

머신러닝 개요

이상현



Contents

001 머신러닝

- 머신러닝이란?
- 머신러닝의 분류

002 머신러닝 알고리즘

- 예측 문제
- 분류 문제
- 군집 문제
- 강화 학습

003 딥러닝

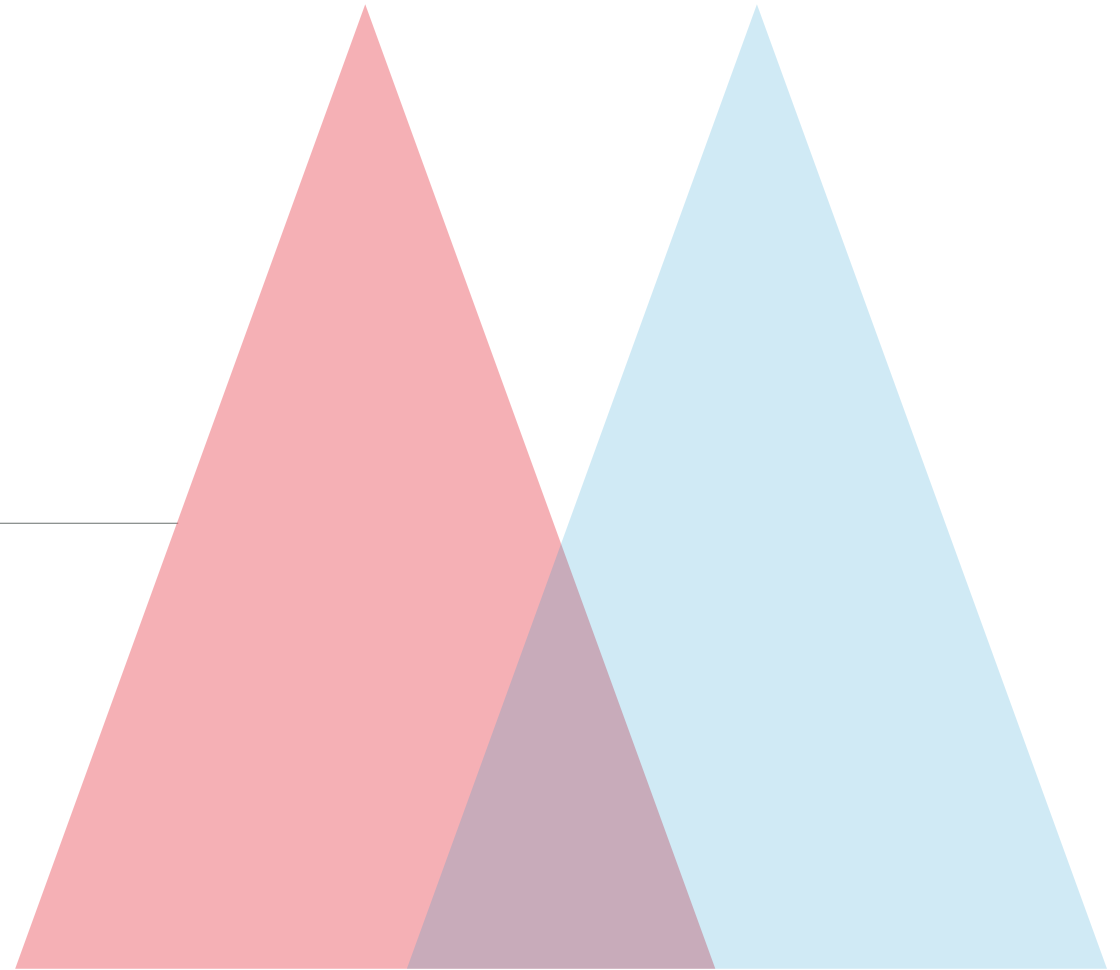
- 딥러닝이란?
- 인공신경망
- 딥러닝 활용 사례

004 참고 문헌

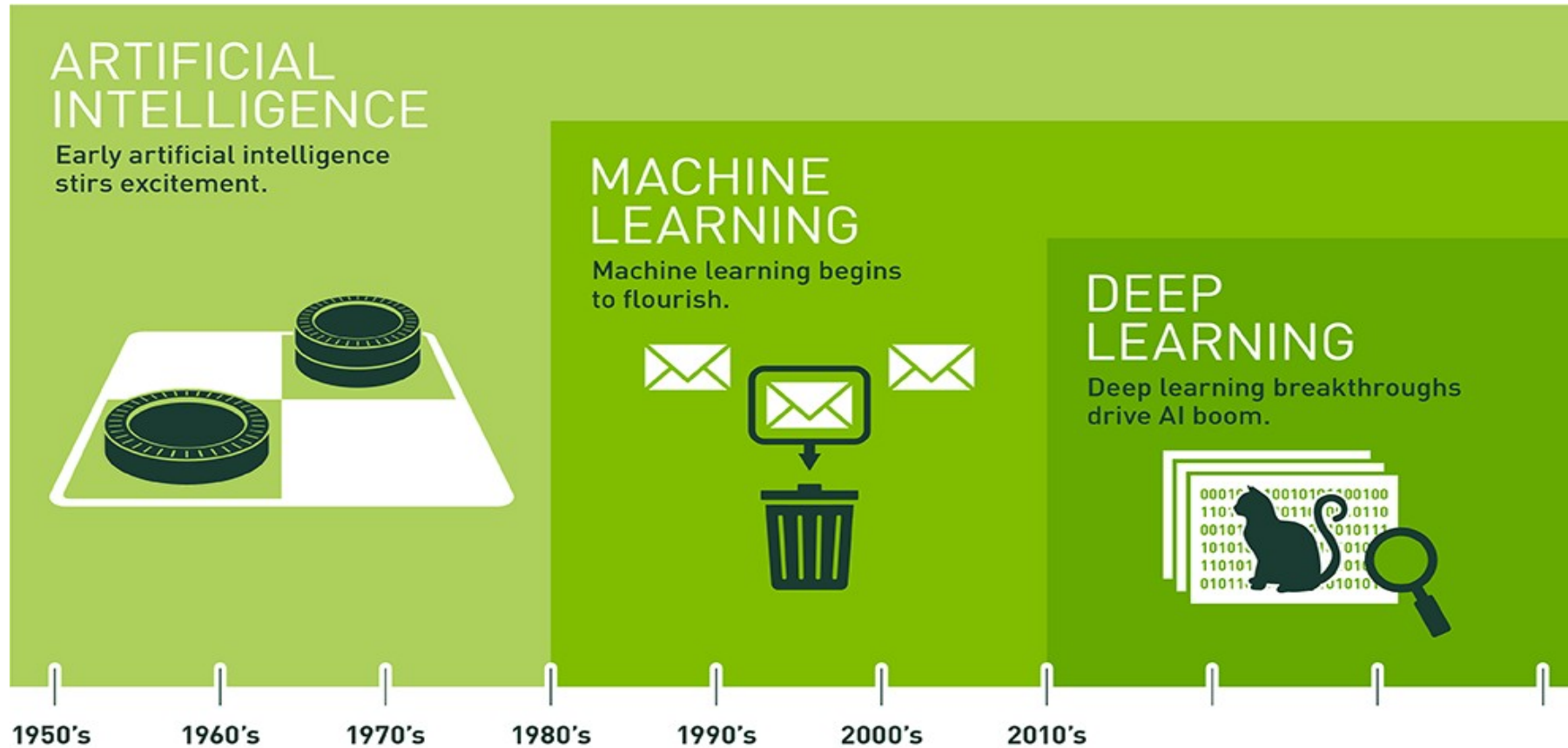
- 참고 문헌

001

머신러닝



머신러닝 이란?



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.

