cortical Neuron Structure>



Dendrite:

이전 뉴런의 전기적 자극을 입력받는 수용체

Nucleus & Cell Body:

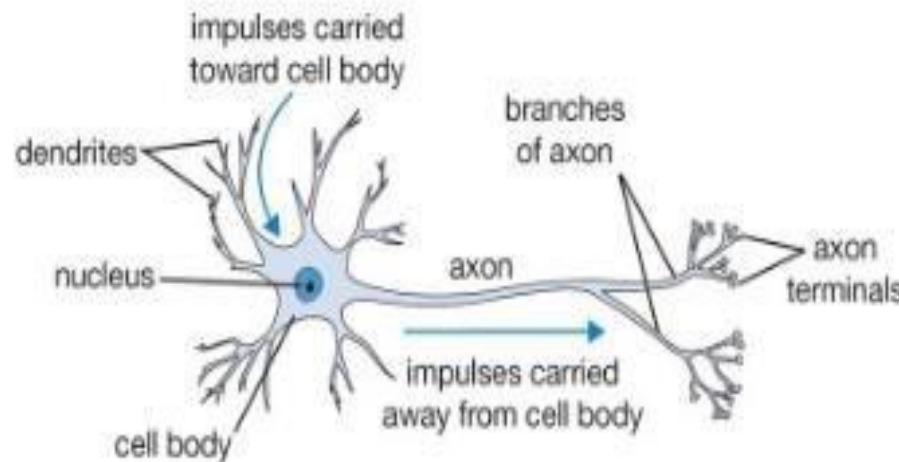
수용된 자극을 바탕으로 다음 뉴런에게 전달한 자극을 생성

Axon:

Cell body로부터 생성된 자극을 발생(Spike)시키고 다음 뉴런에게 전달

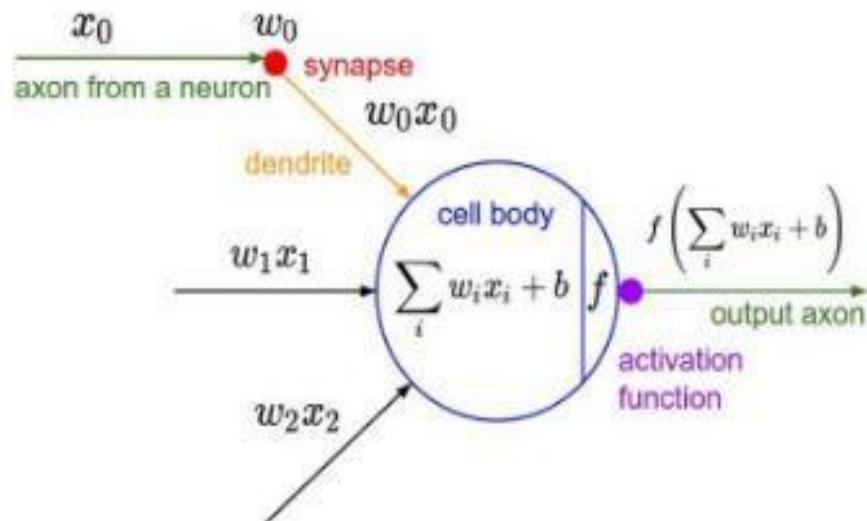
뇌가 처리하는 복잡한 인지적 작용들은 모두 간단한 계산의 뉴런들의 연결을 통해 이루어짐

Principle of Cortical Neuron



인간의 뉴런(neuron)

- 시냅스(synapse)를 통해 뉴런간 신호를 전달
- 각 뉴런은 수상돌기(dendrite)를 통해 입력 신호를 받음
- 입력 신호가 특정 크기(threshold) 이상인 경우에만
- 활성화되어 축삭돌기(axon)을 통해 다음 뉴런으로 전달



인공 뉴런: 노드(node)

- 각 노드는 가중치(weight)가 있는 입력 신호를 받음
- 입력신호는 모두 더한 후, 활성화 함수(activation function)을 적용함
- 활성화 함수의 값이 특정 값 이상인 경우에만, 다음 노드의 입력값으로 전달

Awesome Characteristics of Neuron

- 인간의 뇌에 있는 뉴런의 작동 및 계산 방식은 컴퓨터 프로그램 등의 방식에 비해 다양한 장점이 있음
- 1) 높은 효율/효과의 계산 속도와 2) 모듈 단위의 적응(Adaptation)이 대표적인 장점으로 꼽힘

<효율적인 계산 속도>

뉴런의 연결인 시냅스는 매우 적은 전력으로 신호를 주고 받음

인간의 뇌는 약 10^15 개의 뉴런을 가지고 있으며, 한 뉴런당 평균 10^4 개의 시냅스를 가지고 있음
이는 매우 높은 차원의 Neural Network 모델의 Weight 개수보다 훨씬 많은 개수임

Human's Synapse

10^{15}

Neural Network's Synapse (Max)

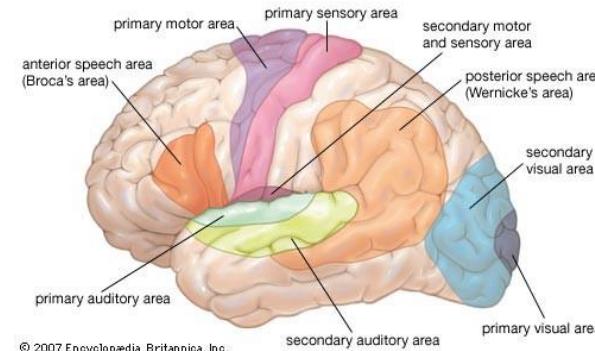
>

10^{11}

<모듈 단위의 적응(Adaptation)>

Brain Learning Process Experiment

- 대뇌 피질의 시각을 담당하는 부분(Primary visual area)가 손상되어 시신경(Optic nerve)를 Auditory cortex에 연결
- 뉴런의 적응 이 후, 시각 자극에 대해 기존 Auditory cortex 부분이 새롭게 시각 자극에 대한 반응을 학습함

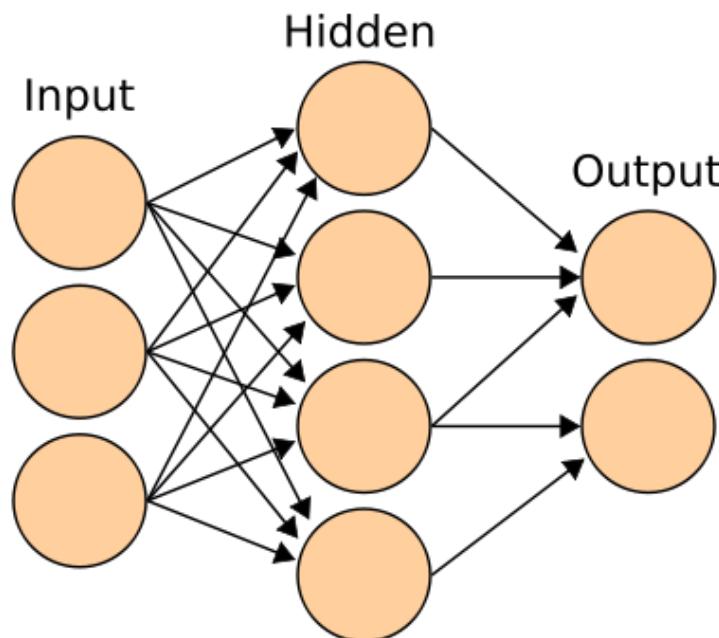


Concept of Neural Network

□ Neural Network (인공신경망)

- 신경세포의 간단하고 효과적인 처리 방식에 착안해 구현된 머신러닝 알고리즘의 한 종류
- 다량의 뉴런(Neuron, Unit)들이 층(Layer)으로 연결되어 간단한 계산과 연결을 통해 복잡한 문제를 해결함

<Neural Network Example>



Neuron: 각 뉴런이 **계산한 결과** 데이터 (Value)를 의미

Synapse : 뉴런 결과값 사이의 **weight** 값을 의미

Input Layer:

초기 입력값을 받는 가장 첫번째 Layer (# of Hidden = 1)

Hidden Layer:

중간 단계의 모든 Layer를 의미함 (# of Hidden >= 1)

Output Layer:

가장 마지막 단계의 Layer로 모델의 출력값을 계산
(# of Output = 1)

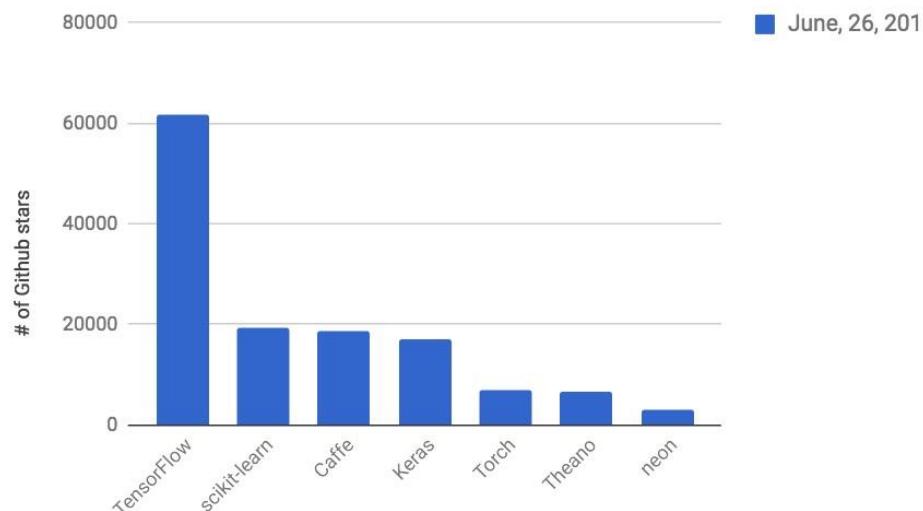
Deep Learning Library

- 머신러닝과 딥러닝을 구현하기 위해 다양한 개발 환경이 존재함
- 모델 구현을 위해 다양한 언어를 사용할 수 있지만, 대개 특정 언어에서 오픈 라이브러리를 이용함
- 주요 머신러닝 모델에 따라 사용하는 딥러닝 라이브러리는 달라질 수 있음

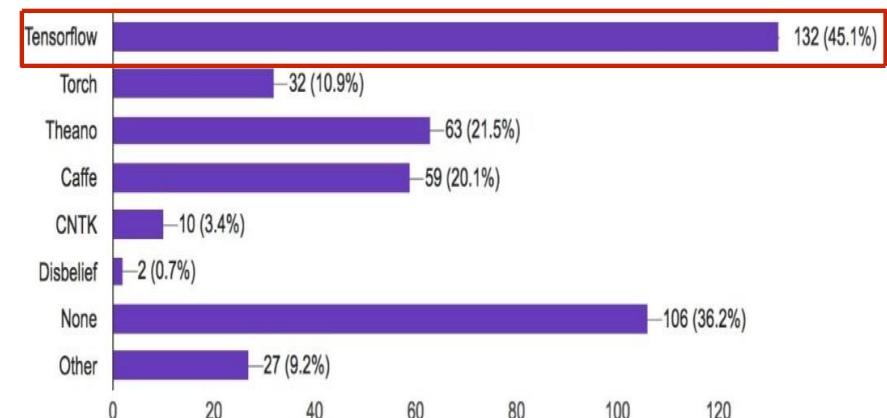
<Various Software and Library for Deep Learning>



Github Stars per Deep Learning Github Repository (as of June 2017)



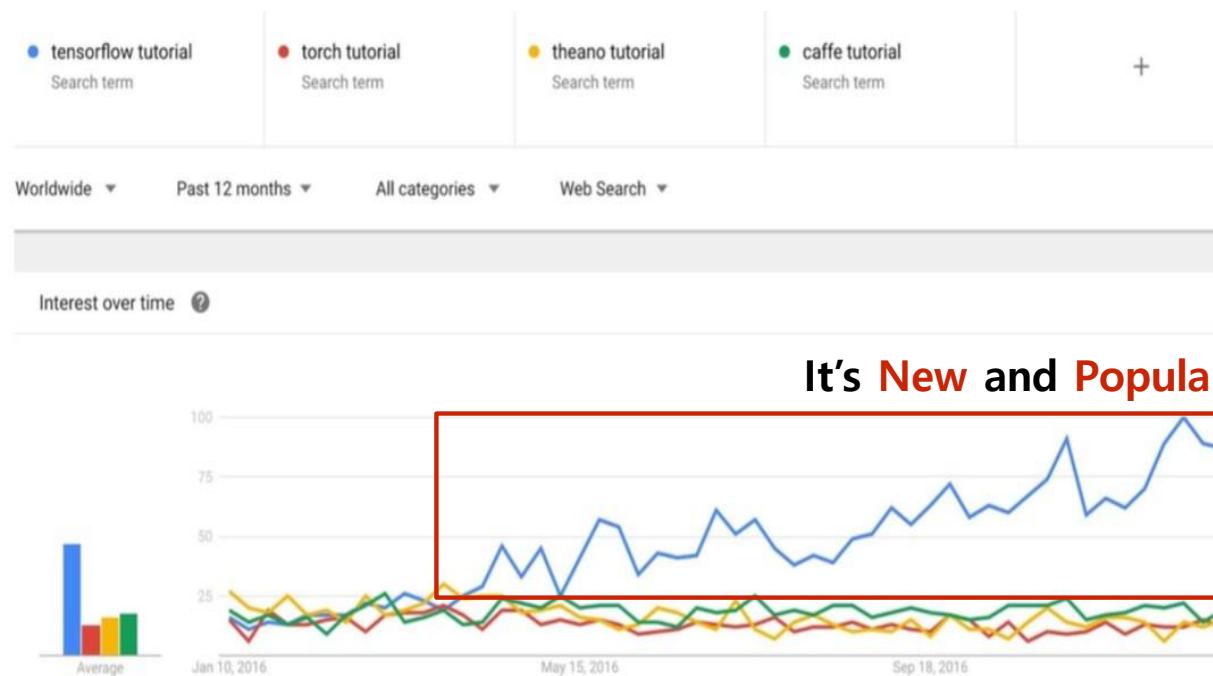
<Stanford Students' Deep Learning Library Usage Experience >



Usage Tendency of TensorFlow

- TensorFlow는 다른 오픈 라이브러리에 비해 늦게 출시되었지만, 현재 가장 유명한 Deep Learning 라이브러리

<Search Trend of Each Keywords>



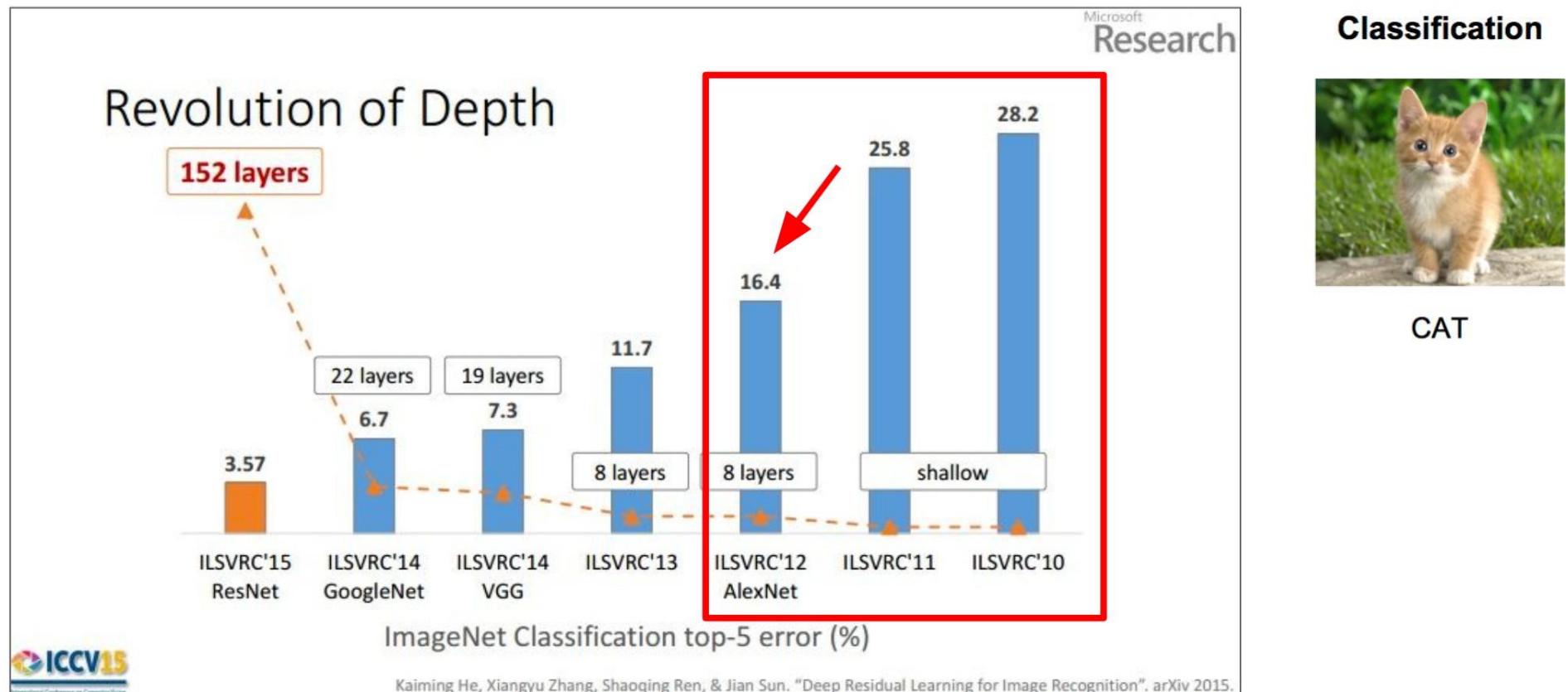
<Companies using TensorFlow>



Then, What is on earth **TensorFlow**?

Applications on Image Classification

- ❑ After AlexNet (2012 Winner), various architectures were designed with Conv. Net.
- ❑ With computational power of GPU, the model tends to be deeper and deeper.

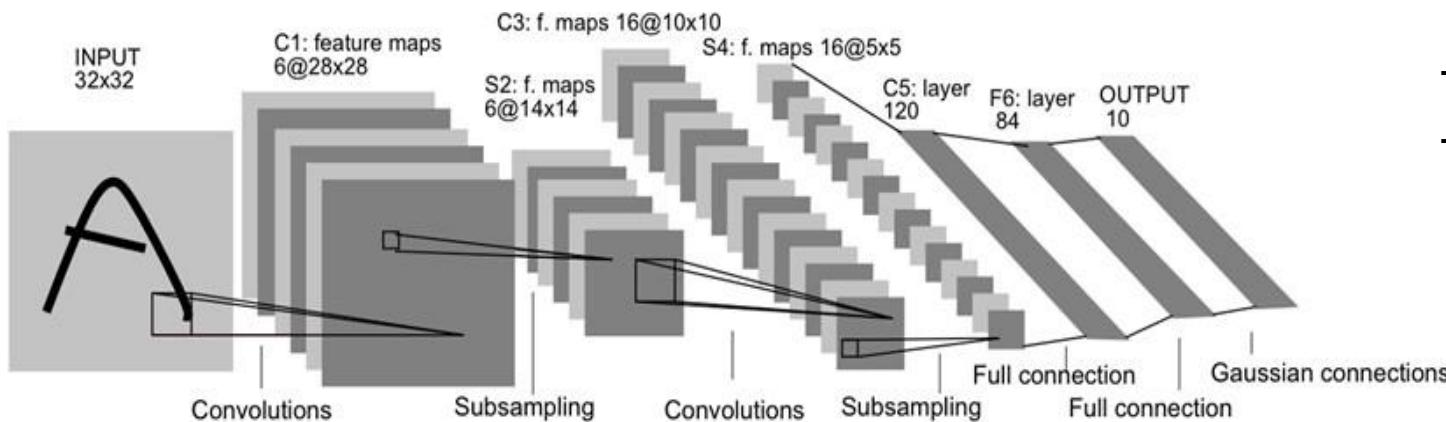


(slide from Kaiming He's recent presentation)

LeNet-5

□ LeNet-5

- Early convolutional neural network architecture (Le Cunn, 1998)
- LeNet-5 was used in order to recognize handwritten digit images.

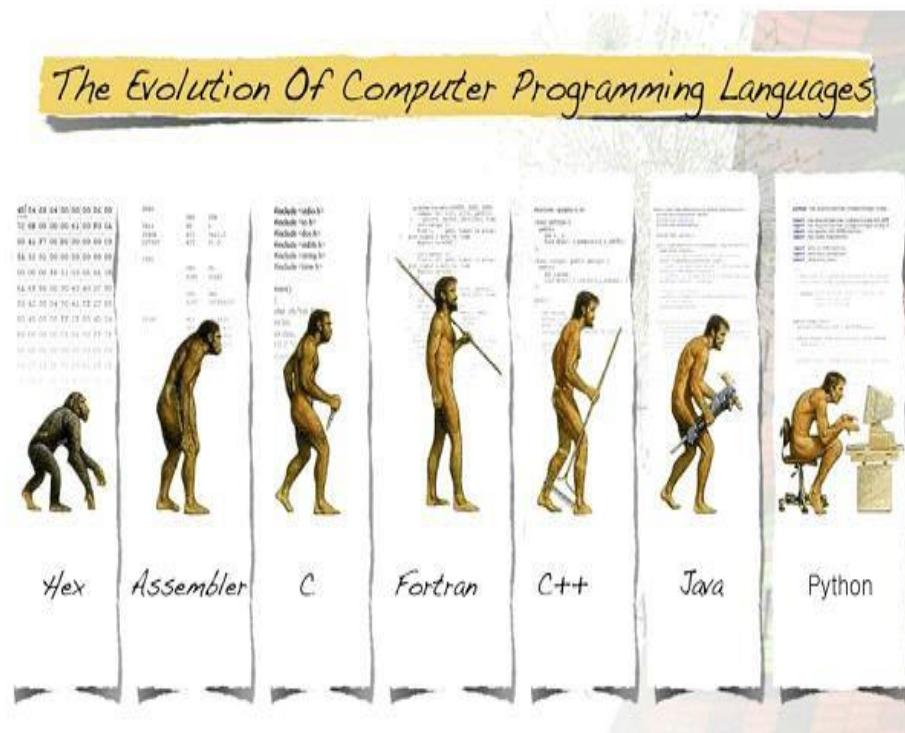


- **Sigmoid Activation**
- **Subsampling means only reduction of size, not max pooling**

An early (Le-Net5) Convolutional Neural Network design, LeNet-5, used for recognition of digits

파이썬 프로그래밍 언어

프로그래밍 언어의 발전사



파이썬 순위 (3위)

	Jul 2020	Jul 2019	Change	Programming Language	Ratings	Change
1	2	3	▲	C	16.45%	+2.24%
2	1	2	▼	Java	15.10%	+0.04%
3	3	1	▼	Python	9.09%	-0.17%
4	4	5		C++	6.21%	-0.49%
5	5	4		C#	5.25%	+0.88%
6	6	7		Visual Basic	5.23%	+1.03%
7	7	8		JavaScript	2.48%	+0.18%
8	20	19	▲	R	2.41%	+1.57%
9	8	9	▼	PHP	1.90%	-0.27%
10	13	12	▲	Swift	1.43%	+0.31%
11	9	10	▼	SQL	1.40%	-0.58%
12	16	17	▲	Go	1.21%	+0.19%
13	12	13	▼	Assembly language	0.94%	-0.45%
14	19	20	▲	Perl	0.87%	-0.04%
15	14	15	▼	MATLAB	0.84%	-0.24%

파이썬의 특징

파이썬 (Python)

1991년 구도 반 로섬 (Guido van Rossum)이 개발
초보자가 쉽게 배울 수 있는 프로그래밍 언어

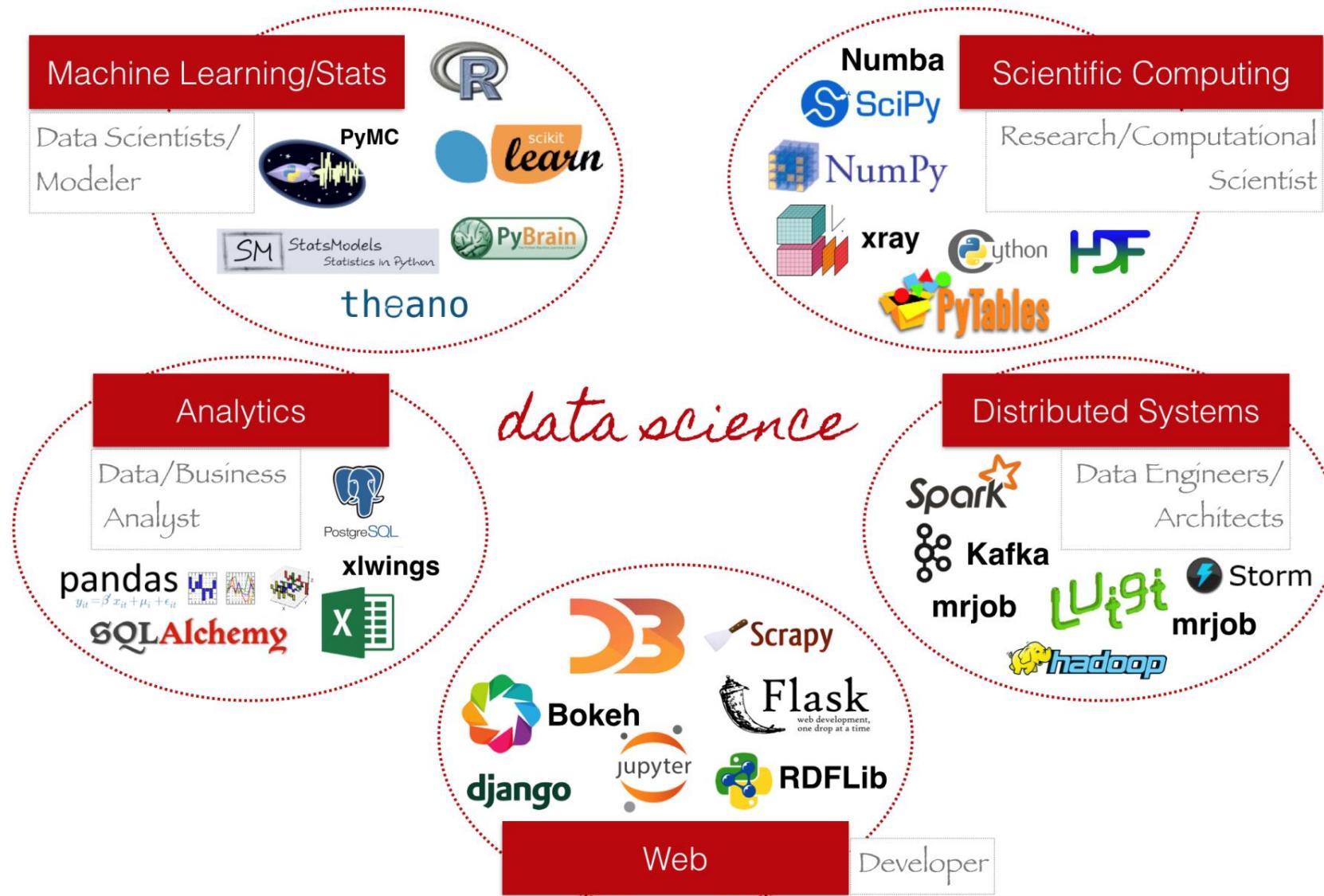
파이썬의 장점

문법이 쉬워 빠르게 학습할 수 있다 (비전공자도 쉽게 배울 수 있음)
간결하다
다양한 분야에서 활용할 수 있음
대부분의 운영체제에서 동일하게 사용됨

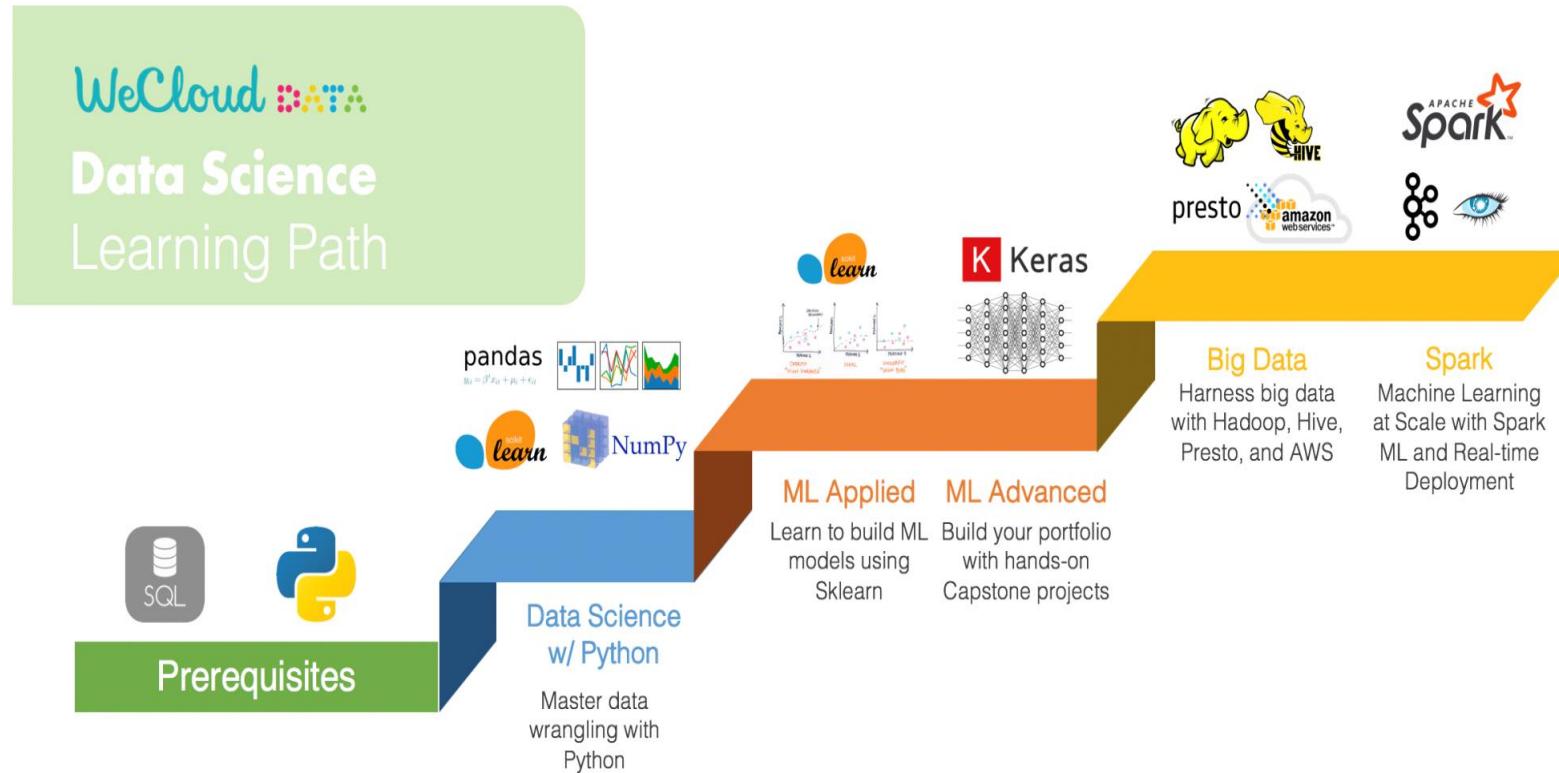
파이썬의 단점

C언어에 비해 일반적으로 느림
최근에는 컴퓨터 성능이 좋아져서 연산이 많이 필요한 프로그램이 아니라면 차이 크게 느낄 수 없음

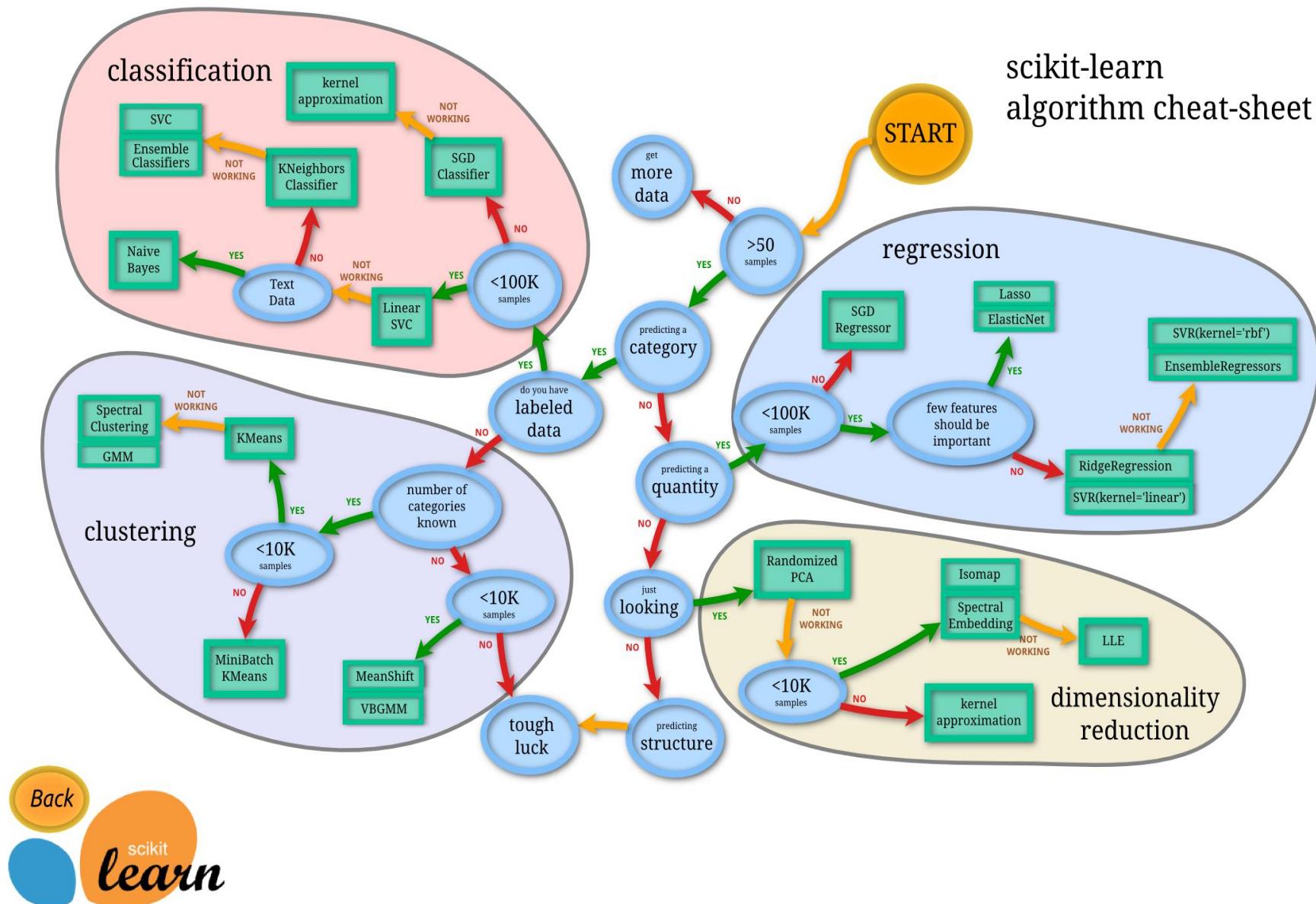
파이썬으로 할 수 있는 일들



파이썬 데이터 사이언스 러닝 패스



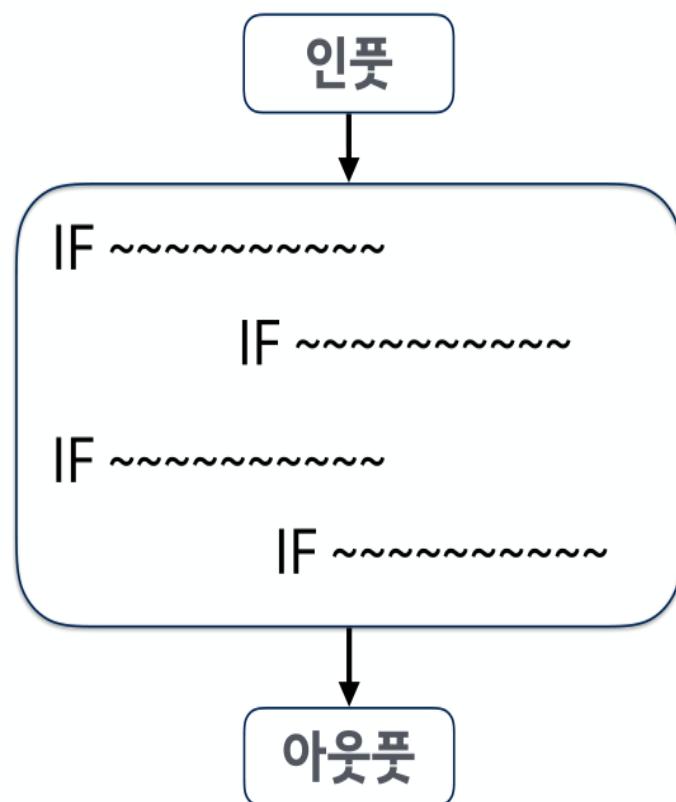
파이썬 머신러닝의 이해



패러다임 쉬프트

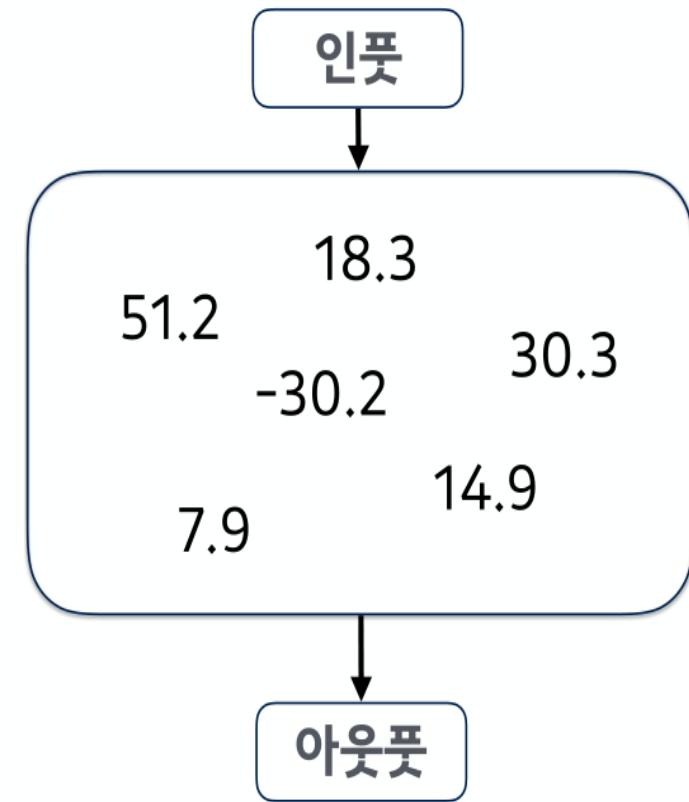
기존의 코딩:

조건을 라인바이라인으로 **긴 코드**로 써내려 가는 일



앞으로의 코딩:

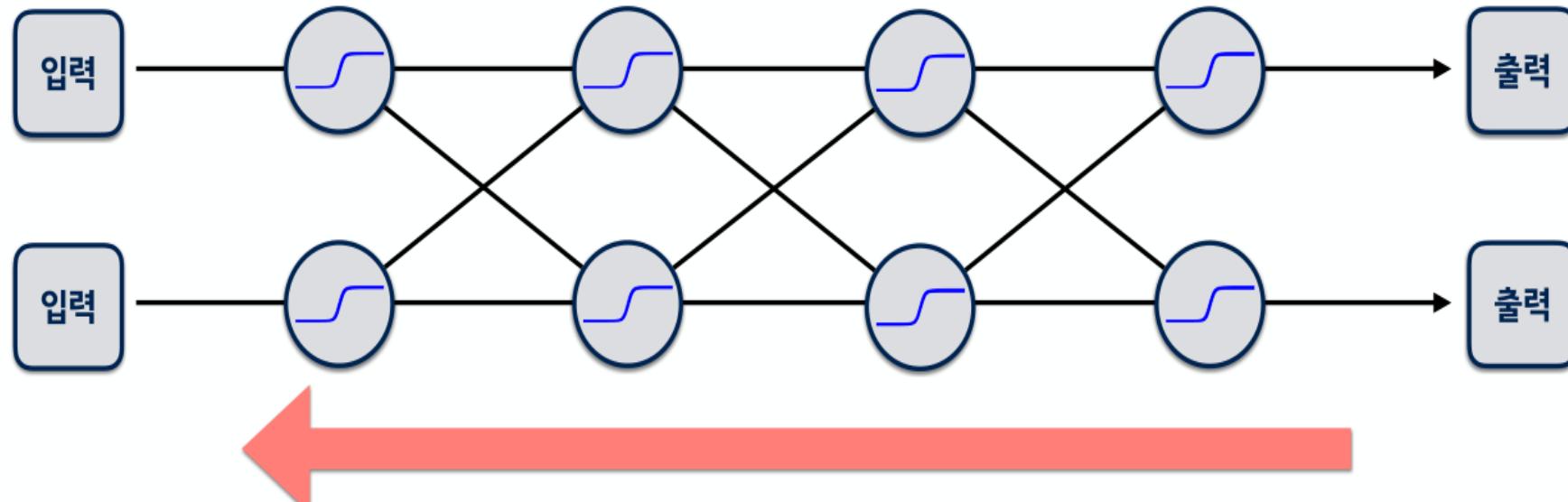
조건을 학습 **모델의 여러 가중치로 변환**하는 일



인공신경망의 학습 방법 : Back propagation

뭐를 전달하는가?

