Machine Learning

LG 인화원 교육
윤세영
KAIST 김재철AI대학원

KAIST

Important References

Stanford CS231n course
http://cs231n.stanford.edu/index.html
Lecture slides, Youtube video,

Coursera Deep Learning course by Andrew Ng
https://www.deeplearning.ai
Not free if you want to get certifications

PyTorch Deep Learning Mini Course
https://github.com/Atcold/PyTorch-Deep-Learning-Minicourse

Many source codes in Github

목차

인공지능 Artificial Intelligent (AI)

Pieter Abbeel & Dan Klein CS188 Lecture note (UC Berkeley)

Quiz

아래 세가지 기술 분류의 차이를 설명하시오

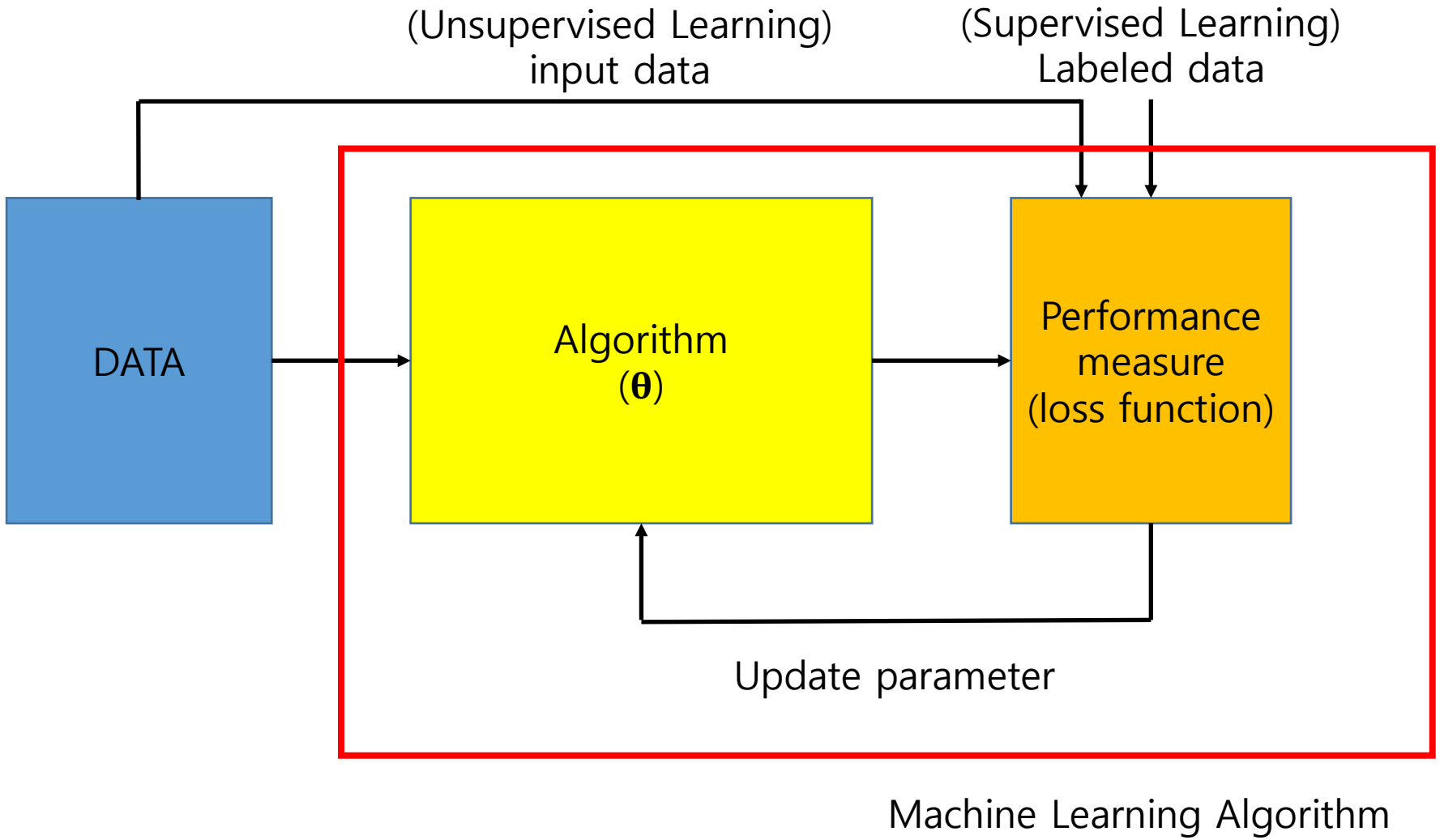
- Artificial Intelligent (AI)
- Machine Learning (ML)
- Deep Learning (DL)