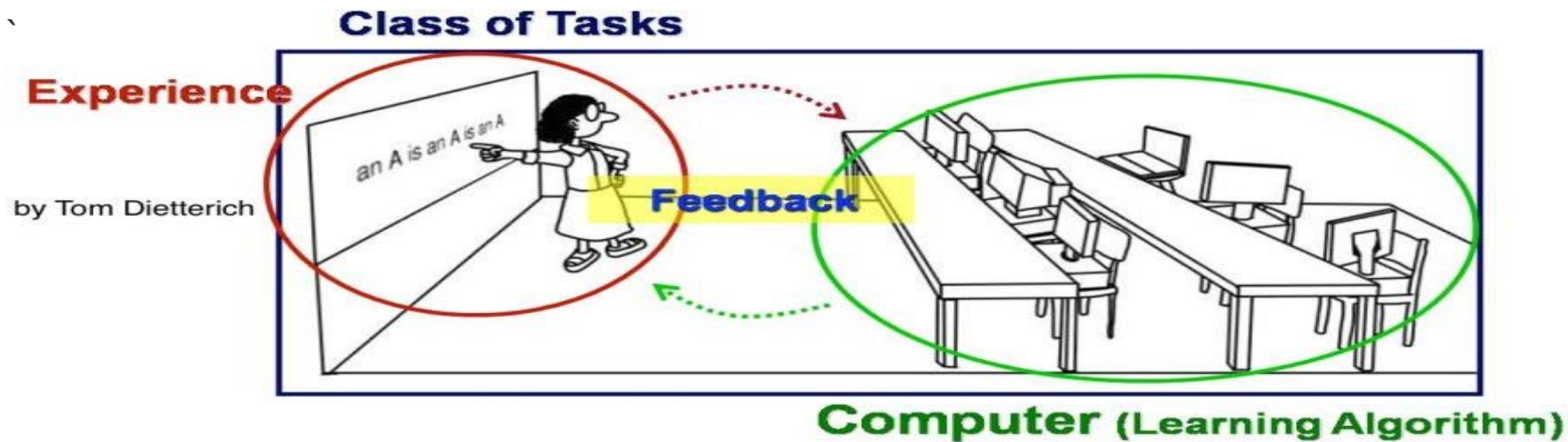
인공지능이 가장 큰 개념이며 그 다음은 머신러닝이고, 가장 작은 개념은 딥 러닝이다.

머신러닝 이란?

- 기계학습은 인공지능의 한 분야로 기계 스스로 대량의 데이터로부터 지식이나 패턴을 찾아 학습하고 예측을 수행하는 것이다

- **Tom Mitchell** (카네기멜론대 교수)

“만약 컴퓨터 프로그램이 특정한 태스크 T 를 수행할 때 성능 P 만큼 개선되는 경험 E 를 보이면, 그 컴퓨터 프로그램은 테스트 T 와 성능 P 에 대해 경험 E 를 학습했다” 라고 할 수 있다



머신러닝 이란?

▪ Arthur Samuel (1959)

명시적으로 프로그램을 작성하지 않고 컴퓨터에 학습할 수 있는 능력을 부여하기 위한 연구 분야

"The field of study that gives computers the ability to learn without being explicitly programmed"

▪ Machine Learning의 예 ($y=3x$)

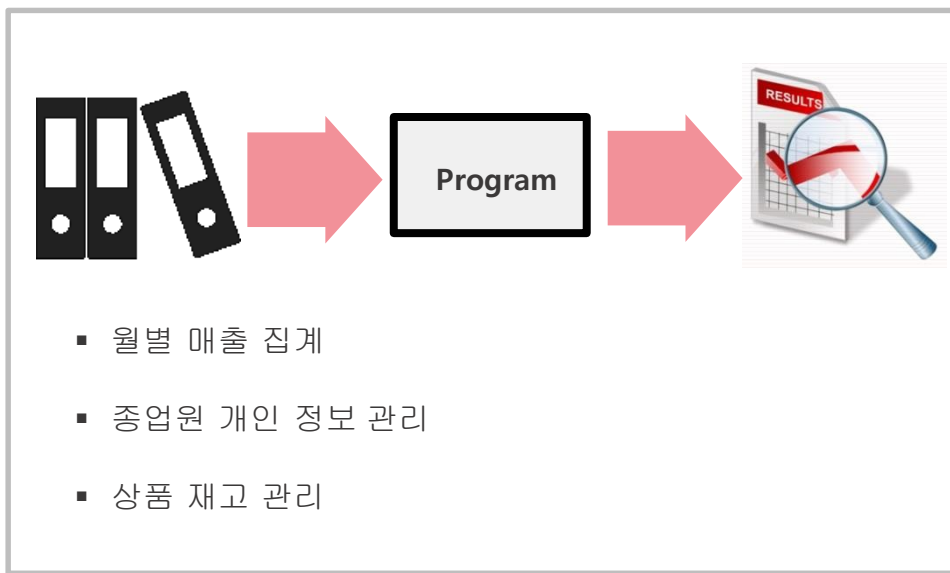
- 학습데이터 (1,3), (3,9), (4,12), (6,18)
- 컴퓨터에 $y=3x$ 의 함수를 프로그래밍하지 않아도 앞의 학습데이터를 학습한 후 (8, ?) (10, ?)

의 질문을 던져도 그 대답을 할 수 있게 만드는 것.

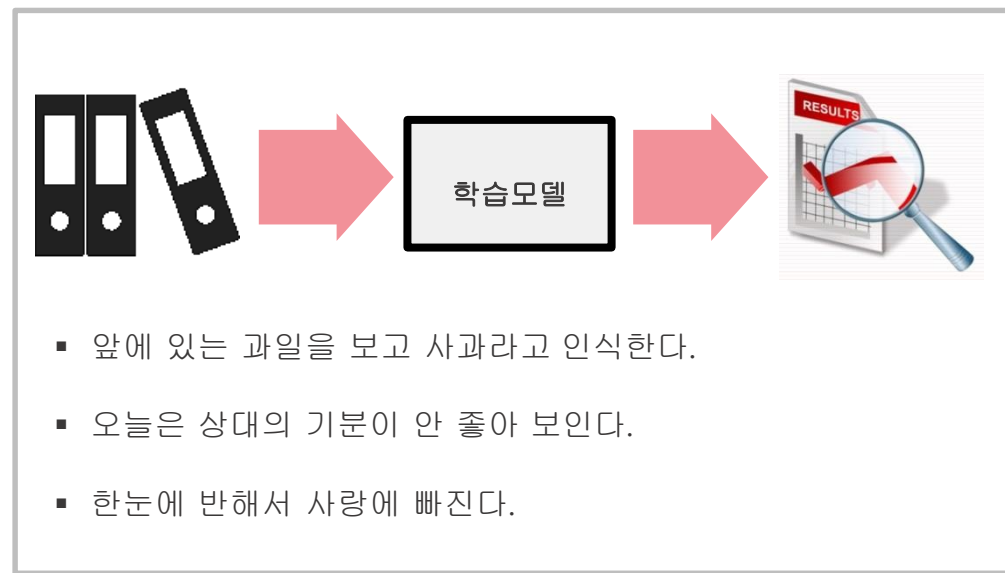
머신러닝 이란?

순서나 이유를 명확하게 설명하지 못하는 일을 처리하기 위한 방법으로 기계 학습을 선택할 수 있다.

【 순차적 처리 】



【 기계 학습 】



▪ Training Set (학습 데이터)

Machine Learning이 학습 모델을 만들기 위해 사용하는 데이터

학습 데이터가 나쁘면 실제 현장의 특성을 제대로 반영하지 못하므로, 학습 데이터 확보 시 실제 데이터의 특성이 잘 반영되고 편향되지 않는 학습 데이터를 확보하는 것이 매우 중요하다.

▪ Model (학습 모델)

Machine Learning에서 구하려는 최종 결과물로 가설(Hypothesis)이라고도 부른다.

머신러닝 이란?

데이터를 이용해서 컴퓨터를 학습시킨다.

학습한 내용을 기반으로 예측을 할 수 있다.

머신 러닝 알고리즘은 크게 세가지로 분류 할 수 있다.