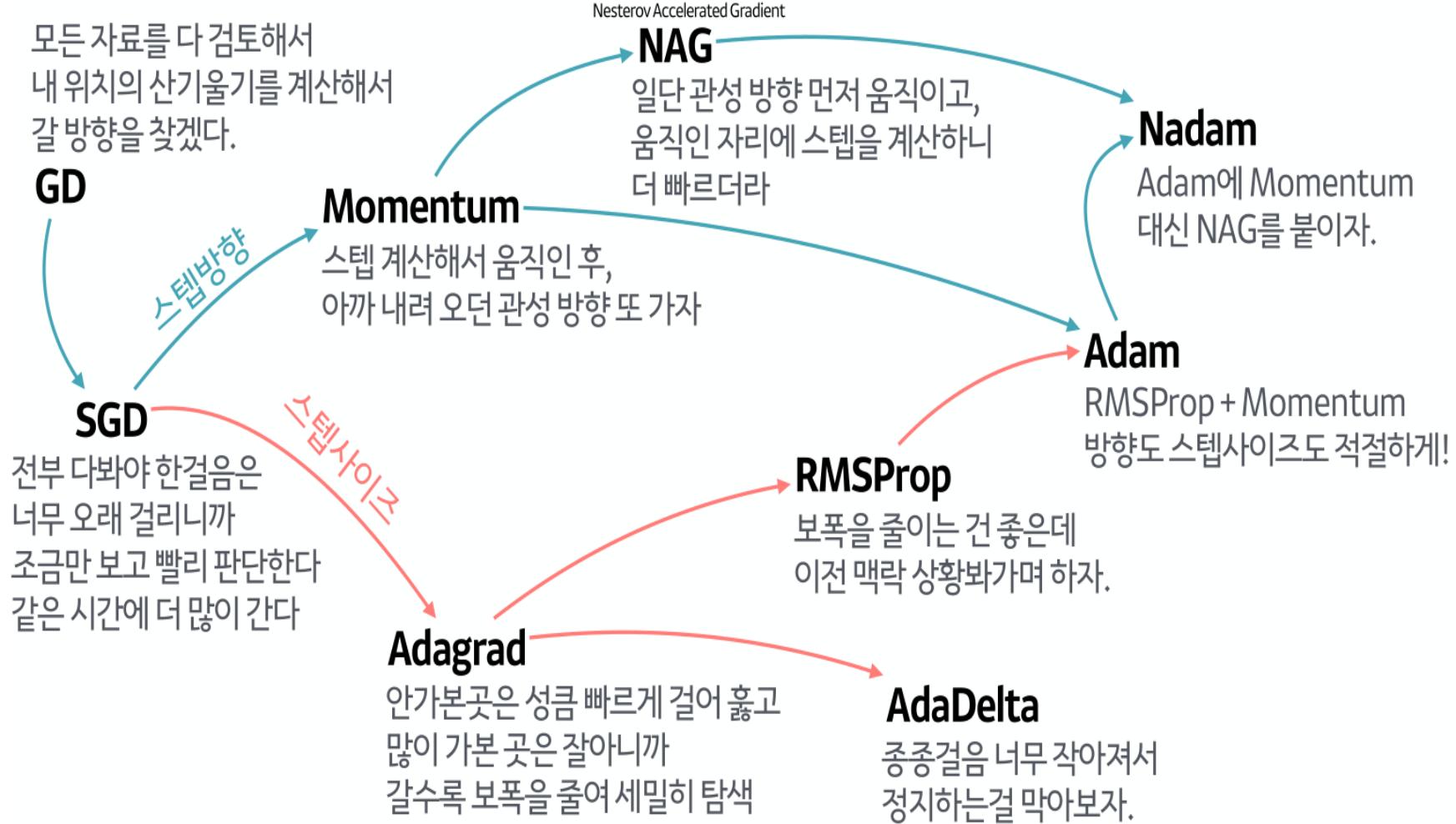
현재 내가 틀린정도를 ‘미분(기울기)’ 한 거



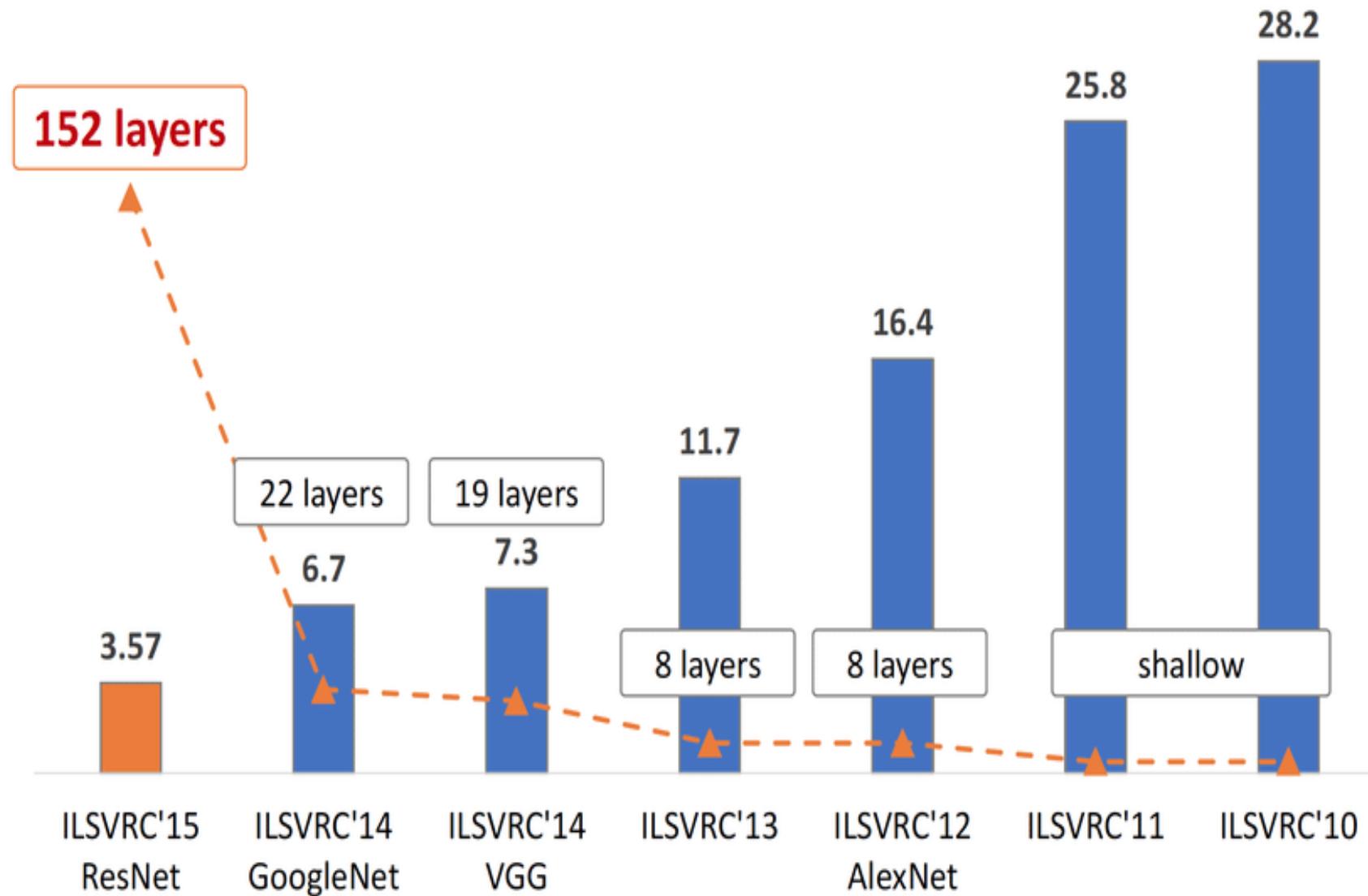
미분하고, 곱하고, 더하고를 역방향으로 반복하며 업데이트한다.

옵티마이저 발달 계보

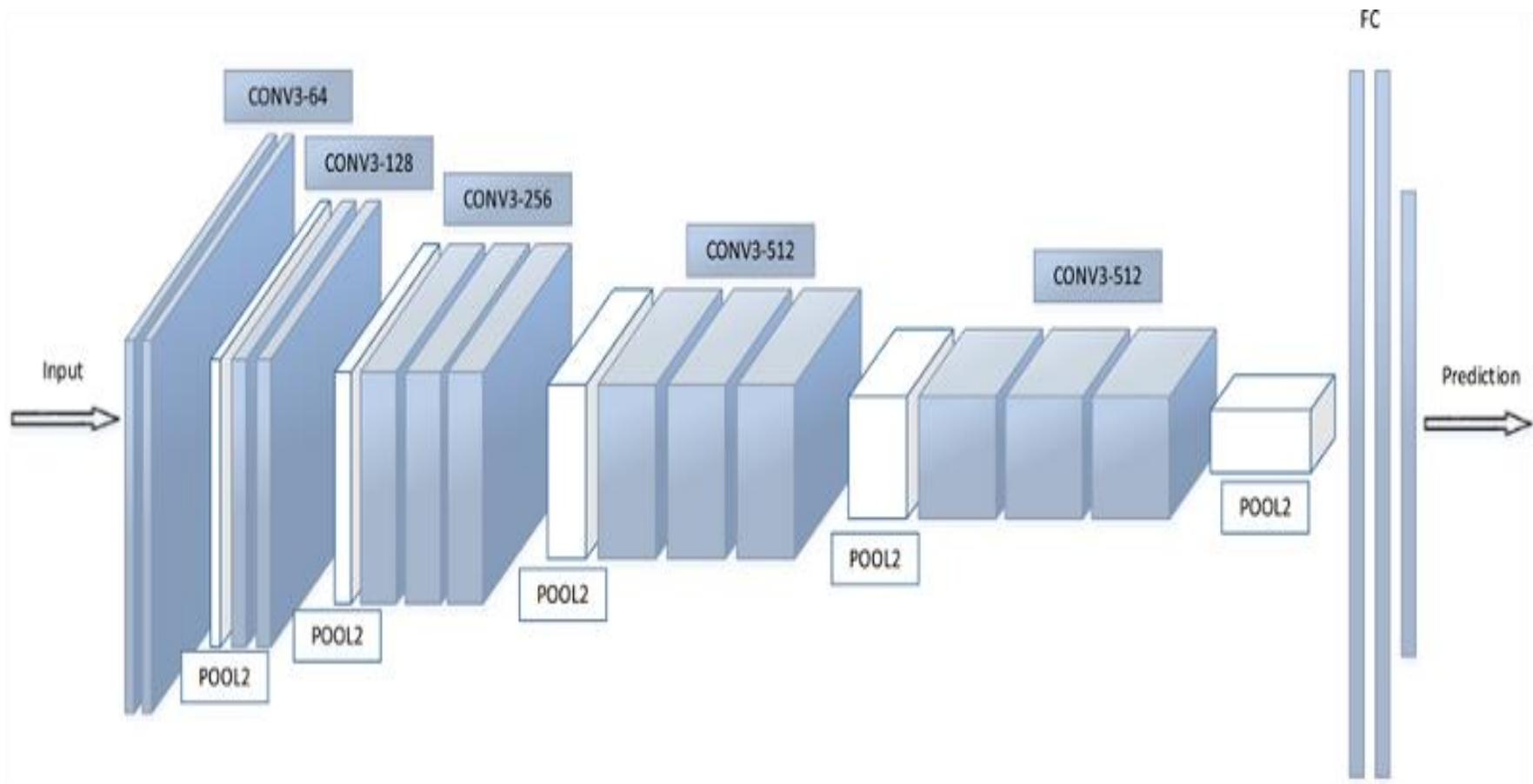
산내려오는 작은 오솔길 잘찾기(Optimizer)의 발달 계보



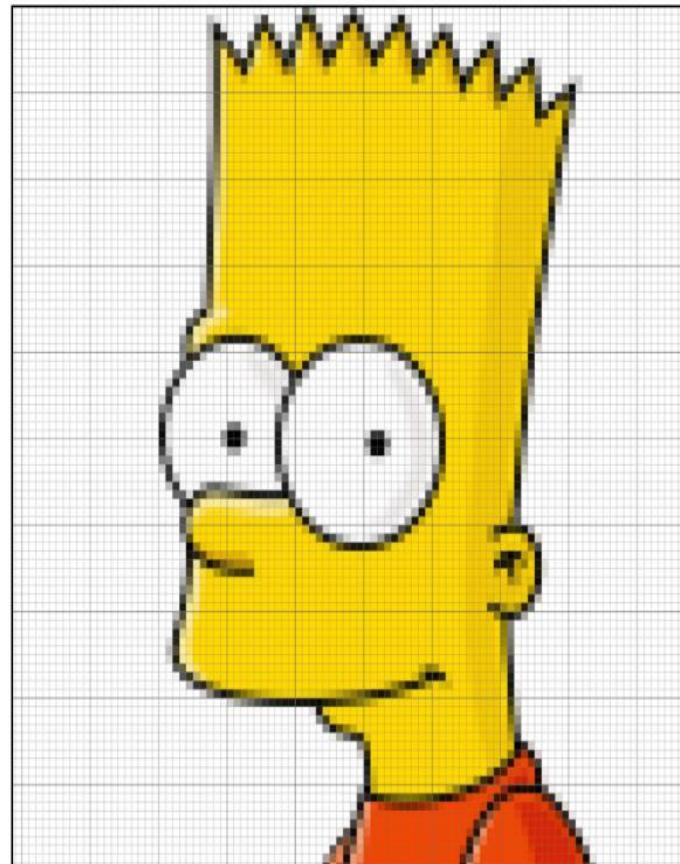
이미지넷 대회 결과



딥러닝 CNN 아키텍처



CNN 추상화한 설명



시작은 이렇게 256*256
픽셀을 다 보았어야 해도

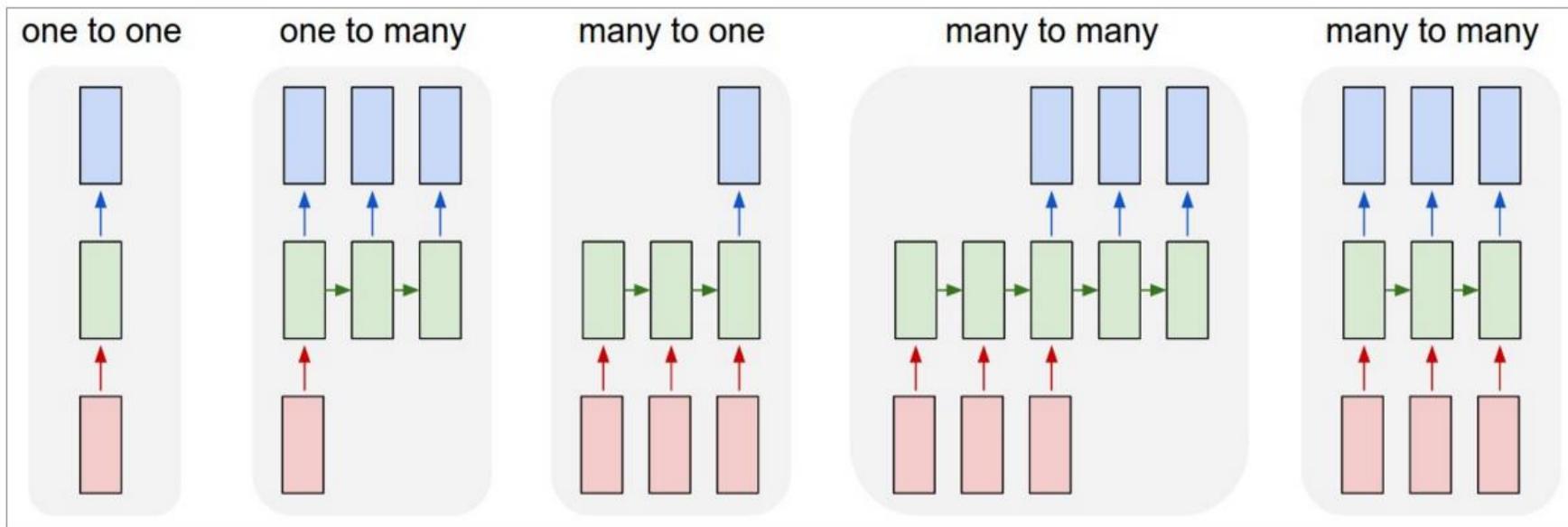


Conv와
MaxPooling
의 반복



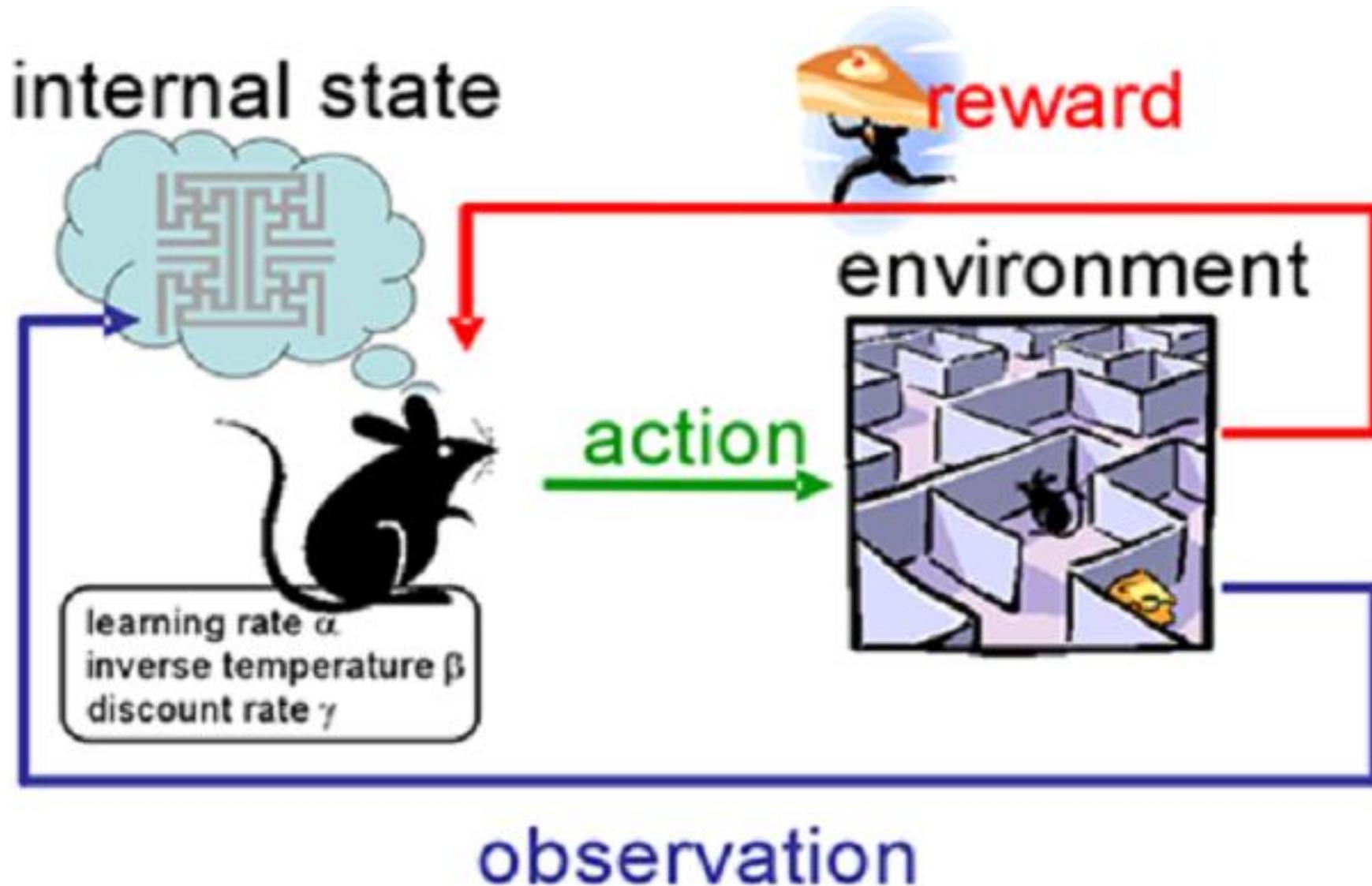
우리는 궁극적으로
이런 녀석을 가지게 된다.

RNN 아키텍처



Each rectangle is a vector and arrows represent functions (e.g. matrix multiply). Input vectors are in red, output vectors are in blue and green vectors hold the RNN's state (more on this soon). From left to right: (1) Vanilla mode of processing without RNN, from fixed-sized input to fixed-sized output (e.g. image classification). (2) Sequence output (e.g. image captioning takes an image and outputs a sentence of words). (3) Sequence input (e.g. sentiment analysis where a given sentence is classified as expressing positive or negative sentiment). (4) Sequence input and sequence output (e.g. Machine Translation: an RNN reads a sentence in English and then outputs a sentence in French). (5) Synced sequence input and output (e.g. video classification where we wish to label each frame of the video). Notice that in every case are no pre-specified constraints on the lengths sequences because the recurrent transformation (green) is fixed and can be applied as many times as we like.

강화 학습





인공지능

AI,

MY NEIGHBORS!

인공지능? 머신러닝? 딥러닝 ?

인공지능(Artificial Intelligence): 특정 분야를 지칭하는 것이 아닌, 지능적 요소가 포함된 기술을 총칭

머신러닝(Machine Learning): ‘데이터’에서 ‘모델’을 스스로 찾아내는 기법

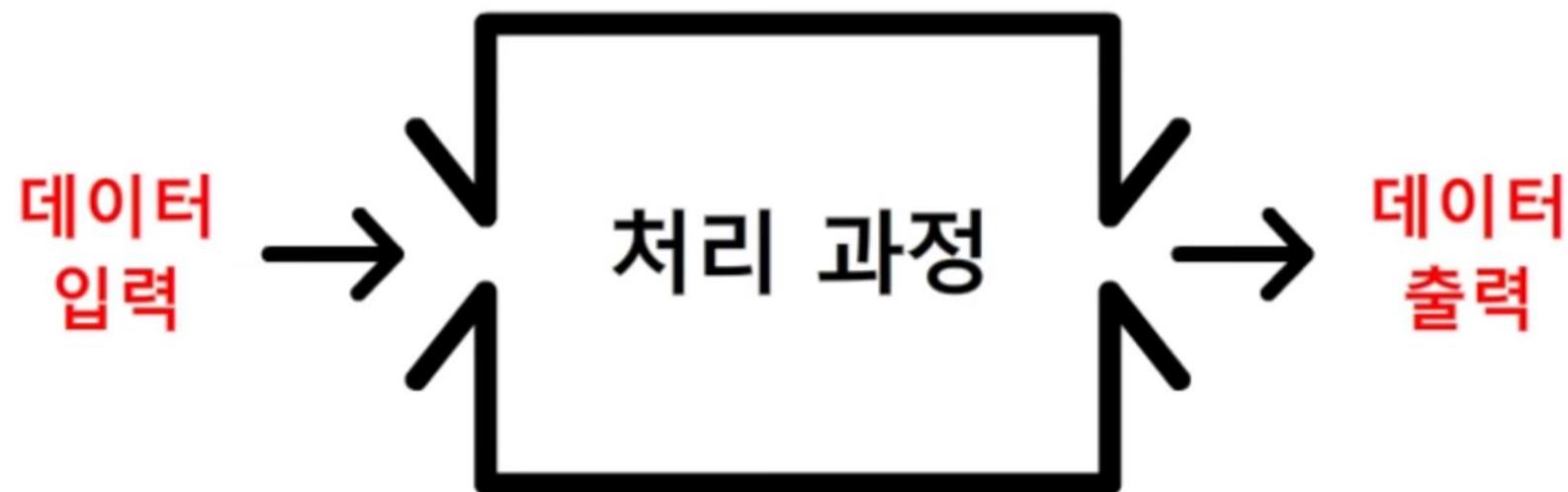
딥러닝(Deep Learning)
:심층 신경망을 이용한 머신러닝 기법

기존의 프로그램



<https://www.youtube.com/watch?v=ZY6eBxyEOq0>

머신 러닝



인공지능

지능 작업을 수행할 수 있는 기계의 능력

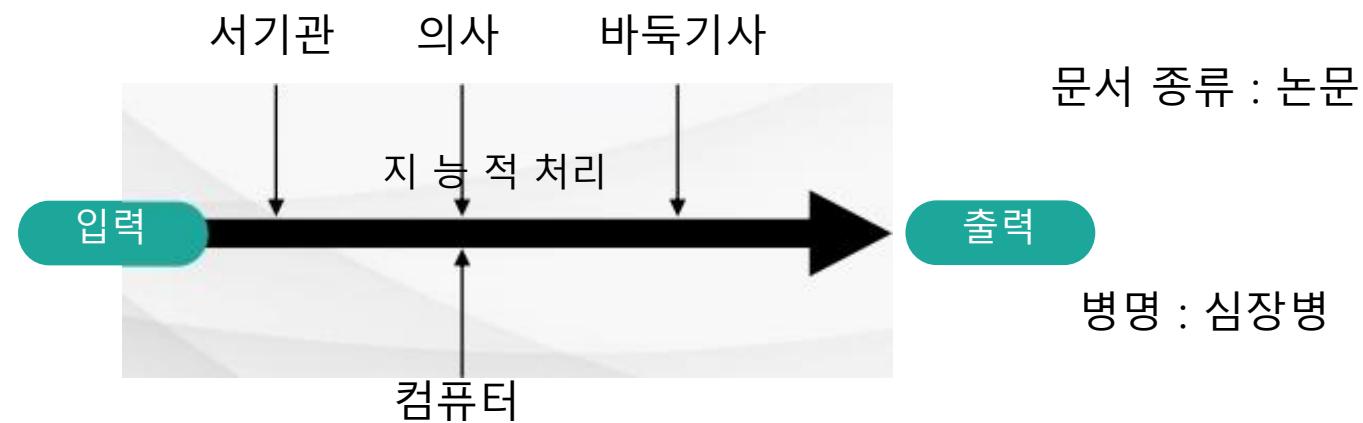
In this paper, we proposed
the video story QA model

...

환자 1 번 기록
나이 : 67
성별 : 남자

가슴통증 종류 : 무증상
혈압 : 160mm Hg
혈중 콜레스테롤 : 286mg/dl
혈당 : < 120mg/dl

...



인공지능의 역사



Alan Turing(1912~1954)

인공지능의 서막을 열다

1950

1960

1970

1980

1990

2006

2016~

낙관의 시대
AI 용어등장 ('56년)

암흑기
('70년중반 ~ '80년초)

암흑기
('80년후반 ~ '90년초)

인공지능,
바둑 챔피언을 이김
4 차 산업혁명 첫 언급
2016

인공지능
퀴즈 챔피언을 이김
2011
딥러닝
인공지능
DARPA 그랜드 챌린지에서 우승
2005
인공지능
체스 챔피언을 이김
1997
전문가 시스템 000

둘러보기

인공지능, 머신러닝, 딥 러닝

- ▶ 인공지능
 - ▶ 외부 관찰자에게 인간처럼 스마트하게 소프트웨어를 작동시키는 폭넓은 방법, 알고리즘 및 기술
 - ▶ 머신러닝, 컴퓨터 비전, 자연어 처리, 로봇 공학 및 그와 관련된 모든 주제를 포괄하는 개념
- ▶ 머신러닝
 - ▶ 더 많은 데이터 축적을 통해 성능을 개선할 수 있도록 하는 다양한 알고리즘과 방법론
 - ▶ 신경망, 서포트 벡터 머신, 결정 트리, 베이지안 신뢰 네트워크, k 최근접 이웃, 자기 조직화 지도, 사례 기반 추론, 인스턴스 기반 학습, 은닉 마르코프 모델, 회귀 기법
- ▶ 딥 러닝
 - ▶ 신경망 (**Neural Network**) 을 부르는 다른 이름
 - ▶ 여러 개의 히든 레이어를 통해 깊게 학습한다고 해서 붙여진 이름

둘러보기



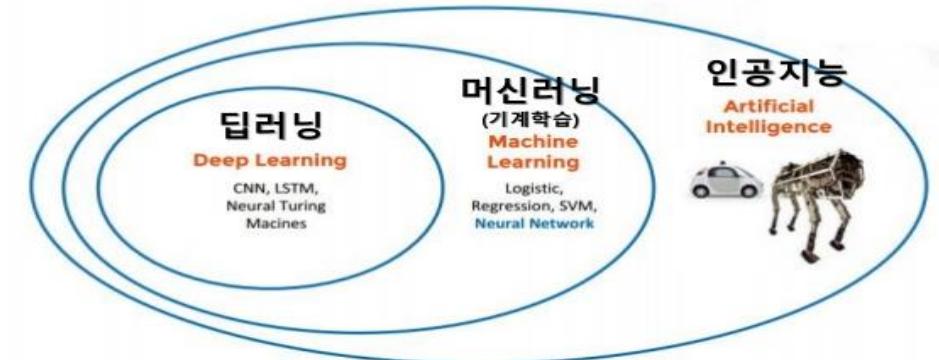
정의 - TOM M. MITCHELL, CARNEGIE MELLON UNIVERSITY

- ▶ A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

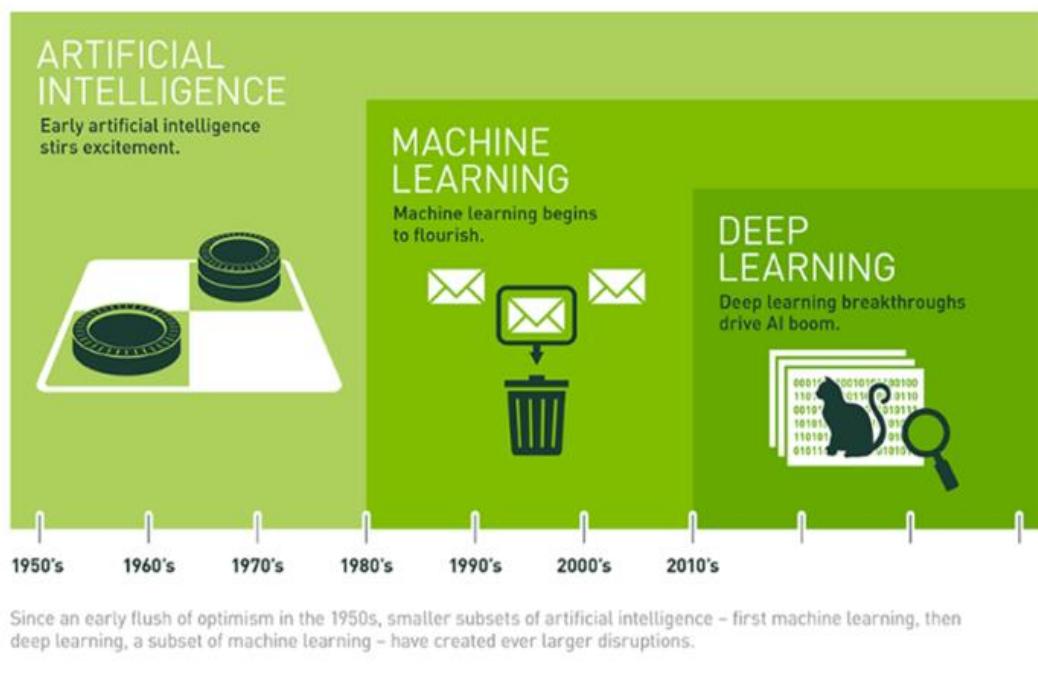
- ▶ 태스크 (T) 에 대해 꾸준한 경험 (E) 을 통해 T 에 대한 성능 (P) 을 높이는 것을 기계학습이라고 한다 .

- ▶ 기계학습에서 가장 중요한 것은 E 에 해당하는 데이터이다 .
좋은 품질의 데이터를 많이 가지고 있다면 보다 높은 성능을 끌어낼 수 있다 .

Intro



Slide by Jiaqiong Qian at DeepNet 2016
<http://www.slideshare.net/SaintGroup/intro-deep-learning-by-jiaqiong-qian-deepnet-2016>



둘러보기

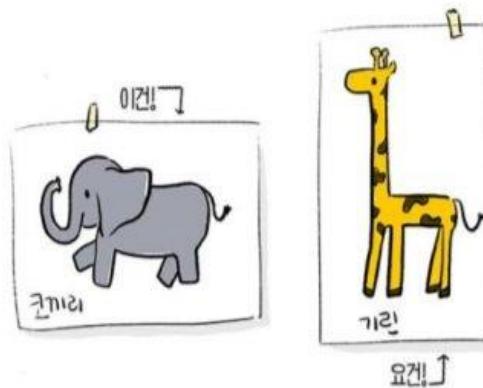
분류

- ▶ 지도 학습 (Supervised Learning)
 - ▶ 회귀 (Regression)
 - ▶ 분류 (Classification)
- ▶ 비지도 학습 (Unsupervised Learning)
 - ▶ 군집화 (Clustering)
 - ▶ 분포 추정 (Underlying Probability Density Estimation)
- ▶ 강화 학습 (Reinforcement Learning)

머신 러닝

지도 학습 (Supervised Learning)

문제와 정답을 모두 알려주고
공부시키는 방법

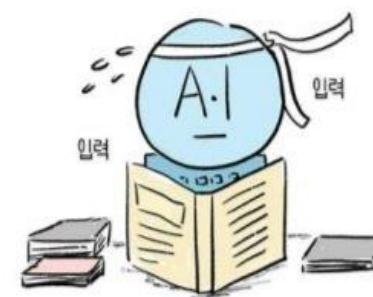


예측, 분류

비지도 학습 (Unsupervised Learning)

답을 가르쳐주지 않고
공부시키는 방법

비지도학습은 답을 가르쳐주지 않고 공부를 시키는거야.



연관 규칙, 군집

강화 학습 (Reinforcement Learning)

보상을 통해
상은최대화, 벌은최소화하는
방향으로 행위를 강화하는 학습

강화학습은 일종의 게임 같이 보상해주는거야



보상

머신러닝의 분류

지도학습(Supervised Learning)

데이터와 그에 대응되는 예측결과 값(Label)을 투입하여 서로 간의 관계를 학습하고, 해당 데이터와 일치 또는 유사한 데이터가 입력되었을 때, 학습시킨 관계에 따른 결과 값을 내도록 하는 것.

비지도학습(Unsupervised Learning)

결과 값이 없는 데이터들을 입력하여 각각의 데이터들에 내재된 속성을 기반으로 분류 등의 학습을 하고, 새로운 데이터가 입력되었을 때 해당 데이터의 내재된 속성에 따라 학습된 결과를 도출하는 것.

강화학습(Reinforcement Learning)

에이전트(agent)가 특정 상태에 대한 반응으로서의 행동(action)을 내보내면, 이에 따른 보상(reward) 또는 벌칙(penalty)을 주어 달성하고자 하는 목표 결과(action)를 내보내도록 학습하는 것.

SUPERVISED LEARNING



입력된 레이블(Label)과
결과값(Output)의 차이를 최소화

UNSUPERVISED LEARNING



데이터의 내재된 특성(Feature)의
유사성, 관련성을 바탕으로 학습

REINFORCEMENT LEARNING



에이전트와 환경 간의 주고 받음(action//state, reward)
을 통해 조성/Shaping하는 것

Training dataSet

$$y = ax + b$$

feature label

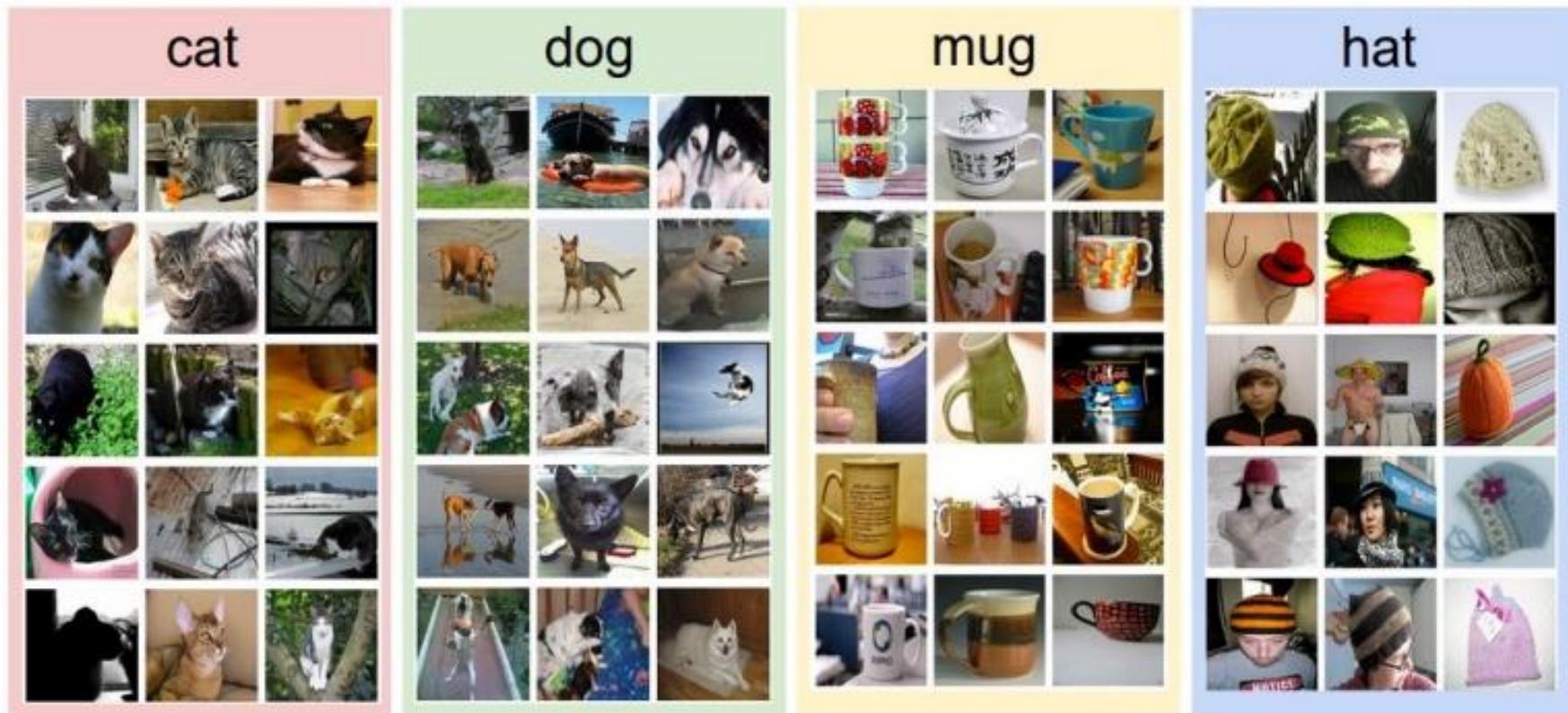
x	y
1	2
2	4
3	6
4	8
5	10
10	20
20	40

**Test dataset** $5 \rightarrow 10$

$$\begin{aligned} a &= 2 \\ b &= 0 \end{aligned}$$

Supervised learning

An example training set for four visual categories.



둘러보기

SUPERVISED LEARNING (1)

▶ 서포트 벡터 머신 (support vector machine)

패턴 인식, 자료 분석을 위한 지도 학습 모델이며, 주로 분류와 회귀 분석을 위해 사용한다. 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모델을 만든다.

만들어진 분류 모델은 데이터가 사상된 공간에서 경계로 표현되는데 SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다.

▶ 은닉 마르코프 모델 (Hidden Markov model)

통계적 마르코프 모델의 하나로, 시스템이 은닉된 상태와 관찰가능한 결과의 두 가지 요소로 이루어졌다고 보는 모델이다.

관찰 가능한 결과를 야기하는 직접적인 원인은 관측될 수 없는 은닉 상태들이고, 오직 그 상태들이 마르코프 과정을 통해 도출된 결과들만이 관찰될 수 있기 때문에 '은닉'이라는 단어가 붙게 되었다.

음성 인식, 필기 인식, 동작 인식, 품사 태깅, 약보에서 연주되는 부분을 찾는 작업, 부분 방전, 생물정보학과 같이 시간의 영향을 받는 시스템의 패턴을 인식하는 작업에 유용한 것으로 알려져 있다.

둘러보기

SUPERVISED LEARNING (2)

▶ 회귀 분석 (Regression)

통계학에서, 회귀분석 (regression analysis)은 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한뒤 적합도를 측정해 내는 분석 방법이다. 회귀분석은 시간에 따라 변화하는 데이터나 어떤 영향, 가설적 실험, 인과 관계의 모델링등의 통계적 예측에 이용될 수 있다.

인공신경망은 생물학의 신경망 (특히 뇌)에서 영감을 얻은 통계학적 학습 알고리즘이다.

인공신경망은 시냅스의 결합으로 네트워크를 형성한 인공 뉴런 (노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜, 문제 해결 능력을 가지는 모델 전반을 가리킨다. 좁은 의미에서는 역전파법을 이용한 다층 퍼셉트론을 가리키는 경우도 있지만, 인공신경망은 이에 국한되지 않는다.

특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종으로 1950년대 이후 광범위하게 연구되고 있다.

텍스트 분류에 사용됨으로써 문서를 여러 범주 (예 : 스팸, 스포츠, 정치) 중 하나로 판단하는 문제에 대한 대중적인 방법으로 남아있다.

둘러보기

UNSUPERVISED LEARNING

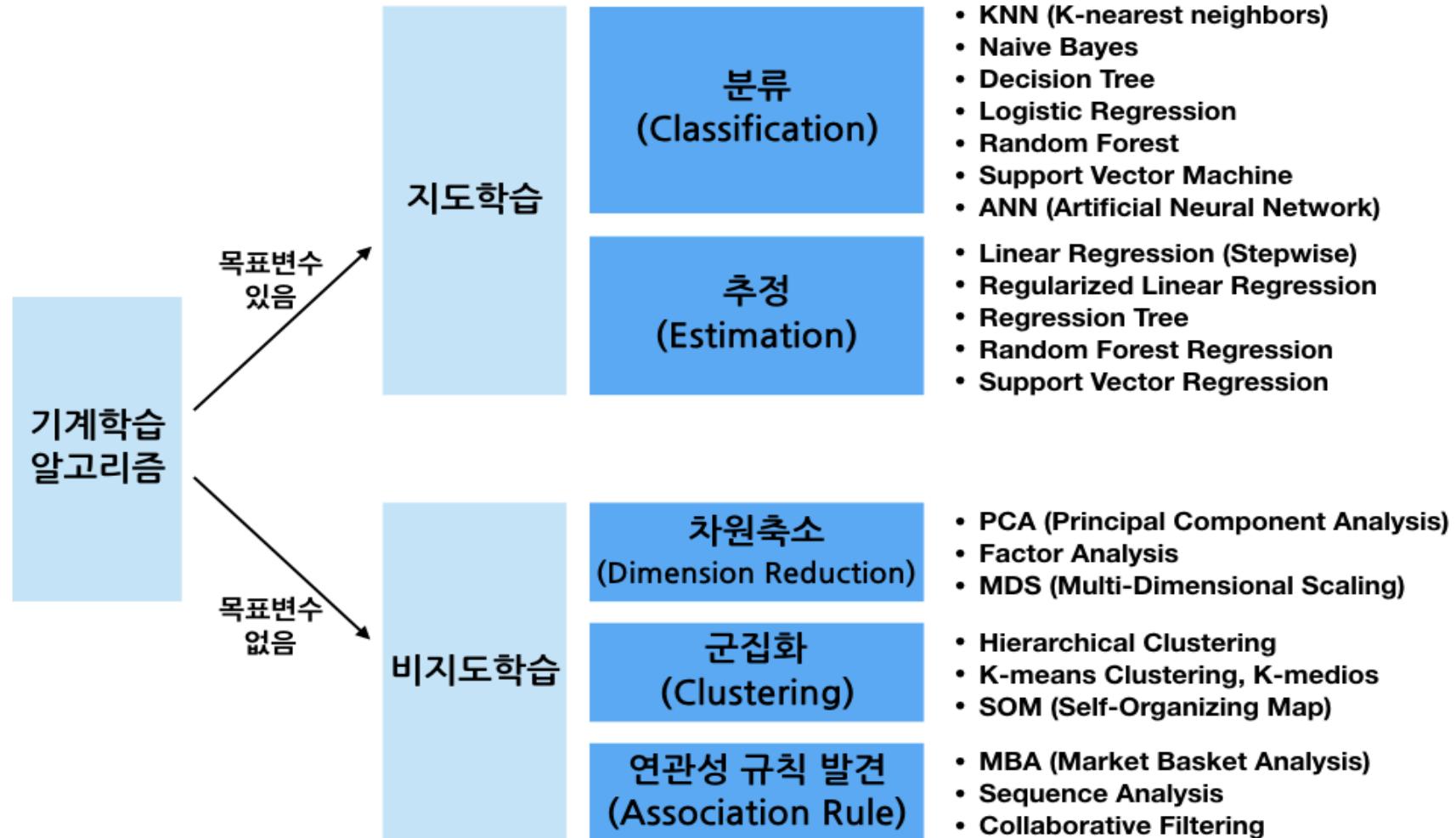
- ▶ 클러스터링 (Clustering)

개체를 다수의 메트릭스에서 상호 유사한 세그먼트 또는 클러스터로 그룹화하는 기법.
고객세분화가 클러스터링의 실제 예다 . 클러스터링 알고리즘은 무척 다양한데 ,
가장 널리 사용되는 것이 k- 평균 (k-means) 이고
비슷한 알고리즘으로 Gaussian Mixture Model 도 있다 .

- ▶ 독립 성분 분석 (Independent Component Analysis)

다면량의 신호를 통계적으로 독립적인 하부 성분으로 분리하는 계산 방법이다 .
각 성분은 비가우스 성 신호로서 서로 통계적 독립을 이루는 성분으로 구성되어 있다 .
독립 성분 분석은 블라인드 신호를 분리하는 특별한 방법이다 .
독립 성분 분석의 전형적인 알고리즘은 복잡성을 줄이기 위한 전 단계로서 중심화
(centering), 백색화 (whitening), 차원 감소 (dimensionality reduction) 등의 과정이 필요하다.
백색화와 차원 감소는 주 성분 분석 (Principal Component Analysis) 과
특이 값 분해 (Singular Value Decomposition) 로 한다 .