정답

- Artificial Intelligent (AI)
  - 인간의 지능적 행위를 기계가 행하는 것 (계획, 언어 소통, 객체 인식, 음성 인식, 문제 풀이)
  - General AI: 사람처럼 지능을 가지고 사람이 하는 것 모든 것을 할 수 있음
  - Narrow AI: 특정 task에 대한 지능을 가지도록 만든 기계 (예: 이미지 인식은 가능하지만 음성 인식은 불가능)
- Machine Learning (ML)
  - AI를 달성하기 위한 방법. 하지만 AI가 ML을 반드시 이용하는 것은 아니다.
  - 데이터를 통하여 알고리즘을 학습 시켜서 성능을 향상 시키는 과정
- Deep Learning (DL)
  - Machine Learning 의 한 분류
  - 인간의 신경망에서 영감을 받은 알고리즘. Artificial Neural Network (ANN) 이라는 이름으로도 불린다.
  - 계층적 구조로 이루어져있으며, 일반적으로 계층의 수가 많아서 (구조가 깊어서) Deep Learning이라고 불림

목차

Machine Learing



Supervised Learning

기계학습: 주어진 데이터를 이용하여 parameter들을 학습한다. (edge weight, bias 값들을 학습)

- 학습의 방향은 목적함수를 최적화
- 목적함수는 기계학습의 종류에 따라 다양하게 설정 가능

기계학습은 크게 Regression 과 Classification 문제로 구분 가능함

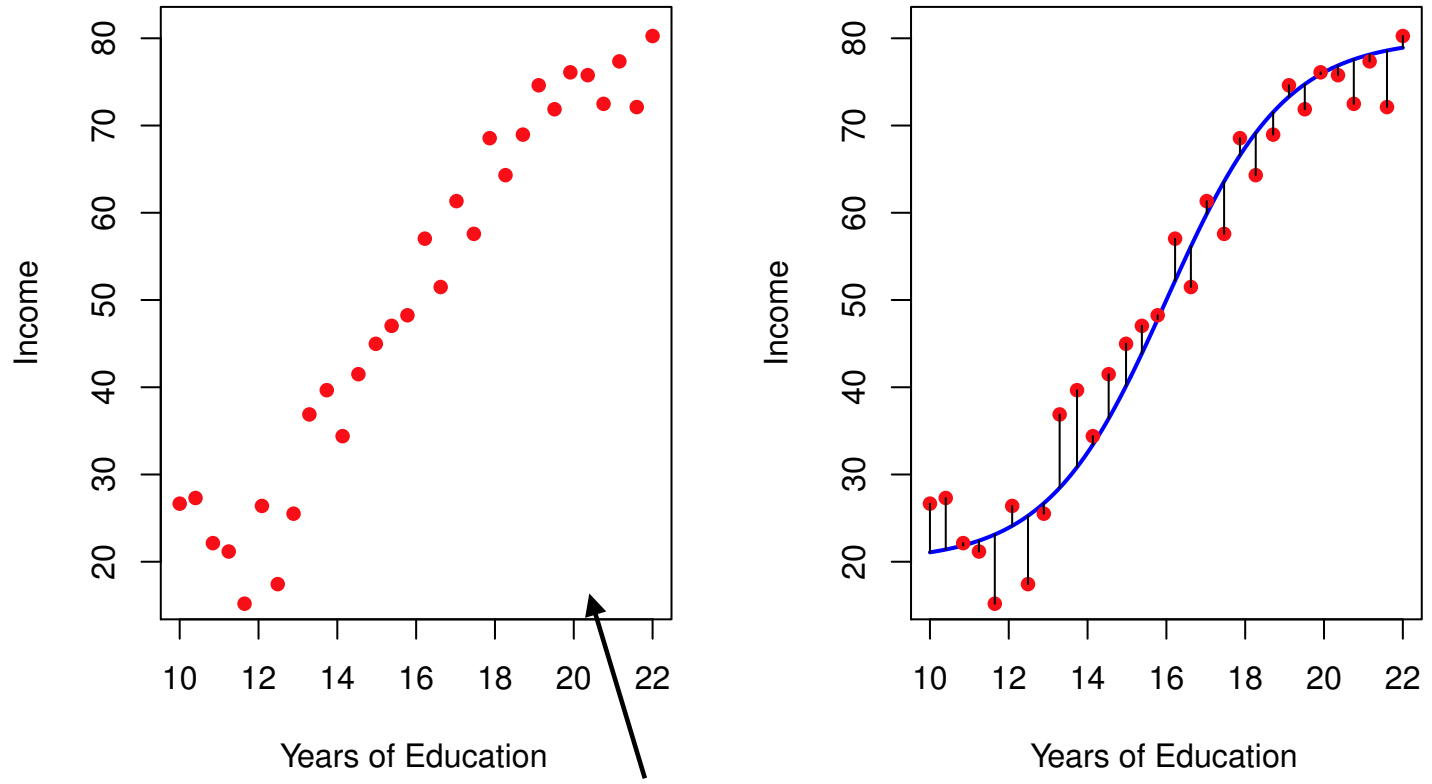
- Regression: 최대한 정확하게 입력에 대응되는 출력 값을 예측 (예: 집값 예측)
- Classification: 입력으로 들어오는 데이터들을 정확하게 분류 (예: 고양이 개 사진 구분)