머신러닝의 분류(지도 학습)

학습 데이터에 레이블(Label)이 있는 경우 지도 학습

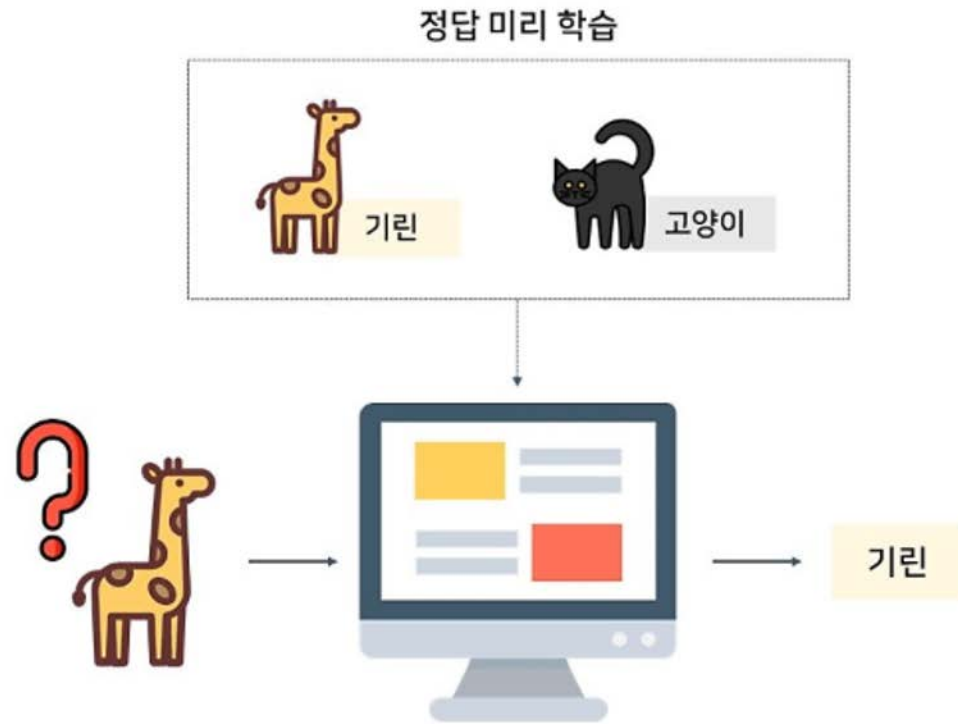
학습 데이터의 속성을 분석하고자 하는 관점에서 정의
사물을 구분하는 태스크



정답이 주어진 상태에서 학습하는 알고리즘

머신러닝의 분류(지도 학습)

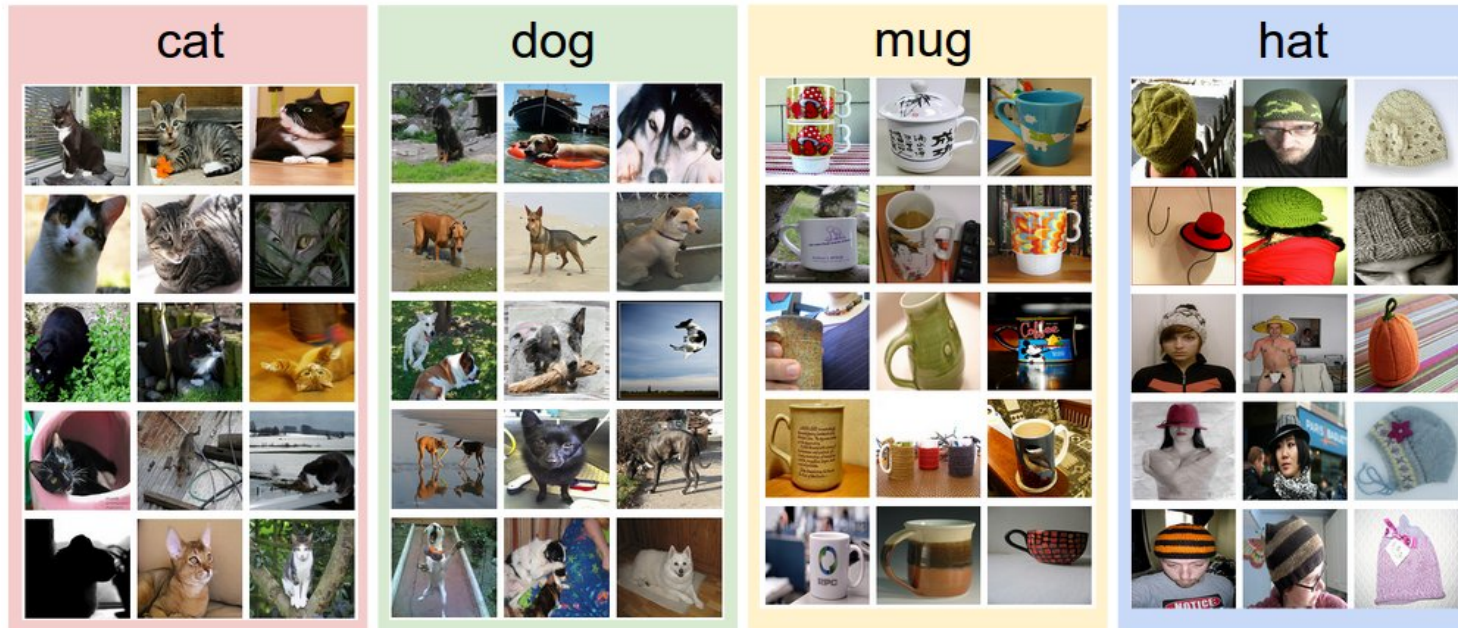
예를 들어 수많은 고양이와 기린의 사진을 주고 각 사진이 고양이인지, 기린인지 정답을 알려줍니다.



머신러닝의 분류(지도 학습)

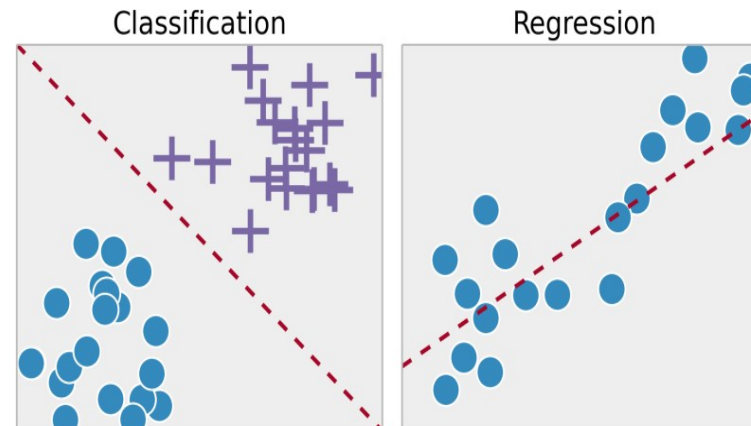
【 지도학습의 Training Set의 예】

Label이 있는 학습 데이터(Training Set)를 이용해서 학습.



■ 분류와 회귀의 비교

	분류 (Classification)	회귀 (Regression)
결과	학습데이터의 레이블 중 하나를 예측 (discrete)	연속된 값을 예측 (Continuous)
예제	학습데이터가 A, B, C 인 경우 결과는 A, B, C 중 하나다. 예) 스팸메일 필터	결과 값이 어떠한 값도 나올 수 있다. 예) 중고차 가격 예측



머신러닝의 분류(지도 학습)

Types	Tasks	Algorithms
지도학습 (Supervised Learning)	분류 (Classification)	<ul style="list-style-type: none">▪ KNN : k Nearest Neighbor▪ SVM : Support Vector Machine▪ Decision Tree (의사결정 나무)▪ Logistic Regression
	회귀 (Regression)	<ul style="list-style-type: none">▪ Linear Regression (선형 회귀)

머신러닝의 분류(비지도 학습)