둘러보기





인공지능 사례

AI,
example

FORWARD

FORWARD HOSPITAL



- ▶ <https://goforward.com>
- ▶ 2017년 1월 샌프란시스코
- ▶ 엔지니어들이 설립한 AI 병원
- ▶ 월정액 149\$
- ▶ 24시간 건강 검진 및 상담 무제한
- ▶ Design Your Health

FORWARD

진료 순서

1. 병원 방문
2. 아이패드로 로그인
3. 바디스캐너로 건강 진단
4. 의사 상담
5. 24 시간 상담 가능 (스마트폰 앱 연동)



ALEXA

ALEXA : KID SKILLS

- ▶ Sesame Street
<https://www.amazon.com/Sesame-Workshop-Street/dp/B073XBBWRQ>
- ▶ The SpongeBob Challenge
- ▶ Amazon Storytime
- ▶ NASA Mars
- ▶ Word of the Day



ALEXA

ALEXA : KID SKILLS

- ▶ 1-2-3 Math

<https://www.amazon.com/Shanthan-Kesharaju-1-2-3-Math/dp/B01AVQLZQ0>

- ▶ This Day In History
- ▶ Jeopardy!
- ▶ Animal Game
- ▶ The Magic Door



今日頭條

진르터우탸오 (오늘의 헤드라인)

- ▶ 모바일 뉴스앱
- ▶ 인공지능 활용해 개인별 맞춤 콘텐츠 제공
- ▶ 7,800 만명 구독 - 3년새 200 배 성장
- ▶ 누적 사용자 6 억명 , 하루 이용자 수 1 억명
- ▶ 하루 평균 이용 시간 76 분
- ▶ 기업가치 110 억 \$(약 12 조원)



今日頭條

下载APP | 北京 晴 15° / 29°

登录 反馈 侵权投诉 头条

今日头条

推荐 热点 视频 图片 段子 社会 娱乐 游戏 体育 汽车 财经 搞笑

文在寅第五次与特朗普通电话 商讨半岛局势

程序员被骗婚自杀案和王宝强离婚案代理律师张起淮：曾代理婚内诈骗案，罪犯被判10年

中国之声 · 2520评论 · 刚刚

社保交满15年，退休能领多少钱？

头条问答 · 8分钟前

大家都在搜：乔任梁逝世一周年

搜索

要闻 社会 娱乐 体育 军事 明星

网上有害信息举报专区
举报电话：12377

淘宝网 taobao.com 抢 ￥330 销量：64

24小时热闻 中超-权健2-2国安全场集锦：帕托

ZUMEPIZZA

ZUMEPIZZA

- ▶ <https://www.zumepizza.com/>
- ▶ 매장에서 먹는 것과 같은 배달 피자
- ▶ 2016년 4월
- ▶ 200여평 주방에서 1시간에 288개 생산
- ▶ 1분에 4.5개 완성
- ▶ 로봇으로 인건비 줄이고, 유기농 재료 사용



ZUMEPIZZA

매장

- ▶ 주문이 들어오면 주방의 스크린으로 전송
- ▶ 사람이 도우를 사람이 얇게 펼친다 .
- ▶ 존과 페페가 메뉴에 따라 적당량의 토마토 소스를 뿌린다 .
- ▶ 천장에 매달린 마르타가 소스를 고르게 바른다 .
- ▶ 사람이 토픽을 얹는다 .
- ▶ 브루노가 토픽이 끝난 피자를 오븐에 넣어서 굽는다 .
- ▶ 1 차로 구워진 피자를 빈센치오가 배달 트럭에 싣는다 .

ZUMEPIZZA

배달

- ▶ 56 개의 오븐을 갖춘 트럭
- ▶ GPS 로 목적지에 도착하기 4 분 전에 초벌 피자를 굽는다 .
- ▶ 고객이 받았을 때 , 가장 맛있는 상태의 피자 완성



ZUMEPIZZA

목표

- ▶ 주문 -> 생산 -> 도착을 5 분 이내로 단축
- ▶ 피자배달 평균은 45 분 , ZumePizza 는 22 분 .
- ▶ 주문 즉시 배달트럭이 출발하고 배달 중에 완성
- ▶ 고객 성향을 파악해 재료 예측 후 배달 중에 주문 접수
- ▶ 샐러드 로봇 , 요구르트 로봇 , 탄산음료 로봇 , 볶음요리 로봇 개발
- ▶ 자율주행 배달

ZUMEPIZZA

Automation done right

Alongside our pizzaiolos, we also employ pizza-making robots, Pepe, Giorgio, Marta, Bruno, and Vincenzo, that work together in our kitchen to craft each and every pie. Our co-bots perform low-skill, repetitive, and dangerous tasks, giving human employees more opportunities to do creative, high-skill jobs at Zume.



ZUMEPIZZA



INKITT

INKITT

- ▶ <https://www.inkitt.com/>
- ▶ 2016년 여름 첫 번째 책을 출간한 출판사
- ▶ 24 권 출간해서 22 권이 아마존 베스트셀러
- ▶ 99.99%의 확률로 베스트셀러를 만드는 것이 목표
- ▶ e북 인세는 25%, 종이책 인세는 51%
- ▶ 22 권의 책을 쓴 16 명의 신예 베스트셀러 작가 배출

Inkitt

1. 누구나 자유롭게 글을 올린다

저자 4 만명, 진행 중인 내용 15 만개

2. 독자는 구독하고 평가한다

선호 장르를 선택하면 스토리 추천

객관식과 주관식의 다양한 형태로 평가

3. 인공지능이 독자 반응을 분석해서 베스트셀러 가능 여부 판단

페이지에 머문 시간, 몰입도, 재접속 후 다시 읽었는지 등의 데이터 종합 평가

4. 출판사가 작가에게 출판을 제의하고 작가는 수정한다

저자는 독자 평가를 바탕으로 수정 및 보완

표지 디자인은 3 개의 후보 중에서 독자들 반응으로 판단

평가

- ▶ 객관식 평가

구성 (plot), 문체 (writing style), 문법 , 전반적 느낌 (overall) 각각에 별점 부여

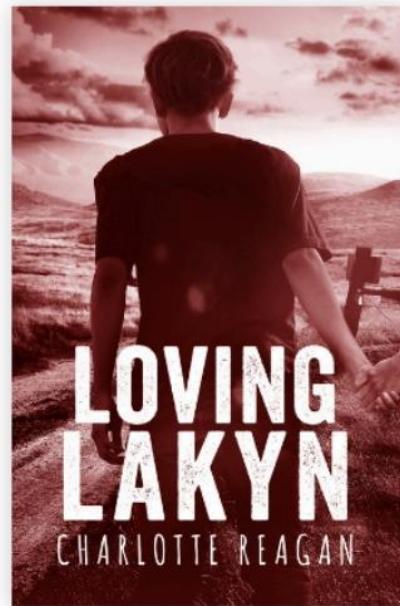
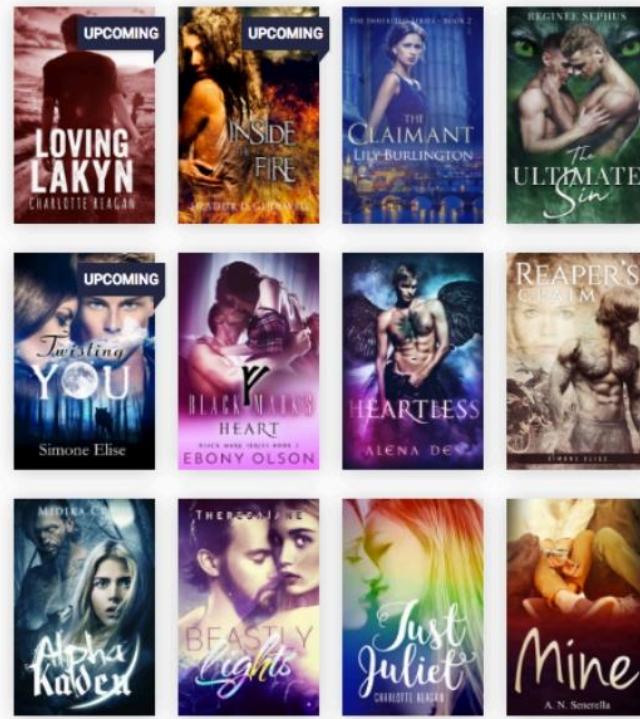
- ▶ 주관식 평가

특별했거나 교훈이 됐던 점

출간되면 구입할지 , 그 이유는

다른 사람에게 추천할건지 , 이유는

스토리가 어떻게 바뀌었으면 좋을지



Loving Lakyn

Lakyn James is sixteen years old and hating every second of it. He was supposed to be done, he'd tapped out. End of story, unsubscribe here. Suicide "attempt", they said. His intentions had no "attempt" in them.

Re-entering normal life after 'trying' to take his own is weird. Especially when the world keeps going like it never happened. He still has to eat breakfast, go to school, and somehow convince a cute boy that he's too damaged to date.

Scott White comes with his own problems,

INKITT

Inkitt Groups Writing Contests Published Books Free Novel Writing Contest

Writing Contests in 2017: The Definitive Guide

A comprehensive and regularly updated list of writing contests in 2017. Whether you are an aspiring or established author, here you will find a complete list of short story, novel, poetry, and essay competitions where you can submit your fiction and non-fiction masterpieces for the chance to win great prizes and receive notoriety for your work.

Is your writing contest not listed? [Let us know.](#)

TOP PRIZE
Publishing Deal
[VIEW CONTEST](#)

Inkitt Writing Competition 2017



- A Dedicated Marketing Team
- Professional Editing & Cover
- 25% royalties from ebook sales and 51% royalties from print sales

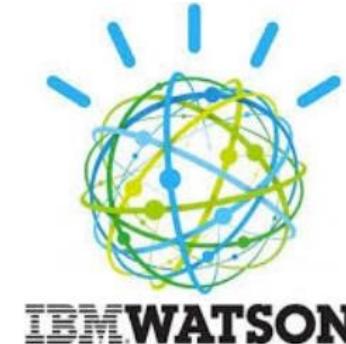
DEADLINE:
September 30, 2017

ENTRY FEE:
Free

WATSON

IBM WATSON

- ▶ IBM 에서 개발
- ▶ 2011 년 퀴즈쇼 제퍼디 참가
- ▶ 인터넷에 연결되지 않은 상태에서 우승
- ▶ 의료 , 금융 , 세금 규정 , 법학 , 소비자 서비스 등의 다양한 영역 제공
- ▶ Bluemix 플랫폼에서 Watson Developer Cloud 서비스 제공



WATSON

- ▶ 금융

정치 , 사회적 위험요소를 파악해 리스크 분산 및 포트폴리오 적용 .
현대카드 , 신한금융그룹 도입

- ▶ 방송

공포영화 예고편을 학습해서 '모건'에 대한 예고편 완성

- ▶ 의학

가천대 길병원 , 건양대 / 부산대 / 대구가톨릭대 병원 , 계명대 동산병원 , 중앙보훈병원
조선대 병원 예정 .

WATSON

▶ 교육

교원에서 수학 디지털 교과서에 왓슨 적용방안 협의 중

▶ 쇼핑

롯데그룹에서 백화점, 마트, 편의점, 면세점 대고객 서비스 (지능형 쇼핑 어드바이저)

▶ 스마트 홈스피커

SM 엔터테인먼트에서 스마트 스피커에 한국어 왓슨 채택

▶ 한국어

SK 그룹의 AIBRIL



Service

- 01 대화
- 02 자연어분류
- 03 언어 번역
- 04 검색 및 평가
- 05 문서 변환
- 06 성향 분석
- 07 이미지 인식
- 08 자연어 이해

WATSON

의료 영역

- ▶ 암 환자 1,000 명 대상의 IBM Watson 진료 성적 공개 (2017. 03.05)
<http://www.yoonsupchoi.com/2017/03/05/manipal-watson-for-oncology>
- ▶ 닥터 왓슨과 의료진 항암처방 엇갈리면 ... 환자 " 왓슨 따를게요 " (2017. 01. 12)
http://news.chosun.com/site/data/html_dir/2017/01/12/2017011200289.html



WATSON

DON'T JUST MEET CUSTOMERS' EXPECTATIONS. EXCEED THEM.

MEET IBM WATSON ENGAGEMENT ADVISOR.

CUSTOMERS TODAY WANT MORE. DEMAND MORE. EXPECT MORE. YOURS INCLUDED.

COMPANIES THAT SATISFY MORE, WIN MORE. A ONE POINT CHANGE IN CUSTOMER SATISFACTION = 4.6% CHANGE IN YOUR MARKET VALUES. ACCORDING TO JOURNAL OF MARKETING, JANUARY 2006.

COGNITIVE TECHNOLOGY THAT THINKS AND BRIDGES THE GAP BETWEEN WHAT YOUR CUSTOMERS EXPECT AND THE SERVICES THAT YOU PROVIDE

WATSON ENGAGEMENT ADVISOR CAN:

- COMMUNICATE WITH CUSTOMERS IN NATURAL LANGUAGE
- LEARN FROM CUSTOMERS WITH EACH NEW INTERACTION

TO HELP YOUR BUSINESS:

- ENGAGE CUSTOMERS IN WAYS THEY LIKE
- EMPOWER CUSTOMERS AT THE POINT OF ACTION

WATSON'S THOUGHT PROCESS:

- IT UNDERSTANDS THE REQUEST IN CONTEXT
- IT GENERATES AND EXPLORES HYPOTHESES AGAINST THE DATA
- IT EVALUATES THOSE HYPOTHESES BASED ON MORE DATA
- IT LEARNS FROM ITSELF AS ONLY WATSON CAN DO
- IT THEN SHARES WITH YOU WHAT IT "THINKS" ALL IN SECONDS, ALL IN PLAIN ENGLISH

EXAMPLES OF HOW IT CAN WORK:

BANK CUSTOMERS CAN USE WATSON TO BETTER UNDERSTAND THINGS LIKE RETIREMENT PLANNING.

"WE'RE EXCITED TO EXPLORE HOW WATSON CAN HELP OUR EMPLOYEES PROVIDE BETTER ADVICE FASTER, WITHOUT HAVING TO NAVIGATE THROUGH A LARGE DATABASE SEARCHING FOR ANSWERS." - ROYAL BANK OF CANADA

MOBILE CUSTOMERS ON THE GO CAN HAVE ISSUES RESOLVED FASTER WHEN REPS USE WATSON TO TROUBLESHOOT PROBLEMS.

CLIENTS AT COMPANIES SUCH AS NIELSEN CAN CREATE MORE EFFICIENT AND EFFECTIVE MEDIA PLANS.

WATSON ENGAGEMENT ADVISOR. HOW MIGHT WATSON EMPOWER YOUR CUSTOMERS - AND YOU?

IBM

IBMWATSON.COM

인공지능 50선

- [블로터 11th] 알아두면 쓸데있는 신기한 인공지능 50 선 (2017. 09. 17.)

<http://www.bloter.net/archives/289626>



인공지능 50선

1. 바둑 기사

알파고 , 딥마인드

2. 스피커

에코 , 구글홈 , 홈팟 , 인보크 , 누구 , 기가지니 , 웨이브 , 카카오미니

3. 자살 예방 상담사

문자메시지 기반 24 시간 위기 상담 서비스

크라이시스 텍스트 라인 (CTL) - 고위험군 필터링

4. 오이 분류

9 등급 자동분류 시스템

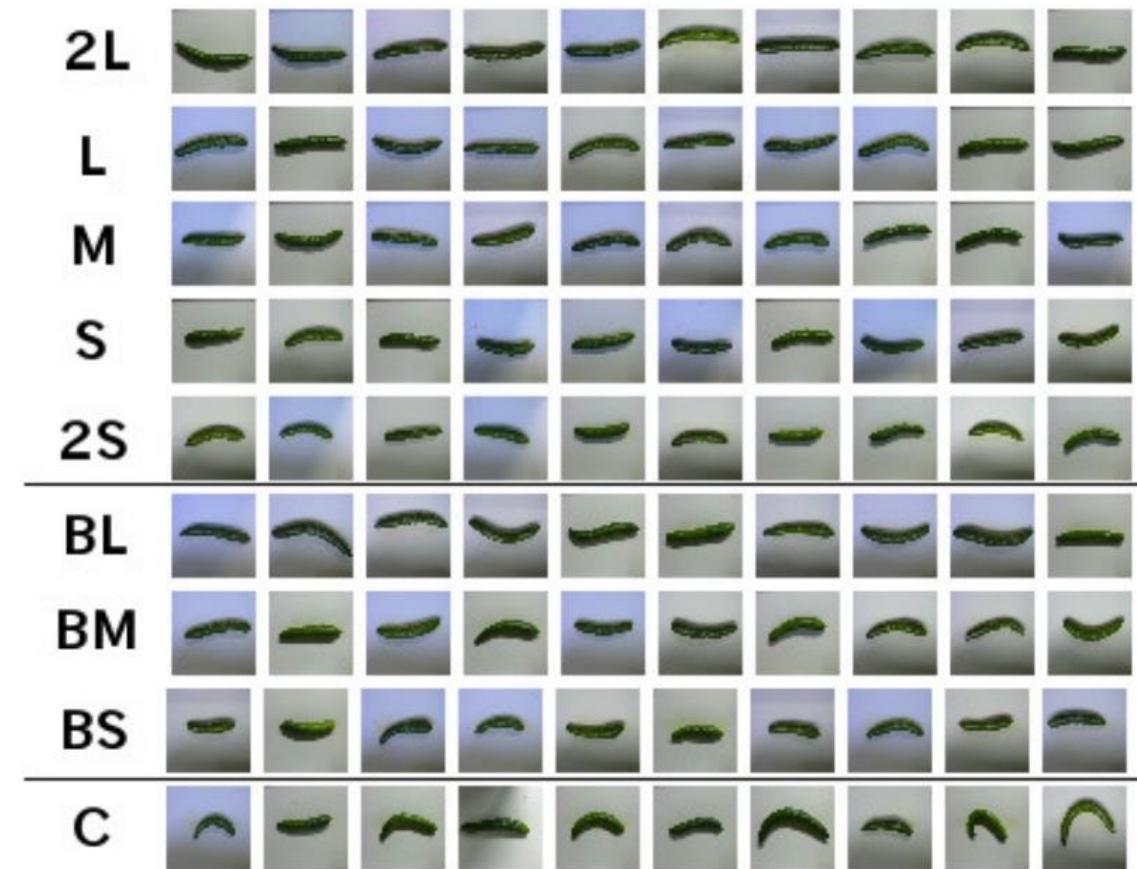
5. 승무원

KLM(Koninklijke Luchtvaart Maatschappij) 네델란드 항공사

인공지능 챗봇 서비스 - 일정 확인 , 체크인 , 발권 , 예약 변경 업무 수행

인공지능 50선

스피커 , 오이



인공지능 50선

6. 아케이드 게이머

팩맨 인간 최고점 266,330 점 .

마이크로소프트 말루바 999,999 점으로 만점

7. 흑백 사진을 컬러로

구글과 와세다 대학 공동 연구 - 흑백 사진을 컬러로 변환

8. 쇼핑 도우미

아웃도어 브랜드 노스페이스 - 왓슨 활용 . ' 플루이드 리테일 ' 개발

9. 보험 상담사

AIA 생명 한국지점의 인공지능 콜센터 서비스 .

고객과 대화 - 계약 정보 확인 및 확정

10. 돌고래 언어 해석

스웨덴 스타트업 ' 가비가이 AB ' - 2021 년을 목표로 프로젝트 돌입

인공지능 50선

말루바



인공지능 50선

흑백사진 변환

SIGGRAPH 2016



인공지능 50선

11. 그림 도우미

구글의 '오토드로우' - 펜으로 그린 그림을 멋진 그림으로 변환

12. 포르노 비평가

'얼빠진 해커톤'에 등장한 프로젝트. 포르노를 학습하고 해석

13. 멸종위기동물 보호

바다소 탐지기 - 드론 항공촬영, 텐서플로우 자동 판별 (80%)

14. 변호사

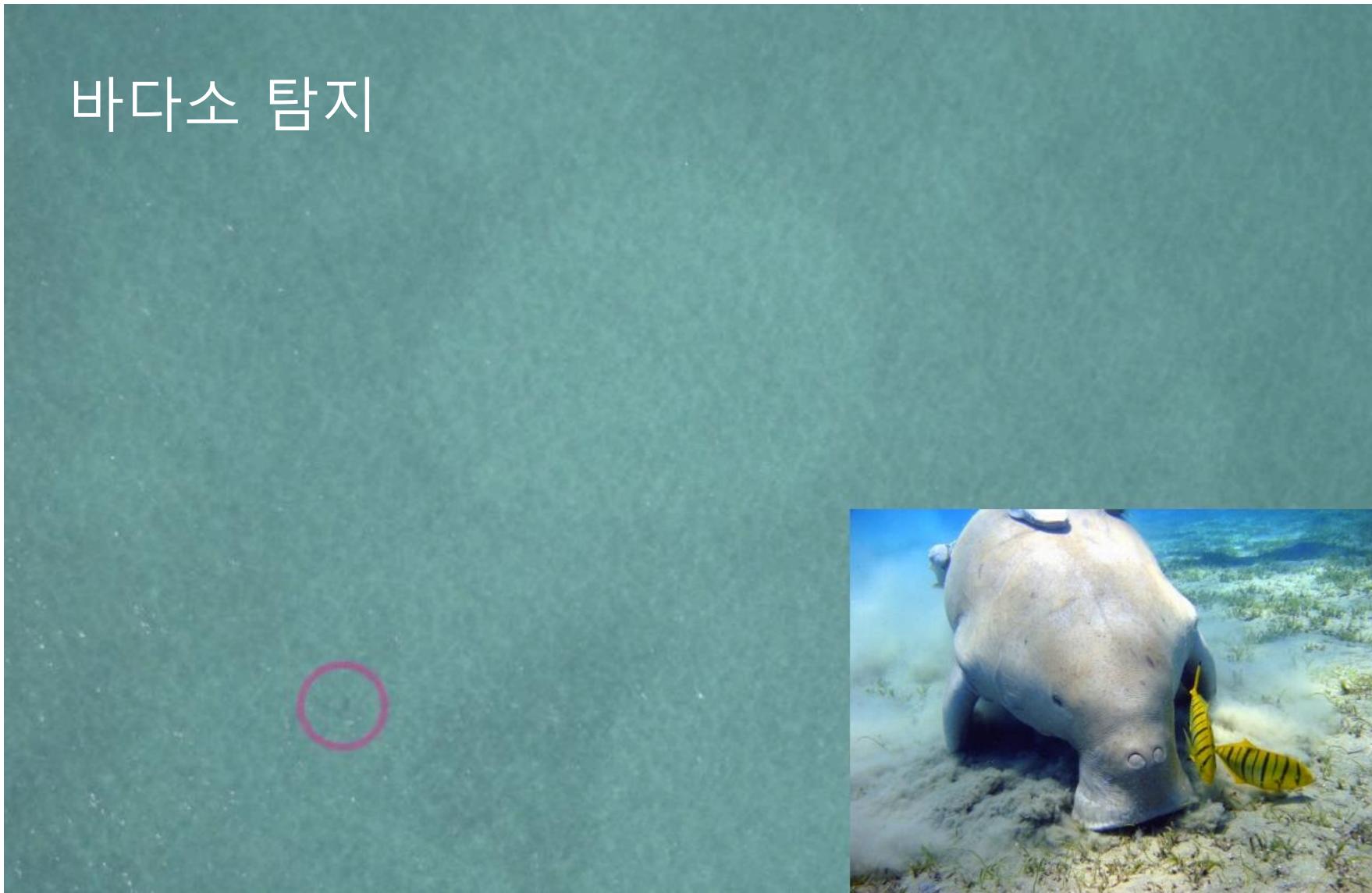
베이커앤호스테틀러 법무법인 - 파산 분야에 왓슨 기반의 '로스' 배치

15. 법률상담 서비스

19 살 조슈아 브라우더 - 주차 딱지 취소에 필요한 채팅봇 '두낫페이' 개발

인공지능 50선

바다소 탐지



인공지능 50선

16. 기자

2012년 개발된 LA 타임스의 '퀘이크봇' - 지진 탐지 자동 기사 작성

17. 고문서 번역

시스트란 인터네셔널과 미래창조과학부 - 고전문헌 자동번역 시스템 구축
첫 번째 프로젝트로 국보 303호 '승정원 일기' 선택 - 완역 45년 소요

18. 반려동물 장난감

반려동물 케어 플랫폼 '고미랩스' - 놀이패턴, 견종, 나이, 성별 분석

19. 사물 감별사

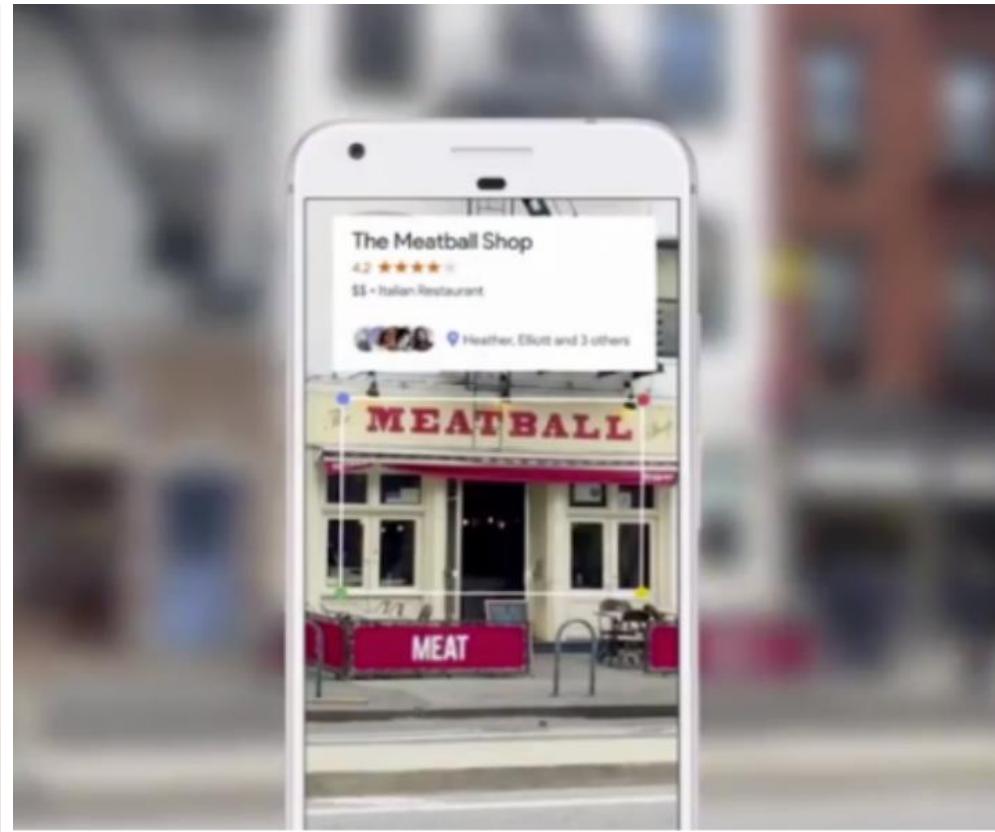
구글 렌즈 - 사물을 보여주면 이해하고 정보 전달. 구글 어시스턴트와 포토에 도입

20. 대선 뉴스 큐레이터

19 대 대통령 선거 - 메인 화면과 뉴스 섹션에 '루빅스' 적용

인공지능 50선

고미 , 구글렌즈



인공지능 50선

21. 난민 심리치료

스타트업 'X2AI' - 정신적 고통을 겪는 난민을 위한 챗봇 '카림' 개발

22. CCTV

국내 스타트업 '마인드셋' - '마인드아이': 하드웨어 없이 상황, 물체 식별

23. 영화 예고편 제작

'모건' 예고편 - 왓슨 100 여편의 공포영화 학습.

제작 기간을 1 개월에서 24 시간으로 단축

24. 경주용 차

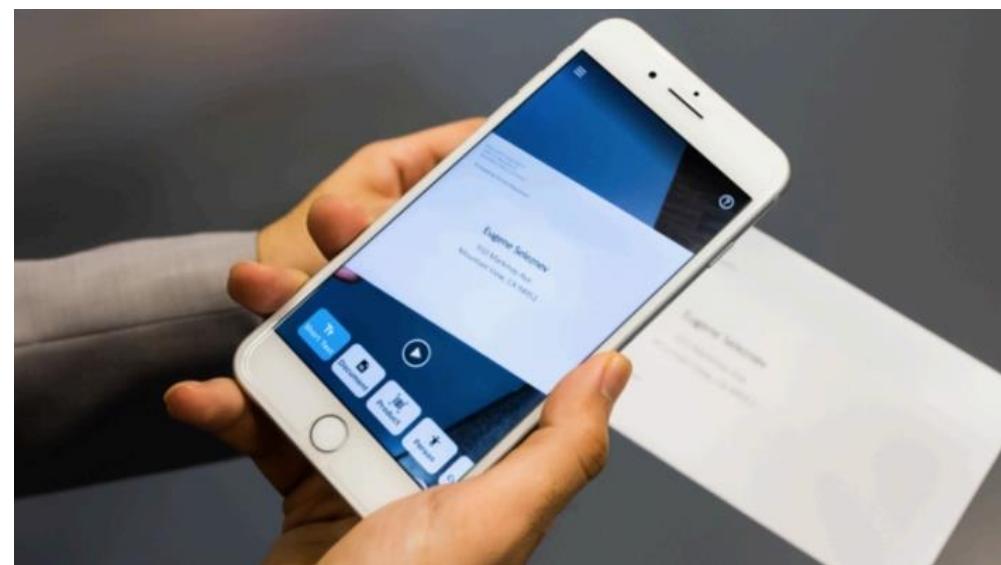
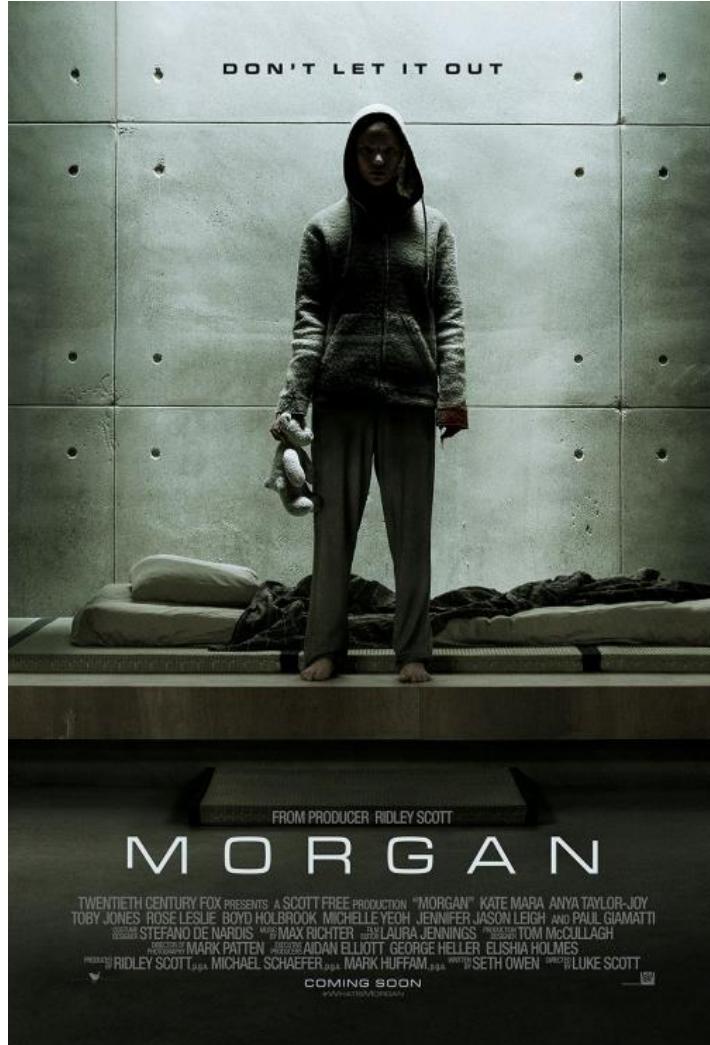
무인 경주용 차량 '로보카' - 드라이버가 탑승하지 않는 '로보레이스' 출전

25. 시각장애인의 눈

마이크로소프트 '씨잉 AI' - 시각장애인에게 주변 환경, 인물, 사물 설명

인공지능 50선

모건 , 로보카 , SEEING AI



인공지능 50선

26. 명품가방 판별

스타트업 '엔트루피' - 3 만여종의 핸드백 지갑을 98% 정확도로 판별

27. 이유식 재료 선정

일본 식료품 업체 '큐피' - 400 개 이상의 5 톤 식재료에서 재료 판별

28. 자연재해 예측

오재호 부경대 교수팀 - 기상변화 예측 '알파멧' 개발, 한국지 지형 데이터 활용

29. 매장 레이아웃 개선

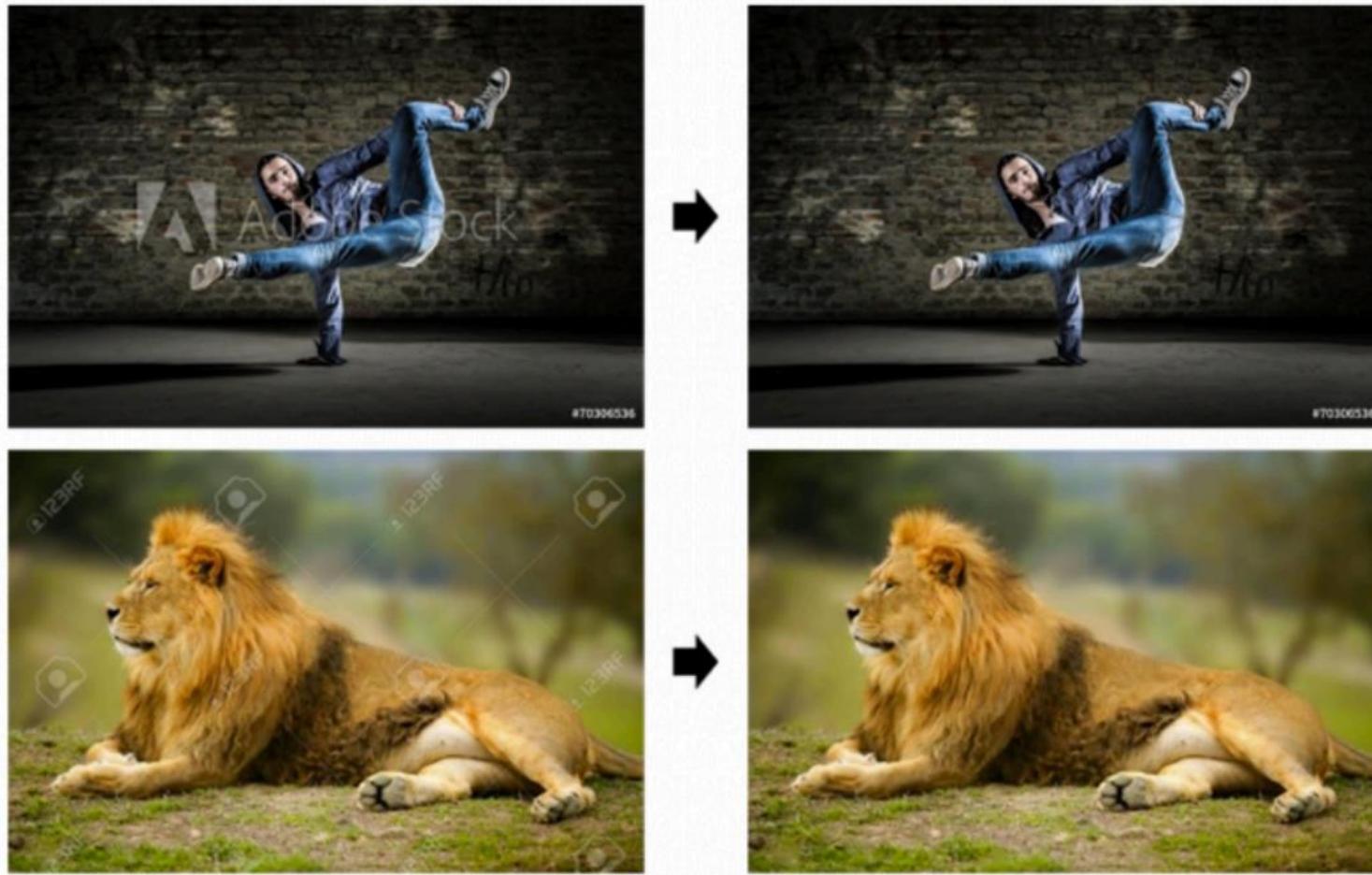
인컨텍스트 솔루션스 - 인공지능과 VR 을 조합해서 매장 레이아웃 구성

30. 워터마크 제거

구글에서 워터마크를 제거하는 논문 공개

인공지능 50선

워터마크 제거



인공지능 50선

31. 음란물 필터

네이버 음란물 필터 기술 '네이버 X-eye' - 모든 이미지에 대해 적용 (98.1%)

32. 꽃가루 알레르기 위험 지수 관리

기상청 '꽃가루 농도위험지수'에 AI 적용 - 15.9%에서 69.4%로 개선

33. 항만 관리

일본 국토교통성 - 공장 출하, 도로 / 항만 혼잡도, 선박 도착시간 처리

34. 상어 감지

무인항공기 업체 '리틀 리퍼' - 인공지능 드론 사용 . 20%에서 90%로 향상

35. 폐기물 분류

행정안전부 - AI 객체인식 기반 대형 폐기물 처리시스템 구축 사업 추진 (은평구)

인공지능 50선

상어 감지



인공지능 50선

36. 치매 예측

캐나다 맥길대 - 치매 발생 2년 전에 예측 가능 (84% 정확도)

37. 심정지 예측

호흡수, 심장박동수, 산소포화도, 혈압 등의 데이터 학습

현재 기술로는 30분 전에 예측 - 인공지능은 24시간 전에 가능 (70% 이상)

38. 작곡

구글 '마젠타 프로젝트' - 기계가 예술을 창조할 수 있는지 알아보는 프로젝트

엔신스 - 1천개의 악기, 30만개의 데이터베이스로부터 새로운 소리 및 음악 생성

39. 시신경 질환 예측

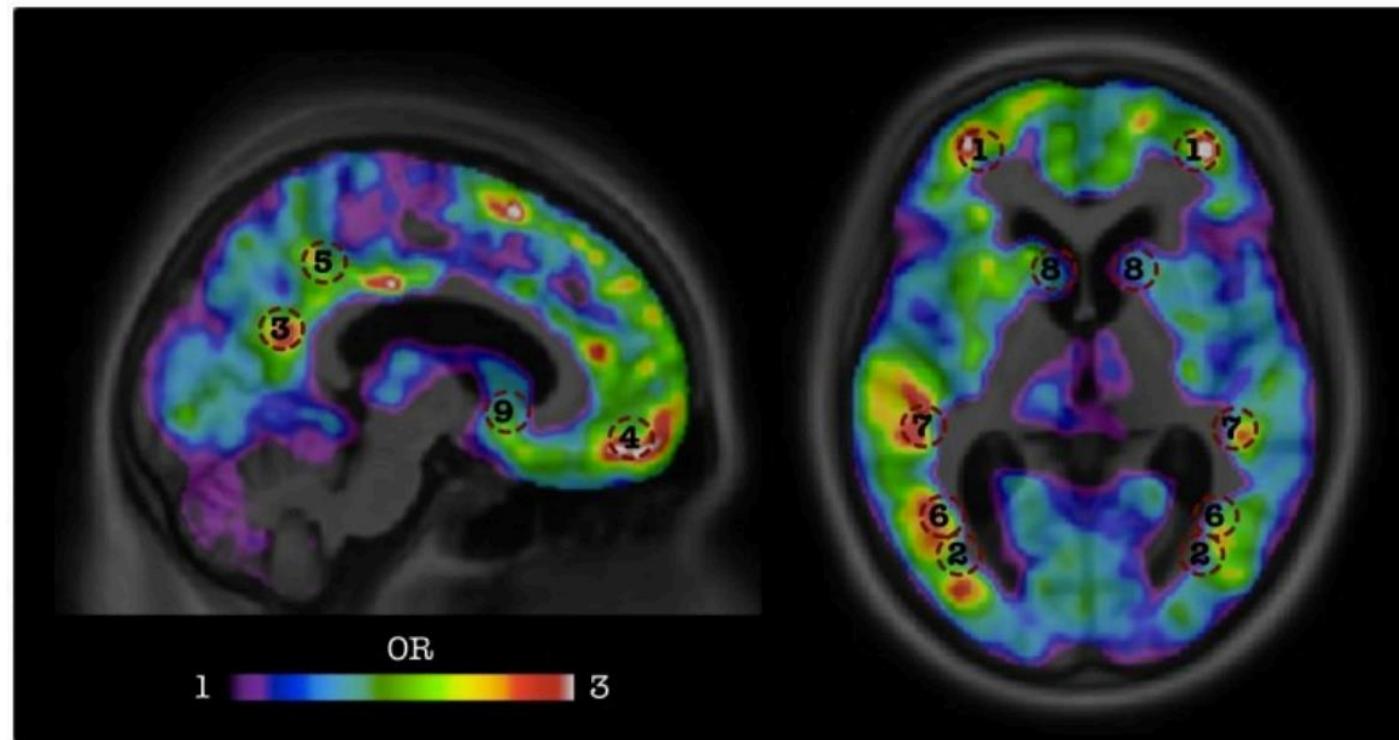
김안과 병원 - 시신경 질환 예측 연구. 녹내장 진단 100% 정확도 달성

40. 저작권 침해 예방

한국저작권보호원 - 불법복제 영상 유통 차단에 활용

인공지능 50선

치매 예측, 마젠타



인공지능 50선

41. 졸음운전 예측

다이아이치교통산업 - 운전자의 심박수, 운전자 태도, 주행 모습 데이터 수집

42. 다이어트 분야

비만치료 전문기업 '365cm 네트웍스' - 인공지능 흡입기술 'MAIL 시스템'
지방흡입술 집도의의 전체 수술 동작 저장 및 분석

43. 영상 조작

워싱턴 대학교 - 음성에 맞춘 '립싱크' 개발. 음성만 같은 다른 영상 제작

44. 드레스 제작

'마르케사' - IBM 과 협업. 인공지능을 감정을 표현한 드레스 제작

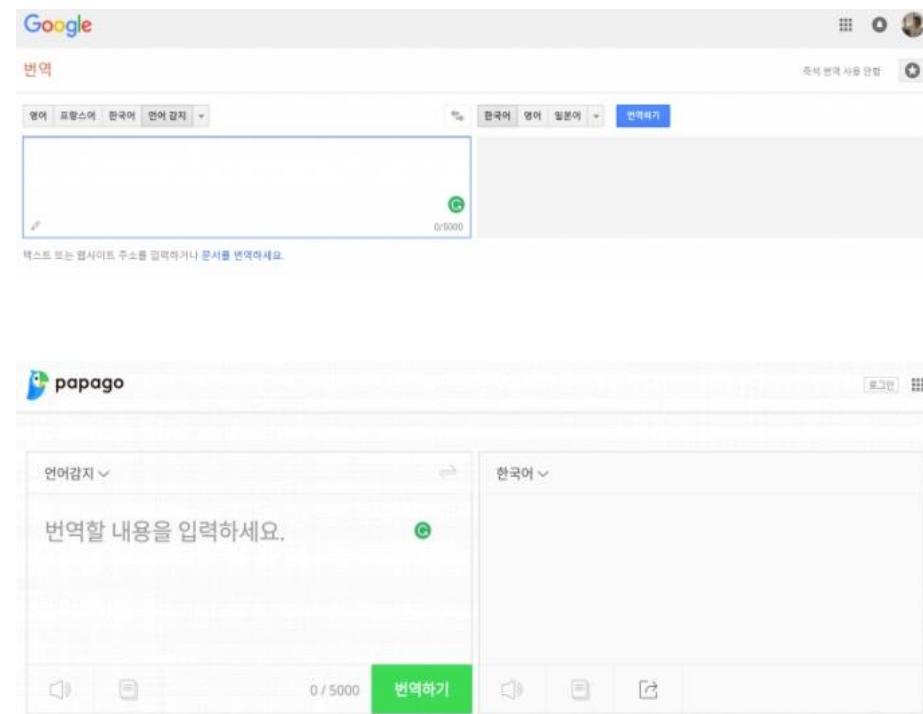
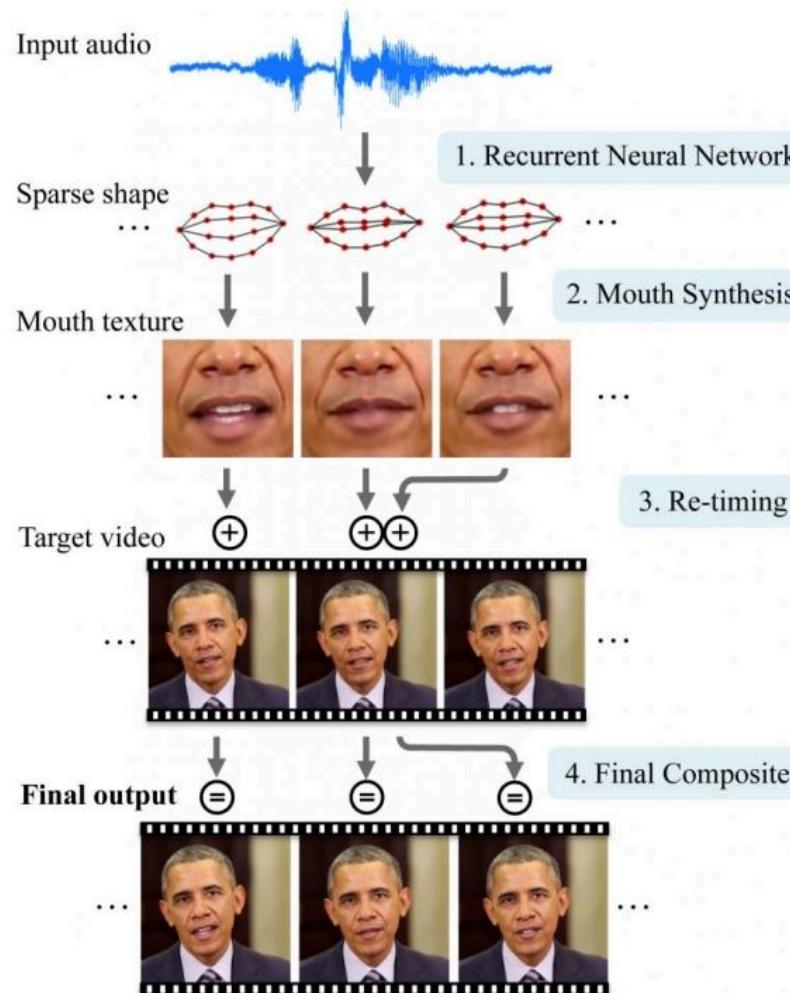
45. 번역기