가장 대표적인 목적 함수

- Regression: MSE, L1
- Classification: Cross Entropy (with softmax)

$$L(W) = \frac{1}{N}\sum_{i=1}^{N} L_i(x_i, y_i, W) + \lambda R(W)$$

Regression



Training data

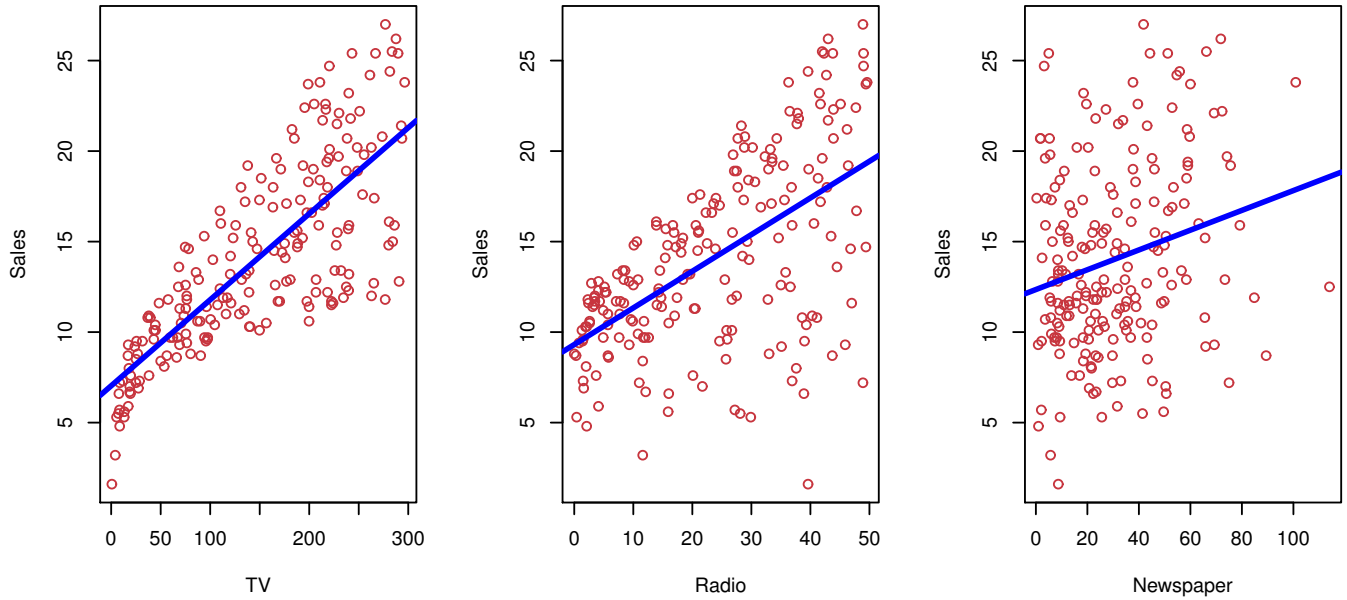
- Training data 를 잘 설명하는 함수를 구한다
- Q: 어떤 함수??