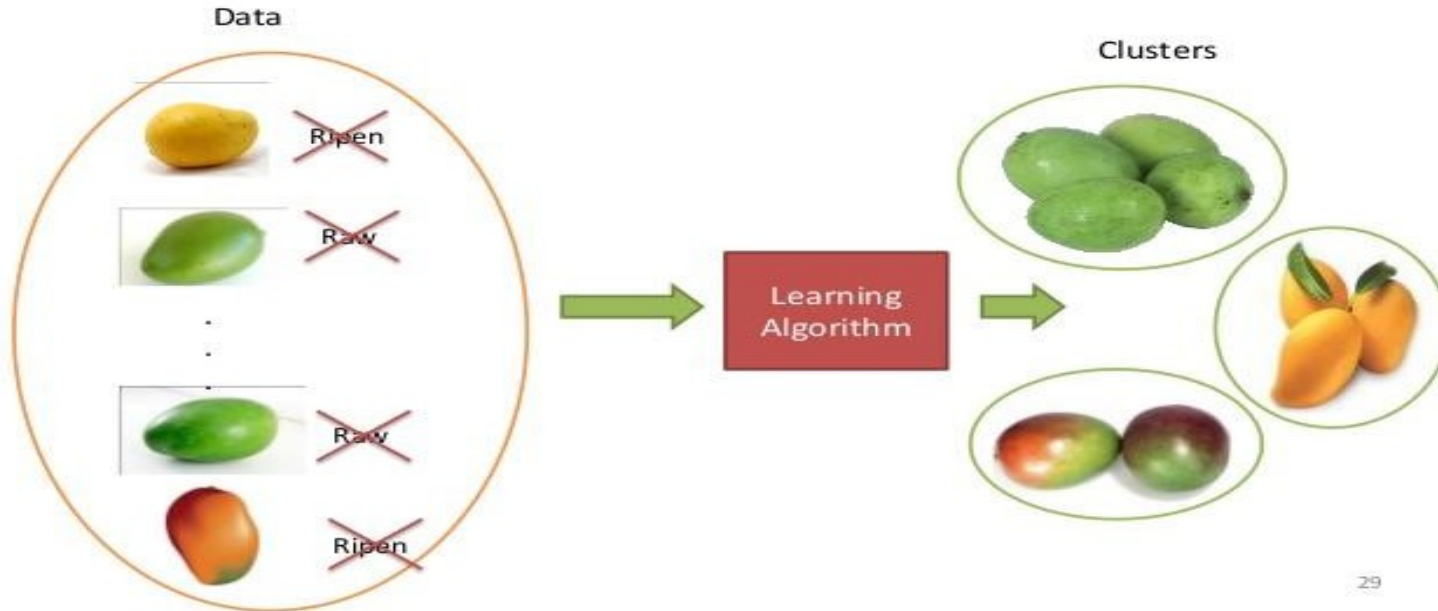


[비지도학습의 학습방식]

머신러닝의 분류(비지도 학습)

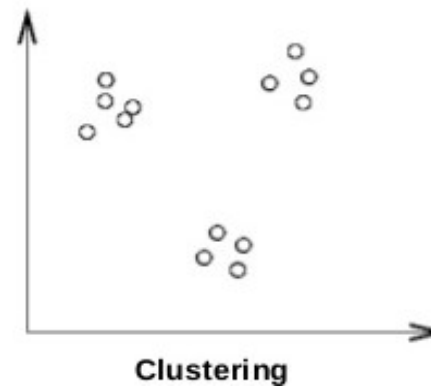
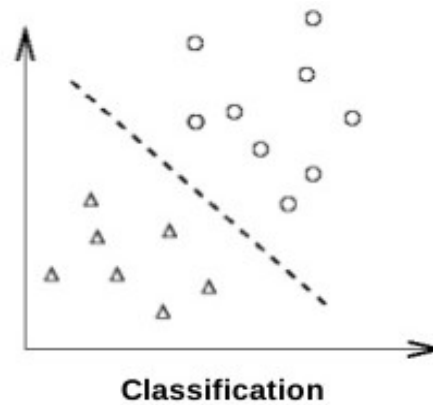
【비지도학습의 Training Set의 예】

Label이 없는 학습 데이터(Training Set)를 이용해서 학습.



분류와 군집의 비교

	분류 (Classification)	군집 (Clustering)
공통점	입력된 데이터들이 어떤 형태로 그룹을 형성하는지가 관심사	
차이점	레이블이 있다.	레이블이 없다. 예) 의학 임상실험 환자군 구별 예) 구매자 유형 분류



머신러닝의 분류(비지도 학습)

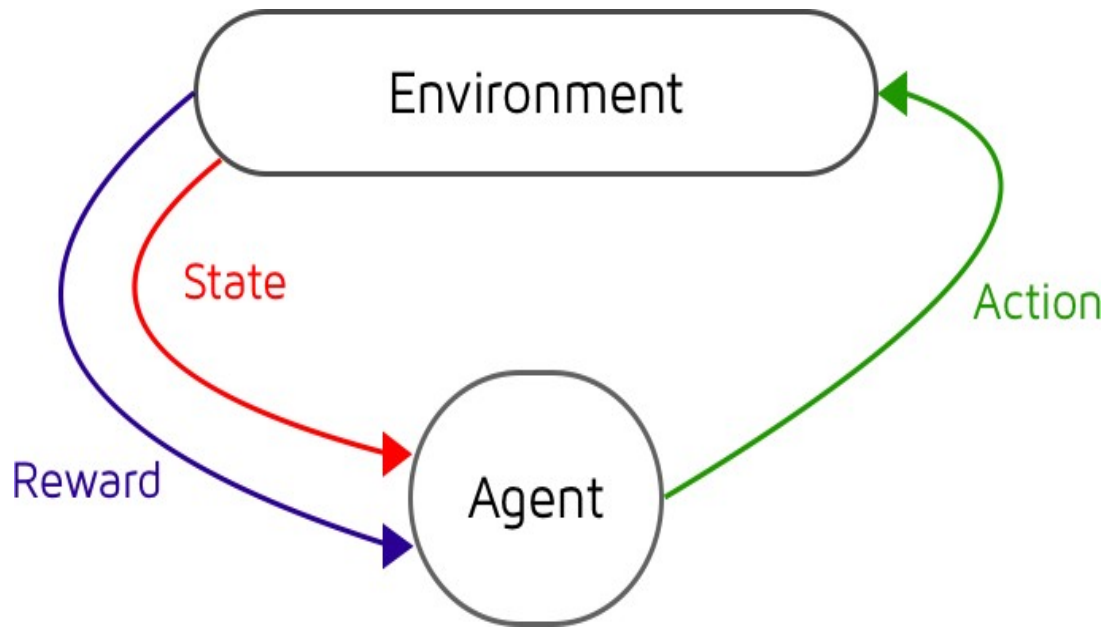
Types	Tasks	Algorithms
비지도학습 (Unsupervised Learning)	군집 (Clustering)	<ul style="list-style-type: none">▪ K-Means Clustering▪ DBSCAN Clustering▪ Hierarchical Clustering (계층형 군집)

머신러닝의 분류(강화 학습)

보상을 통해 상은 최대화, 벌은 최소화하는 방향으로 강화한다.

시행착오를 겪으며 경험이 쌓이고, 옳고 그른지 판단을 하게 됨.

머신러닝의 분류(강화 학습)