구글 번역, 네이버 파파고, 시스트란

인간 vs. 인공지능 번역 대결 - 인간이 승리했지만, 평가와 번역 환경 공정성 논란

인공지능 50선

영상 조작, 번역기



인공지능 50선

46. 채용 도우미

리쿠르트 '헤이스' - 인력 정보에 바탕해서 헤드 헌터의 업무 경감

47. 목소리 재현

스타트업 '라이어버드' - 60 초의 음성 데이터로 목소리 재현 . 감정 표현 가능

48. 신용카드 거래 승인

마스터카드 - '디시전 인텔리전스' . 고객 개별 거래 평가 , 점수 , 학습

모든 거래를 분석하고 산출된 정보를 바탕으로 승인 여부 결정

49. 영화 선호도 예측

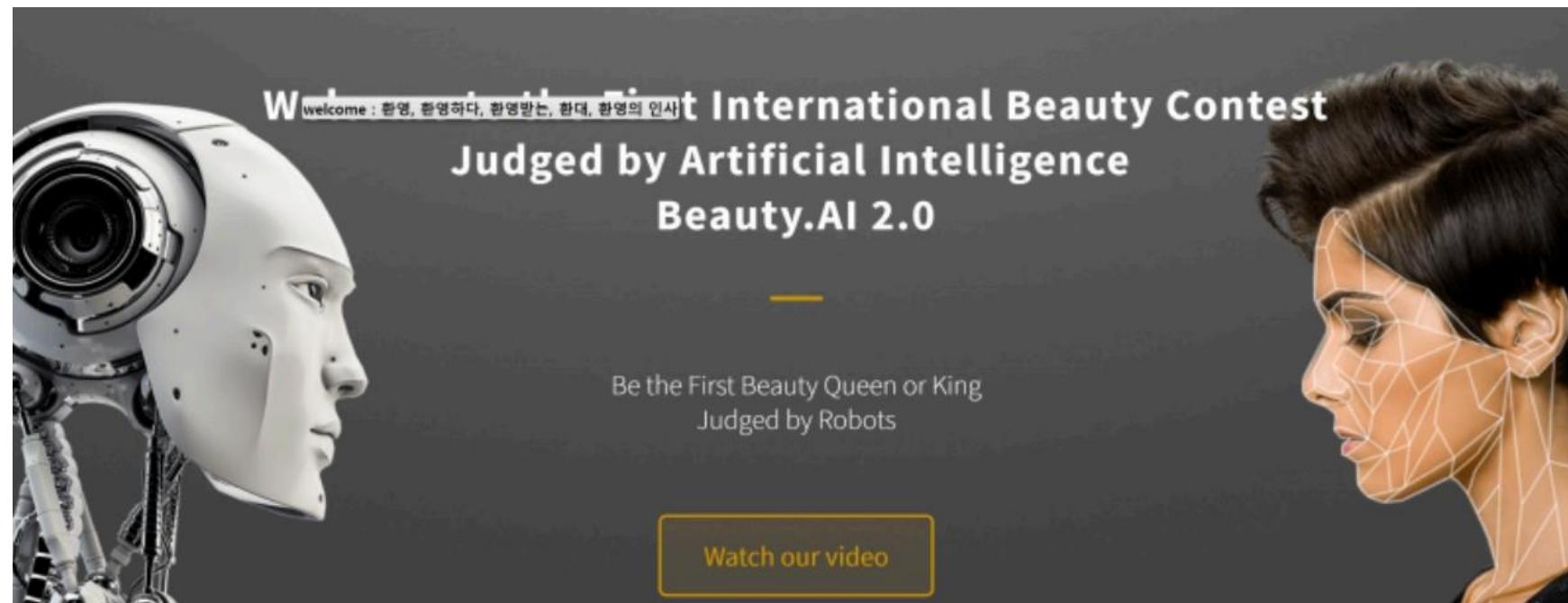
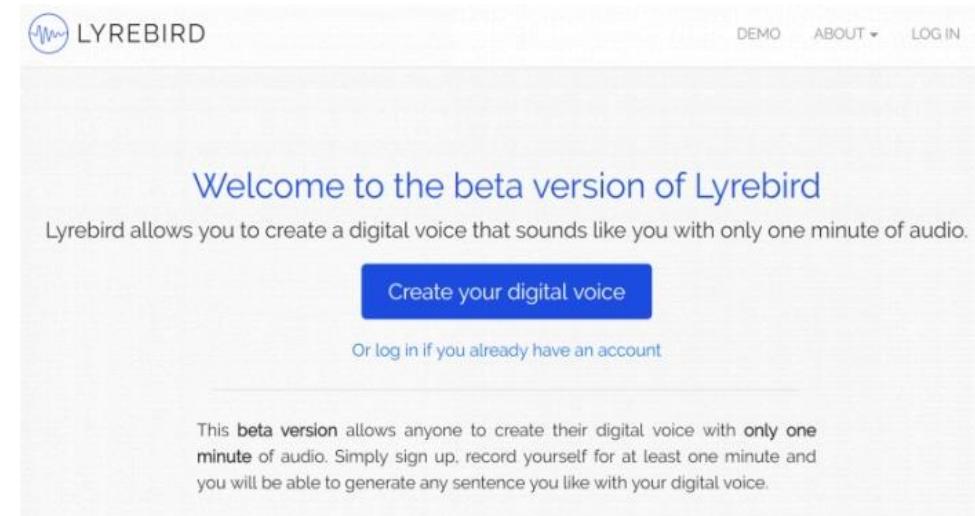
디즈니 리서치팀 - 단편 이야기를 평가할 수 있는 신경망 연구

50. 미인대회 심사위원

'뷰티닷에이아이' - 로봇 판정단으로만 구성된 미인대회 . 백인 편향 알고리즘 논란

인공지능 50선

미인대회, 목소리 재현





DEEP LEARNING OVERVIEWS

둘러보기

머신러닝의 확산 - 폭발적인 증가

- ▶ 무어의 법칙 (Moore's Law) 으로 컴퓨팅 비용이 급격히 낮아져 지금은 최소한의 비용으로 강력한 컴퓨팅 성능을 폭넓게 이용할 수 있다 .
- ▶ 새롭고 혁신적인 알고리즘이 더욱 빠른 결과를 제공한다 .
- ▶ 데이터 과학자들이 머신러닝을 효과적으로 적용하기 위한 이론과 실무지식을 축적했다.

둘러보기

PROGRAMMING LANGUAGES - 나무위키

▶ R 언어

현재 가장 많이 쓰이는 통계 기반 프로그래밍 언어이다 .

R 언어의 강점은 간단한 코딩으로 충분히 인지 가능할 정도로 시각화된 데이터를 얻어내는 것에 있다 . 이를 통해 개발 파이프라인을 단축시키는데 많은 기여를 한다 .

▶ 파이썬

R 언어에 이어 이 분야에서 두 번째로 많이 쓰이는 언어이며 , numpy 라이브러리를 써서 기계학습 알고리즘을 코딩하도록 도와준다 . 비록 R 언어보다 코딩에 다소 시간이 걸리지만, 파이썬 언어의 대표적 장점인 이식성이 있어 다양한 분야에서 사용되어지며 , scipy 라이브러리 등을 추가하여 내부 변수의 계산을 하거나 Cython 등을 이용하여 알고리즘의 속도를 빠르게 하기에도 용이하다 .

▶ Matlab

수학적 정밀도가 어느 정도 보장되는 언어이다보니 주로 연구실 등에서 사용되어진다 .

둘러보기

ELON MUSK VS. MARK ZUCKERBERG

The image shows a Twitter post from Darren Cunningham (@dcunni) dated July 25, 2017. The post features a photo of Mark Zuckerberg in profile, looking towards the right. The caption reads: "Zuckerberg blasts @elonmusk warnings against artificial intelligence as 'pretty irresponsible' bizjournals.com/sanjose/news/2... @svbizjournal #ai". Below the post, there is a reply from Elon Musk (@elonmusk) stating: "I've talked to Mark about this. His understanding of the subject is limited." The Twitter interface includes a sidebar for Elon Musk's profile, which shows his bio as CEO of Tesla, SpaceX, OpenAI & Neuralink, and his location as Boring. It also shows his follower count at 1,541 and the date of the tweet as July 25, 2017, at 12:07.

둘러보기

MARK ZUCKERBERG

- ▶ 내가 AI 에 낙관적인 이유 중 하나는 (AI 가) 다양한 분야의 시스템을 개선하기 위한 기초 연구를 더 잘할 수 있게 한다는 점이다 . 질병 진단에서부터 건강 유지 , 안전한 자율주행차 , 당신의 뉴스피드에 더 나은 콘텐츠를 보여주는 것 , 더 연관성이 높은 검색 결과를 제공하는 것까지 여러분야의 시스템을 개선할 수 있다 . 우리가 우리의 AI 방법을 향상시킬때마다 매번 이 모든 시스템이 더 나아진다 . 이는 세상을 더 좋게 만들 가능성이 있다 .



DEEP LEARNING

TERMS

용어 정리

1. 회귀분석 ([HTTP://MATH7.TISTORY.COM/118](http://MATH7.TISTORY.COM/118)에서 발췌)

- ▶ 점들이 퍼져있는 형태에서 패턴을 찾아내고, 이 패턴을 활용해서 무언가를 예측하는 분석 .
- ▶ 새로운 표본을 뽑았을 때 평균으로 돌아가려는 특징이 있기 때문에 붙은 이름
- ▶ 회귀 (回歸 돌회, 돌아갈 귀)라는 용어는 일반적으로 '돌아간다'는 정도로만 사용하기 때문에 회귀로부터 '예측'이라는 단어를 떠올리기는 쉽지 않다 .

용어 정리

2. LINEAR REGRESSION

- ▶ 2 차원 좌표에 분포된 데이터를 1 차원 직선 방정식을 통해 표현되지 않은 데이터를 예측하기 위한 분석 모델.
- ▶ 머신러닝 입문에서는 기본적으로 2 차원이나 3 차원까지만 정리한다.
- ▶ 여기서는 편의상 1 차원 직선으로 정리하고 있다.
 xy 축 좌표계에서 직선을 그렸다고 생각하면 된다.

용어 정리

3. HYPOTHESIS

- ▶ **Linear Regression**에서 사용하는 1 차원 방정식을 가리키는 용어로 ,
우리말로는 가설이라고 한다 . 수식에서는 $h(x)$ 또는 $H(x)$ 로 표현된다 .
- ▶ 최저점 (**minimize cost**) 이라는 정답을 찾기 위한 가정이기 때문에 가설이라고 부를 수 있다 .
- ▶ $H(x) = Wx + b$
 $\Rightarrow x$ 에 대한 1 차 방정식

용어 정리

4. COST (비용)

- ▶ 앞에서 설명한 **Hypothesis** 방정식에 대한 비용 (**cost**) 으로 방정식의 결과가 크게 나오면 좋지 않다고 얘기하고 루프를 돌 때마다 **w** 와 **b** 를 비용이 적게 발생하는 방향으로 수정하게 된다 .
- ▶ 미분을 사용해서 스스로 최저 비용을 찾아간다 .
- ▶ **Gradient Descent Algorithm** 을 사용해서 최저 비용을 찾는다 .

용어 정리

5. COST 함수

- ▶ **Hypothesis** 방정식을 포함하는 계산식
- ▶ 현재의 기울기 (W) 와 절편 (b) 에 대해 비용을 계산해 주는 함수
- ▶ W 와 b 가 변함에 따라 반드시 **convex**(오목) 한 형태로 설계되어야 하는 것이 핵심.
- ▶ **convex** 하지 않다면 , 경사를 타고 내려갈 수 없기 때문에 최저점 계산이 불가능 해질 수 있다 .
- ▶ **Linear Regression** 을 비롯한 머신러닝 전체에서 최소 비용을 검색하기 위한 역할 담당

용어 정리

6. GRADIENT DESCENT ALGORITHM

- ▶ 딥러닝의 핵심 알고리즘
- ▶ 경사타고 내려가기 , 경사하강법 등의 여러 용어로 번역되었다 .
- ▶ 미분을 사용해서 비용이 작아지는 방향으로 진행하는 알고리즘
- ▶ 생각보다 어렵지 않고 간단한 미분 정도만 이해하면 알고리즘 자체는 너무 단순하다 .
- ▶ 텐서플로우에 포함된 **Optimizer** 는 대부분 **Gradient Descent Algorithm** 에서 파생된 방법을 사용하고 있다 .



DEEP LEARNING

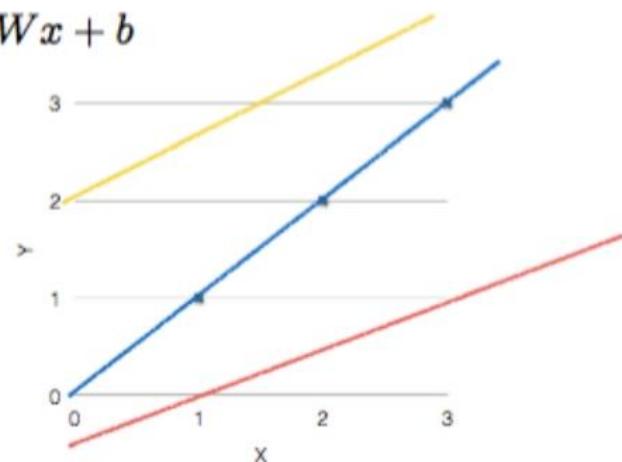
LINEAR REGRESSION

LINEAR REGRESSION

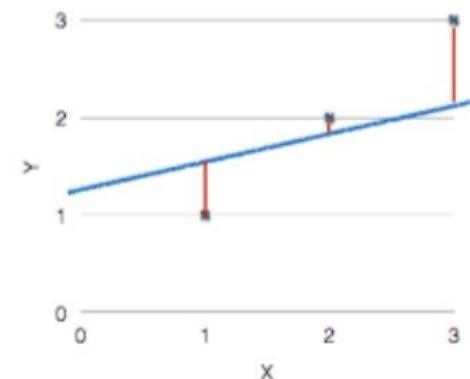
좋은 가설 ?

(Linear) Hypothesis

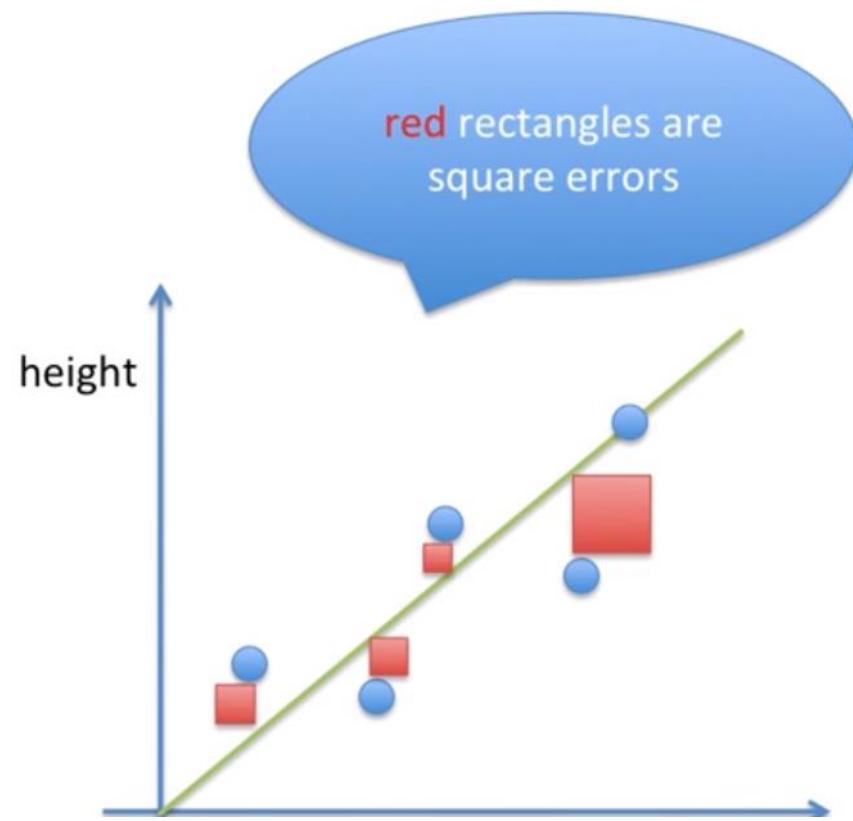
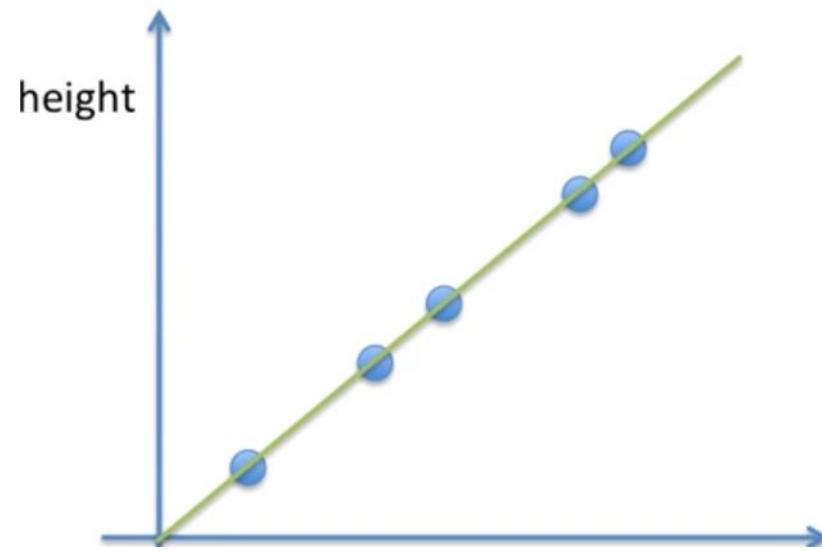
$$H(x) = Wx + b$$



Which hypothesis is better?



Square Error- (difference between prediction and real value)²

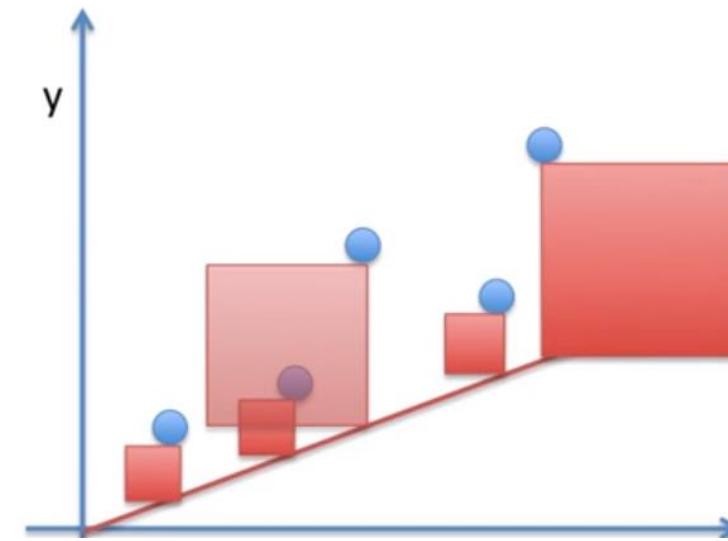
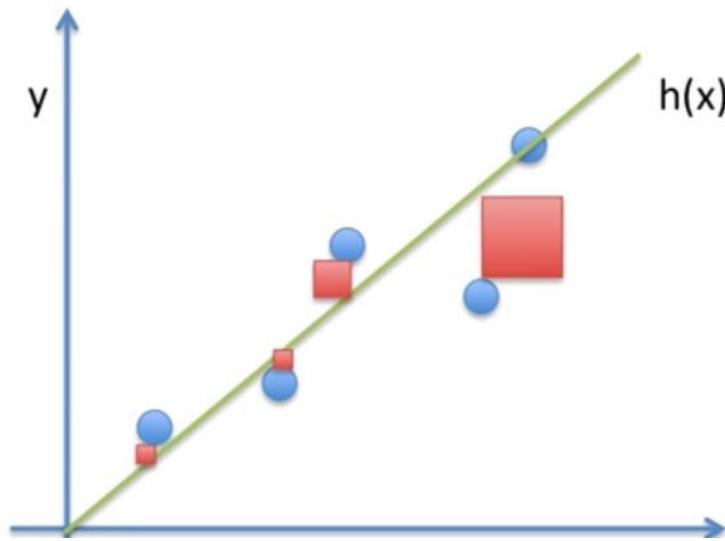


Find linear equation has Least Mean Square (LMS) Error

$$\text{Error} = h(x) - y$$

$$\text{Square Error} = (h(x) - y)^2$$

$$\text{Mean Square Error} = \frac{1}{n} \sum (h(x) - y)^2$$



LINEAR REGRESSION

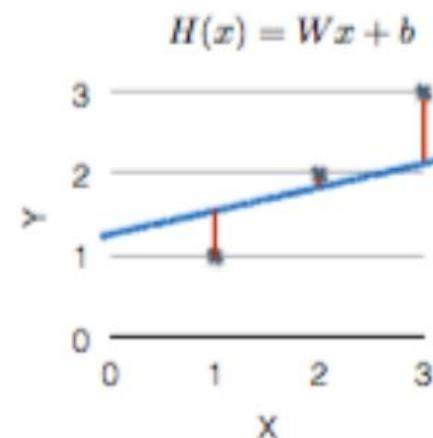
직선으로부터 점까지의 거리 계산

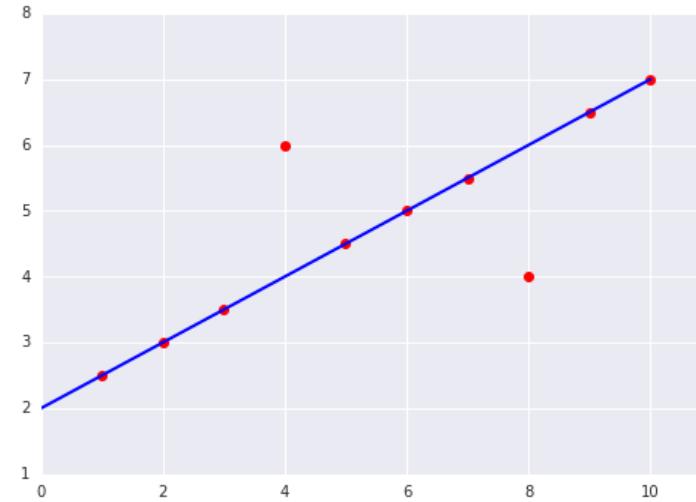
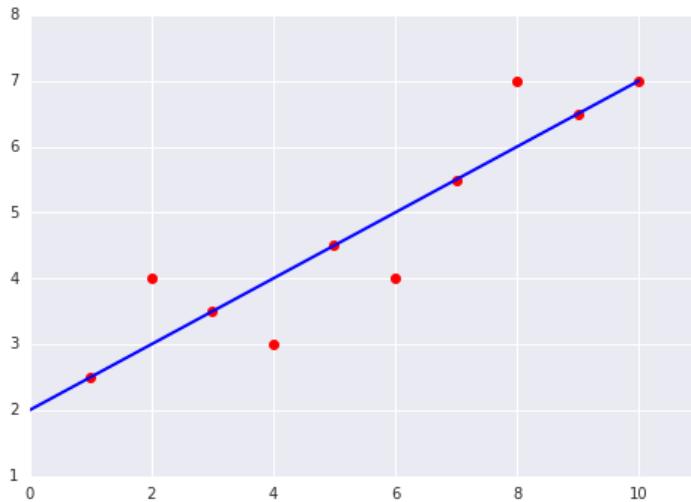
Cost function

- How fit the line to our (training) data

$$\frac{(H(x^{(1)}) - y^{(1)})^2 + (H(x^{(2)}) - y^{(2)})^2 + (H(x^{(3)}) - y^{(3)})^2}{3}$$

$$cost = \frac{1}{m} \sum_{i=1}^m (H(x^{(i)}) - y^{(i)})^2$$





선에 위치한 8개의 점에 발생하는 총 손실은 0입니다. 그러나, 선에서 벗어난 점은 2개밖에 안 되지만, 2개의 점 모두 선에서 벗어난 정도가 왼쪽 그림의 이상점에 비해 2배 더 큽니다. 제곱 손실값의 경우 이러한 차이가 증폭되므로 오프셋 2의 경우 오프셋 1보다 4배 더 큰 손실이 발생합니다.

$$MSE = \frac{0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2}{10} = 0.8$$

LINEAR REGRESSION

COST 수식

$$cost = \frac{1}{m} \sum_{i=1}^m (H(x^{(i)}) - y^{(i)})^2$$

$$H(x) = Wx + b$$

$$cost(W, b) = \frac{1}{m} \sum_{i=1}^m (H(x^{(i)}) - y^{(i)})^2$$

LINEAR REGRESSION

최저 COST 를 만드는 W 와 B 는 ?

- ▶ $H(x) = wx + b$
- ▶ $x = [1, 2, 3]$
 $y = [1, 2, 3]$
- ▶ w 와 b 가 아래와 같다면 ?
 1. $w = 1/2, b = 0$
 2. $w = 0, b = 2$
 3. $w = 1, b = 1/2$
 4. $w = 1, b = 1$

LINEAR REGRESSION

THE GOAL OF LINEAR REGRESSION

- ▶ Cost 를 최소로 만드는 W(Weight) 와 b(bias) 를 찾는 것 .
- ▶ Linear Regression 의 목표는 곧 머신러닝의 목표 !



DEEP LEARNING

LINEAR REGRESSION

GRADIENT DESCENT ALGORITHM

HOW TO HIDE BIAS?

Hypothesis and Cost

$$H(x) = Wx + b$$

$$cost(W, b) = \frac{1}{m} \sum_{i=1}^m (H(x^{(i)}) - y^{(i)})^2$$

Simplified hypothesis

$$H(x) = Wx$$

$$cost(W) = \frac{1}{m} \sum_{i=1}^m (Wx^{(i)} - y^{(i)})^2$$

GRADIENT DESCENT ALGORITHM

COST FUNCTION GRAPH

What $\text{cost}(W)$ looks like?

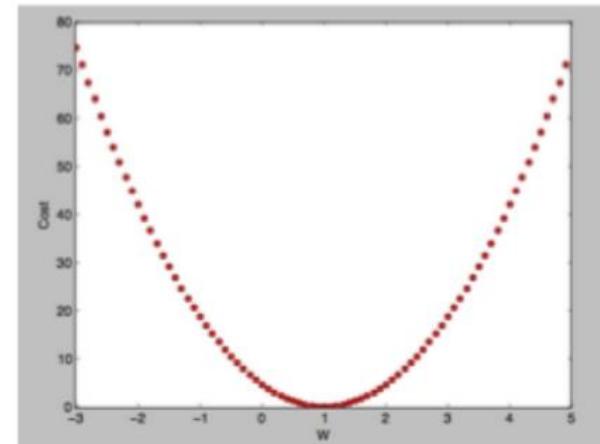
$$\text{cost}(W) = \frac{1}{m} \sum_{i=1}^m (Wx^{(i)} - y^{(i)})^2$$

x	y
1	1
2	2
3	3

- $W=1, \text{cost}(W)=0$
 $\frac{1}{3}((1*1-1)^2 + (1*2-2)^2 + (1*3-3)^2)$
- $W=0, \text{cost}(W)=4.67$
 $\frac{1}{3}((0*1-1)^2 + (0*2-2)^2 + (0*3-3)^2)$
- $W=2, \text{cost}(W)=?$

How to minimize cost?

$$\text{cost}(W) = \frac{1}{m} \sum_{i=1}^m (Wx^{(i)} - y^{(i)})^2$$



GRADIENT DESCENT ALGORITHM

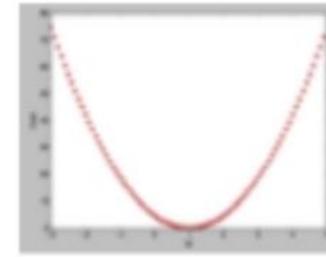
GRADIENT DESCENT ALGORITHM

- Minimize cost function
- Gradient descent is used many minimization problems
- For a given cost function, $\text{cost}(W, b)$, it will find W, b to minimize cost
- It can be applied to more general function: $\text{cost}(w_1, w_2, \dots)$

GRADIENT DESCENT ALGORITHM

HOW IT WORKS?

- Start with initial guesses
 - Start at 0,0 (or any other value)
 - Keeping changing W and b a little bit to try and reduce $\text{cost}(W, b)$
- Each time you change the parameters, you select the gradient which reduces $\text{cost}(W, b)$ the most possible
- Repeat
- Do so until you converge to a local minimum
- Has an interesting property
 - Where you start can determine which minimum you end up



GRADIENT DESCENT ALGORITHM

FORMAL DEFINITION

$$\text{cost}(W) = \frac{1}{m} \sum_{i=1}^m (Wx^{(i)} - y^{(i)})^2 \quad W := W - \alpha \frac{\partial}{\partial W} \frac{1}{2m} \sum_{i=1}^m (Wx^{(i)} - y^{(i)})^2$$

$$\text{cost}(W) = \frac{1}{2m} \sum_{i=1}^m (Wx^{(i)} - y^{(i)})^2 \quad W := W - \alpha \frac{1}{2m} \sum_{i=1}^m 2(Wx^{(i)} - y^{(i)})x^{(i)}$$

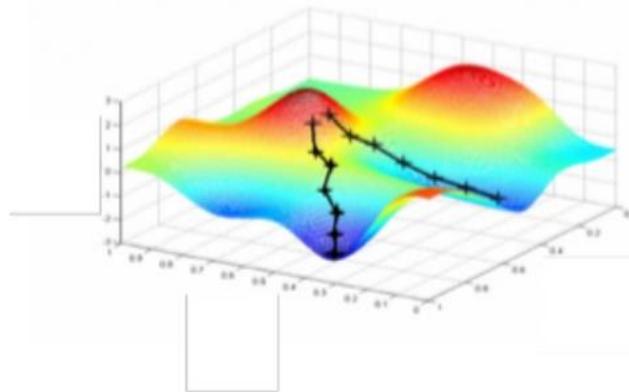
$$W := W - \alpha \frac{\partial}{\partial W} \text{cost}(W)$$

$$W := W - \alpha \frac{1}{m} \sum_{i=1}^m (Wx^{(i)} - y^{(i)})x^{(i)}$$

GRADIENT DESCENT ALGORITHM

CONVEX FUNCTION

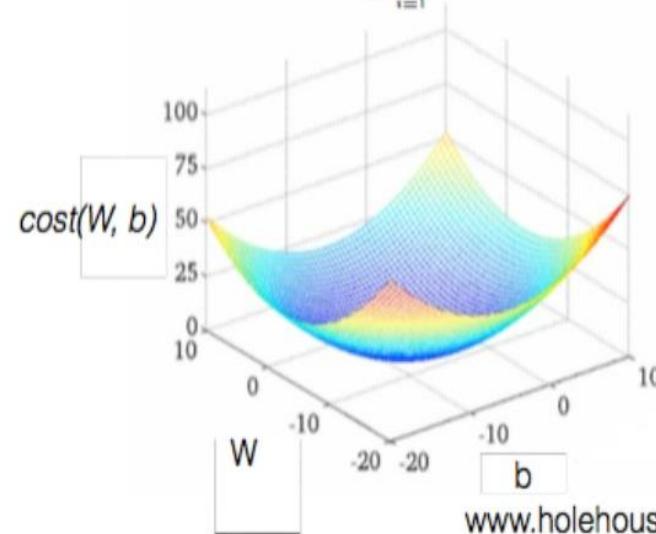
Convex function



www.holehouse.org/mlclass/

Convex function

$$\text{cost}(W, b) = \frac{1}{m} \sum_{i=1}^m (H(x^{(i)}) - y^{(i)})^2$$



www.holehouse.org/mlclass/



DEEP LEARNING

NEURAL NETWORKS

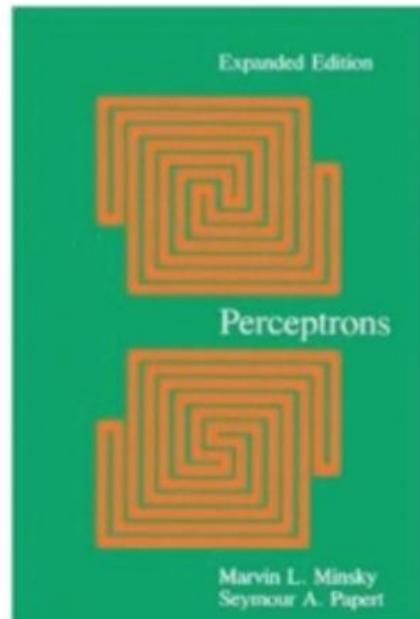
HISTORY

NEURAL NETWORKS HISTORY

PERCEPTRONS

Perceptrons (1969)

by Marvin Minsky, founder of the MIT AI Lab



- We need to use MLP, multilayer perceptrons (multilayer neural nets)
- No one on earth had found a viable way to train MLPs good enough to learn such simple functions.

NEURAL NETWORKS HISTORY

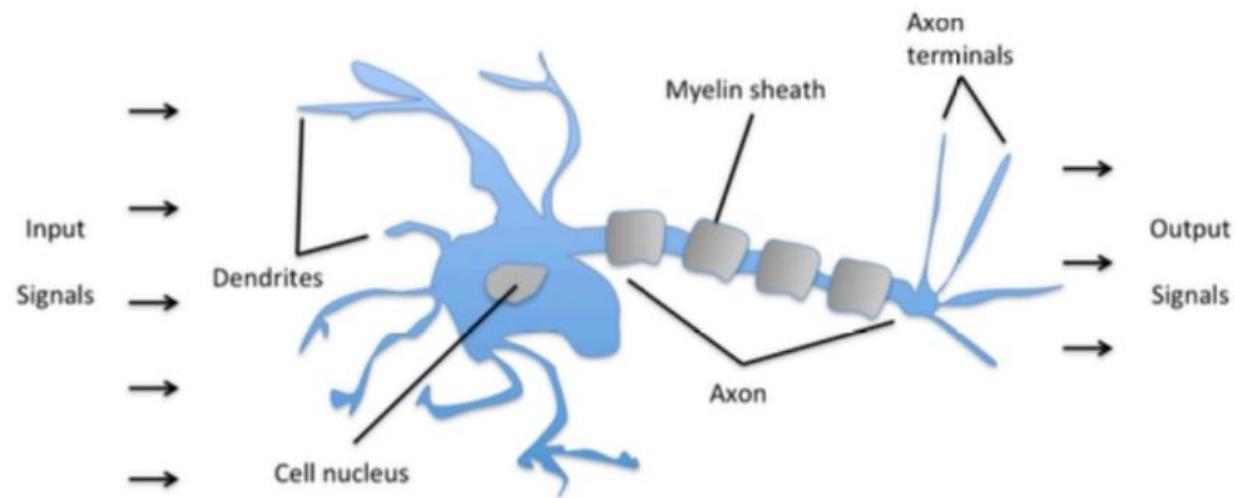
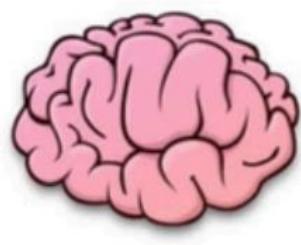
XOR PROBLEM

(Simple) XOR problem: linearly separable?



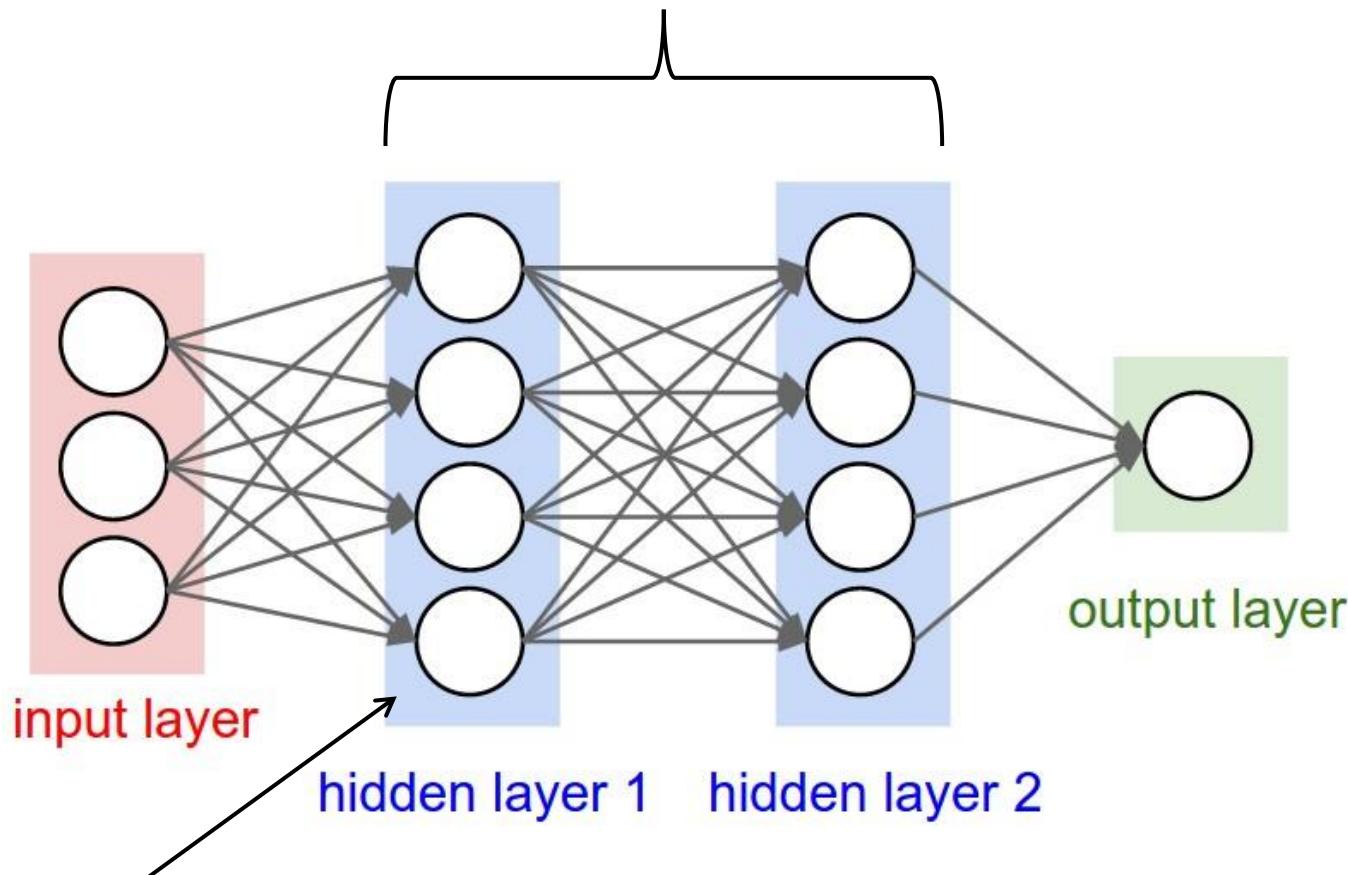
NEURAL NETWORKS HISTORY

SCHEMATIC OF A BIOLOGICAL NEURON



Schematic of a biological neuron.

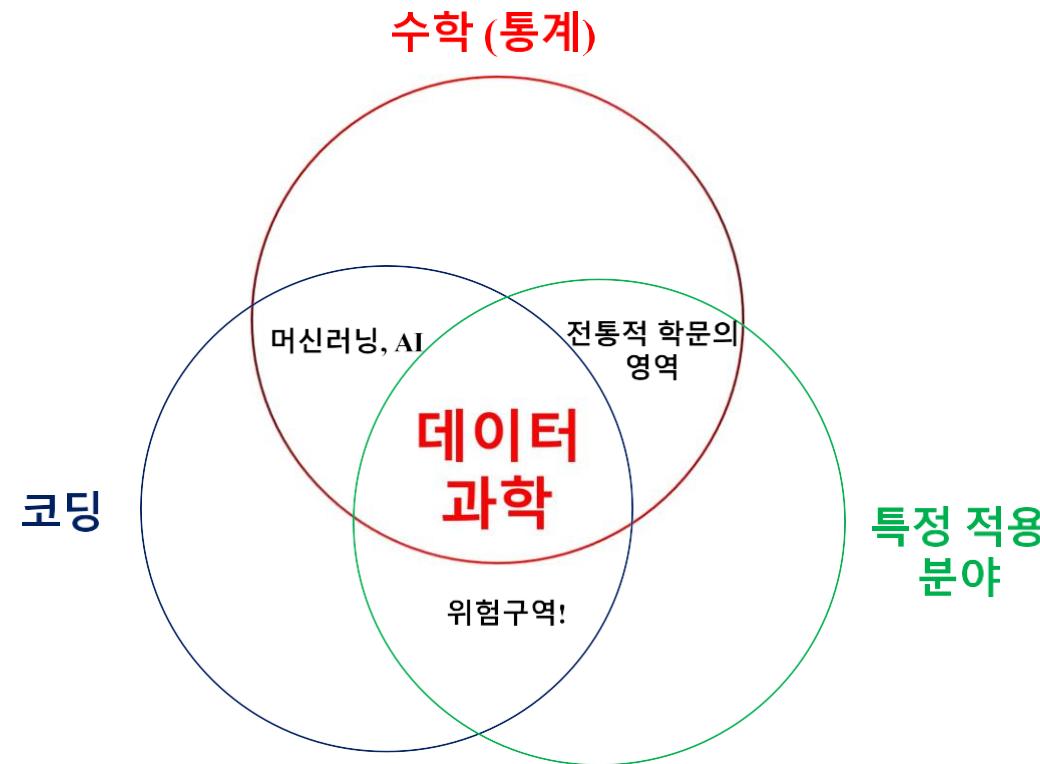
Deep



Wide

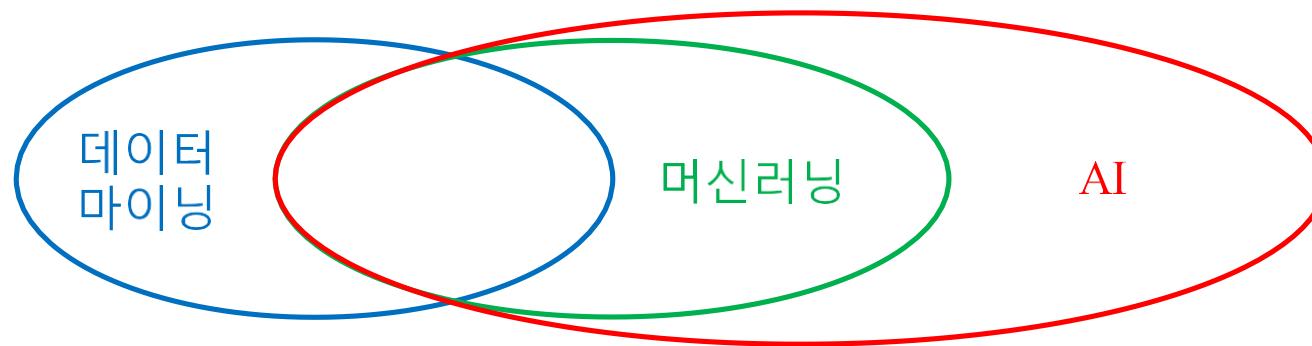
Python 머신러닝

데이터 과학



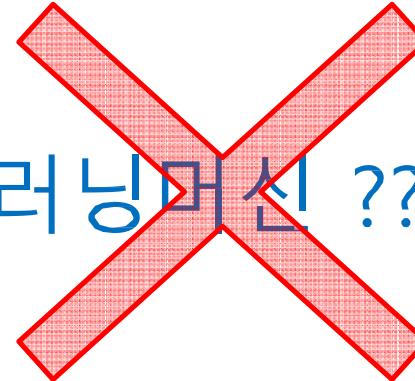
→ 데이터 과학은 빅데이터, 머신러닝, AI (인공지능)을 다루는 학문이다.