Parametric Statistics

Parametric Statistics

- 데이터가 몇개의 parameter로 정의된 특정 분포 그룹에서 발생하였다고 가정하고, 데이터의 값에 따라 paramter 값을 학습
- 예: linear regression (선형 계획법) 은 parameter들과 함께 다음과 같은 모델을 가정한다.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$



Parameter Learning

Define loss function

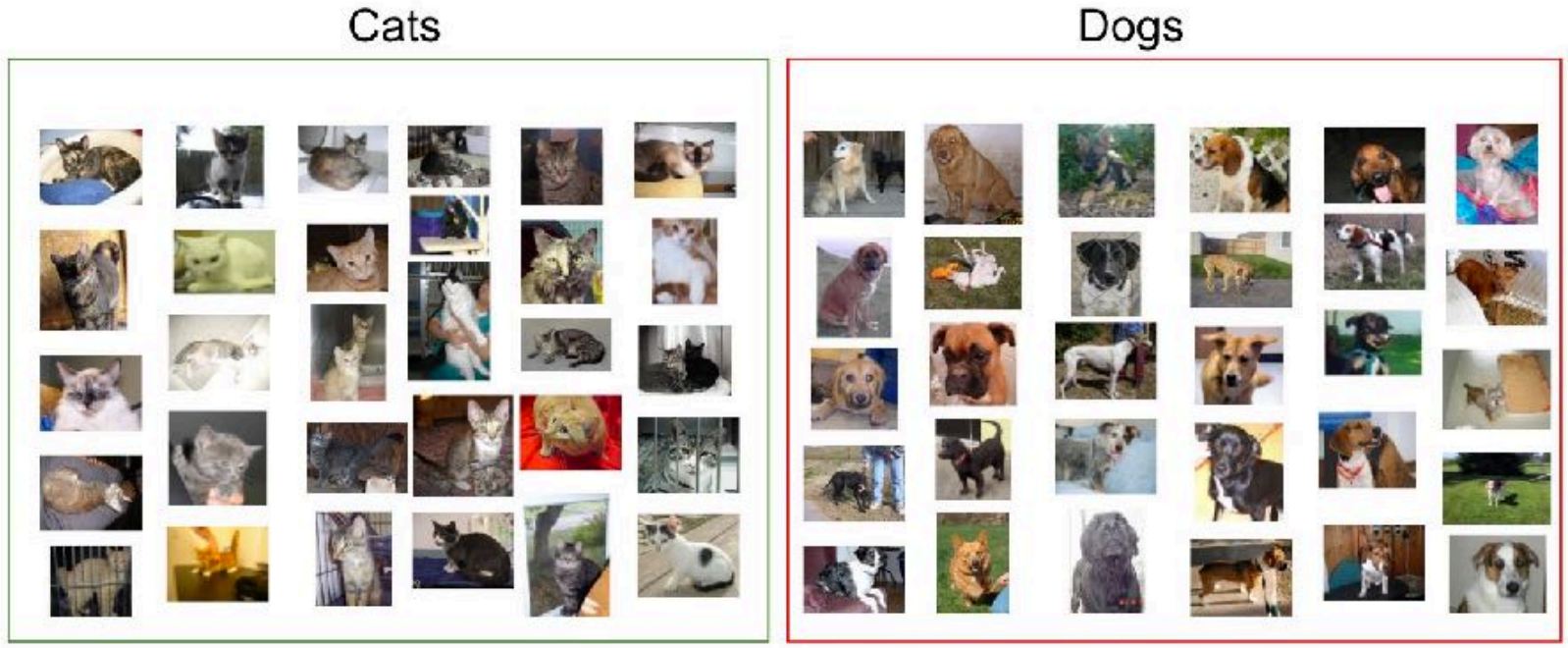
- Residual (error): $e_i = Y_i - \widehat{Y}_i$
- Residual Sum of Squares (RSS)

$$\begin{aligned}\text{RSS} =& e_1^2 + e_2^2 + \cdots + e_n^2 \\ =& \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2\end{aligned}$$