▪ 강화학습

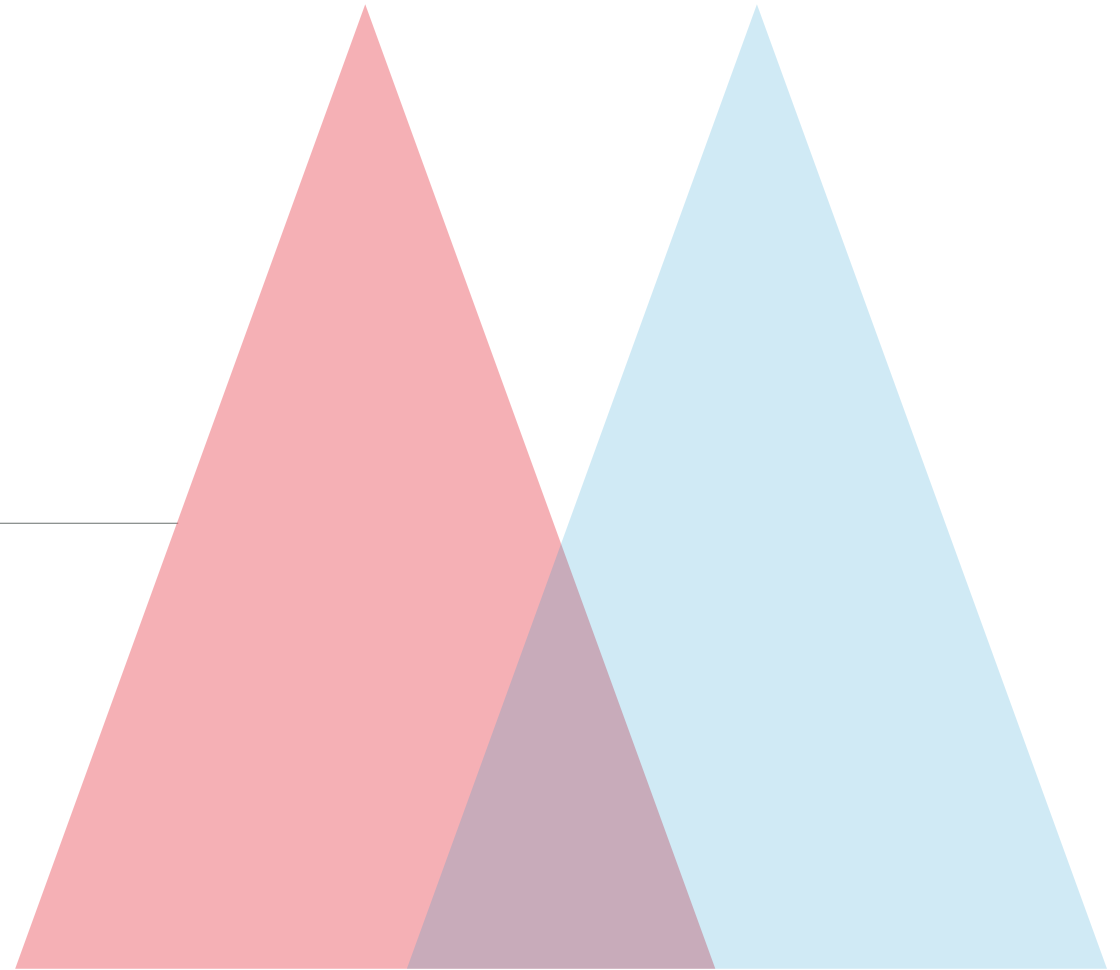
- ✓ 시행착오 과정을 거쳐 학습하기 때문에 사람의 학습방식과 유사
- ✓ Agent는 환경으로부터 상태를 관측하고 이에 따른 적절한 행동을 하면, 이 행동을 기준으로 환경으로부터 보상을 받는다.
- ✓ 관측 – 행동 – 보상의 상호작용을 반복하면서 환경으로부터 얻는 보상을 최대화하는 태스크를 수행하기 위한 일련의 과정.
- ✓ 관측 – 행동 – 보상의 과정을 경험(**Experience**)이라고도 한다.

머신러닝의 분류(강화 학습)

Types	Tasks	Algorithms
지도학습 (Supervised Learning)	분류 (Classification)	<ul style="list-style-type: none">▪ KNN : k Nearest Neighbor▪ SVM : Support Vector Machine▪ Decision Tree (의사결정 나무)▪ Logistic Regression
	예측 (Prediction)	<ul style="list-style-type: none">▪ Linear Regression (선형 회귀)
비지도학습 (Unsupervised Learning)	군집 (Clustering)	<ul style="list-style-type: none">▪ K-Means Clustering▪ DBSCAN Clustering▪ Hierarchical Clustering (계층형 군집)
강화학습 (Reinforcement Learning)		<ul style="list-style-type: none">▪ MDP : Markov Decision Process

002

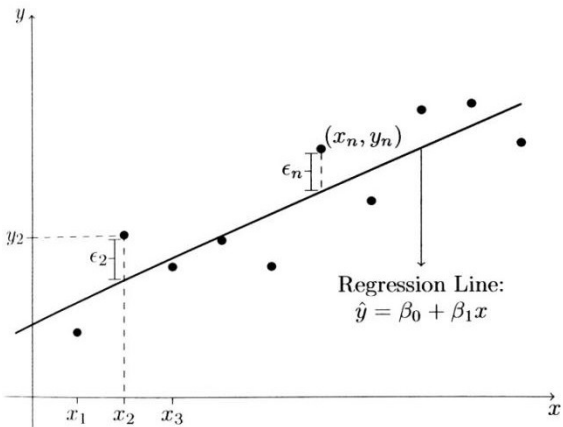
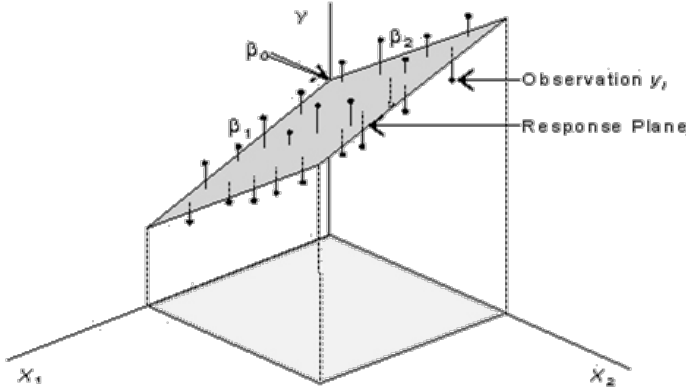
머신러닝 알고리즘



예측 문제

Linear Regression은 독립변수와 종속변수의 관계를 설명하여 예측 문제를 해결하는 데 사용된다.

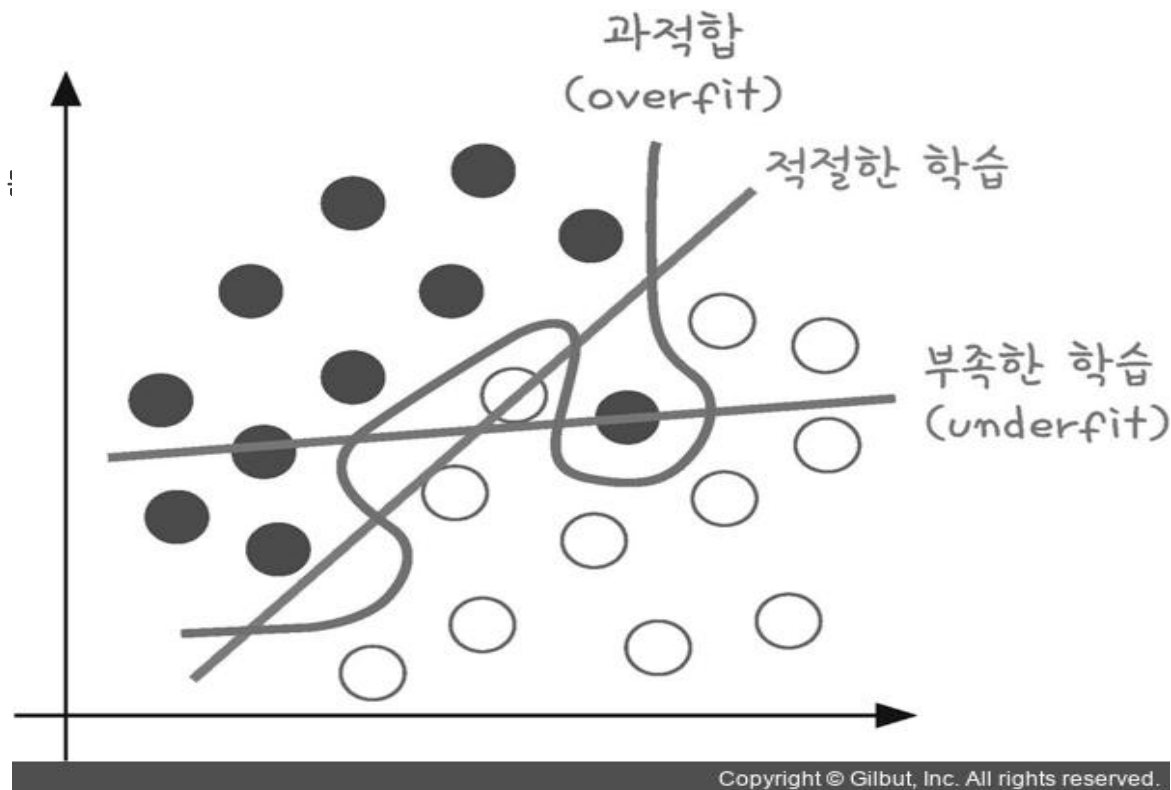
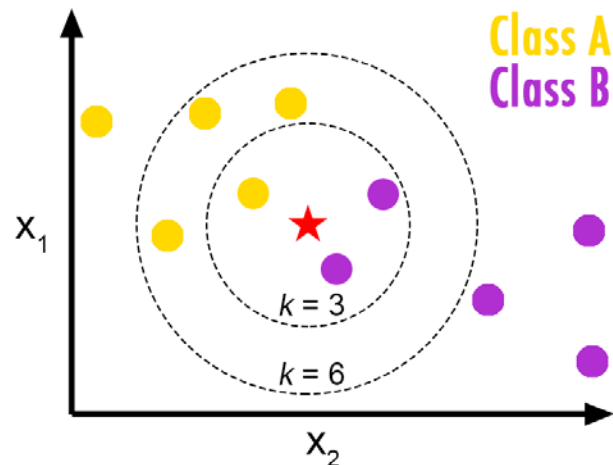
선형 회귀의 구분