Overview



기계학습

러닝머신 ???



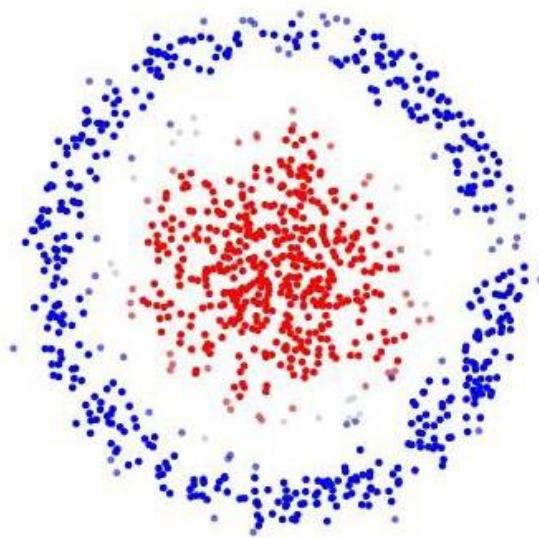
기계학습

머신러닝 ! (Machine Learning)

→ 통계 모형과 데이터를 사용한 학습과 예측을 의미한다.

기계학습 : 유형

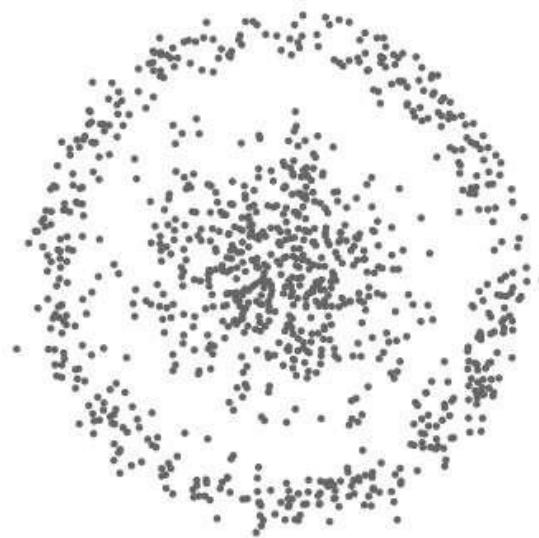
지도학습 (supervised learning):



- 학습목표와 내용이 분명하다. 학습패턴은 주어진다.

기계학습 : 유형

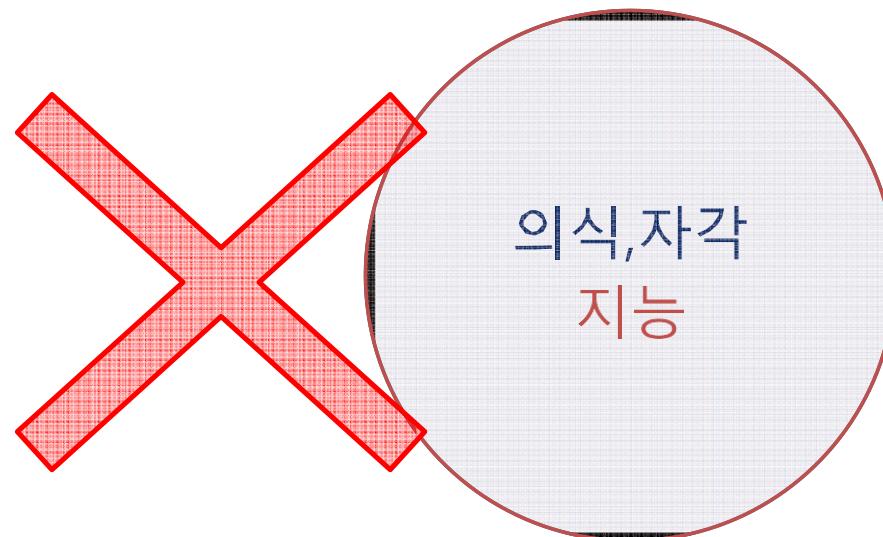
비지도학습 (unsupervised learning):



- 학습목표와 내용이 불분명하다. 스스로 패턴을 찾아 내어야 한다.

인공지능

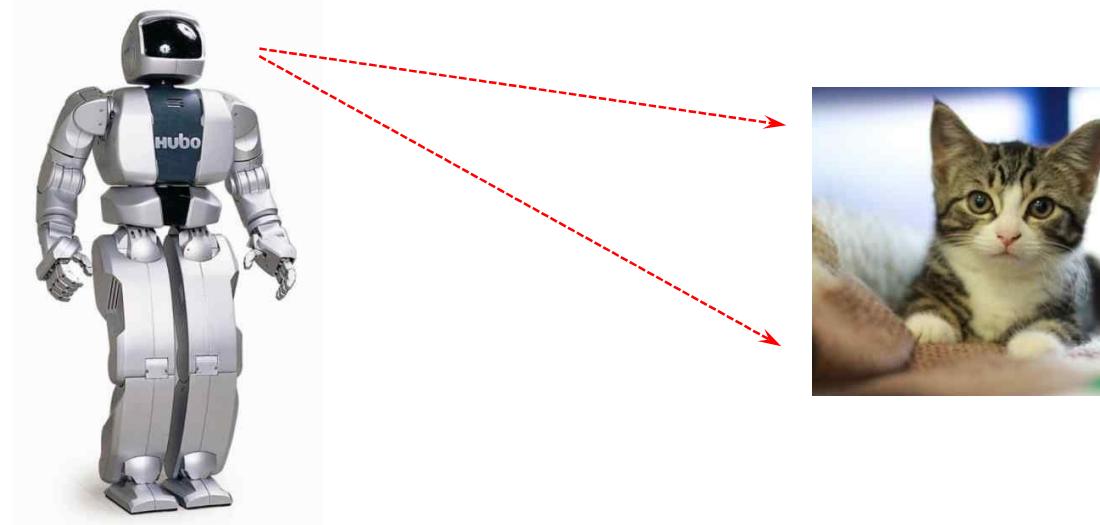
의식과 지능의 분리:



인공지능

인공지능

동물?? ??
고양이|??



신문기사 (2017-07-01)

▶ 데일리한국 > 경제 > 최신기사

국내 개발 인공지능(AI), 금융권서 가장 많이 쓴다

금융분야 26개·국방 23개·미디어 22개·유통 의료 17개 순

조진수 기자 rokmc4390@hankooki.com



사진=유로이미지

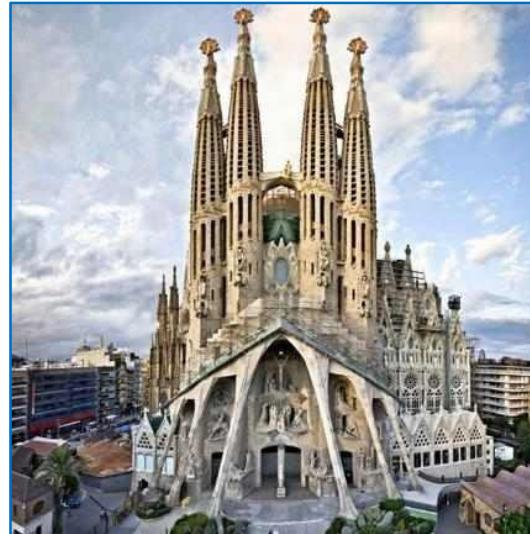
[데일리한국 조진수 기자] 국내 개발 인공지능(AI) 기술은 금융 분야에서 가장 많이 쓰이는 것으로 밝혀졌다.

30일 소프트웨어 정책연구소가 발간한 '국내 지능정보산업 실태에 대한 연구' 보고서에 의하면 35개 인공지능 기술 개발기업의 95개 제품(중복 포함 238개) 중 금융분야에 쓰인 경우가 26개로 가장 높았다.

- 국내 개발 인공지능 (AI), 금융권에서 사용 1위.
- 고객·민원상담 (20.8%), 마케팅 최적화 (18.8%), 이상거래 탐지 (15.8%).
- 머신러닝 (19.8%), 딥러닝·신경망 (18.0%), 자연어 처리 (13.5%), 상황인지 (10.8%).

미완성의 아름다움

스페인 바르셀로나에 위치한 Sagrada Familia 성당



- 건축가 가우디가 1882년 건축 시작.
- 136년째 건축 중. 2028년경 완공 예정.

미완성의 아름다움

- 데이터 분석은 시간과의 싸움입니다.
- 대다수의 경우에는 ‘정답’이 없습니다.
- 100% ‘정답’을 찾기 보다는 80% ‘**쓸만한**’ 답을 빠른 시간에 찾아내는 것이 목표입니다.

데이터 사이언스 도구

데이터 사이언스 도구 (tool):

- 데이터 관리용 도구: 입력, 저장, 보존.
예) MySQL, MongoDB, Hadoop, Excel, 등.
- 데이터 분석용 도구: 변형, 분석, 패턴학습.
→ 머신러닝을 가능케 하는 도구.

데이터 사이언스 도구

데이터 분석용 도구:

- R 프로그래밍 언어: caret, dplyr, tm, 등의 패키지.
- Python 프로그래밍 언어: pandas, numpy, scipy, sklearn, keras, theanos, 등의 패키지.
- 클라우드: 구글, 아마존, MS (Azure).

빅데이터 : 출현 배경

빅데이터의 출현 배경:

- 산업계: 양 ↔ 질 변환 법칙.
- 학계: 거대 데이터 활용 과학 확산. 예) 인간 게놈 프로젝트.
- 관련기술의 발전: +디지털화의 가속, 저장 기술의 발달과 가격 하락.
+인터넷, 모바일, 클라우드 등 관련기술의 보편화.

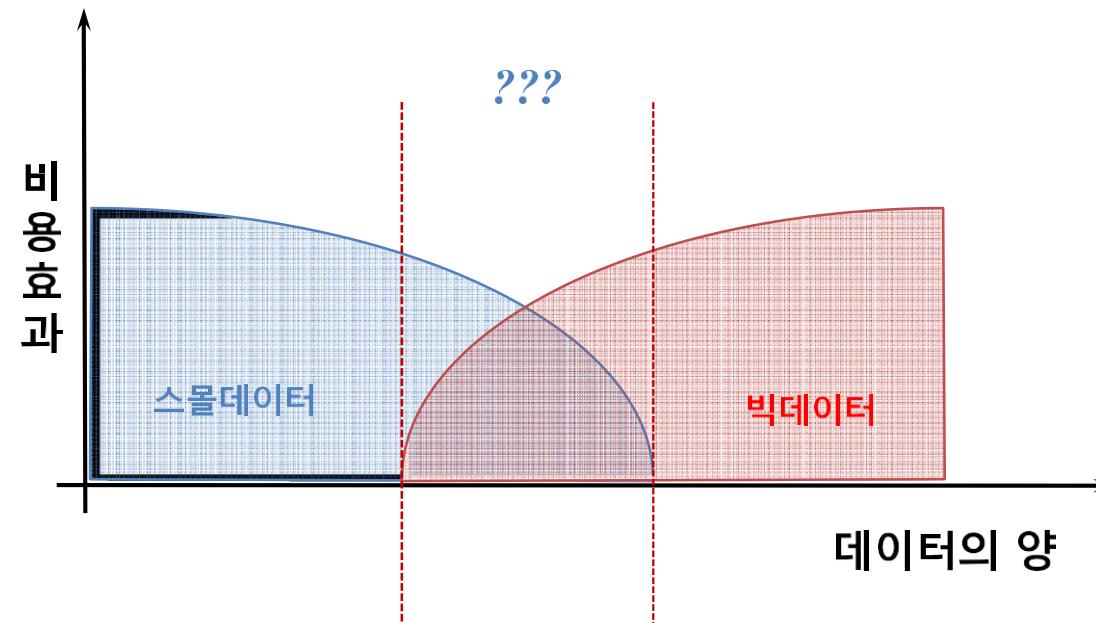
빅데이터 VS 스몰데이터

그러면, ‘스몰데이터’란?

- 사람이 직관적으로 이해할 수 있는 정도의 크기와 형태.
- 쉽게 접근 (accessible), 이해 (understandable), 실행 (actionable)할 수 있는 데이터.
- 일상 생활과 업무에서 자주 접하는 데이터의 유형.

빅데이터 VS 스몰데이터 : 비용효과

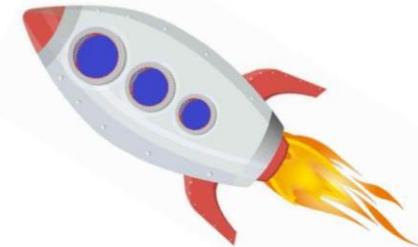
비용효과 비교:



빅데이터 시대의 본질적인 변화

사전처리에서 사후처리로 포커스 이동:

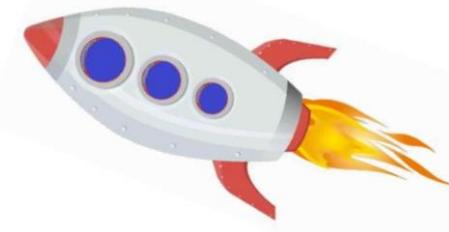
필요한 정보만 수집하는 시스템이 아니고 가능한 한 많은 데이터를 수집하여서 그 데이터를 다양한 방법으로 조합해 숨은 정보를 찾아낸다.



빅데이터 시대의 본질적인 변화

표본조사에서 전수조사로:

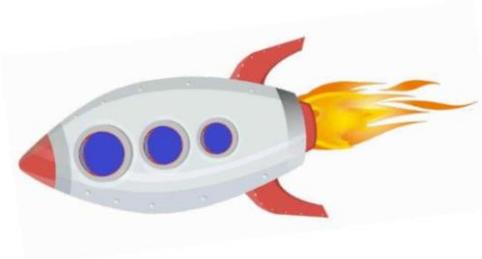
표본조사를 통해서 나타나지 않는 패턴이나 정보를 알아낼 수 있다.



빅데이터 시대의 본질적인 변화

질보다는 양:

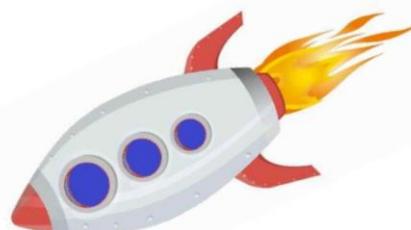
정보가 축척될 수록 오류 정보보다는 양질의 정보가 많아져서 전체적으로 좋은 결과 산출에 긍정적 영향을 미친다.



빅데이터 시대의 본질적인 변화

인과관계에서 상관관계로:

상관관계 분석이 주는 통찰력이 인과관계에 기초해서 할 수 있는 미래예측을 압도해가는 시대가 도래.



빅데이터 Landscape : 2012년도



빅데이터 Landscape : 2016년도



데이터 분석 기술

데이터 분석 기술: 정보 검색 (IR)

- 컴퓨터 시스템을 사용하여 데이터를 색인하고 주제와 관련된 자료를 빠르게 찾아 분석하는 기술.

데이터 분석 기술

데이터 분석 기술: 정보 수집 (Crawling)

- 기존의 웹 검색을 위한 수집기보다 매우 발전된 정보 수집 기술이 필요함.
- 예를 들어서, 트위터와 같은 SNS의 실시간 데이터 수집을 위한 스트림 처리 기술.

데이터 분석 기술

데이터 분석 기술: 기계 학습 (Machine Learning)

- 학습 데이터를 사용하여 통계 모형을 만들고 예측에 적용함.

데이터 분석 기술

데이터 분석 기술: 소셜 네트워크 분석

- SNS의 연결 구조 및 강도 등을 바탕으로 사용자의 인지도 및 영향력 측정.

데이터 분석 기술

데이터 분석 기술: 자연언어 처리 (NLP)

- 인간의 언어를 컴퓨터로 처리하기 위한 기술.
- 형태소 분석, 구문 분석, 개체명 인식 등의 기술을 포함.

데이터 분석 기술

데이터 분석 기술: 텍스트 마이닝 (Text mining)

- 비정형 텍스트 데이터에서 통계적, 연관적 특성을 추출하는 기술.

데이터 분석 기술

데이터 분석 기술: 클라우드 컴퓨팅 & NoSQL

- 빅데이터의 저장과 관리 운영을 위한 기술.
- NoSQL은 전통적인 관계형 데이터베이스의 틀을 벗어난 일관성 모델을 이용.
예) MongoDB, Hadoop, Hbase, Cassandra, 등.

데이터 분석 기술

데이터 분석 기술: 통계 기술 (Statistics)

- 빅데이터의 통계적 의미를 찾고, 그 패턴을 분석하기 위해서는 강력한 통계 기능 필요.

데이터 분석 기술

데이터 분석 기술: 시각화 (Visualization)

- 분석된 결과의 통찰력 있는 이해를 돋기 위한 기술.

데이터 분석의 유형

데이터 분석의 유형: 서술적 분석 (Descriptive Analytics)

- 주어진 상황에서 어떤 일이 벌어졌는지 설명하는데 사용.
- 예) 무슨 일이 일어났는가?, 누가 우리의 고객인가?, 고객은 유형별 어떻게 분류할 수 있는가?
- 시각화, 군집분석 (clustering analysis), 등 사용.

데이터 분석의 유형

데이터 분석의 유형: 진단형 분석 (Diagnostic Analysis)

- 어떤 이유로 특정 현상이 발생한지 원인을 밝히는 것이 주된 목적.
- 예) 왜 매달 이탈 고객이 늘어가는가?
- 군집분석 (clustering analysis), 의사결정트리, 등 사용.

데이터 분석의 유형

데이터 분석의 유형: 예측 분석 (Predictive Analytics)

- 과거 데이터를 사용하여 미래에 대한 전망 제시.
- 예) 신용 카드 거래의 부정 여부, 고객이 높은 요금제로 전환할 확률 예측.
- 회귀분석, 몬테카를로 시뮬레이션, 의사결정트리, 랜덤포레스트, 인공신경망, 등 사용.

데이터 분석의 유형

데이터 분석의 유형: 처방적 분석 (Prescriptive Analytics)

- 구체적인 실행 방안과 예상 효과 제시.
- 업무 방법과 예측 모형의 결합.
- 게임이론, 몬테카를로 시뮬레이션, 의사결정트리, 선형 및 비선형 프로그래밍, 등 사용.

~

데이터 분석을 잘 하려면...



먼저, **선입관**과 **편견**을 버린다.

~

데이터 분석을 잘 하려면...



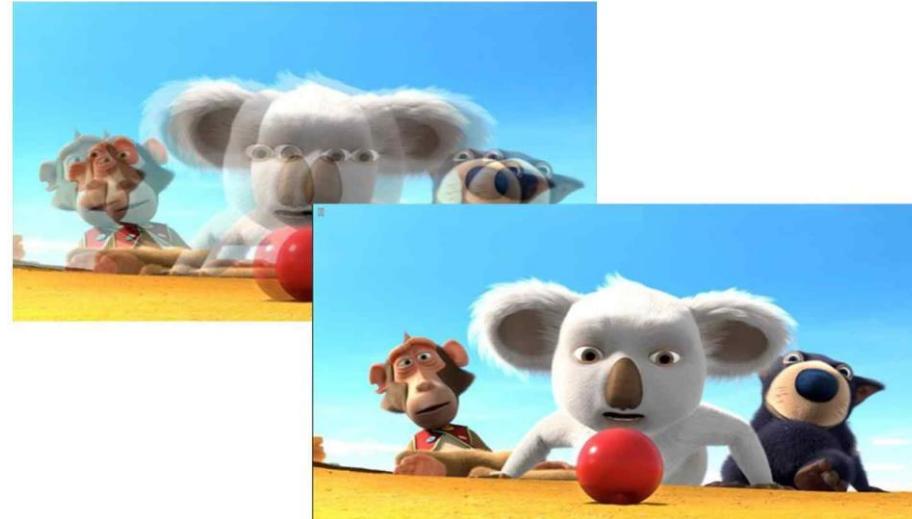
갈색 고양이???

데이터 분석을 잘 하려면...



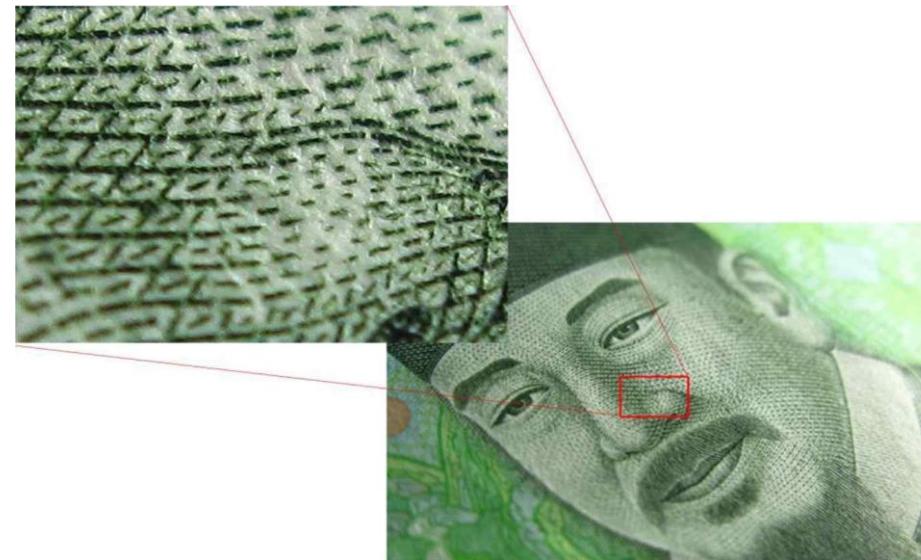
당장 눈앞에 보이는 것이 사실이 아닐 수도 있다.

데이터 분석을 잘 하려면...



사물을 명확하게 꿰뚫어 보려는 노력!

데이터 분석을 잘 하려면...



가끔은 디테일에 신경을 쓰도록 한다.

데이터 분석을 잘 하려면...



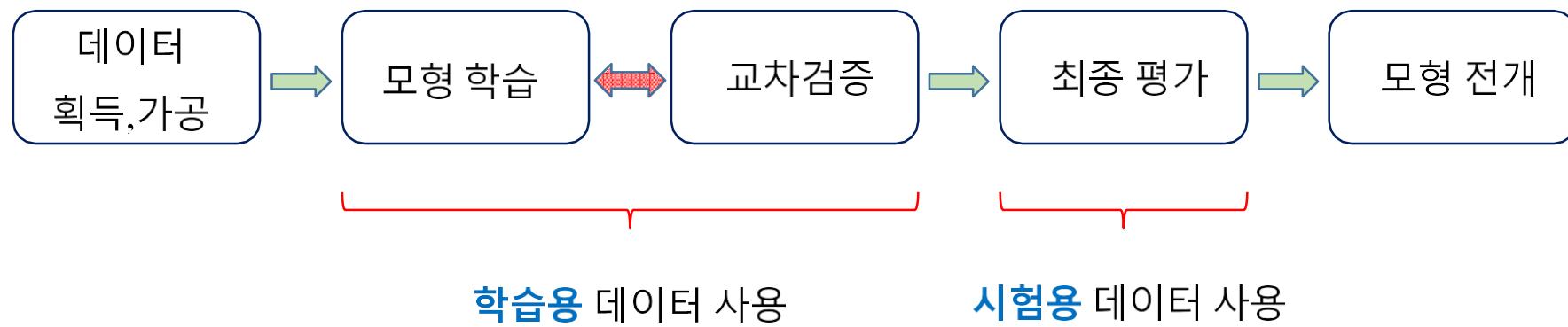
꾸준한 노력의 결과로 귀중한 통찰력을 얻는다.

지도학습

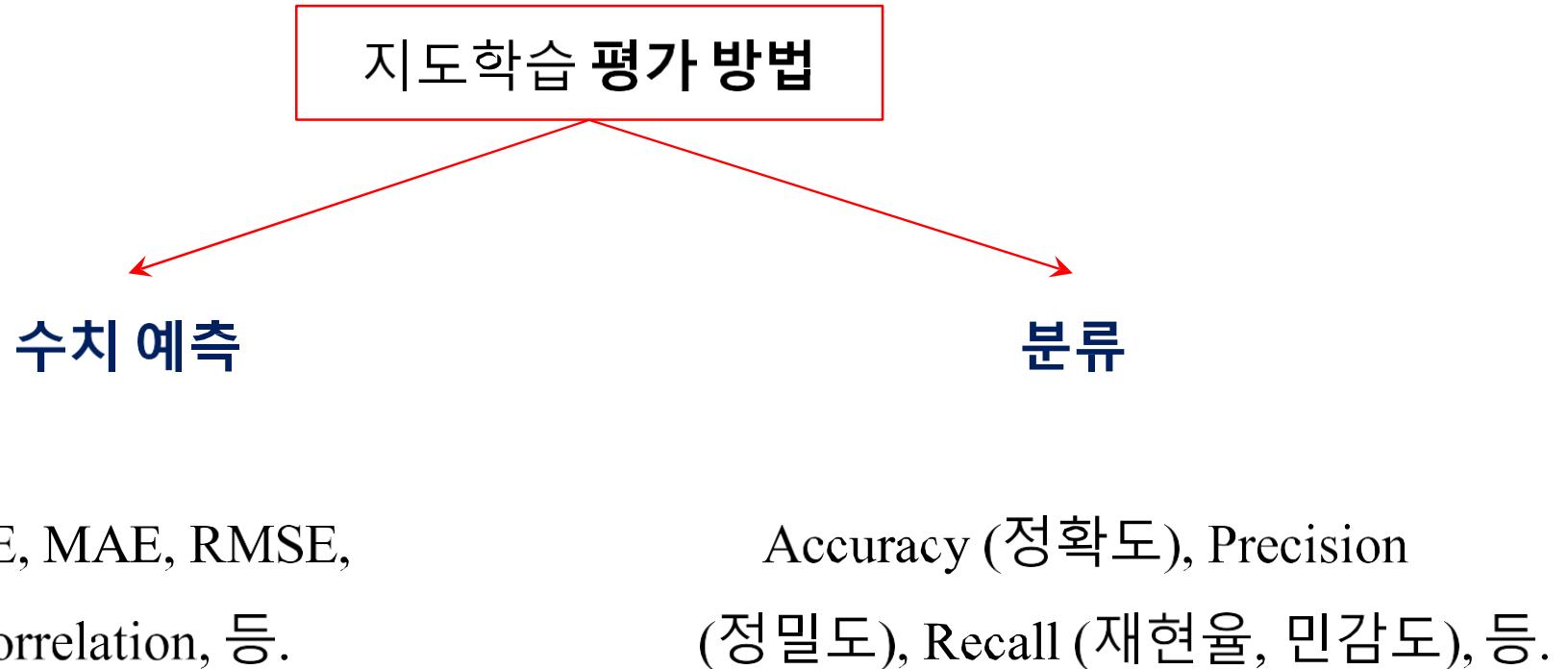
머신러닝의 유형

유형	방법
지도학습 (Supervised Learning)	선형회귀, 로지스틱 회귀
	트리, 랜덤포레스트, 애이디부스트
	Naïve Bayes
	Support Vector Machine (SVM)
	인공신경망
	k-NN
비지도학습 (Unsupervised Learning)	군집분석: k-means, hierarchical, DBSCAN
	주성분 분석 (PCA), 비음수 행렬분해 (NMF)
	t-SNE
	연관성 분석

머신러닝 단계 : 지도학습



머신러닝 : 지도학습 평가 방법



머신러닝 : 지도학습과 시험

지도학습과 시험:

- 진정한 의미의 학습은 암기한 내용을 똑같이 되새김하는 것만이 아니다.

→ 학습 후에는 현실에 적용할 수 있도록 일반화가 검증되어야 한다.

- 학습 (train)과 시험 (test).

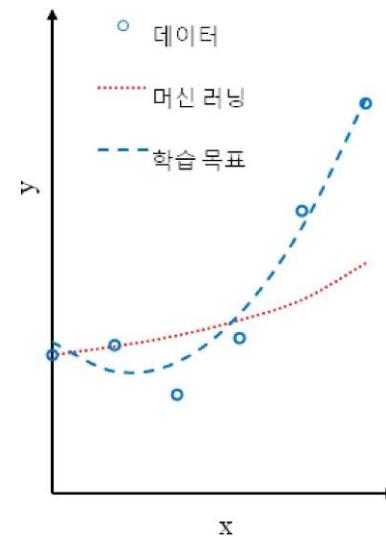
→ 일반화를 검증하기 위해서는 학습 (train)된 모형을 시험(test) 한다.

→ 시험의 결과는 “오류”로 평가하게 된다.

머신러닝 : 오류의 유형

오류의 유형: 편향 오류 (bias error) or 과소적합 오류 (underfitting error)

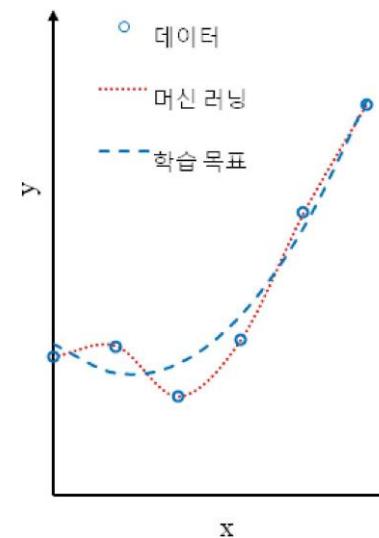
- 모형이 편향적 즉 과하게 단순해서 발생하는 오류의 유형이다.
- 모형의 복잡성이 증가할 수록 줄일 수 있다.



머신러닝 : 오류의 유형

오류의 유형: 분산 오류 (variance error) or 과적합 오류 (overfitting error)

- 모형이 과하게 복잡해서 발생하는 오류이며 많은 학습 데이터로 줄일 수 있다.
- 매개변수 최적화의 어려움으로 표출되는 오류이다.



In-Sample 오류는 작지만
Out-Of-Sample 오류는 큰
경우이다.

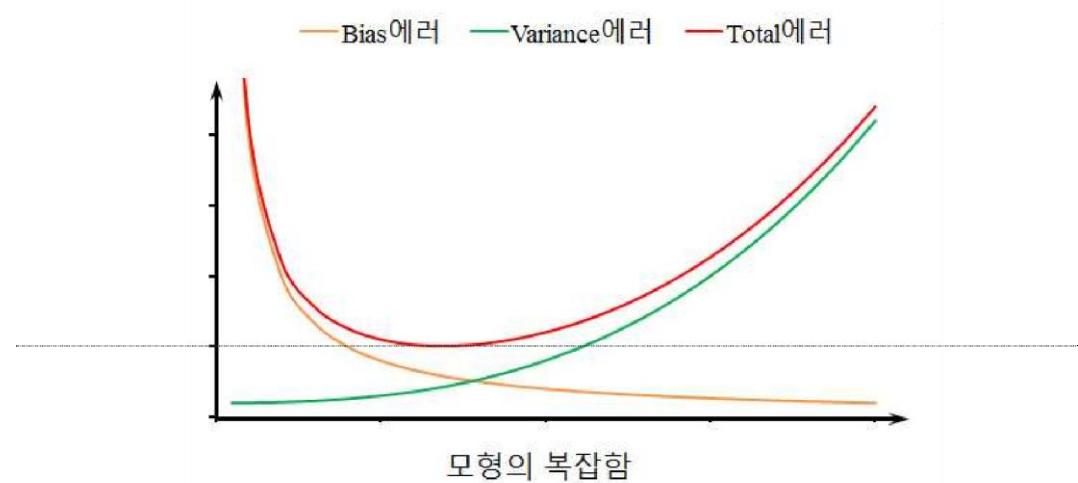
머신러닝 : 오류의 유형

오류의 유형:

토탈 오류 = 편향 오류 + 분산 오류 + 상수

머신러닝 : 시험 오류의 최소화

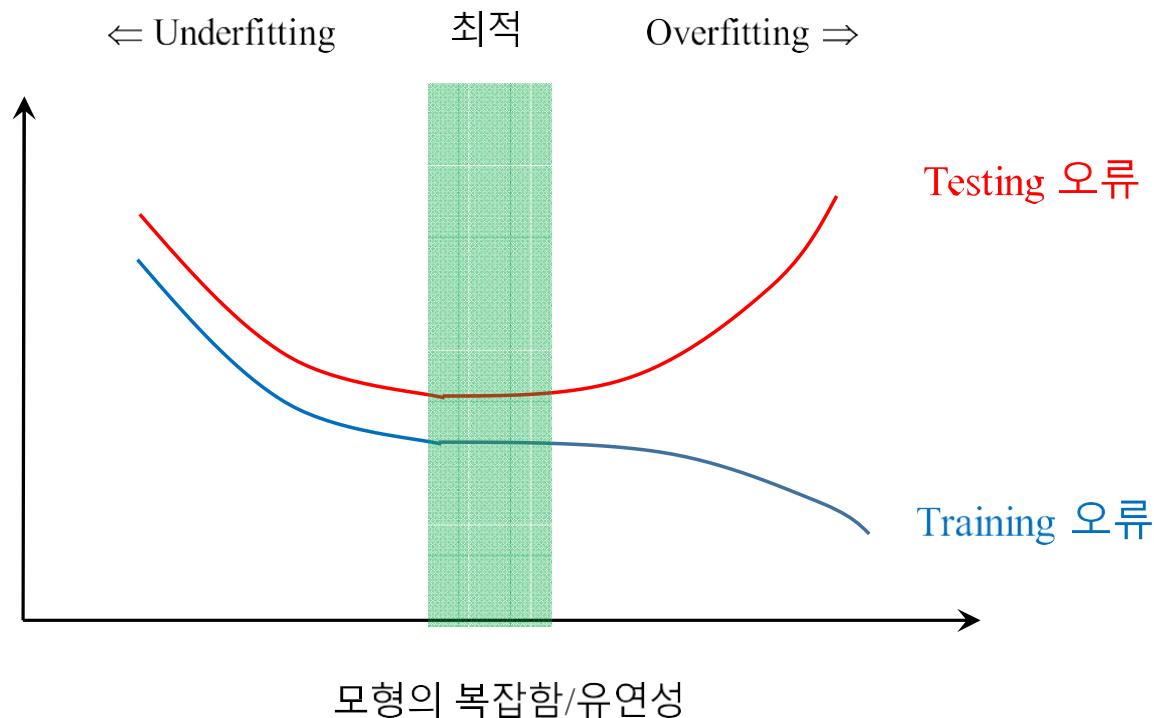
Out-of-Sample 시험 오류의 최소화:



→ 모형의 복잡함 (model complexity)에는 최적점(optimal point)이 있다.

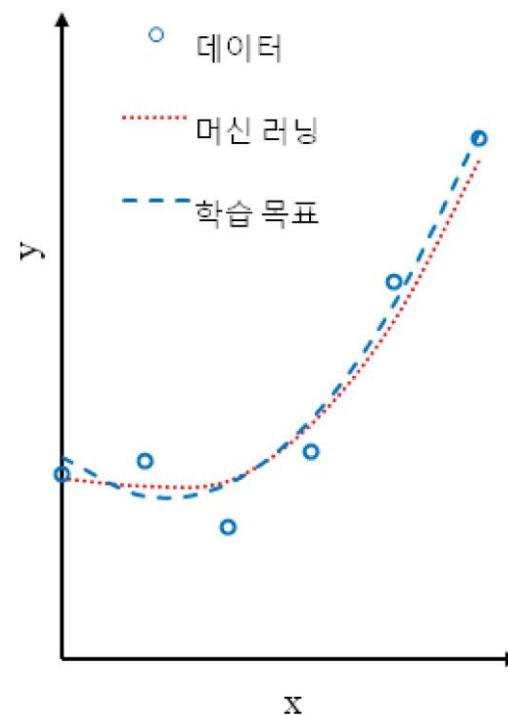
머신러닝 : 시험 오류의 최소화

모형의 최적화:



머신러닝 : 시험 오류의 최소화

오류의 최소화:



Python 의 Scikit-Learn 패키지

Scikit-Learn 패키지:

- Scikit-Learn은 Python의 대표적인 머신러닝 패키지이다.
- 모형 가져오기: from sklearn.<family> import <model>
예). from sklearn.linear_model import LinearRegression
- 모형의 파라미터는 객체 생성시 설정:
예). myModel = LinearRegression(fit_intercept=False, normalize=False)
- 지도학습 모형 학습 (트레이닝): <myModel>.fit(X_train, Y_train)
- 비지도학습 모형 학습 (트레이닝): <myModel>.fit(X_train)
- 지도학습 예측 (테스팅): <myModel>.predict(X_test)

선형회귀

선형회귀의 구성 요소:

- 한 개 이상의 독립변수 (설명변수). X_1, X_2, \dots, X_K
- 한 개의 종속변수. Y
- 선형 조합 (β 계수 표기). $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K + \varepsilon$
- 지도 학습.

선형회귀

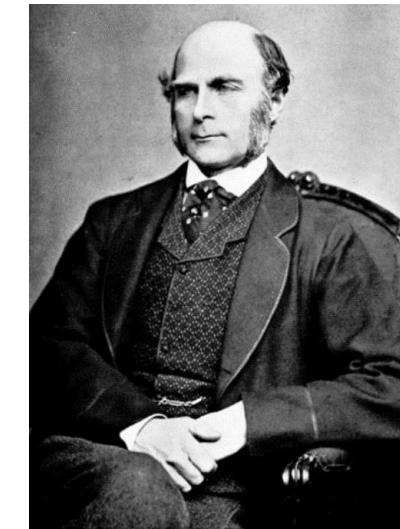
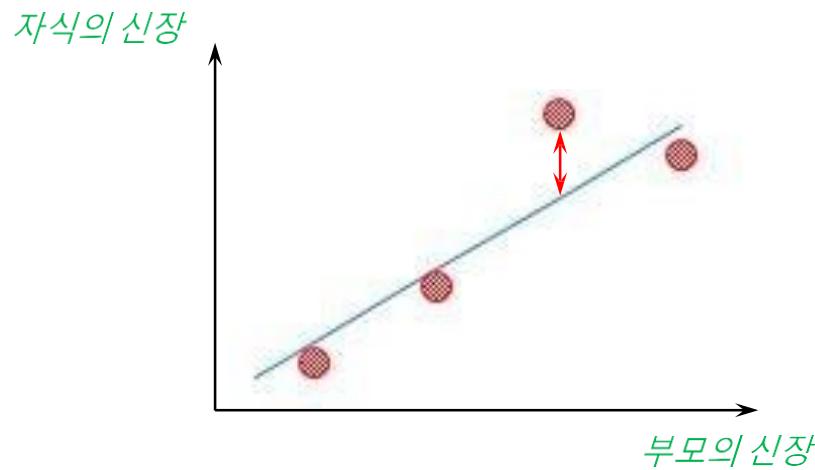
선형회귀의 목적:

- 종속변수를 설명하는 독립변수를 밝혀냄.
→ 예) 아파트의 시세는 평수, 지하철 역까지의 거리, 방의 수 등으로 설명할 수 있다(?)
- 독립변수 값의 변동에 따른 종속변수의 변동을 예측함.
→ 예) 데이터로 선형회귀 모형을 학습시키고 아직 아무에게도 알려져 있지 않은
아파트의 적정 가격을 알아 맞춘다.

선형회귀

역사적 배경:

- 19세기 영국의 우생학자인 Francis Galton이 평균으로 돌아간다라는 의미의 “회귀”라는 용어를 처음 사용함.
→ 신장에 있어서 부모와 자식 사이의 유전적 관계를 연구함.



선형회귀

선형회귀: 장단점

장점	단점
<ul style="list-style-type: none">✓ 수치예측의 가장 보편적인 방법.✓ 통계적, 이론적 배경이 견고하다.✓ 트레이닝이 빠르다.	<ul style="list-style-type: none">✓ 여러가지 가정을 전제한다: 선형성, 정상성, 독립성, 등.✓ 외상치 (outlier)에 비교적 민감하다.✓ 다중공선성의 문제가 쉽게 발생한다.

선형회귀

선형회귀의 전제:

- 종속변수는 **독립 변수**의 선형 조합이어야 한다.
- 다중공선성이 존재하지 않거나 거의 **없어야** 한다.
- 오차의 평균은 0이며 **정규분포**를 따라야 한다.
- 오차항은 **등분산성**을 가져야 한다.
- 오차항에는 상관관계가 **없어야** 한다 (추세 **X**).



잔차분석

선형회귀의 원리

선형 모형:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

선형회귀의 원리

선형 모형의 예 #1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$



선형회귀의 원리

선형 모형의 예 #1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

$$\begin{array}{c} \uparrow \quad \uparrow \\ \text{MPG} \quad \text{N\# of} \\ \text{Cylinders} \end{array}$$

선형회귀의 원리

선형 모형의 예 #1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

↑ ↑ ↑

MPG N# of HP
Cylinders

선형회귀의 원리

선형 모형의 예 #1:

선형회귀의 원리

선형 모형의 예 #1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

↑ ↑ ↑ ↑ ↑
MPG N# of HP Weight Auto or
Cylinders Manual

원리에 대해서 자세히
알아보겠습니다.



선형회귀의 원리

오차 변수:

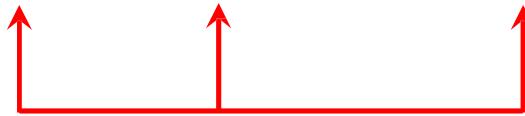
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

↑
평균 $[\varepsilon] = 0$
표준편차 $[\varepsilon] = \sigma_\varepsilon$

선형회귀의 원리

공선성을 피해야 함:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$

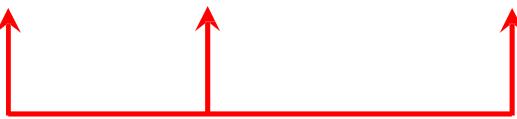


$$\text{Cor}(X_i, X_j) \approx 0$$

for $i \neq j$

선형회귀의 원리

공선성을 피해야 함:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$$


공선성은 계수의 “분산 인플레”

문제를 일으킵니다.

선형회귀의 원리

공선성을 피해야 함:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$


예) 공선성 문제 있음 :

X_1 = 몸무게.

X_2 = 체지방률.

선형회귀의 원리

공선성을 피해야 함:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

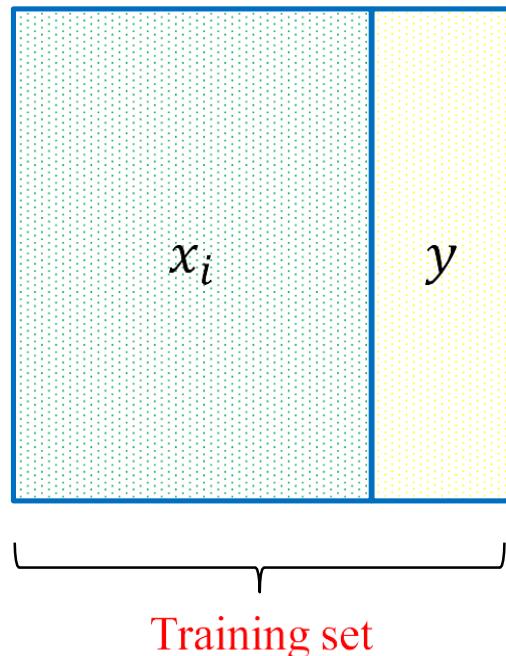

예) 공선성 문제 없음 :

X_1 = 신장.