- How to find parameters?
  - Optimize RSS

Classification

정확하게 Data를 분류



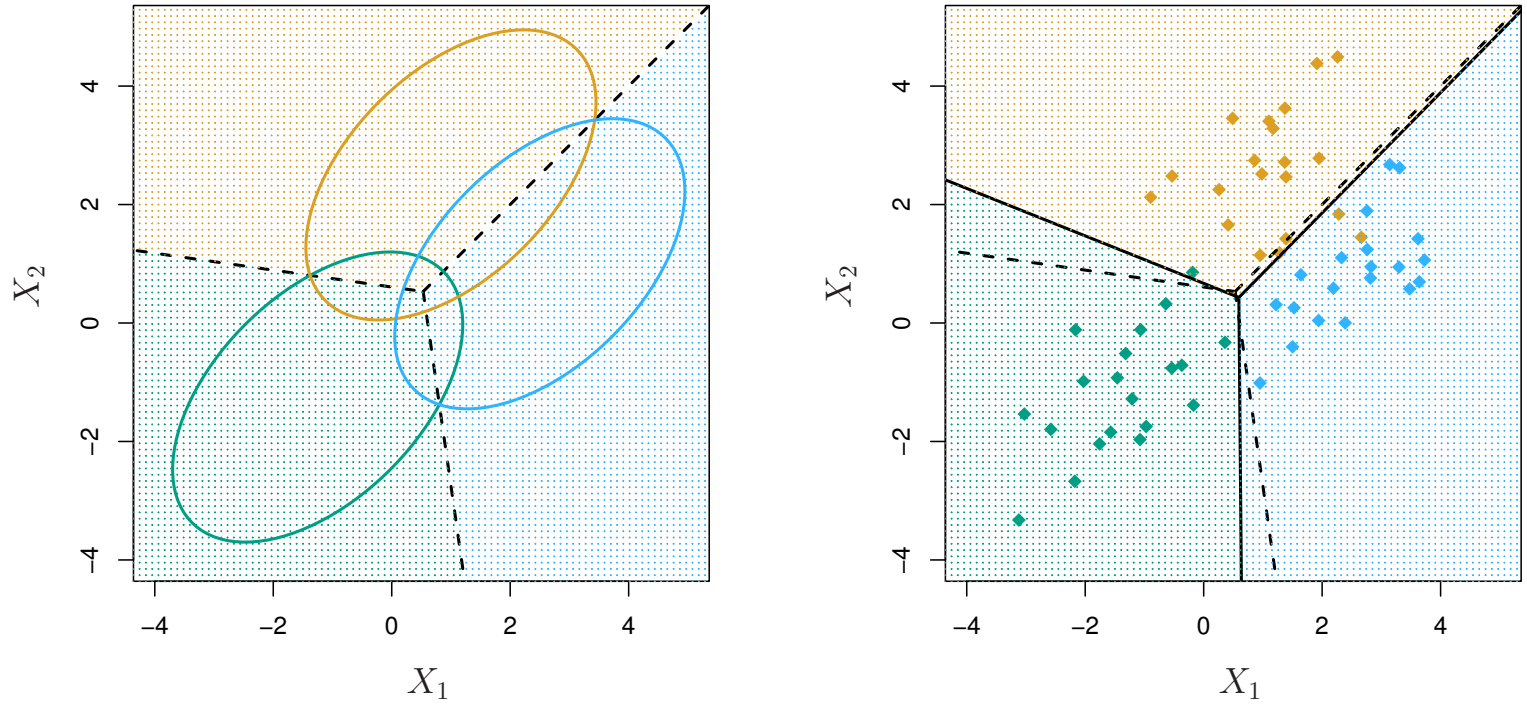
Sample of cats & dogs images from Kaggle Dataset