단순 선형 회귀 (Simple Linear Regression)	다중 선형 회귀 (Multiple Linear Regression)
 <ul style="list-style-type: none"> 독립변수 x가 하나 예) 키와 몸무게의 상관관계 	<p>$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots$</p>  <ul style="list-style-type: none"> 독립변수 x가 하나이상 예) 수면시간, 운동시간, 라면 먹는 횟수가 몸무게에 미치는 영향

분류 문제

kNN은 새로운 데이터가 어느 그룹에 속하는지

▪ kNN (k-Nearest Neighbor) 원리



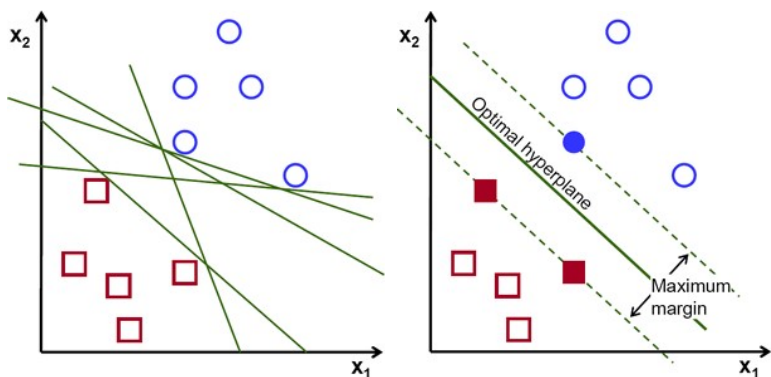
▪ k 값의 선정 (예)

언더피팅 (underfitting)	노멀피팅 (normal fitting)	오버피팅 (overfitting)
<p>Binary kNN Classification (k=1)</p> <p>K=1 학습 에러율 높음 검증 에러율 높음</p> <ul style="list-style-type: none"> 노이즈에 너무 민감하게 반응. 데이터를 판별할 수 있는 특성을 찾지 못함 	<p>Binary kNN Classification (k=5)</p> <p>K=9 학습 에러율 낮음 검증 에러율 낮음</p> <ul style="list-style-type: none"> 데이터의 특성을 잘 대표하도록 학습 됨 	<p>Binary kNN Classification (k=25)</p> <p>K=25 학습 에러율 낮음 검증 에러율 높음</p> <ul style="list-style-type: none"> 의사 결정 경계가 둔감하여 변별력 부족

분류 문제

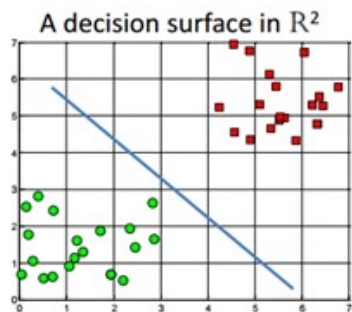
SVM은 두 범주를 갖는 데이터를 분류하는 방법으로 주어진 데이터들을 가능한 멀리 두 개의 집단으로 분리.

▪ SVM (Support Vector Machine) 원리

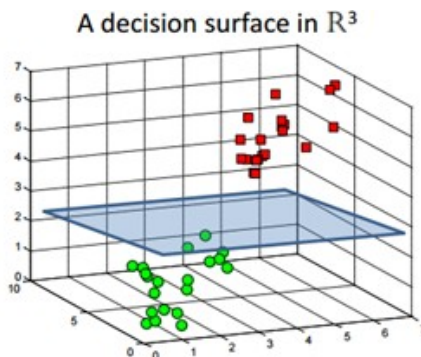


- **Hyperplane** : 데이터를 분류하는 선.
- Support Vector와 Margin을 통해 두 클래스 사이를 분류하는 **최적의 Hyperplane**을 구한다.
- Hard Margin 방법 : 매우 엄격하게 두 개의 그룹을 분리하는 경계식을 구하는 방법으로 몇 개의 노이즈가 있으면 사용이 어렵다.
- Soft Margin 방법 : Support Vector가 위치한 경계선에 약간의 여유 (Slack)을 두는 방식

▪ n차원 공간에서 hyperplane은 n-1차원



2차원 공간에서
hyperplane은 선

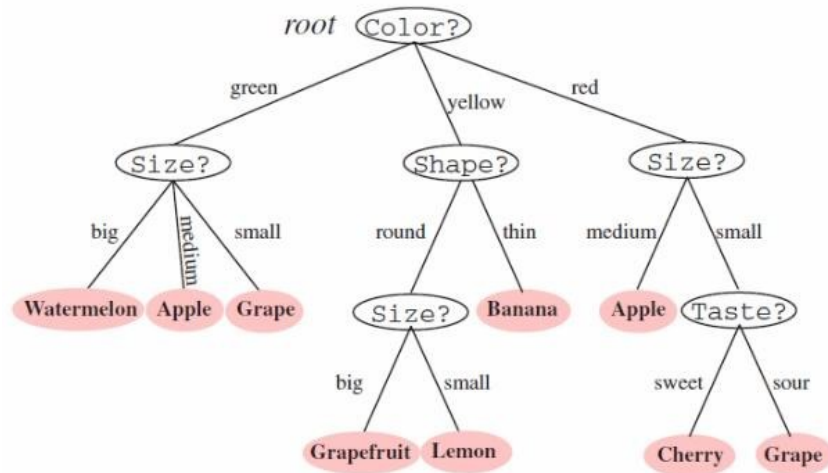


3차원 공간에서
hyperplane은 면

분류 문제

Decision Tree(의사결정 나무)는 학습 데이터를 이용하여 트리 모델을 생성 후 분류 및 예측을 한다.

Decision Tree 원리



level 0

- **Root Node** : 맨 상위의 Decision Node

level 1

- **Attribute** : green, red 같이 분기를 결정하는 값.
- Root → Branch → Leaf 순서로 하향식 의사 결정
- 키가 작고 가지가 별로 없는 모델이 신속한 의사 결정이 됨으로 좋은 모델이 됨.

level 2

- 장점 : 학습 모델 이해가 쉽다.

level 3

단점 : 연속적인 속성의 처리의 문제가 있다.