X_2 = 연간 소득.

선형회귀 : 학습

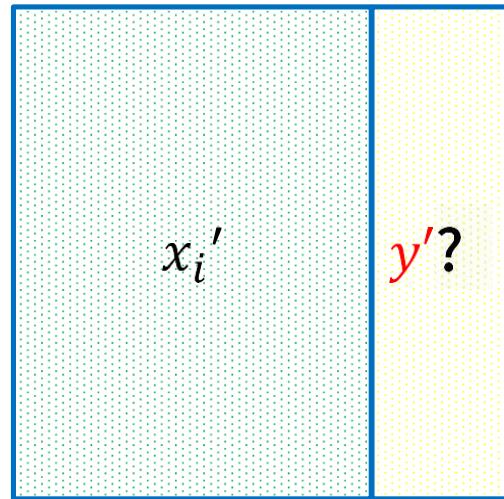
선형회귀에서 학습은:



모형의 파라미터 $\{\beta_i\}$ 를 학습용 데이터를 사용하여 계산해 놓는 것.

선형회귀 : 예측

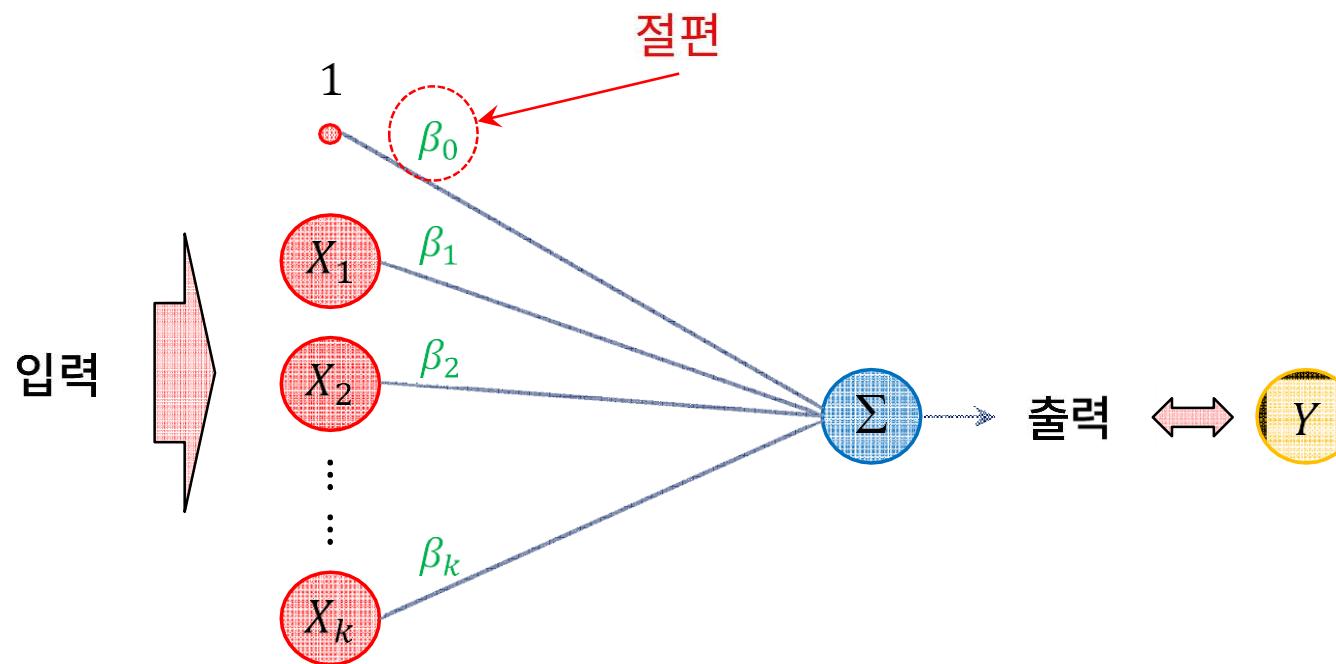
선형회귀에서 예측은:



독립변수의 값 $\{x_i'\}$ 이 새롭게 주어졌을 때,
모르는 상태인 종속변수의 값 y' 을 계산을
통해서 알아낸다.

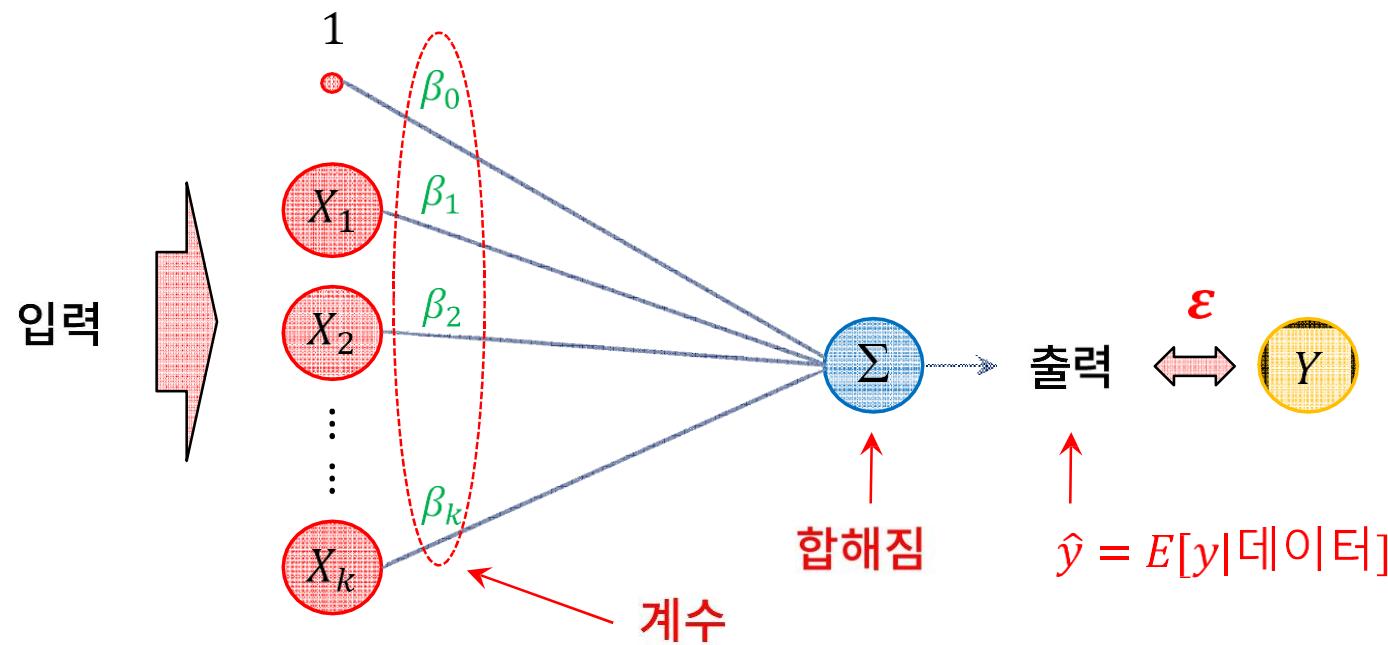
선형회귀 : 예측

그래프로 나타냄:



선형회귀 : 예측

그래프로 나타냄:



로지스틱 회귀

로지스틱 회귀: 장단점.

장점	단점
<ul style="list-style-type: none">✓ 개념과 이론이 단순하다.✓ 노이즈에 비교적 둔감하다.✓ 트레이닝이 빠르다.	<ul style="list-style-type: none">✓ 여러 속성이 동등하게 중요하고 독립적이라는 결함 가정 전제.✓ 범주형 예측에는 적합하나 확률 값 예측에는 정확도가 떨어진다.

로지스틱 회귀의 원리

로지스틱 회귀의 구성 요소:

- k 개의 독립변수 (설명변수)가 있다고 가정한다.
→ 가능한 값에 대해서는 제약이 없다.

$$X_1, X_2, \dots, X_K$$

로지스틱 회귀의 원리

로지스틱 회귀의 구성 요소:

- 그리고 한개의 종속변수가 있다고 가정한다.
→ 그런데 가능한 값은 0과 1.

$$Y = \begin{cases} 1 \\ 0 \end{cases}$$

로지스틱 회귀의 원리

로지스틱 회귀의 구성 요소:

- 즉, 이분법적인 상황이다.

$$Y = \begin{cases} \text{참} (True) \\ \text{거짓} (False) \end{cases}$$

로지스틱 회귀의 원리

로지스틱 회귀의 구성 요소:

- 즉, 이분법적인 상황이다.

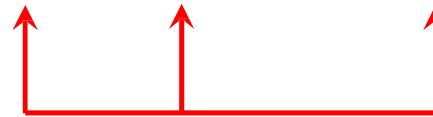
$$Y = \begin{cases} \text{유형 } a \\ \text{유형 } b \end{cases}$$

로지스틱 회귀의 원리

로지스틱 회귀의 구성 요소:

- 이제는 독립변수 $\{X_i\}$ 를 선형조합하여 S 변수 (logit)을 만든다.

$$S = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$



독립변수

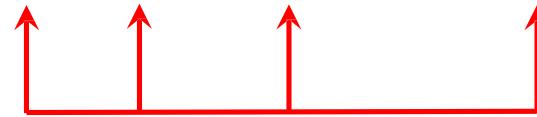
(데이터로 값이 주어짐)

로지스틱 회귀의 원리

로지스틱 회귀의 구성 요소:

- 이제는 독립변수 $\{X_i\}$ 를 선형조합하여 S 변수 (logit)을 만든다.

$$S = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$



계수

(학습을 통해서 밝혀냄)

로지스틱 회귀의 원리

로지스틱 회귀의 구성 요소:

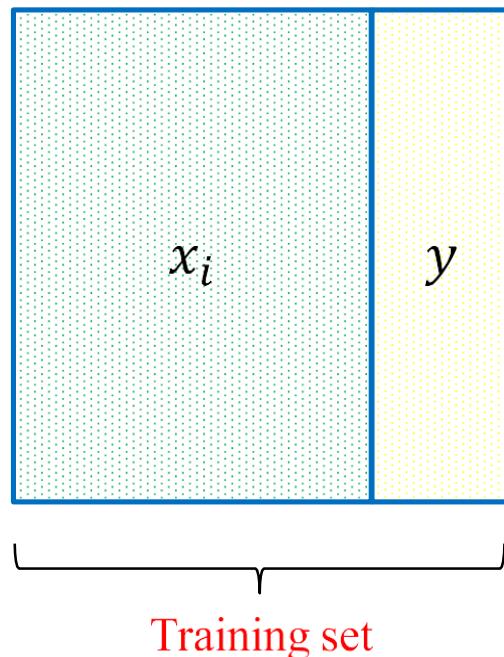
- 종속변수 Y 의 값이 1이 될 조건부 확률 $P(Y = 1 | \{x_i\})$ 은 “로지스틱 함수” 또는 “Sigmoid 함수”를 사용해서 계산된다.

$$f(S) = \frac{e^S}{1 + e^S}$$

→ 인공신경망에서 “활성화 함수” (activation function)의 역할을 함.

로지스틱 회귀 : 학습

로지스틱 회귀에서 학습은:



모형의 파라미터 $\{\beta_i\}$ 를 학습용 데이터를 사용하여 계산해 놓는 것.

로지스틱 회귀 : 예측

로지스틱 회귀에서 예측은:

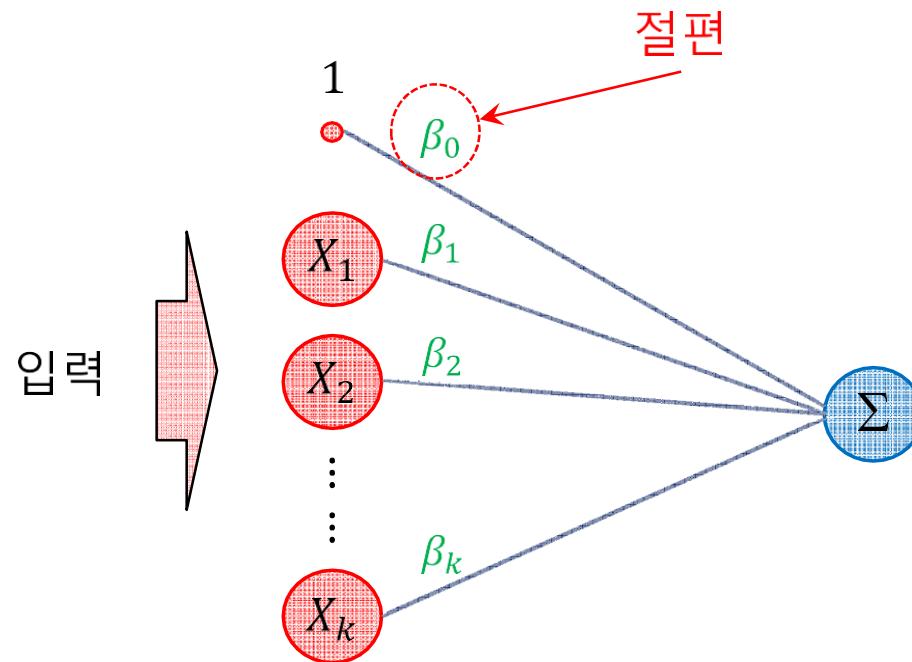
- 관심점에 해당하는 독립변수의 값 $x_1^*, x_2^*, \dots, x_k^*$ 가 주어졌을 때 다음 수식을 사용해서 $P(Y = 1|\text{데이터})$ 를 구한다.

$$S = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \dots + \beta_K x_K^*$$

$$P(Y = 1|\text{데이터}) = \frac{e^S}{1 + e^S}$$

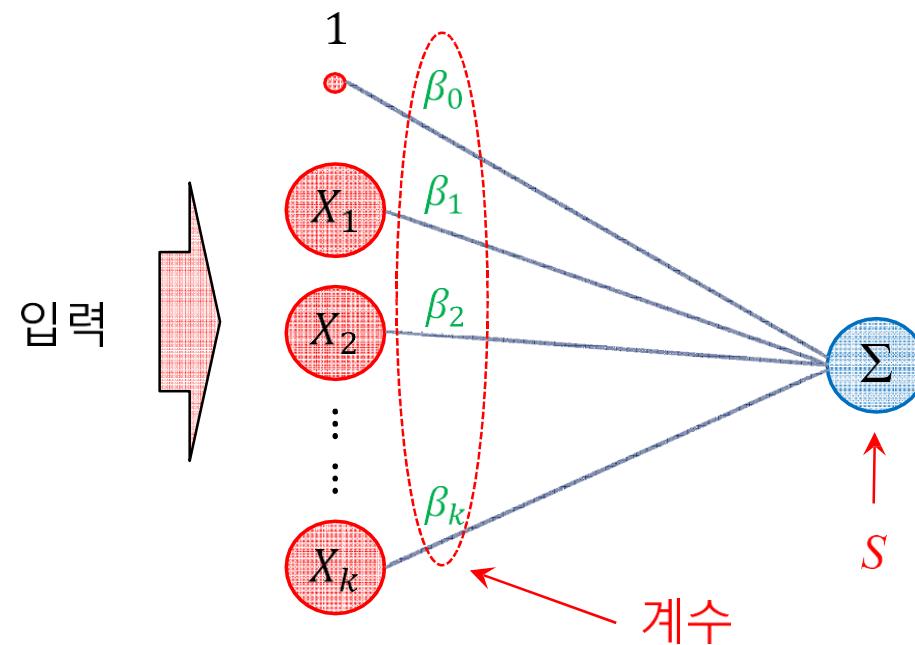
로지스틱 회귀 : 예측

그래프로 나타냄:



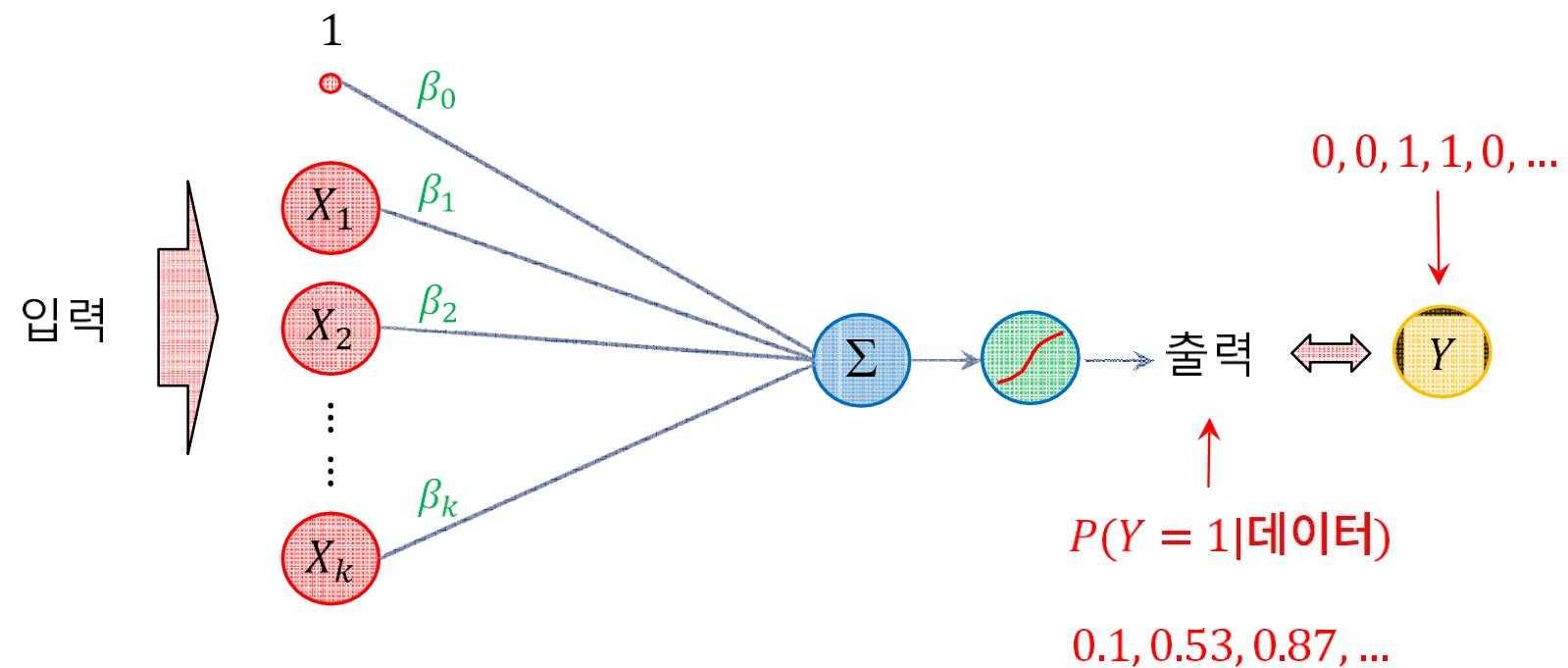
로지스틱 회귀 : 예측

그래프로 나타냄:



로지스틱 회귀 : 예측

그래프로 나타냄:



KNN 알고리즘

KNN (K Nearest Neighbor) 분류 알고리즘:

- 가장 간단한 알고리즘 중의 하나이며 직관적인 이해가 가능하다.
- 주변의 제일 가까운 K개의 데이터 포인트를 찾아서, “다수결”로 값을 정하는 방식.

KNN 알고리즘

KNN (K Nearest Neighbor) 분류 알고리즘: 장단점

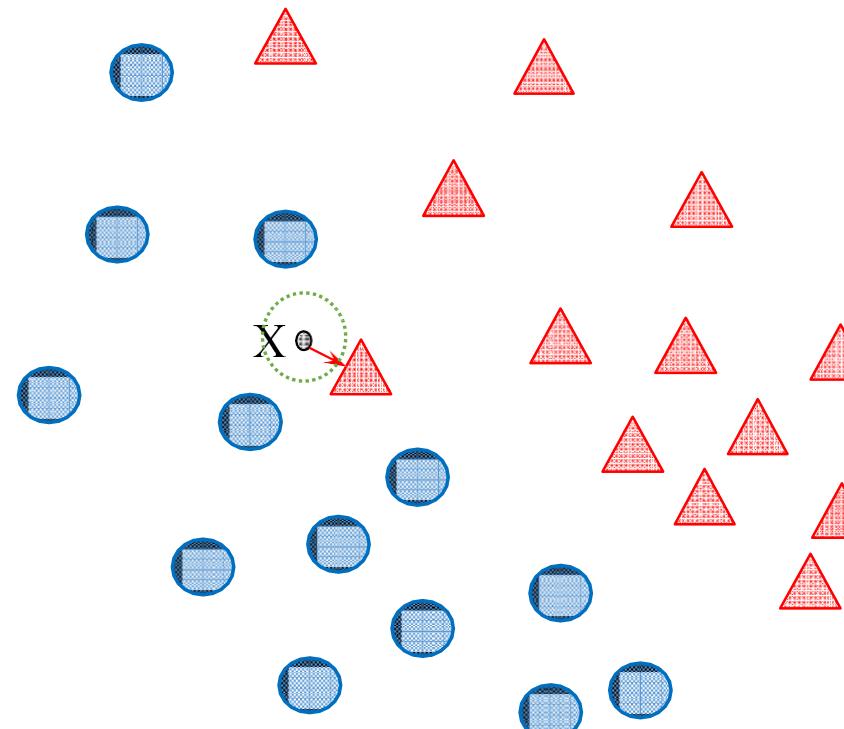
장점	단점
<ul style="list-style-type: none">✓ 간단하며 직관적인 이해가 가능하다.✓ 사전 모형 설정이 필요없고 모수에 대한 추정이 필요없다 (비모수 방법).✓ 트레이닝이 빠르다.	<ul style="list-style-type: none">✓ 예측은 효율적이지 못하다.✓ 모형이 없으므로 구조적 통찰력을 얻기 어렵다.

KNN 알고리즘

KNN (K Nearest Neighbor) 분류 알고리즘:

K = 1인 경우

X는 로 분류.

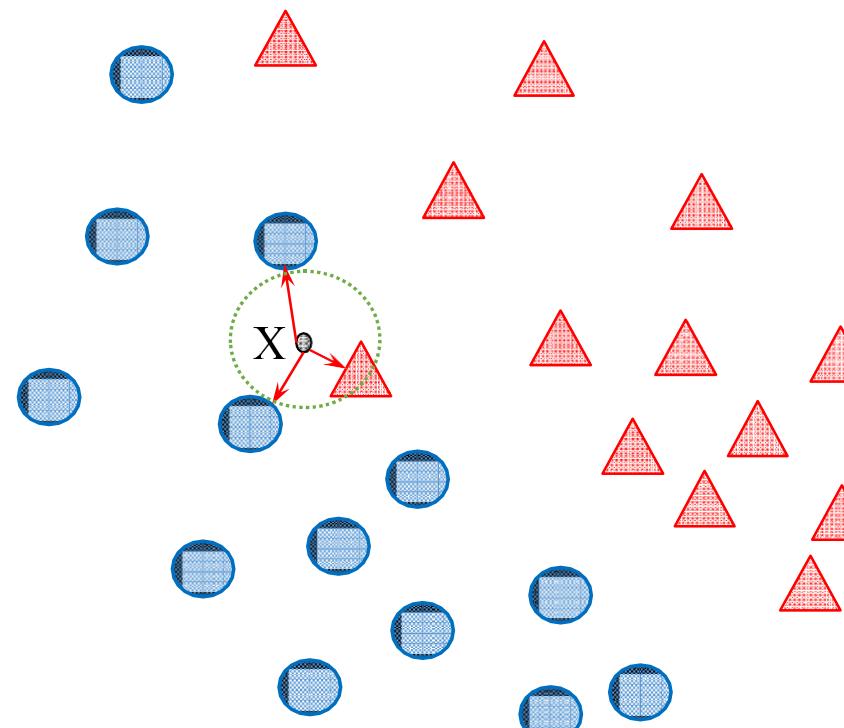


KNN 알고리즘

KNN (K Nearest Neighbor) 분류 알고리즘:

K = 3인 경우

X는 로 분류.



KNN 알고리즘

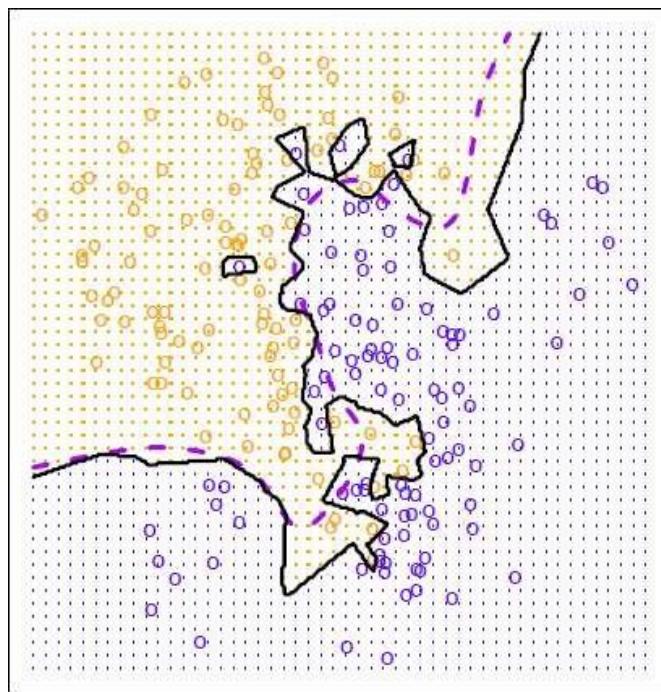
KNN (K Nearest Neighbor) 분류 알고리즘:

- 가장 간단한 기계학습 알고리즘 중의 하나이다.
- 주변의 제일 가까운 K개의 데이터 포인트들 찾아서, “다수결”로 값을 정하는 방식.
- K를 너무 작게 설정하면 노이즈에 민감해지고 과적합 (overfitting) 위험이 있음.
⇒ 분산 오류 증가.
- K를 너무 크게 설정하면 디테일에 둔감해지고 과소적합 (underfitting) 위험이 있음.
⇒ 편향 오류 증가.

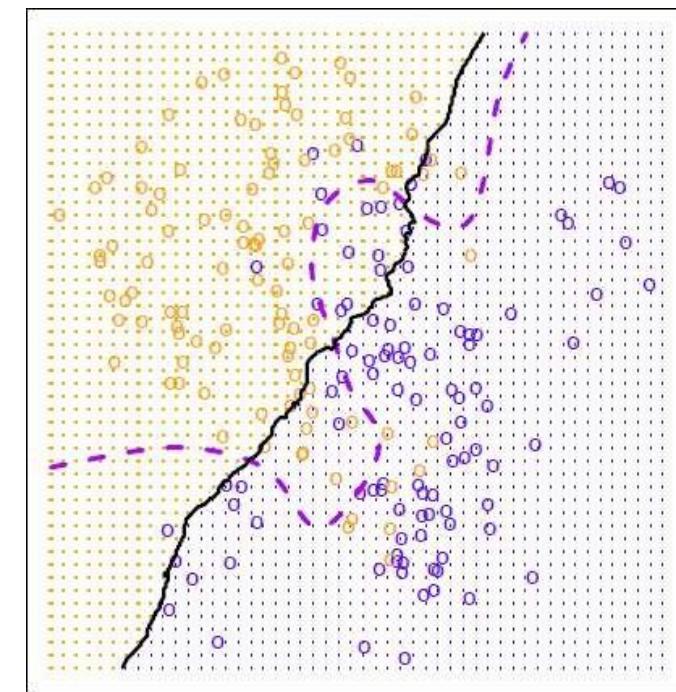
KNN 알고리즘

KNN (K Nearest Neighbor) 분류 알고리즘:

K = 1



K = 100



트리 알고리즘

트리 알고리즘:

- 트리 알고리즘은 크게 **분류형**과 **회귀형**으로 구분된다.
- 트리 알고리즘은 비교적 성능이 좋지 않다.
- Boosting, Bagging, Random Forest 등은 트리를 기반으로 한 고성능 알고리즘이다.

트리 알고리즘

트리 알고리즘: 장단점

장점	단점
<ul style="list-style-type: none">✓ 간단하며 직관적인 이해가 가능하다.✓ 사전 모형 설정이 필요없고 모수에 대한 추정이 필요없다 (비모수 방법).	<ul style="list-style-type: none">✓ 성능이 좋지는 못하다.✓ 해석이 어렵다.✓ 과적합 현상이 쉽게 일어난다.

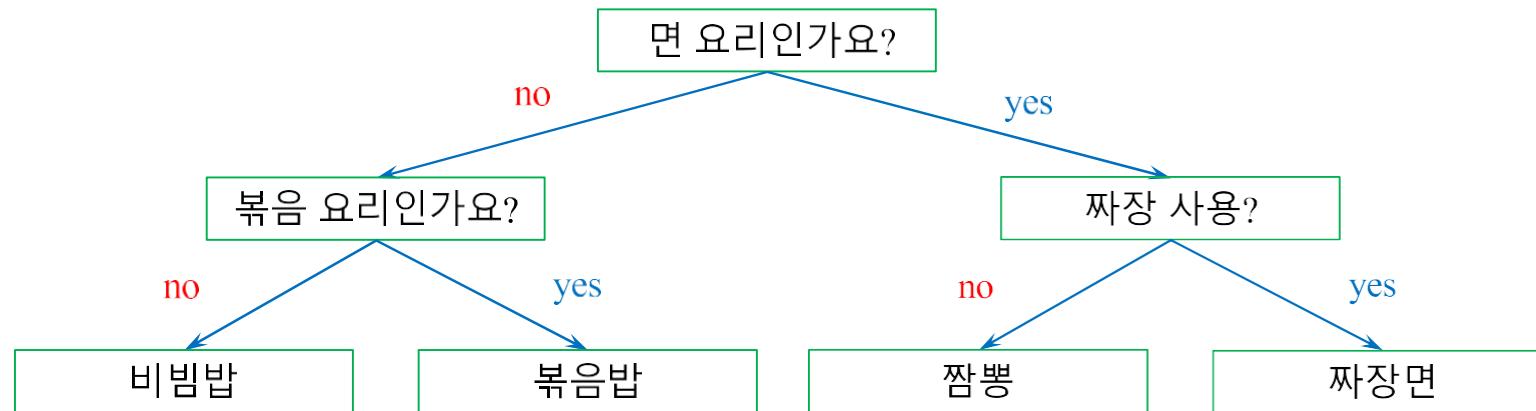
분류형/결정 트리

분류형/결정 트리 (Decision Tree) 알고리즘:

- 명목형 변수인 경우 적용 가능.
- 쉽게 이해할 수 있다는 장점이 있다.
- 쉽게 과적합 (overfitting) 문제가 발생한다 \Rightarrow “가지 치기” 필요 (교차검증).

분류형/결정 트리

분류형/결정 트리 (Decision Tree) 알고리즘:



- 거꾸로 뒤집어 놓은 나무 (트리)의 형상을 보이며 뿌리부터 따라서 내려가다 보면 답이 자연스럽게 구해지는 구조.
- 분기점 (node)의 질문 (조건)들은 데이터를 통한 학습으로 만들어 진다.

분류형/결정 트리

분류형/결정 트리 (Decision Tree) 알고리즘: 키 포인트

- 분기점에서의 조건의 목적은 다음 지니 불순도 (Gini impurity)를 최소화 하는 것이다.

$$G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk})$$

$$= 1 - \sum_{k=1}^K \hat{p}_{mk}^2$$

→ 여기에서 \hat{p}_{mk} 는 m 구역에서 종속변수 값으로 나타나는 k 유형의 비율이다.

→ 유형의 전체 가짓수는 K 이다.

- 지니 불순도 G 는 특정 구역의 순도가 높을 수록 작다.
- 가장 빈도수가 높은 유형이 특정 구역에 해당하는 예측이다.
- 지니 불순도 대신에 엔트로피 (entropy)를 사용할 수 있다.

$$\text{엔트로피} = -\sum_{k=1}^K \hat{p}_{mk} \log(\hat{p}_{mk})$$

분류형/결정 트리

분류형/결정 트리 (Decision Tree) 알고리즘: 절차

- 결정 트리의 기본형을 만든다.
- 과적합 (overfitting) 문제를 피하기 위해서 분별력이 약한 가지들은 잘라낸다 (pruning).
- 교차검증 (cross validation)을 실시한다.
- 정제된 모형을 가지고 평가 (test)나 예측을 한다.

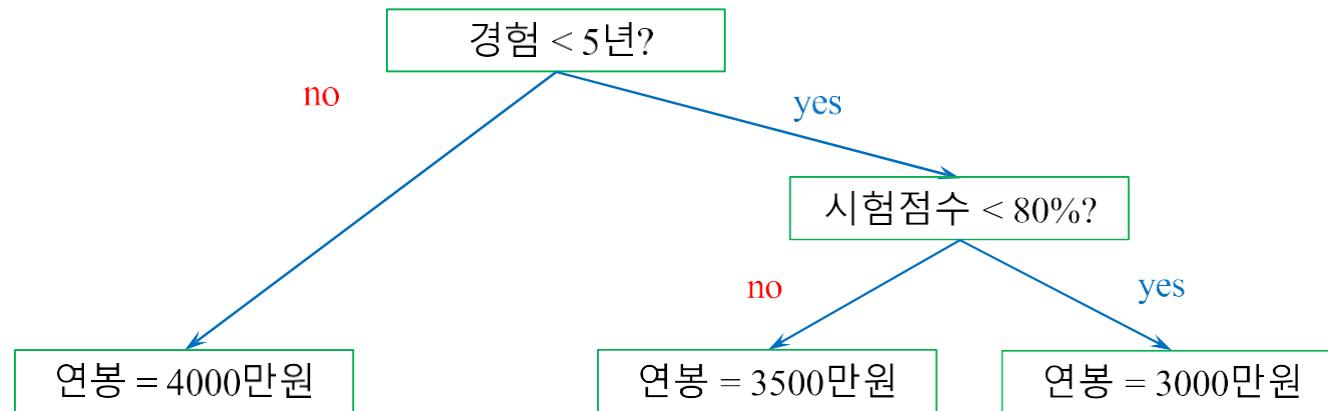
회귀형 트리

회귀형 트리 (Regression Tree) 알고리즘:

- 수치형 변수의 경우에 적용 가능.
- 분류형 트리와 마찬가지로 분기점의 질문에 대한 답에 따라서 갈라져 내려감.
- 설명 변수의 공간을 쪼개서 같은 구역에 속하는 관측값에 대해서는 동일한 예측을 함.
 ⇒ 정확성이 높지는 않다.

회귀형 트리

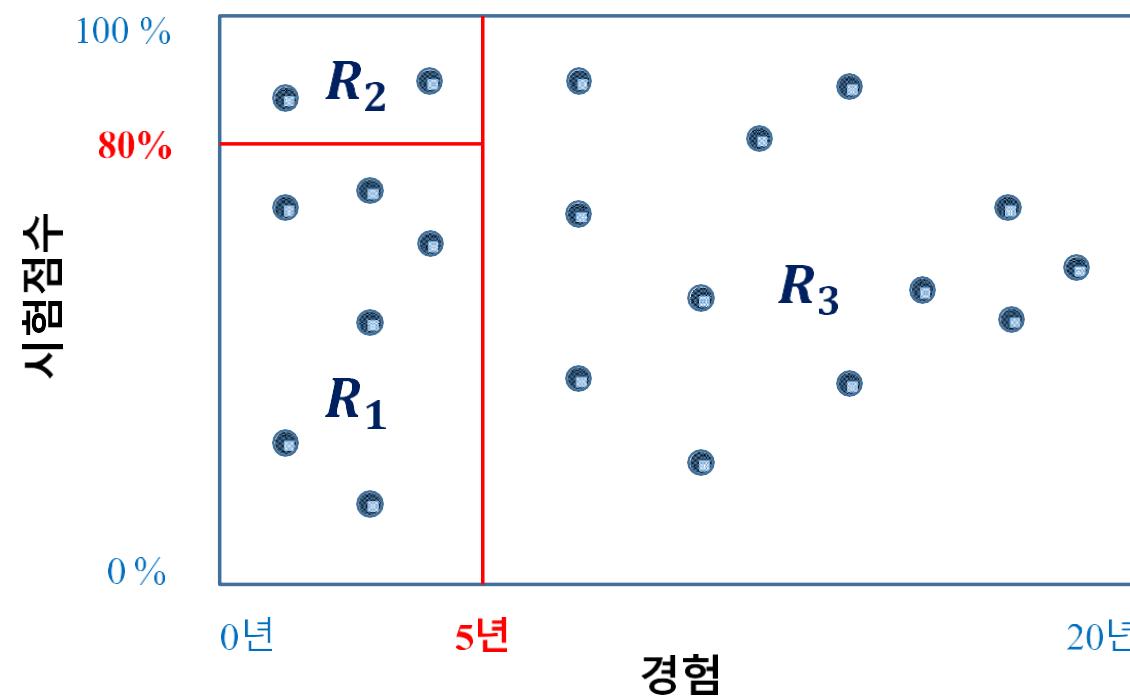
회귀형 트리 (Regression Tree) 알고리즘:



- 분류형 트리와 마찬가지로 거꾸로 뒤집어 놓은 나무의 형상을 보인다.
- 분기점 (node)의 질문 (조건)들은 설명 변수를 두개의 구역으로 쪼개는 역할을 한다.

회귀형 트리

회귀형 트리 (Regression Tree) 알고리즘:



회귀형 트리

회귀형 트리 (Regression Tree) 알고리즘: 키 포인트

- 설명변수의 공간을 겹치지 않는 구간으로 분할한다: $\{R_1, R_2, \dots, R_J\}$
→ 분할의 조건은 다음 RSS를 최소화 하는 것이다.

$$RSS = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

여기에서 \hat{y}_{R_j} 는 j 구역에서의 종속변수 평균값이다.

- 특정 구간 R_i 에 속하는 모든 관측값에 대해서는 같은 예측을 한다. (종속변수 평균값)
- 분류형 트리와 마찬가지로 과적합에 대한 고려가 수반되어야 한다.

랜덤포레스트

랜덤포레스트 (Random Forest) 알고리즘 :

- 비교적 “약한” 학습기인 트리를 기반으로한 앙상블(ensamble)을 만드는 방법이다.
- 분류 또는 평균값 예측(회귀) 문제에 적용 가능하다.

랜덤포레스트

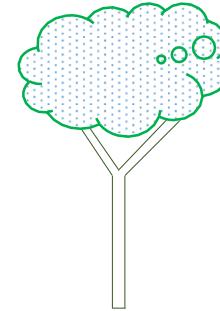
랜덤포레스트 (Random Forest) 알고리즘 : 장단점

장점	단점
<ul style="list-style-type: none">✓ 월등히 높은 정확도.✓ 과적합 문제도 거의 없다.✓ 변수소거 필요없고, 많은 갯수의 입력변수를 다룰 수 있다.	<ul style="list-style-type: none">✓ 계산 시간이 길다.

랜덤포레스트

랜덤포레스트 (Random Forest) 알고리즘: 원리 (이진분류)

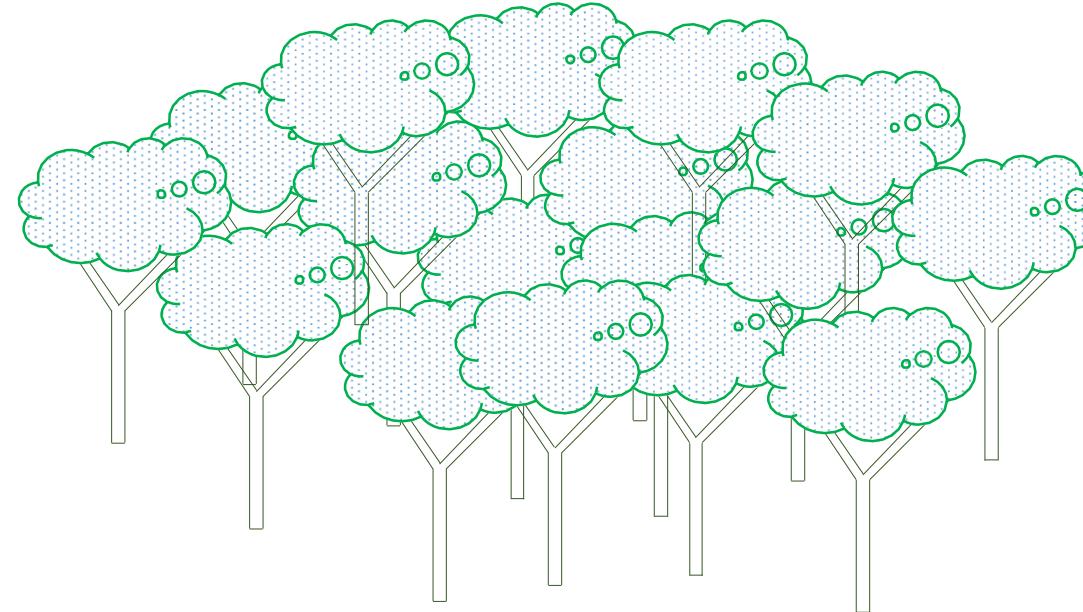
- 하나의 “결정 트리” 또는 “결정 그루터기” (decision stump).



랜덤포레스트

랜덤포레스트 (Random Forest) 알고리즘: 원리 (이진분류)

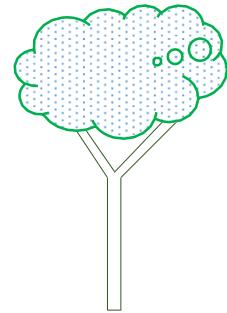
- 여러개의 나무 \Rightarrow “ensemble” 또는 숲(forest).



랜덤포레스트

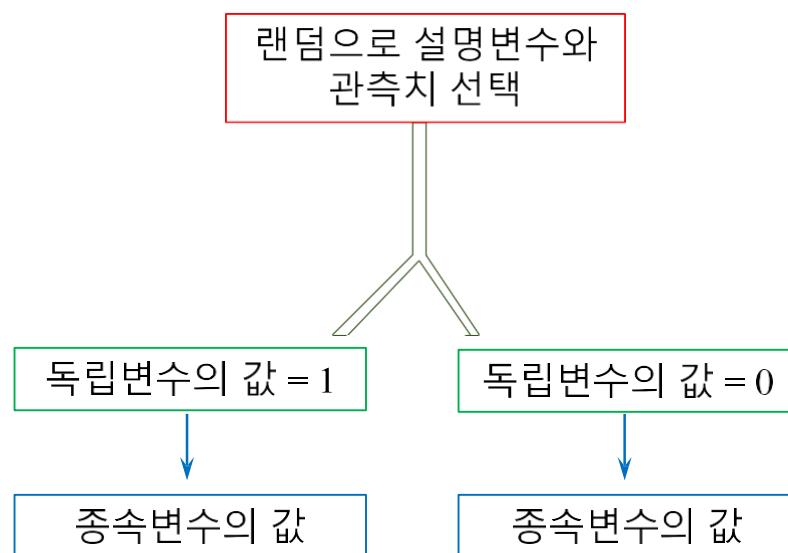
랜덤포레스트 (Random Forest) 알고리즘: 원리 (이진분류)

- 나무 하나 하나를 자세히 살펴보겠습니다.