? 0 = dog
1 = cat

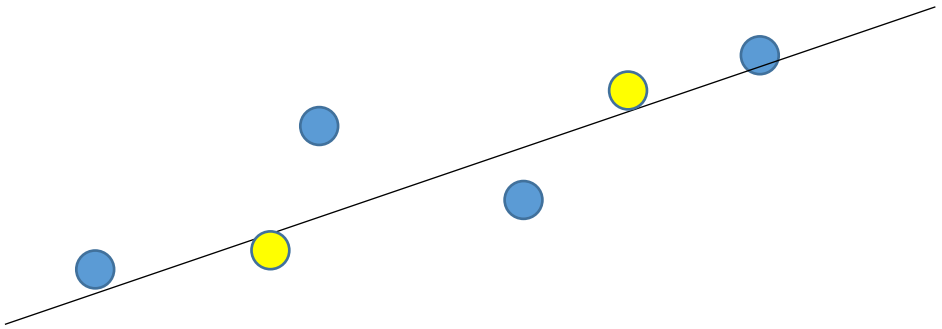
Parametric Learning

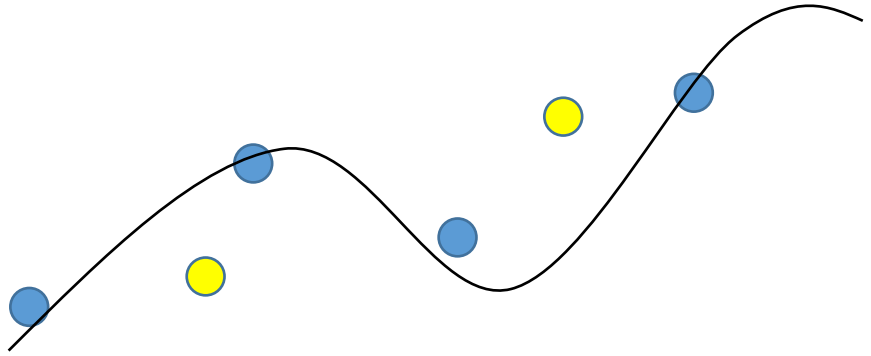
Parametric Statistics

- 데이터를 특정 데이터 분포 모델로 가정하고 해당 parameter들을 학습
- 예: Linear Discriminant Analysis (LDA) 는 다음과 같은 선형 decision boundary를 형성