? Pruning

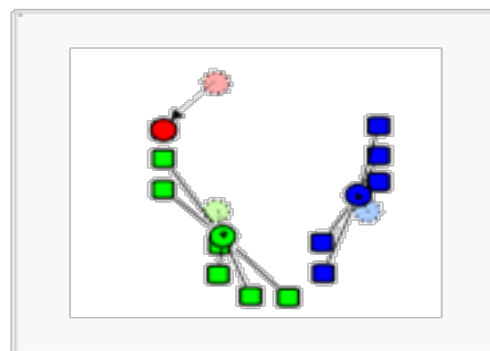
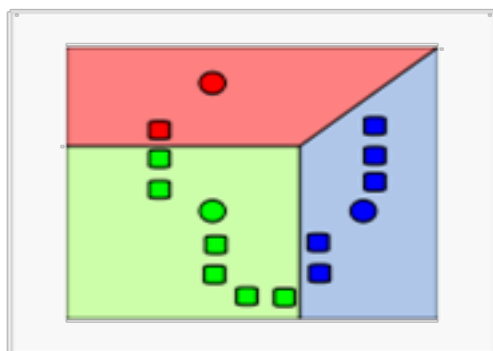
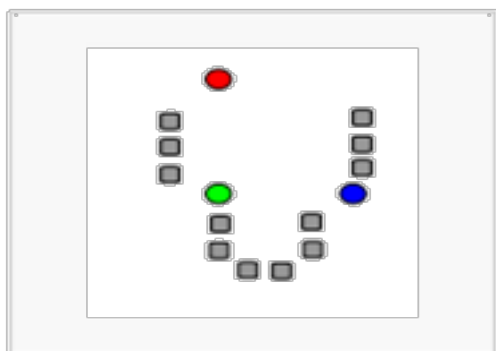
- ID3 알고리즘 기반으로 의사결정 트리 모델을 완성 한 후
- 의사결정 노드에 있는 줄기를 효율적으로 제거하고 합치는 기법



군집 문제

K-means clustering은 중심값을 선정하고, 중심값과 다른 데이터 간의 거리를 이용하여 분류를 수행하는 비지도 학습 (군집)

▪ K-means clustering 알고리즘 수행 절차



Step 1

- Cluster 수인 k를 정의
- 초기 k개 군집 중심 임의 지정 (initial centroid)
- 위 그림에서는 k=3

Step 2

- 모든 데이터들의 거리 계산 후 가장 가까운 Centroid로 Clustering

Step 3

- 각 Cluster마다 계산하여 새로운 중심 계산

Step 4

- Step 2, 3 을 반복
- 데이터가 자신이 속하는 Cluster를 변경하지 않으면 학습 완료

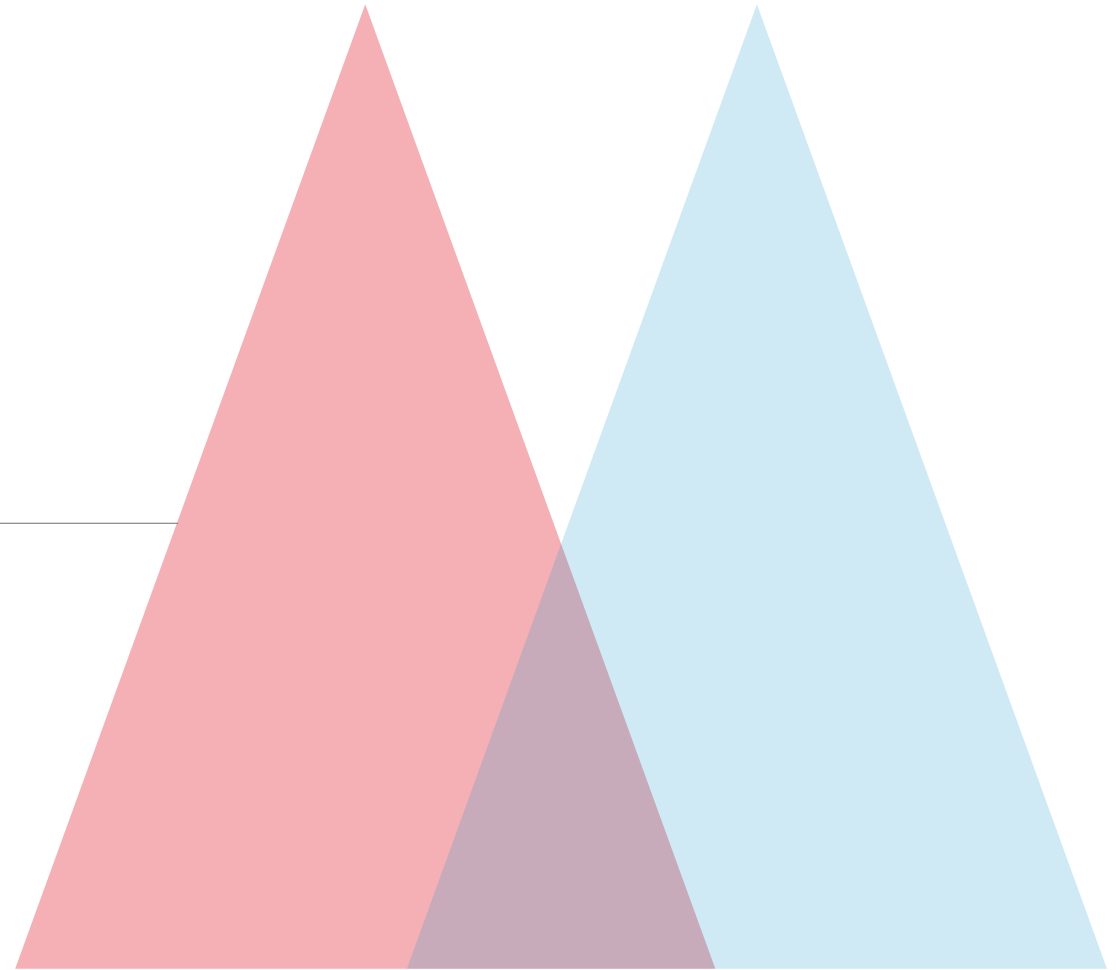
강화 학습

읽을만한 자료

<https://brunch.co.kr/@kakao-it/73>

003

딤러닝

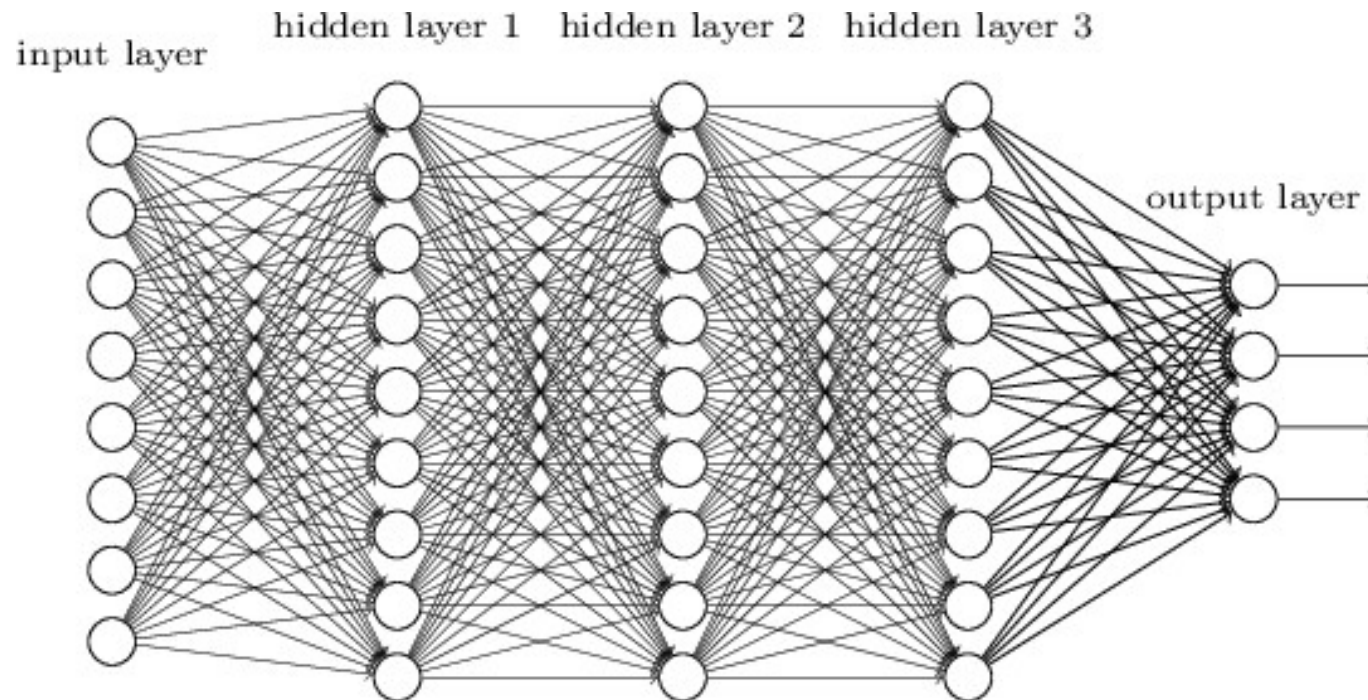


딥러닝이란?