랜덤포레스트

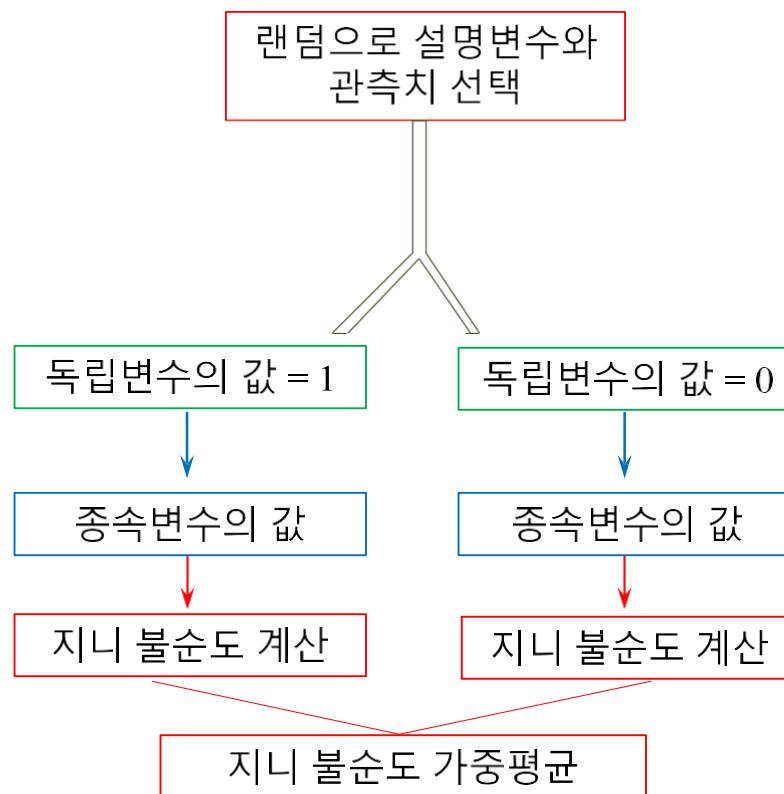
랜덤포레스트 (Random Forest) 알고리즘: 원리 (이진분류)



- 독립변수와 종속변수가 범주형이어야 합니다.
예) {0,1}, {true, false}, {동,서,남,북}, {"a", "b"}, 등.

랜덤포레스트

랜덤포레스트 (Random Forest) 알고리즘: 원리 (이진분류)



- 종속변수 값의 발생 빈도에 따라서 개개 가지의 지니불순도를 구하고 전체 가중평균을 구합니다.

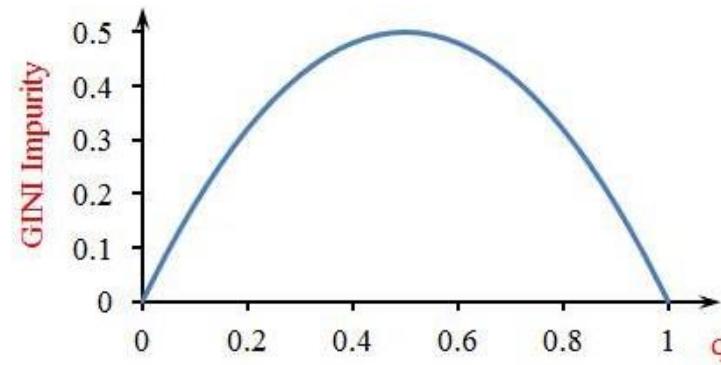
랜덤포레스트

그러면 지니 불순도란 무엇인가요?

1. 트레이닝 데이터에서 표본을 채취했다고 가정합니다.
2. 이 표본에서 종속변수=“a”의 확률은 p_a 그리고 종속변수=“b”의 확률은 p_b 입니다.
3. 이제는, 위의 확률을 그대로 적용해서 표본을 무작위로 labeling 합니다.
4. 결과적으로 “a”레이블이 정확하게 적용되는 확률 = $p_a \times p_a = p_a^2$.
발생 확률과 labeling 확률이 같으므로 제곱이 됩니다.
5. 또한 “b”레이블이 정확하게 적용되는 확률 = $p_b \times p_b = p_b^2$.
6. 그러므로 무작위 labeling이 **부정확**하게 적용될 확률 = $1 - p_a^2 - p_b^2$.
이것이 바로 지니 불순도입니다.

랜덤포레스트

7. p_a 를 q 로 표기하면 p_b 는 $(1 - q)$ 와 같습니다.



8. q 가 0 또는 1에 가까울수록 불순도가 낮습니다.
9. 즉, “a” 또는 “b” 구분하지 않고 어느쪽으로 순도가 높으면 불순도가 낮다는 의미입니다.
10. $q = 0.5$ 이면 불순도가 최고입니다. 즉, 분별력이 없다는 의미입니다.

랜덤포레스트

랜덤포레스트 (Random Forest) 알고리즘: 스텝

1. 트레이닝 데이터의 일부를 랜덤으로 채취합니다.
2. 불순도가 가장 낮은 decision stump (독립변수/설명변수)만을 저장합니다.
3. 위의 스텝1로 돌아가서 특정 횟수만큼 사이클을 반복합니다.
4. 이로서 트레이닝이 완료됩니다.
5. 예측을 하기 위해서는 이미 저장된 decision stump를 사용해서 “투표”하도록 합니다.

SVM

SVM (Support Vector Machine) 분류 알고리즘:

- 분류 정확성과 분류 마진을 최대로 하는 알고리즘.

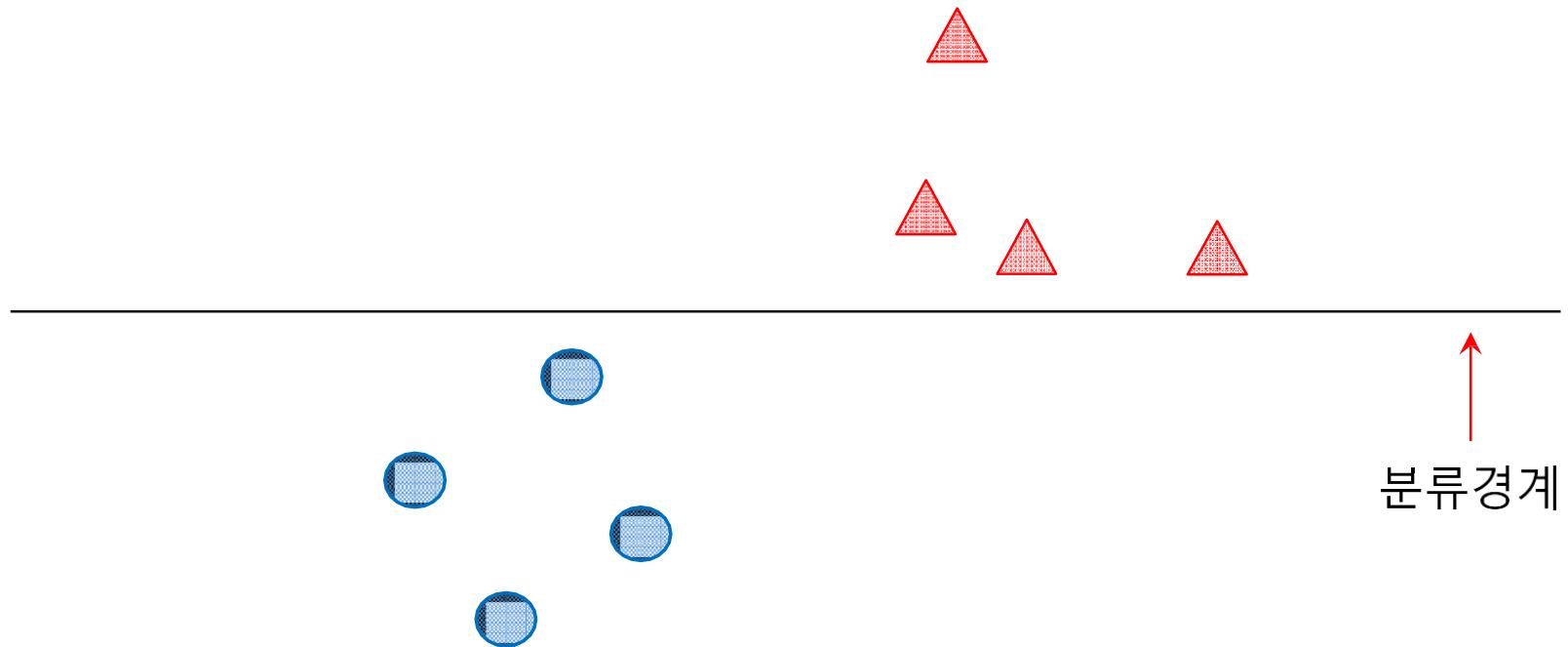
SVM

SVM (Support Vector Machine) 분류 알고리즘: 장단점

장점	단점
<ul style="list-style-type: none">✓ 노이즈 데이터의 영향을 크게 받지 않고 과적합 문제가 적다.✓ 일반적으로 분류 성능이 높다.	<ul style="list-style-type: none">✓ 최적 분류를 위해서는 커널 함수 및 매개변수에 대한 반복적인 테스트가 필요하다.✓ 알고리즘과 결과 해석이 어렵고 트레이닝에 시간이 많이 소요된다.

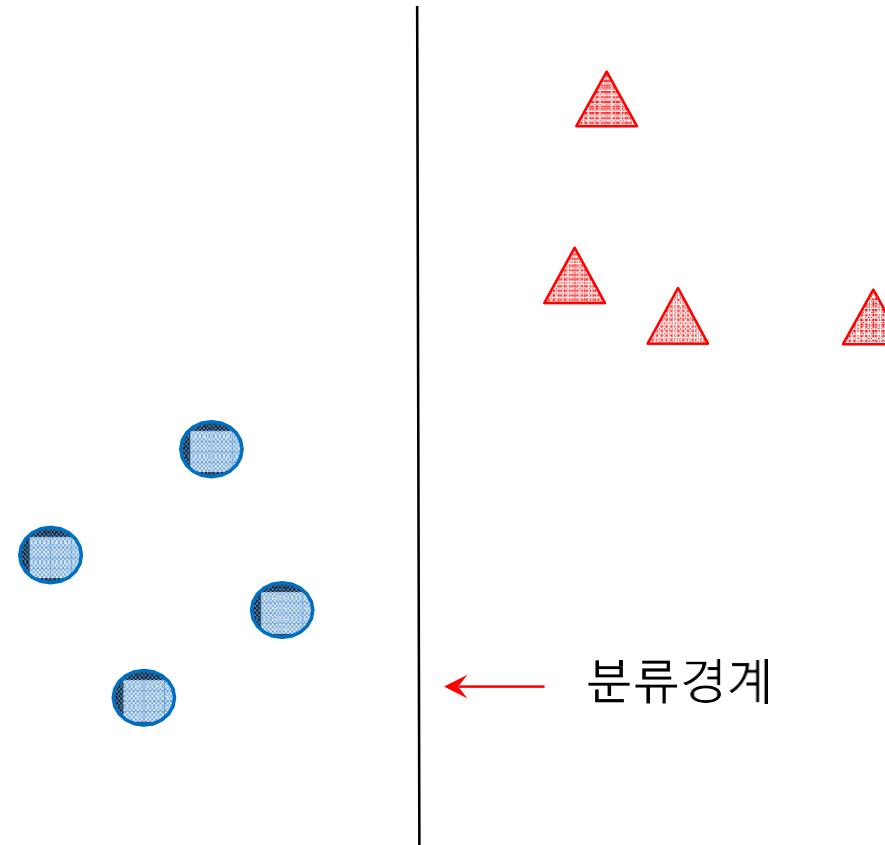
SVM

SVM (Support Vector Machine) 분류 알고리즘:



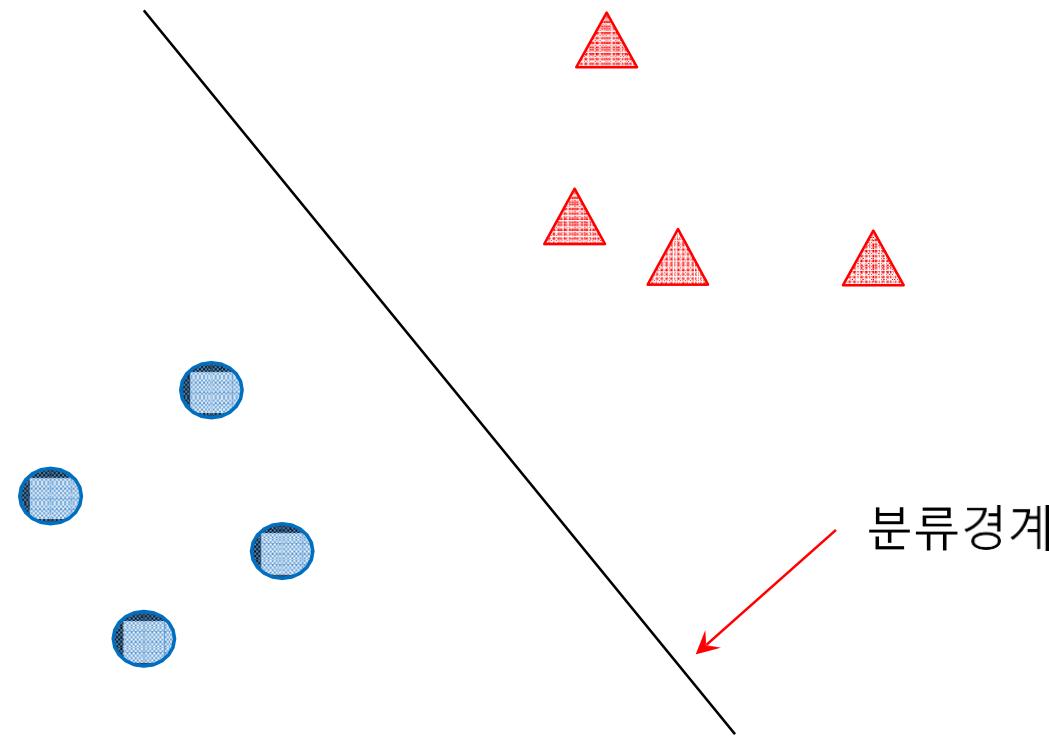
SVM

SVM (Support Vector Machine) 분류 알고리즘:



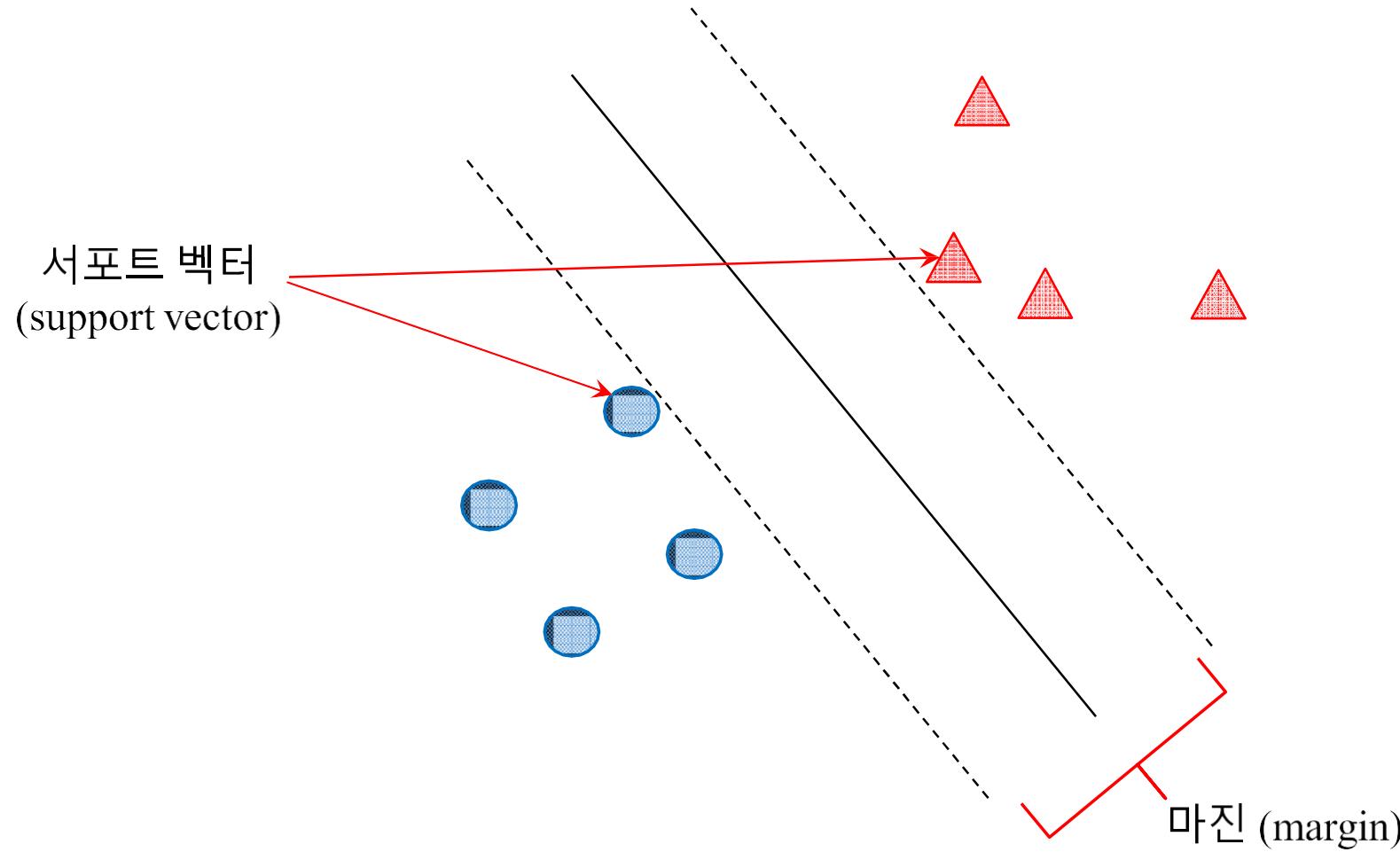
SVM

SVM (Support Vector Machine) 분류 알고리즘:



SVM

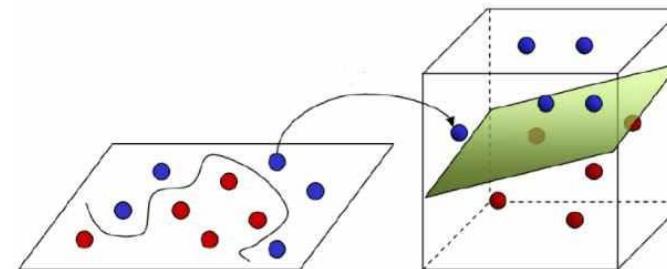
SVM (Support Vector Machine) 분류 알고리즘:



SVM

SVM (Support Vector Machine) 분류 알고리즘:

- 분류 정확성과 분류 마진을 최대로 하는 알고리즘.
- 마진의 종류는 소프트 마진과 하드 마진으로 구분됩니다.
 - 소프트 마진 : 노이즈 등의 이유로 확실한 경계선과 마진을 둘 수 없는 경우.
 - 하드 마진 : 경계선을 기준으로 확실한 구분이 가능한 경우.
- 커널 (kernel) : 마진이 최대화된 **초평면** 공간으로 좌표를 매핑해주는 함수.



SVM

초평면:

- n 차원 공간에서 초평면은 $n - 1$ 차원의 평면이다.
- 예를 들어서 2차원 공간의 초평면은 2차원을 반으로 가르는 직선이며 다음 수식을 따른다.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

→ 위 초평면은 2차원 공간을 위, 아래 영역으로 가르고 관측값은 이 영역 중 한 곳에 속한다:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 < 0$$

SVM

최대 마진 분류기:

- 그러므로 최대 마진 분류기의 최적화 문제는 다음과 같다:

→ 최대화 대상 : M

→ 최적 파라미터 : $\beta_0, \beta_1, \beta_2$.

$$\text{조건 1: } \sum_{i=1}^n \beta_i^2 = 1$$

$$\text{조건 2: } Y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_n X_{in}) \geq M \quad , \quad i = 1, 2, \dots, m$$

$$Y_1, Y_2, \dots, Y_m \in \{-1, 1\}$$

SVM

SVM 분류기:

- SVM 분류기는 최대 마진 분류기의 확장 버전이다. 정확한 분류가 불가능한 경우, 허용된 범위내의 오차를 용인하며 최적화를 시도한다.

→ 최대화 대상 : M

→ 최적 파라미터 : $\beta_0, \beta_1, \beta_2$.

조건 1: $\sum_{i=1}^n \beta_i^2 = 1$

조건 2: $Y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_n X_{in}) \geq M(1 - \varepsilon_i)$, $i = 1, 2, \dots, m$

$Y_1, Y_2, \dots, Y_m \in \{-1, 1\}$

조건 3: $\varepsilon_i \geq 0$ and $\sum_{i=1}^n \varepsilon_i \leq C$

SVM

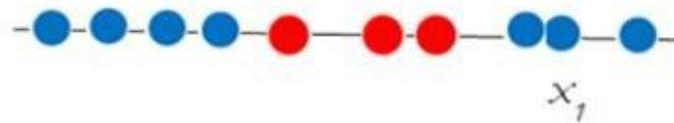
SVM 분류기:

- SVM 분류기는 최대 마진 분류기의 확장 버전이다. 정확한 분류가 불가능한 경우, 허용된 범위내의 오차를 용인하며 최적화를 시도한다.
- C 값에 따라서 오차의 허용도가 조절된다.
 - 높은 C 값은 모델을 더 유연하게 만들어 준다.
 - 낮은 C 값은 위반이 적어져 모델을 더 안정적으로 만들어 준다.
 - C 값은 SVM의 튜닝 파라미터이다.

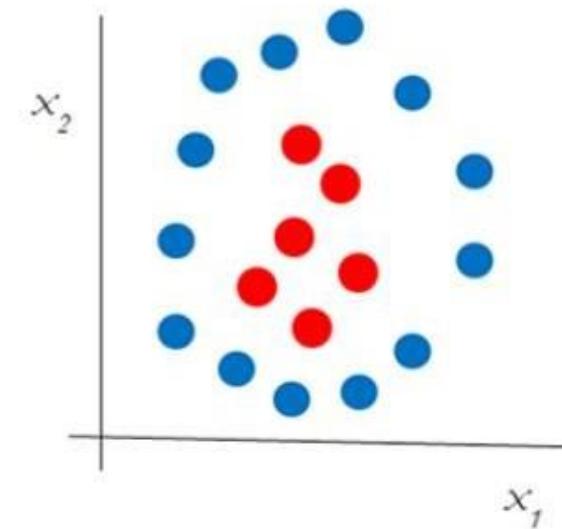
SVM

비선형 경계: 다음은 어떠한 비용 C 를 사용해도 선형 초평면을 구할 수 없다.

1 차원

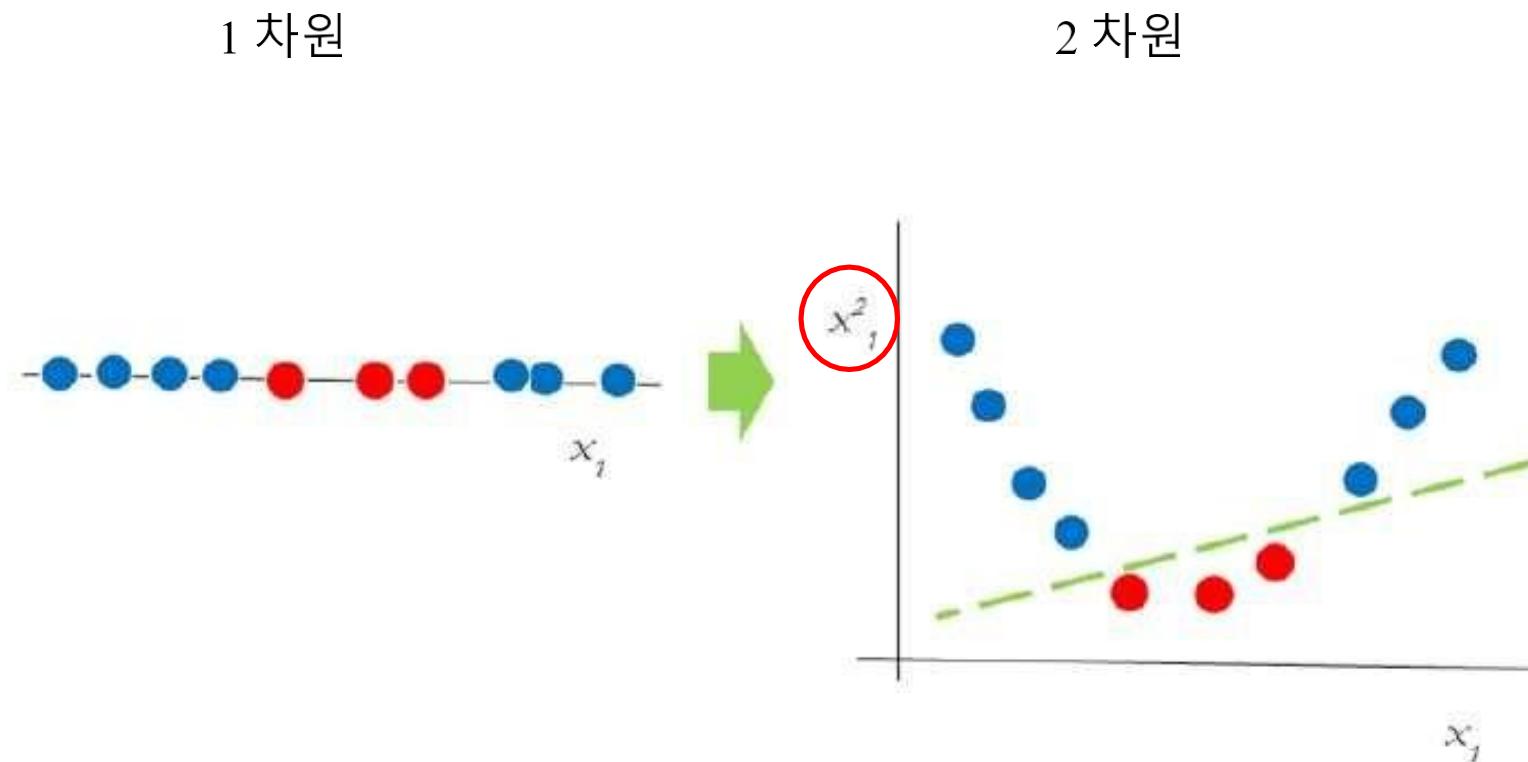


2 차원



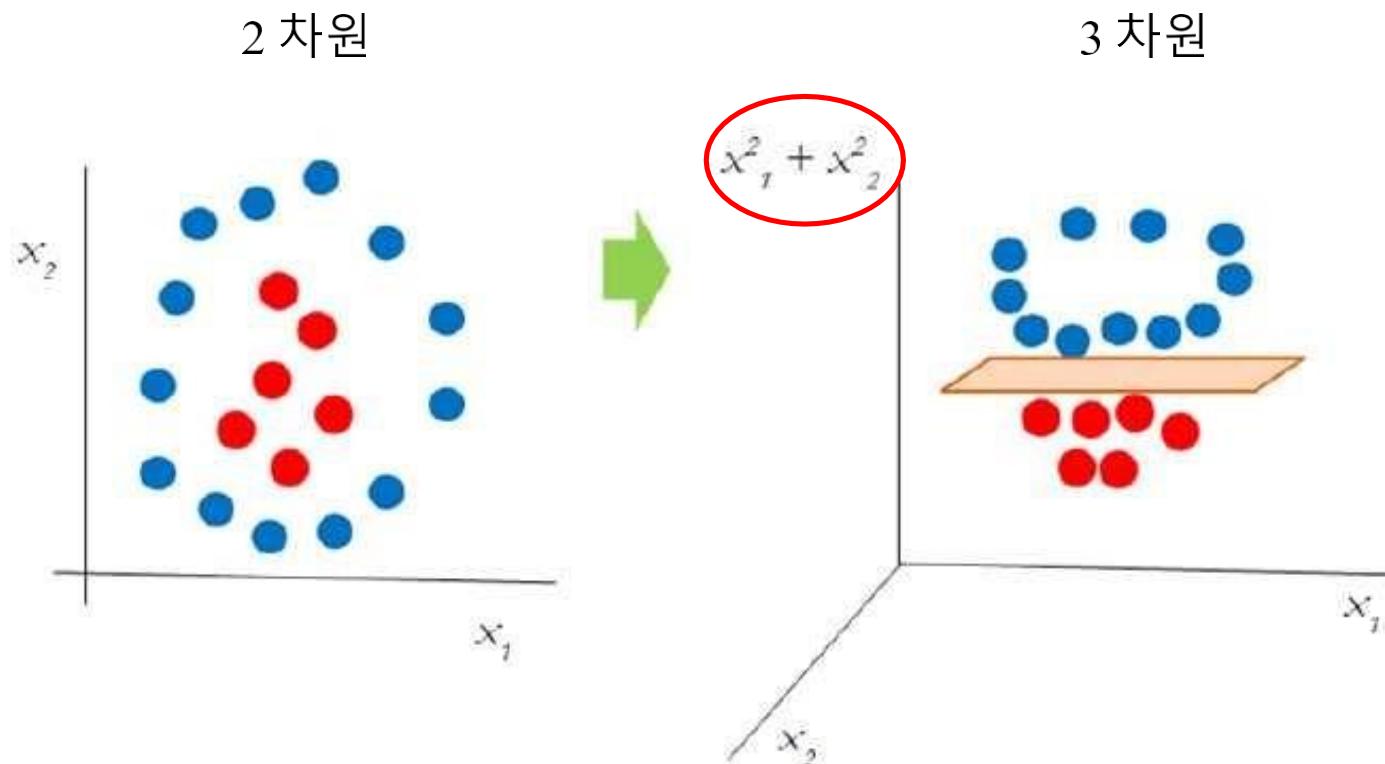
SVM

비선형 경계: 다항 커널을 적용해서 1차원에서 2차원으로 변환.



SVM

비선형 경계: 다항 커널을 적용해서 2차원에서 3차원으로 변환.



SVM

커널함수 (kernel):

- 커널함수는 변환된 벡터 (좌표) 사이의 내적을 구해주는 함수이다.
- 커널함수를 사용하면 벡터를 고차원 공간으로 변환한 후 직접 내적을 구하는 수고를 덜어준다.
- 특정 커널함수를 사용한다는 것은 특정 방식의 변환 (매핑)의 의미를 내포한다.
- 커널의 종류:
 - Polynomial Kernel (다항식 커널).
 - Radial Basis Function Kernel (RBF, 래디얼 함수 커널). 감마 (γ) 파라미터 사용.
감마가 클수록 높은 편중과 작은 분산의 의미.
 - Gaussian Kernel (가우스 커널).

비지도학습

머신러닝의 유형

유형	방법
지도학습 (Supervised Learning)	선형회귀, 로지스틱 회귀
	트리, 랜덤포레스트, 애이디부스트
	Naïve Bayes
	Support Vector Machine (SVM)
	인공신경망
	k-NN
비지도학습 (Unsupervised Learning)	군집분석: k-means, hierarchical, DBSCAN
	주성분 분석 (PCA), 비음수 행렬분해 (NMF)
	t-SNE
	연관성 분석

비지도학습의 활용

비지도학습의 활용:

- 학습 데이터의 구조적 이해를 위해서 활용할 수 있다.
- 새로운 데이터의 분류의 목적으로 활용할 수 있다.

Red? Blue?

⋮

K-means 클러스터 알고리즘

k-means 클러스터 알고리즘:

- 비지도학습.
- 목적은 관측값을 k 개의 군집 (클러스터)으로 분류하는 것.
- 군집에는 centroid 라고도 불리우는 중심점이 있음.
- 반복적 수렴 알고리즘 (Lloyd의 표준 알고리즘).
- 연속적 변수를 사용하며 거리의 개념이 필요하다.

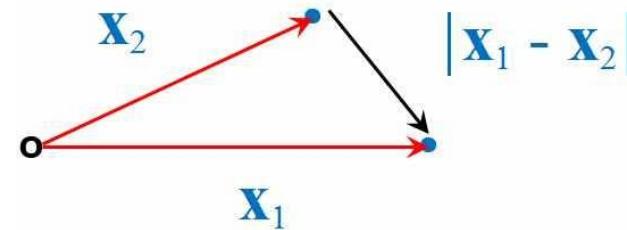
K-means 클러스터 알고리즘

k-means 클러스터 알고리즘: 장단점

장점	단점
<ul style="list-style-type: none">✓ 직관적인 이해가 가능함.✓ 모수에 대한 검정이 필요 없다.✓ 쉽게 적용할 수 있다.	<ul style="list-style-type: none">✓ 노이즈의 영향을 많이 받는다.✓ 외상치 (outlier)에 비교적 민감하다.

거리의 척도 : 유클리드 거리

유클리드 거리 (Euclidian distance):



$$\text{Euclidean distance} = \sqrt{(x_{11} - x_{21})^2 + \cdots + (x_{1m} - x_{2m})^2}$$

거리의 척도 : 정의

다른 거리의 정의:

- 수치 변수인 경우:
 - 유클리드 거리.
 - 표준화 거리.
 - 마할라노비스 거리.
 - 체비셰프 거리.
 - 캔버라 거리.
 - 맨하탄 거리.
 - 민코우스키 거리.
- 유형 변수인 경우:
 - 자카드 거리, 등.

K-means 클러스터 알고리즘의 원리

다음과 같이 N 개의 좌표로 나타내는 데이터 세트를 가정해 봅니다.

$$x_1, x_2, \dots, x_N$$

두개 ($k = 2$) 의 클러스터를 목표로 설정해 봅니다.

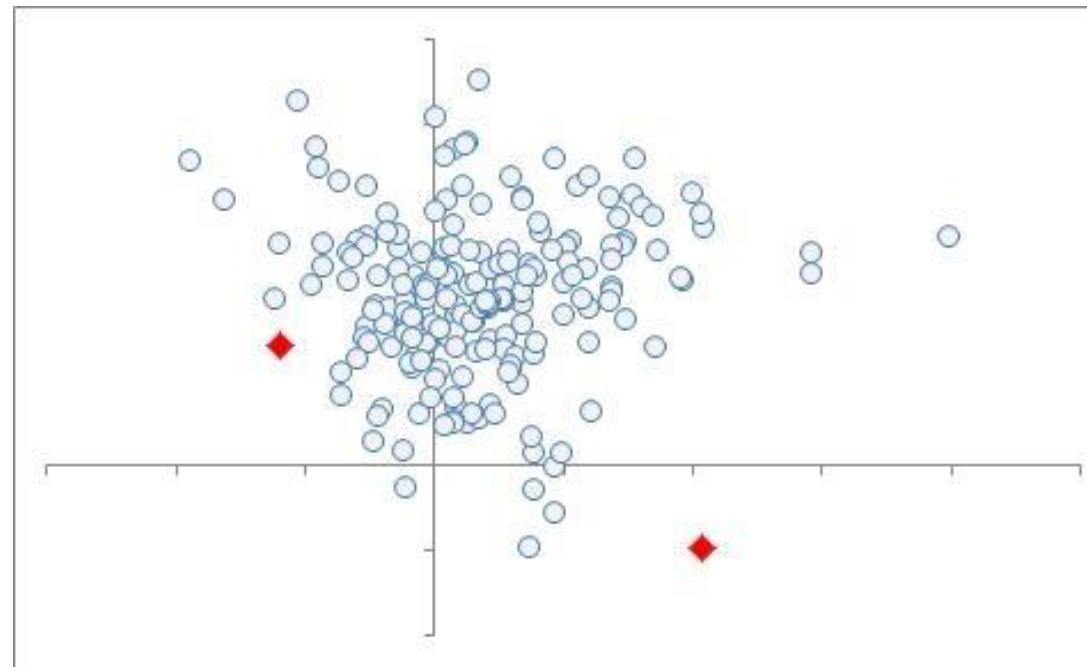
$$C_1 \text{ and } C_2$$

K-means 클러스터 알고리즘의 원리

두개의 centroid의 위치를 랜덤으로 초기화 합니다.

μ_1 and μ_2

K-means 클러스터 알고리즘의 원리



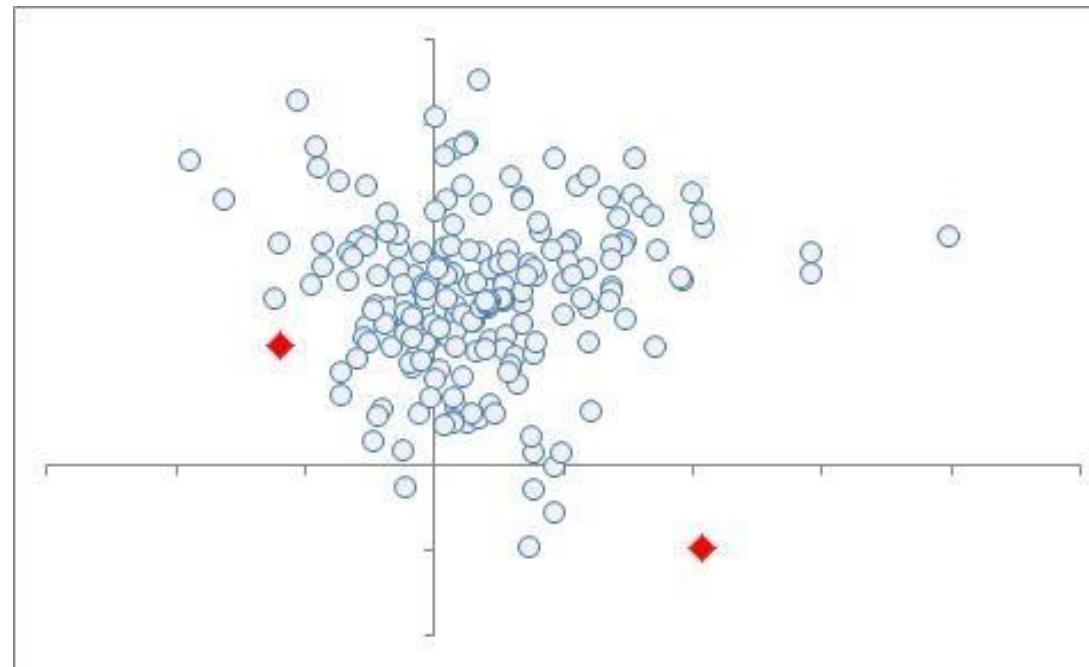
K-means 클러스터 알고리즘의 원리

그리고서는, centroid와 데이터 좌표 사이의 제곱 거리(*)가 최소화 되는 방향으로 centroid의 위치를 갱신 시킵니다.

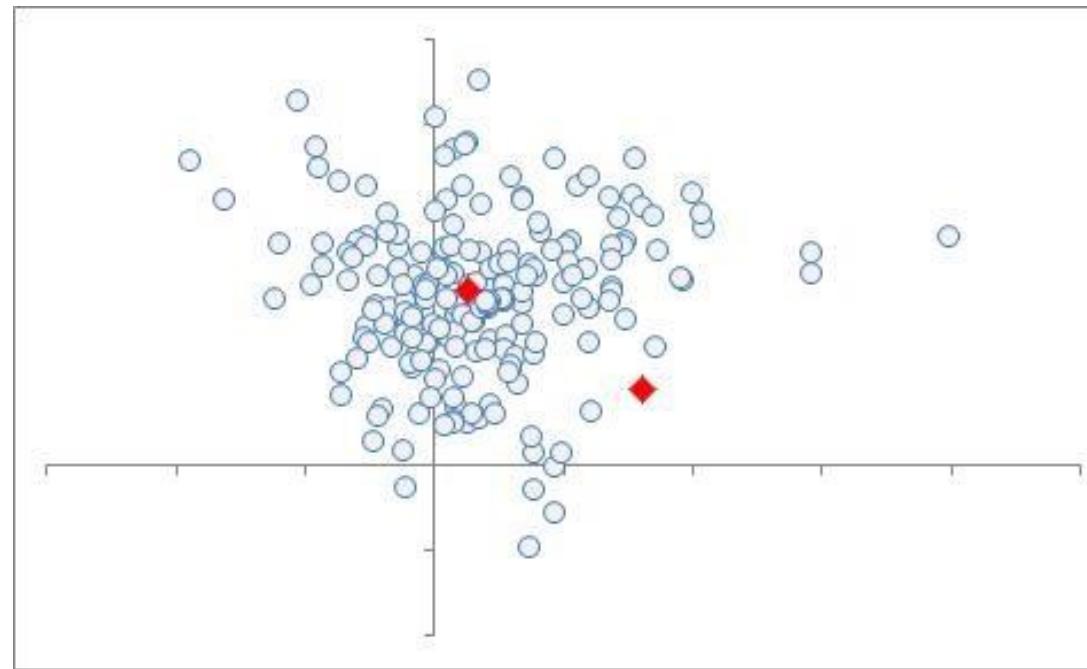
$$\text{minimize} \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

(*) “Sum of Square Distance Within”이라고 불리웁니다.

K-means 클러스터 알고리즘의 원리

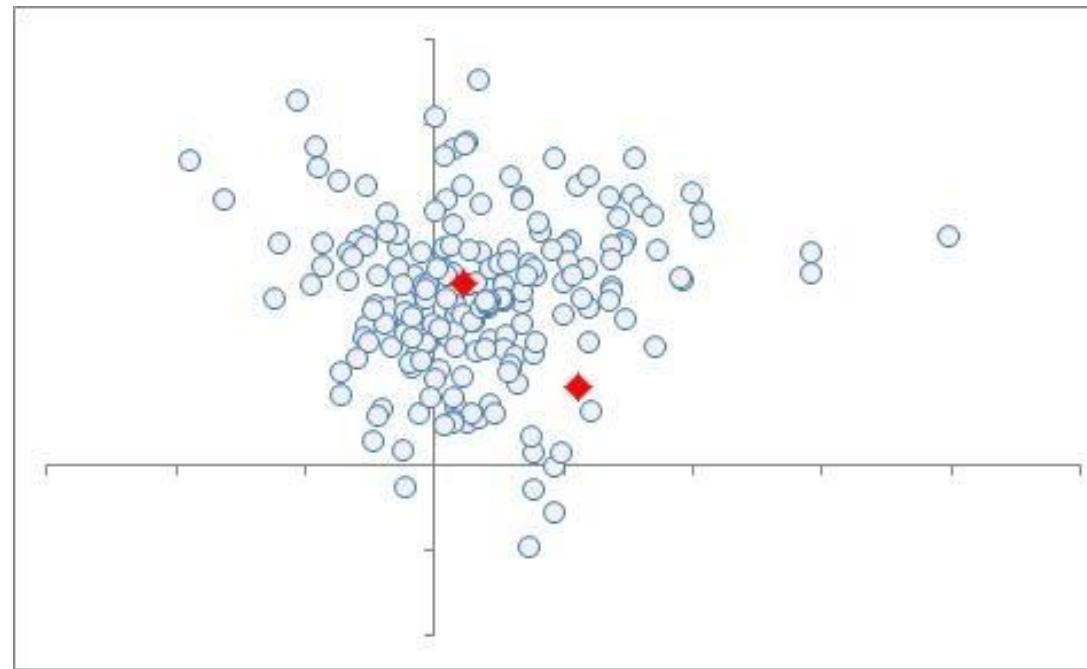


K-means 클러스터 알고리즘의 원리



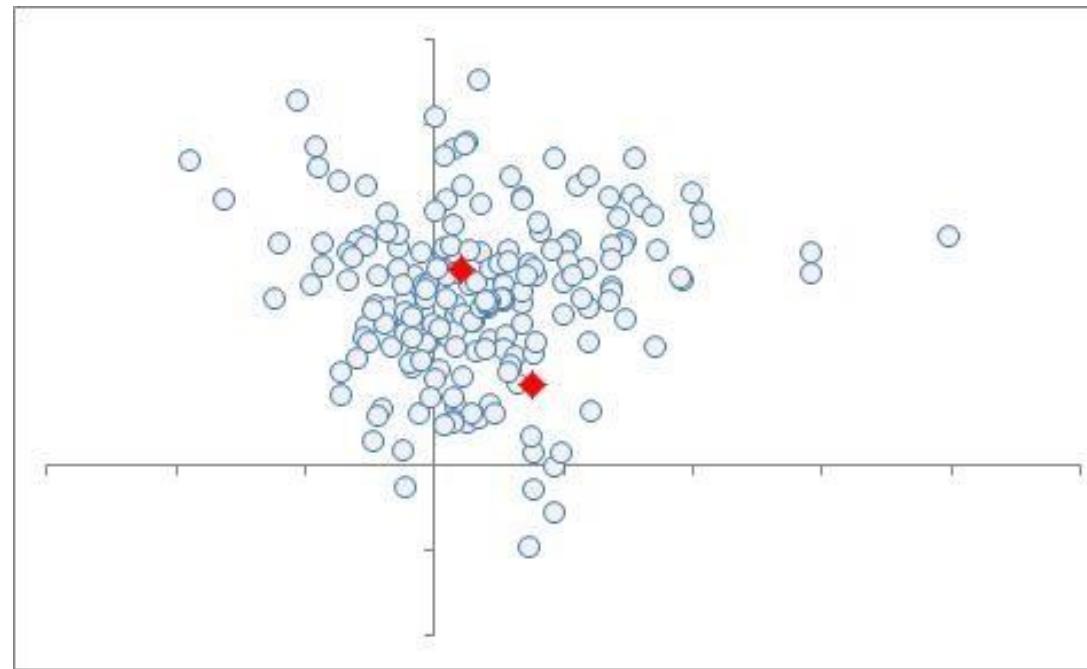
반복적 갱신.