컴퓨터가 스스로 학습할 수 있게 하기 위해 인공 신경망을 기반으로 하는 기계 학습 기술

▪ 딥러닝이란?

인간의 신경망(Neural Network) 이론을 이용한 인공 신경망 (ANN, Artificial Neural Network)의 일종으로, 계층 구조 (Layer Structure)로 구성되면서 입력층(Input Layer)과 출력층(Output Layer) 사이에 하나 이상의 은닉층(Hidden Layer)을 가지고 있는 심층 신경망(DNN, Deep Neural Network)이다.



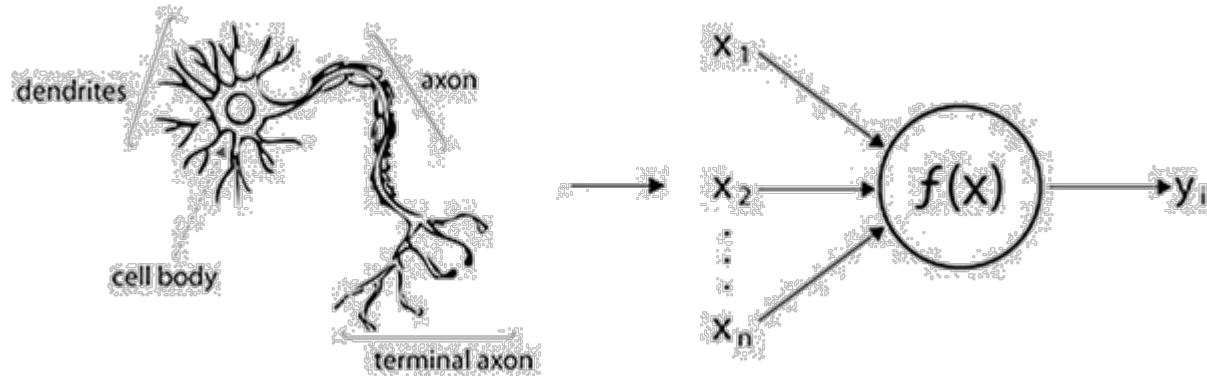
인공신경망

인공신경망(ANN)은 인간의 뇌 구조를 모방하여 모델링 한 수학적 모델이다.

■ 신경세포 (Neuron, 뉴런)

Neuron의 입력은 다수이고 출력은 하나이며, 여러 신경세포로부터 전달되어 온 신호들은 합산되어 출력된다.

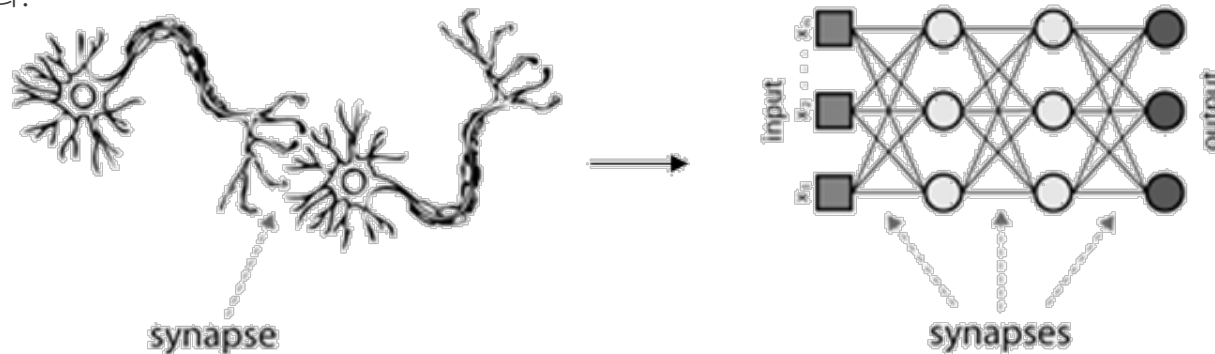
합산된 값이 설정 값(Threshold) 이상이면 출력 신호가 생기고 이하이면 출력 신호가 없다.



생물 신경망	인공 신경망
세포체 (Cell Body)	노드 (Node)
수상돌기 (Dendrites)	입력 (Input)
축삭 (Axon)	출력 (Output)

■ 연결 : Synapse - Weight

다수의 Neuron이 연결되어 의미 있는 작업을 하듯, 인공신경망의 경우도 노드들을 연결시켜 Layer를 만들고 연결 강도는 가중치로 처리된다.



생물 신경망	인공 신경망
Synapse	가중치 (Weight)

딥러닝 성과

이미지 인식 분야에서 성과 (MNIST)

- **MNIST**(Mixed National Institute of Standards and Technology) 란?
 - 기계학습 분야에서 사용되는 손글씨 숫자 이미지 데이터 셋
 - 0부터 9까지 숫자 이미지로 구성되어 있다
 - 각 이미지가 실제 의미하는 숫자가 Label되어 있다.
 - 훈련 이미지가 6만장, 시험 이미지가 1만장으로 준비되어 있다.

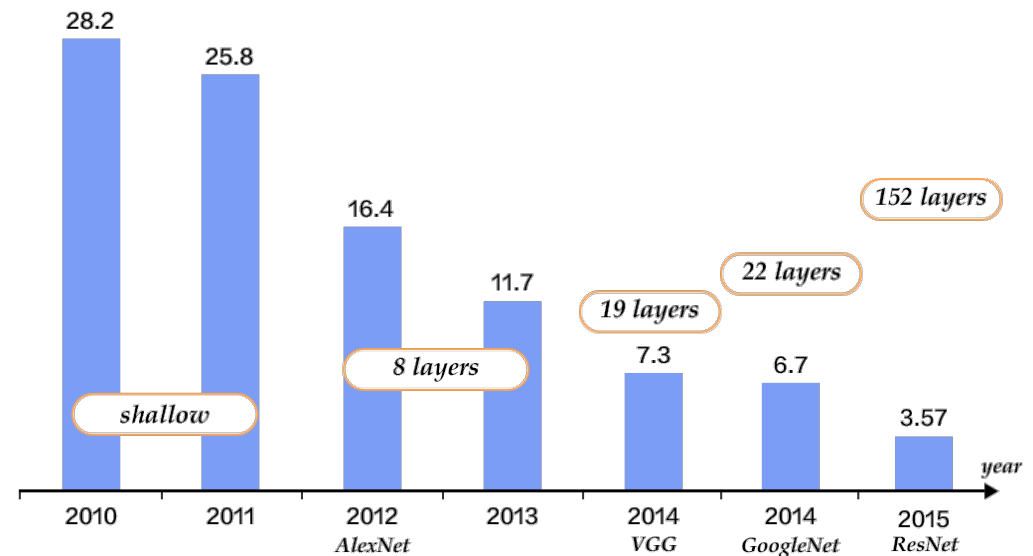


딥러닝 성과

이미지 인식 분야의 성과 (ILSVRC)

▪ ILSVRC (ImageNet Large Scale Visual Recognition Challenge)

- ImageNet이 주관하는 이미지 인식 콘테스트 (매년)
 - ✓ 이미지넷이 보유한 레이블된 이미지를 학습데이터로 하여 1,000가지 종류로 분류하는 대회
- 2012년 CNN을 적용한 팀이 우승 후 2013년 부터 딥러닝을 이용해 참가한 팀이 급증
- 기존의 이미지 인식 기법인 BoF(Bag of Features)와 비교해 정밀도가 압도적으로 높아짐



딥러닝 성과

딥러닝을 이용한 예술 - 그림 그리기

- A neural algorithm of artistic style (2015) - <https://arxiv.org/abs/1508.06576>
 - 독일 튀빙겐 대학교 연구팀에서 딥러닝(CNN)을 이용해 유명 화가의 화풍을 학습한 후 이를 이용하여 그림을 그린다.



윌리엄 터너
'수송선의 난파'



에드바르 뭉크
'절규'



참고 문헌