K-means 클러스터 알고리즘의 원리



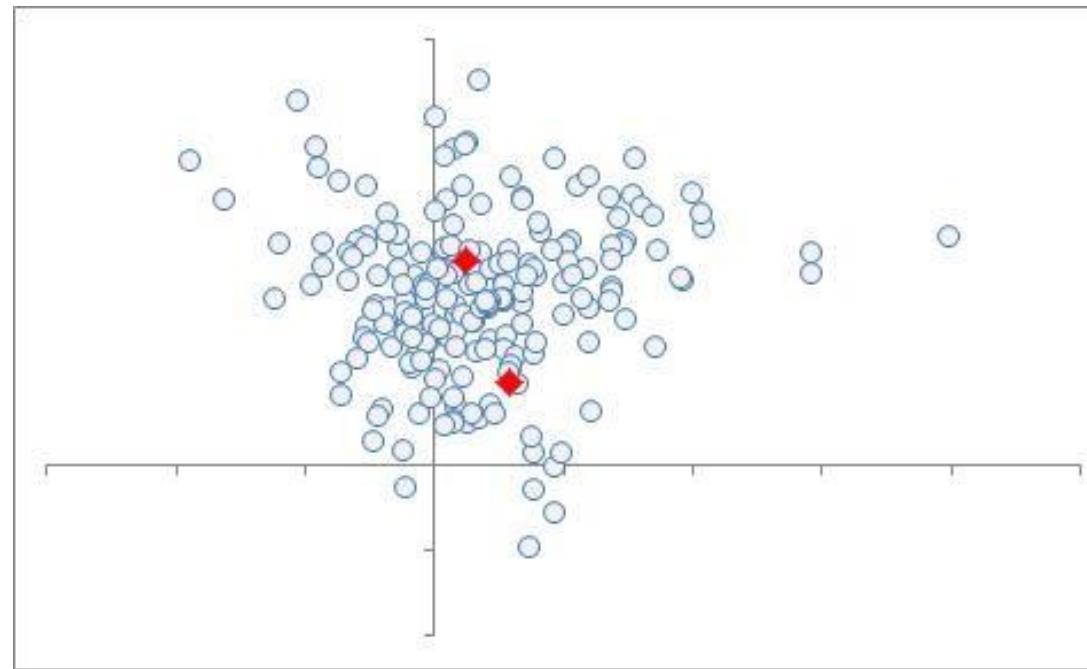
반복적 갱신.

K-means 클러스터 알고리즘의 원리



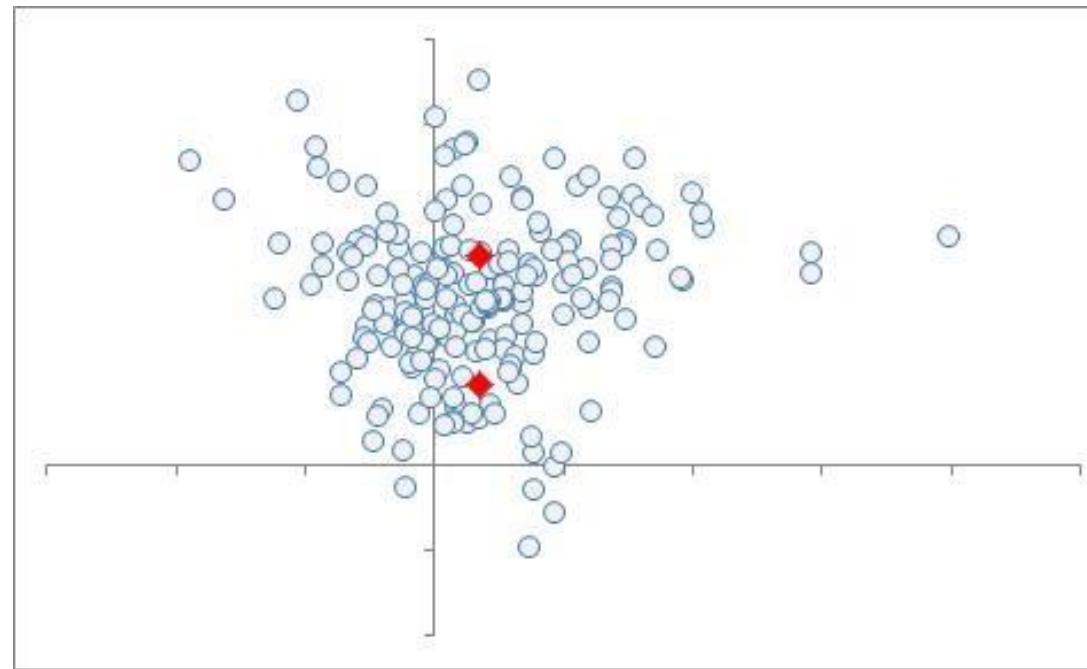
반복적 갱신.

K-means 클러스터 알고리즘의 원리



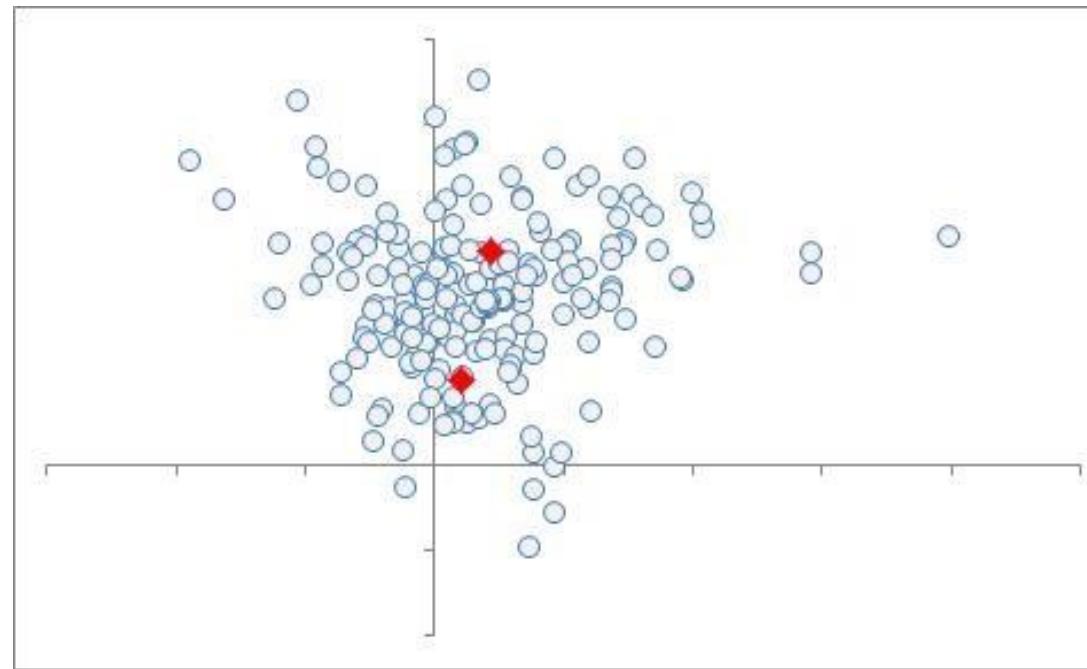
반복적 갱신.

K-means 클러스터 알고리즘의 원리



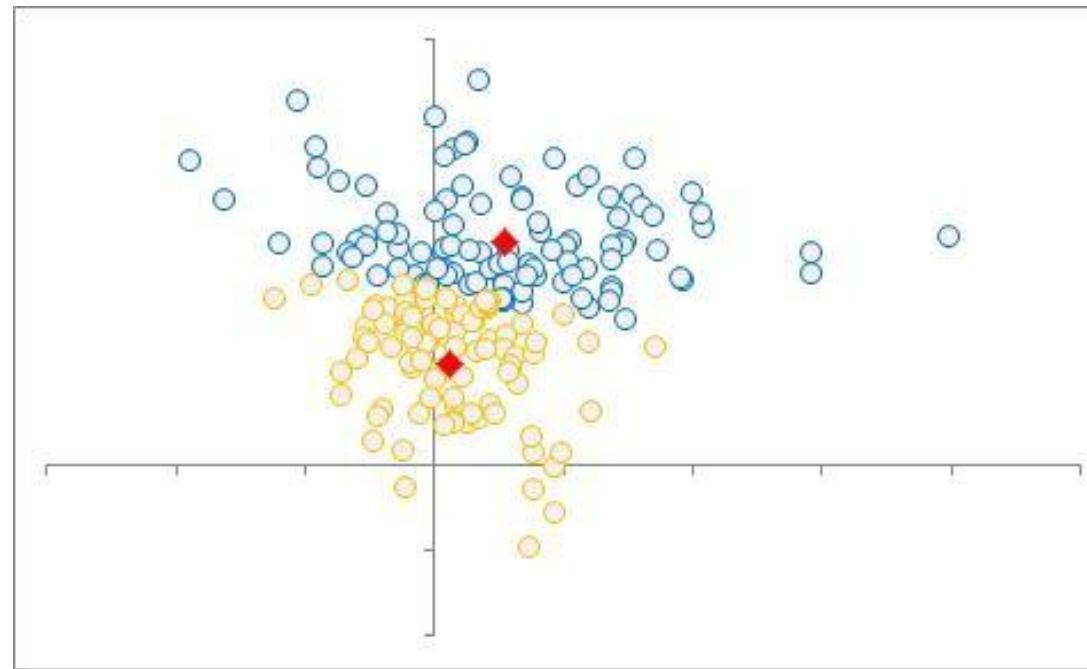
반복적 갱신.

K-means 클러스터 알고리즘의 원리



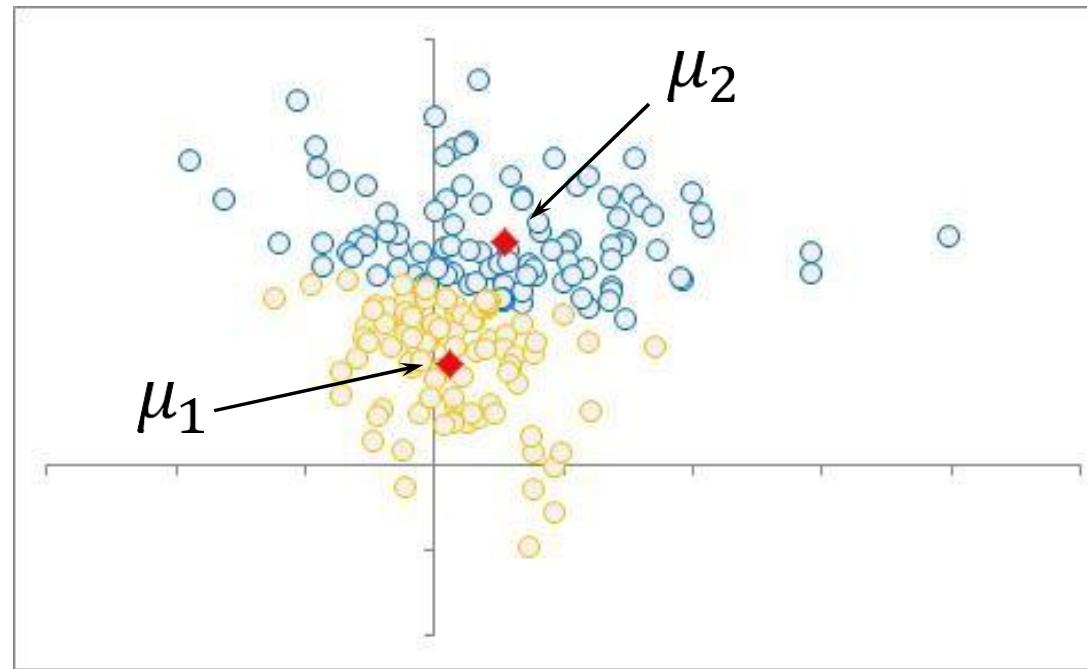
반복적 갱신.

K-means 클러스터 알고리즘의 원리



드디어 수렴!

K-means 클러스터 알고리즘의 원리



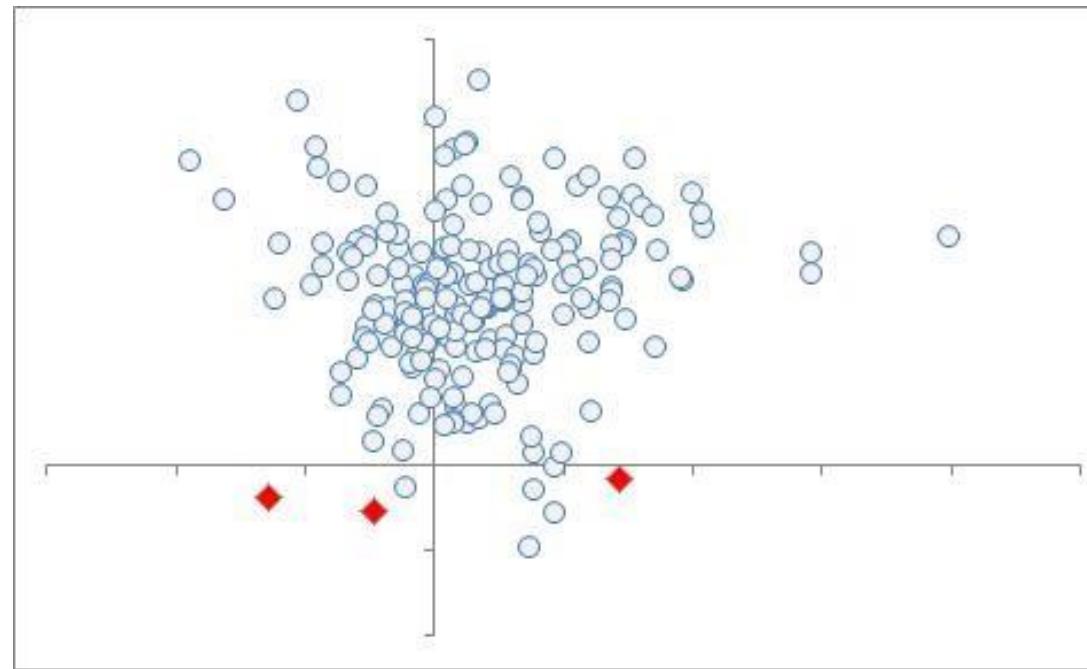
Centroid의 최종 위치.

K-means 클러스터 알고리즘의 원리

이제는 세개 ($k = 3$) 의 클러스터를 목표로 설정해 봅니다.

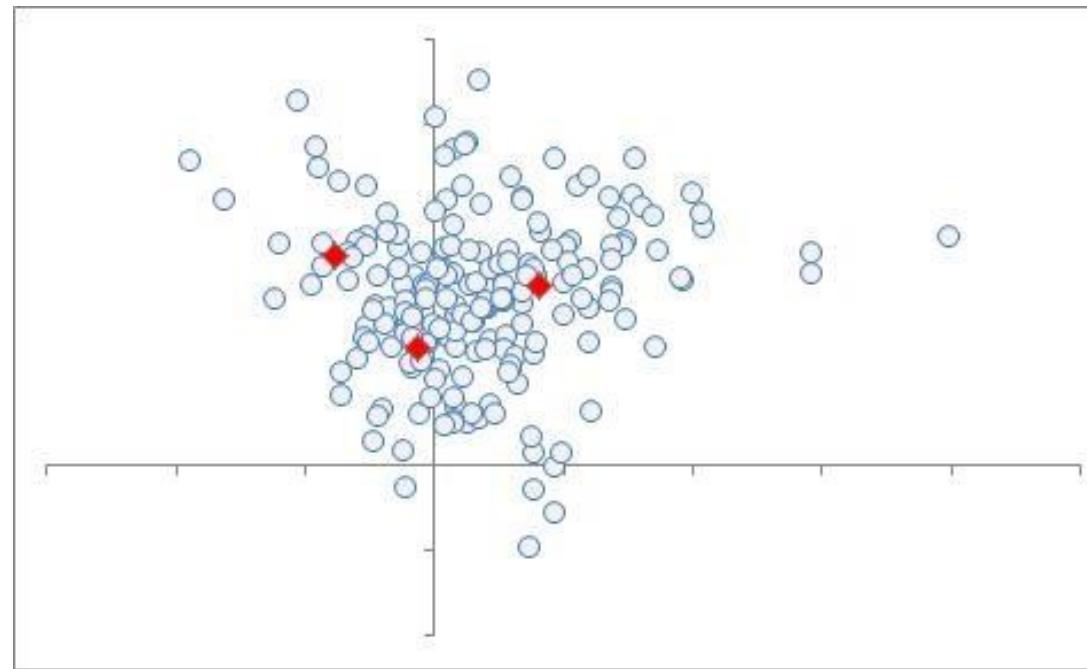
C_1 , C_2 and C_3

K-means 클러스터 알고리즘의 원리



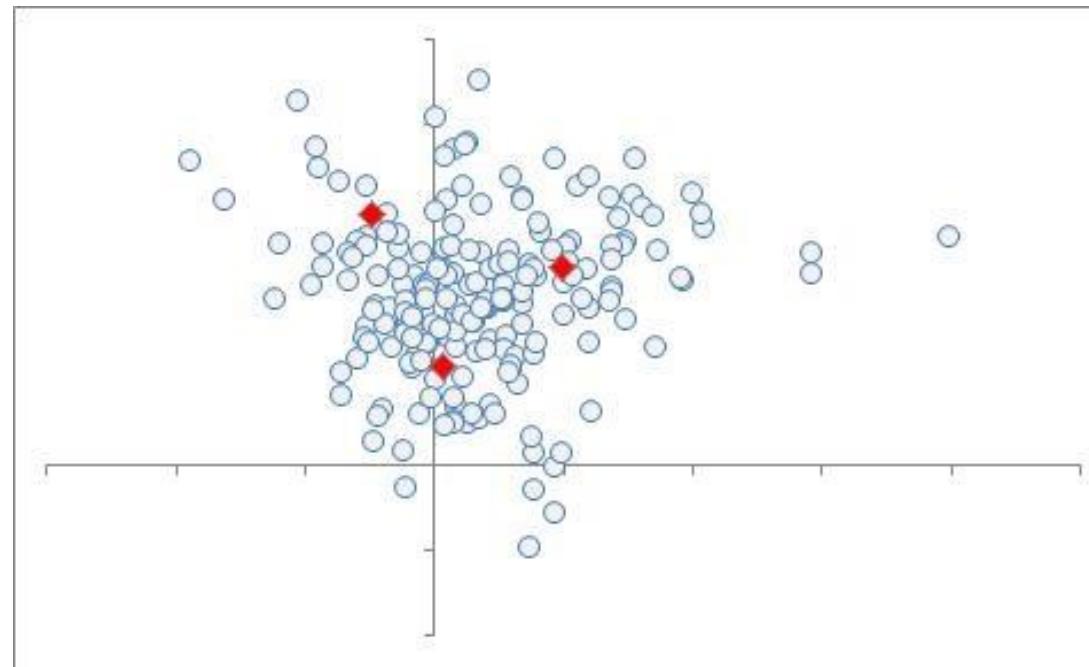
랜덤 초기화.

K-means 클러스터 알고리즘의 원리



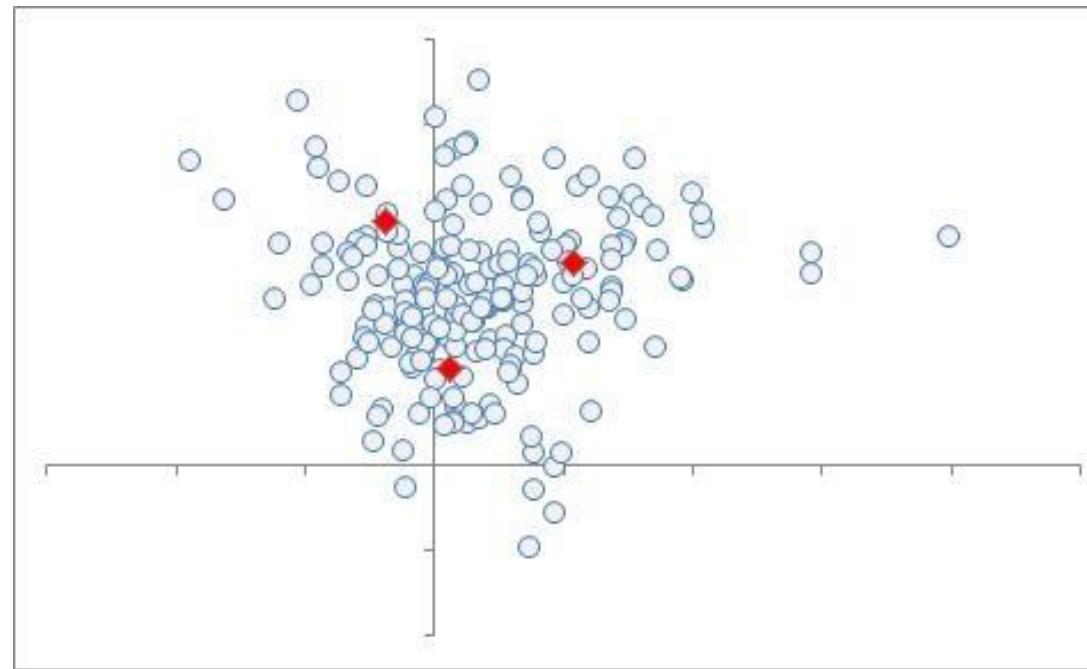
반복적 갱신.

K-means 클러스터 알고리즘의 원리



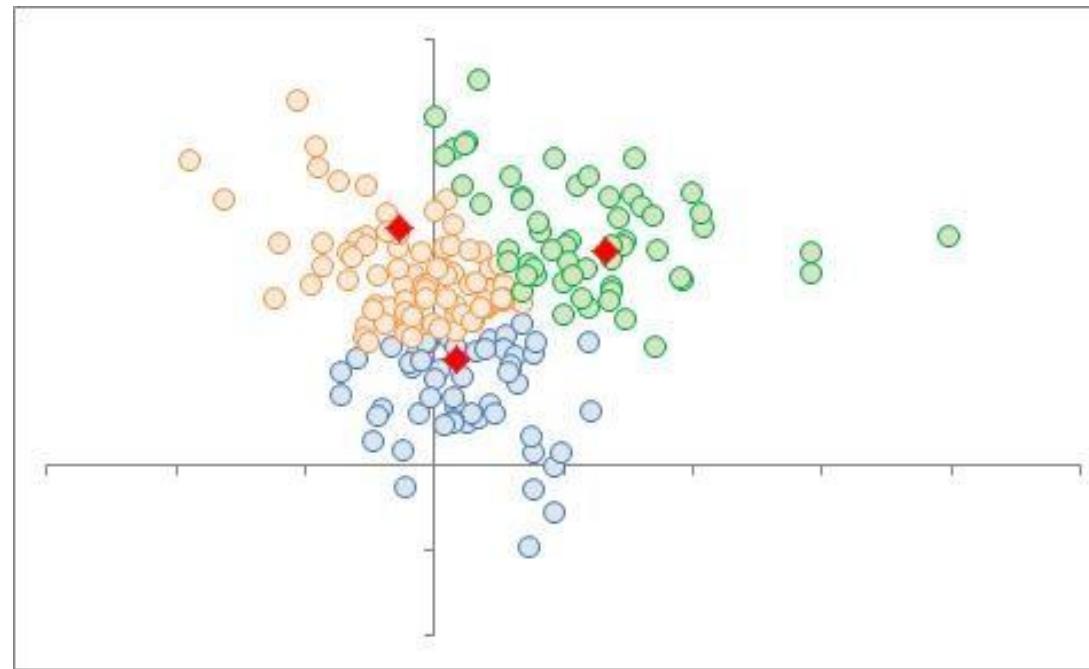
반복적 갱신.

K-means 클러스터 알고리즘의 원리



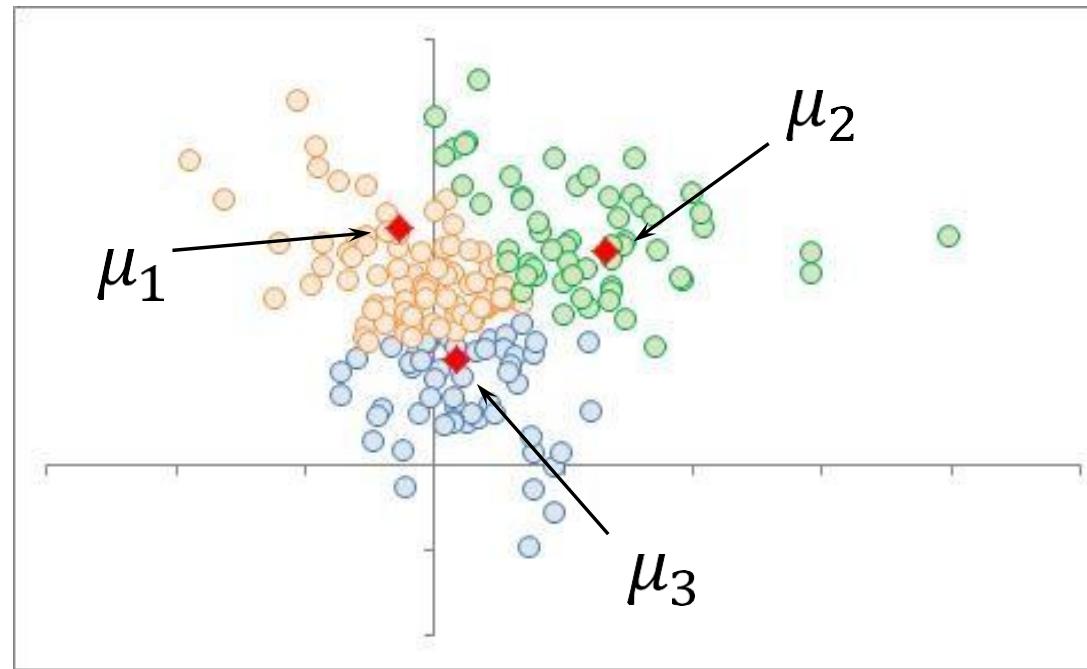
반복적 갱신.

K-means 클러스터 알고리즘의 원리



드디어 수렴!

K-means 클러스터 알고리즘의 원리



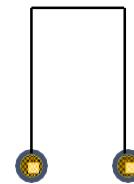
Centroid의 최종 위치.

계층적 군집화

계층적 군집화 (hierarchical clustering):

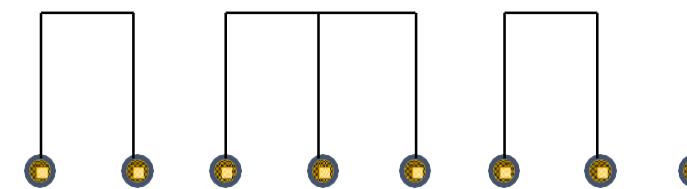
- 자율 학습.
- 병합군집 (agglomerative clustering) 알고리즘의 대표.
- 가까운 아이템끼리 순서대로 뭉쳐가는 형식.
- 위아래가 역전된 나무(tree)의 형상을 보임.

계층적 군집화



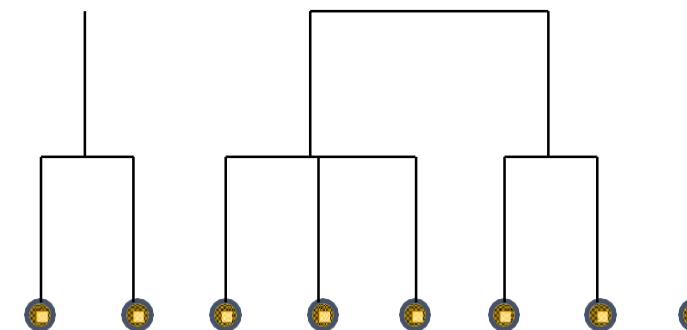
가장 가까운 아이템을 연결.

계층적 군집화



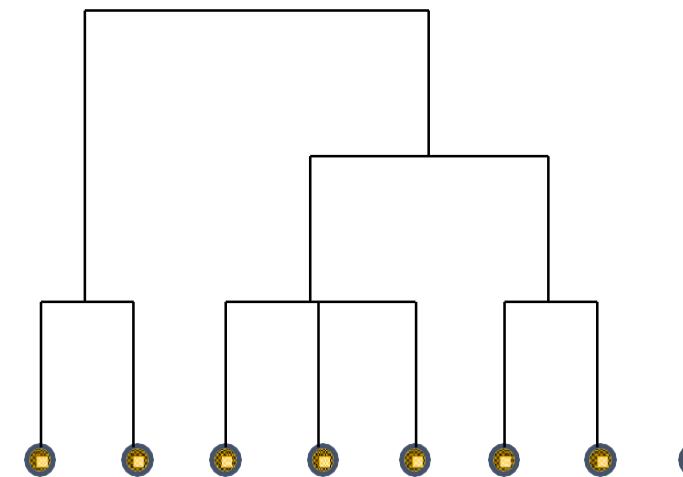
다양한 크기의 군집이 형성됨.

계층적 군집화



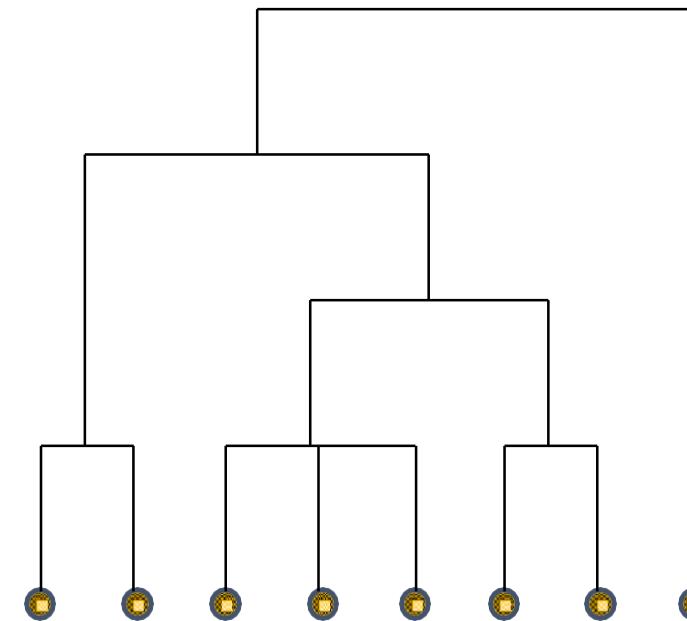
이제는 가까운 군집끼리 연결함.

계층적 군집화



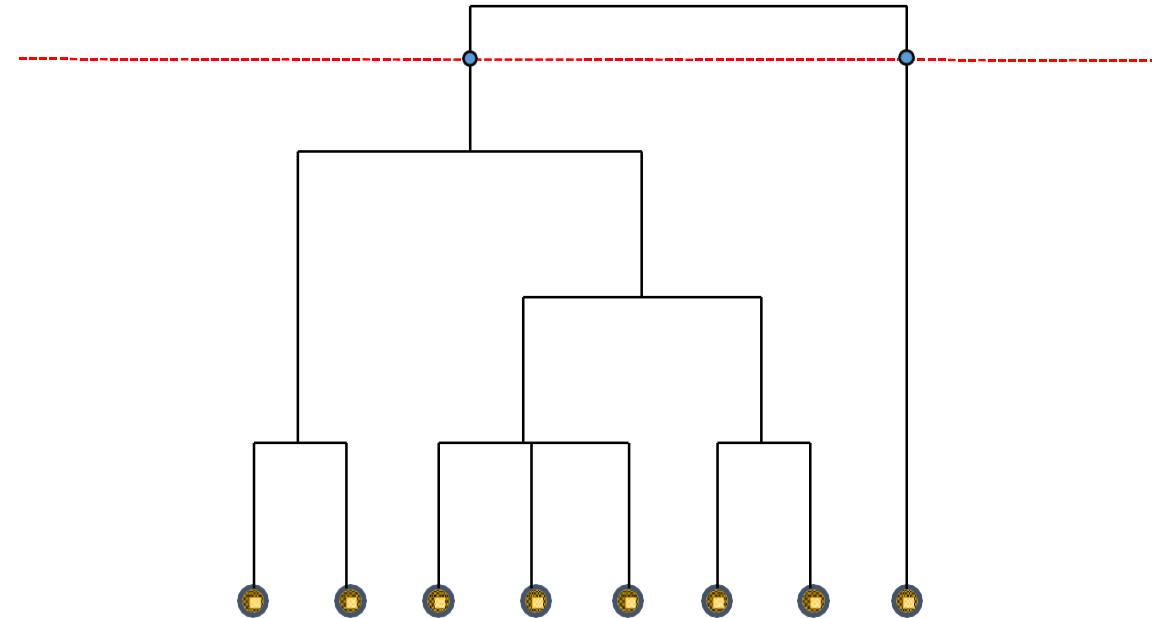
가까운 군집끼리 연결함.

계층적 군집화



궁극적으로는 하나의 둉어리(군집)으로 뭉침.

계층적 군집화



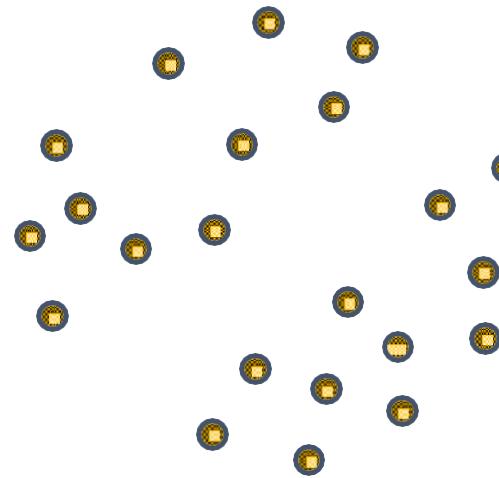
임의의 “높이”를 정하고 살펴봄.

DBSCAN 군집화

DBSCAN 군집화:

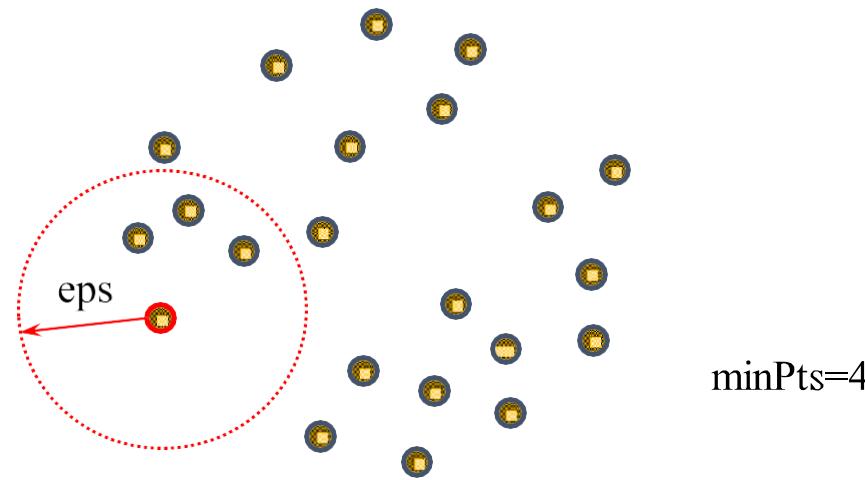
- 자율 학습.
- 1996년에 개발되어서 가장 효과적이고 많이 사용되는 군집분석방법.
- 밀도 (density)에 따라서 군집을 만들어 감.
- 엡실론(eps), 최소밀도(minPts), 등과 같은 파라미터를 정해 주어야 한다.
- 고밀도 지역이 연결되어 있으면 만족스러운 군집화가 어렵다는 단점이 있다.

DBSCAN 군집화



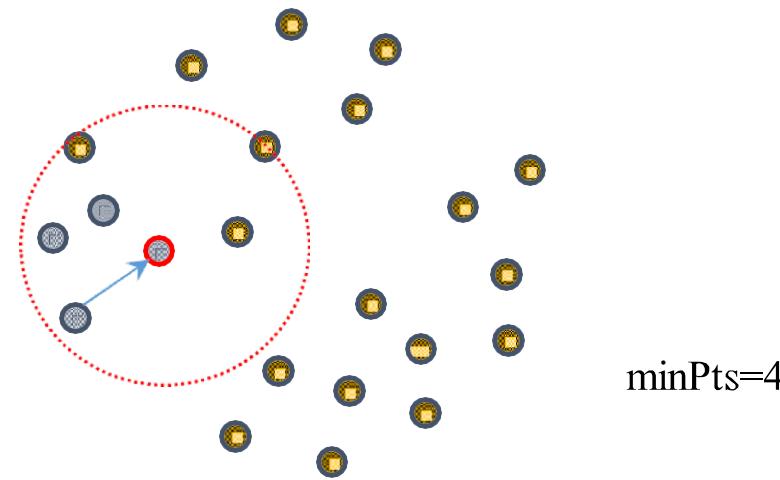
위와 같이 좌표가 분포되어 있다고 가정해 봅니다.

DBSCAN 군집화



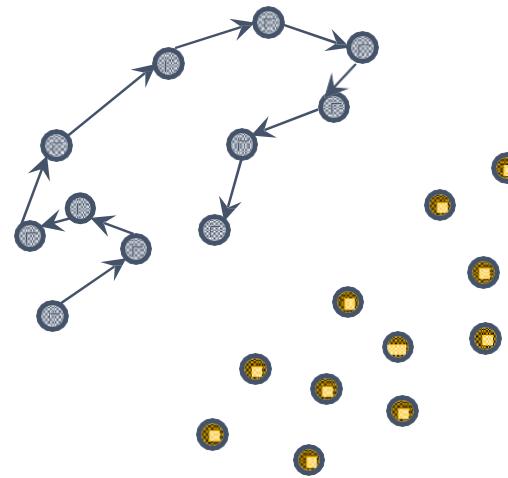
반지름 eps 까지의 거리 안에 minPts 이상의 좌표가 있는지 확인.

DBSCAN 군집화



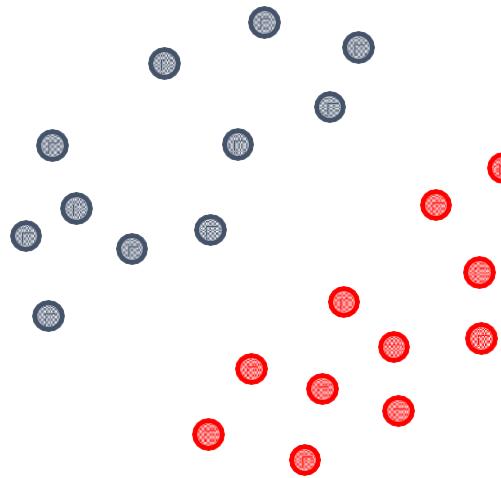
조건이 맞으면 다른점으로 옮겨가서 똑같은 조건 확인.

DBSCAN 군집화



반복적으로 실행해서 끊기지 않고 연결되는 좌표들이 군집을 형성함.

DBSCAN 군집화



완성!

t-SNE

t-SNE manifold learning:

- 주로 비지도 학습의 목적으로 사용된다.
- 군집화를 통해서 시각화를 향상시키는 알고리즘이다.
- 고차원 데이터를 2D 평면에 매핑하게 되는데 단순 투영은 아니다.
- 매핑 과정에서는 가까운 좌표끼리는 뭉치고 먼 좌표끼리는 더 떨어지게 만든다.

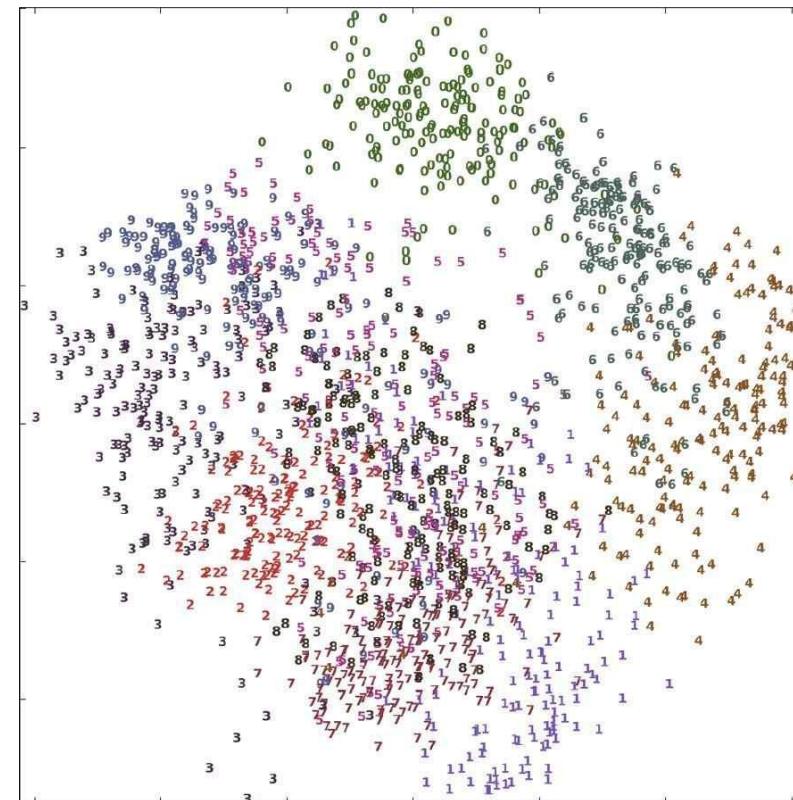
t-SNE

t-SNE manifold learning: 장단점

장점	단점
<ul style="list-style-type: none">✓ PCA 사용 방법보다 더 만족스러운 시각화 결과를 보인다.	<ul style="list-style-type: none">✓ 원 좌표가 변형된다.✓ 좌표 자체의 의미가 퇴색된다.

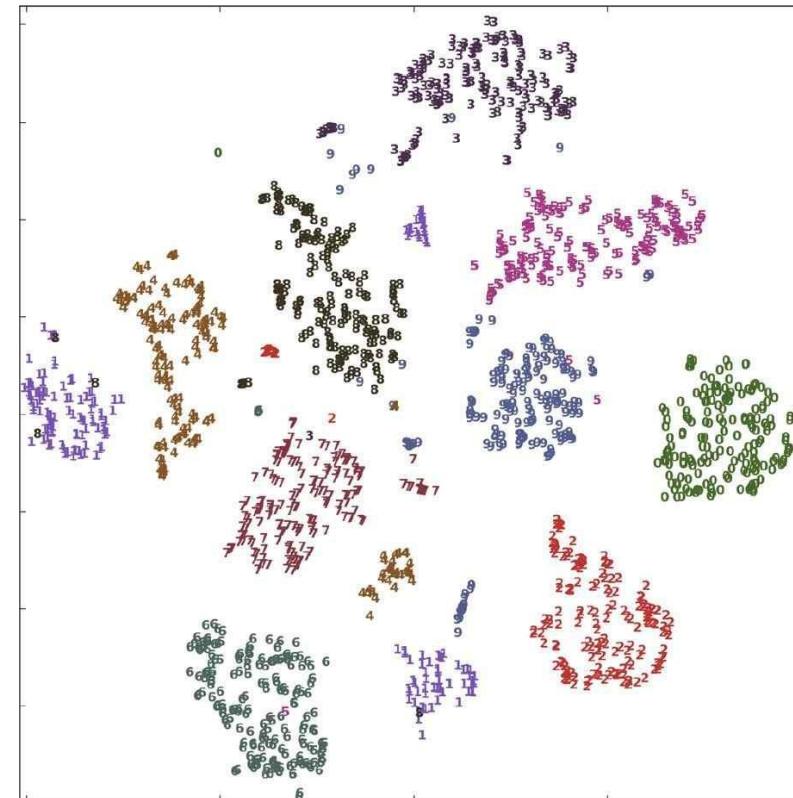
t-SNE : PCA 와 비교

t-SNE 과 PCA 비교: PCA의 주성분에 투영한 경우 군집들이 서로 뒤엉켜있다.



t-SNE : PCA 와 비교

t-SNE 과 PCA 비교: t-SNE를 적용한 경우 군집들을 더욱 또렷히 분간할 수 있다.



국방 인공지능 전문교육
인공지능으로 보다 진보된 대한민국 II 미래를
향해

감사합니다.