Model Selection



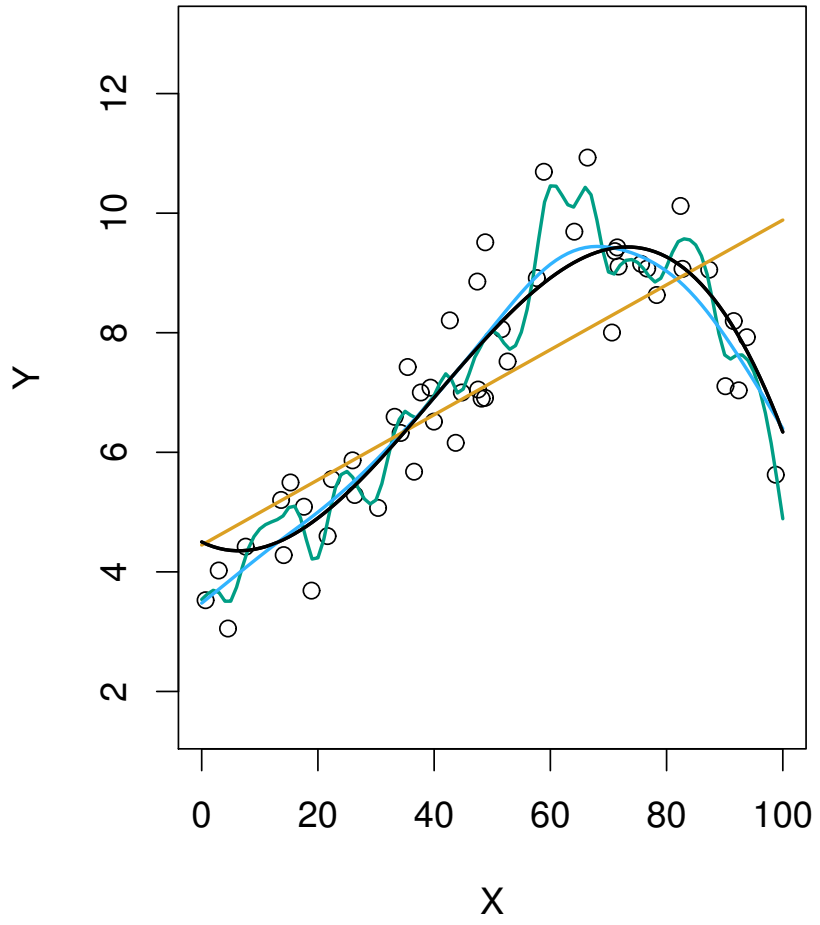


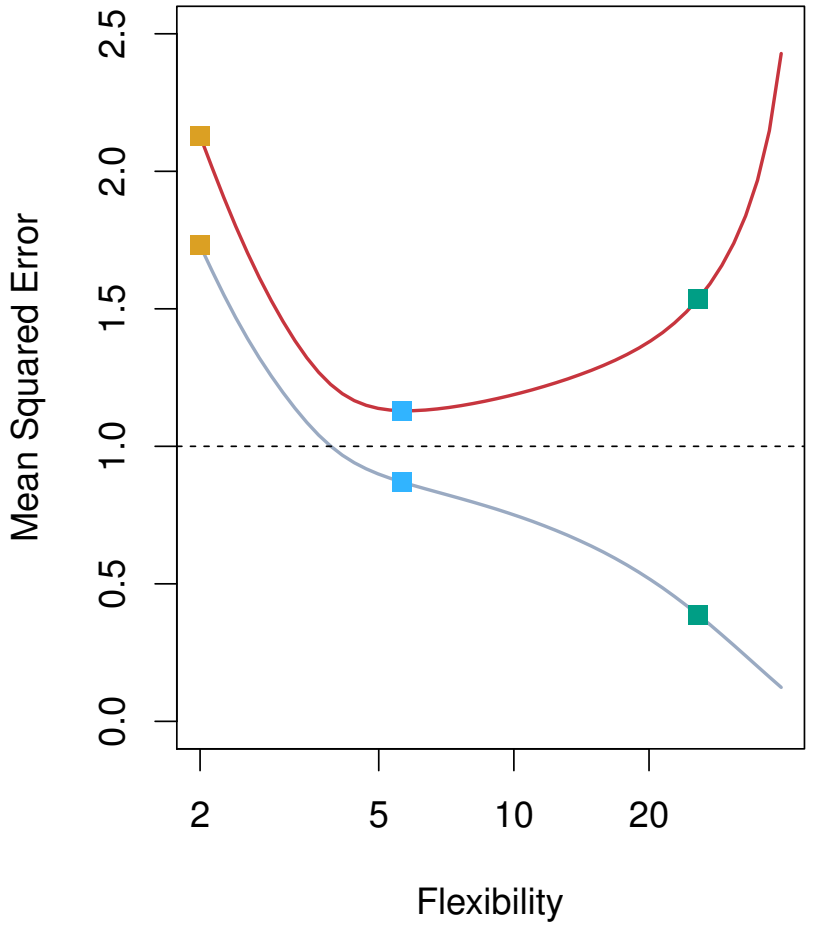
수 많은 함수 형태 중 어떤 함수를 선택해야 하는 것인가?

- 선형 함수 <- 간단하다, 유연성이 적다
- 매우 복잡한 함수 <-복잡하다, 유연성이 크다

Bias-Variance Tradeoff

복잡한 알고리즘은 Training Data를 잘 설명하지만, Overfitting 문제가 존재함





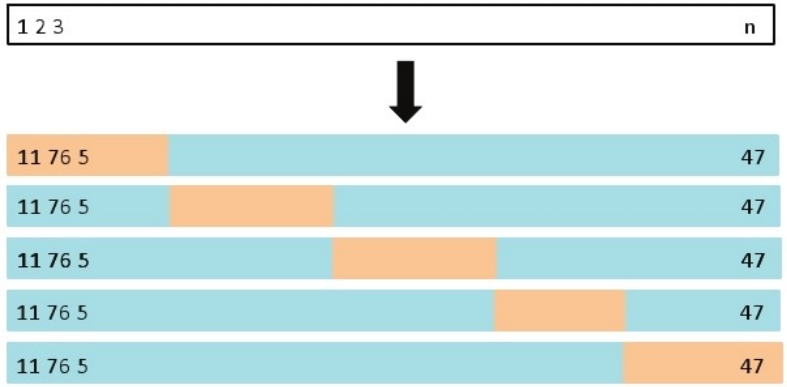
How to Select Model

Training data vs. Test data (cross-validation)

Parameter를 찾는데 사용

학습된 모델을 test

(optional) k-fold cross validation: 많이 사용 되었으나 Deep Learning 에서는 많이 사용 안함



Quiz

Q1. Supervised learning과 unsupervised learning은 어떤 차이가 있나요?

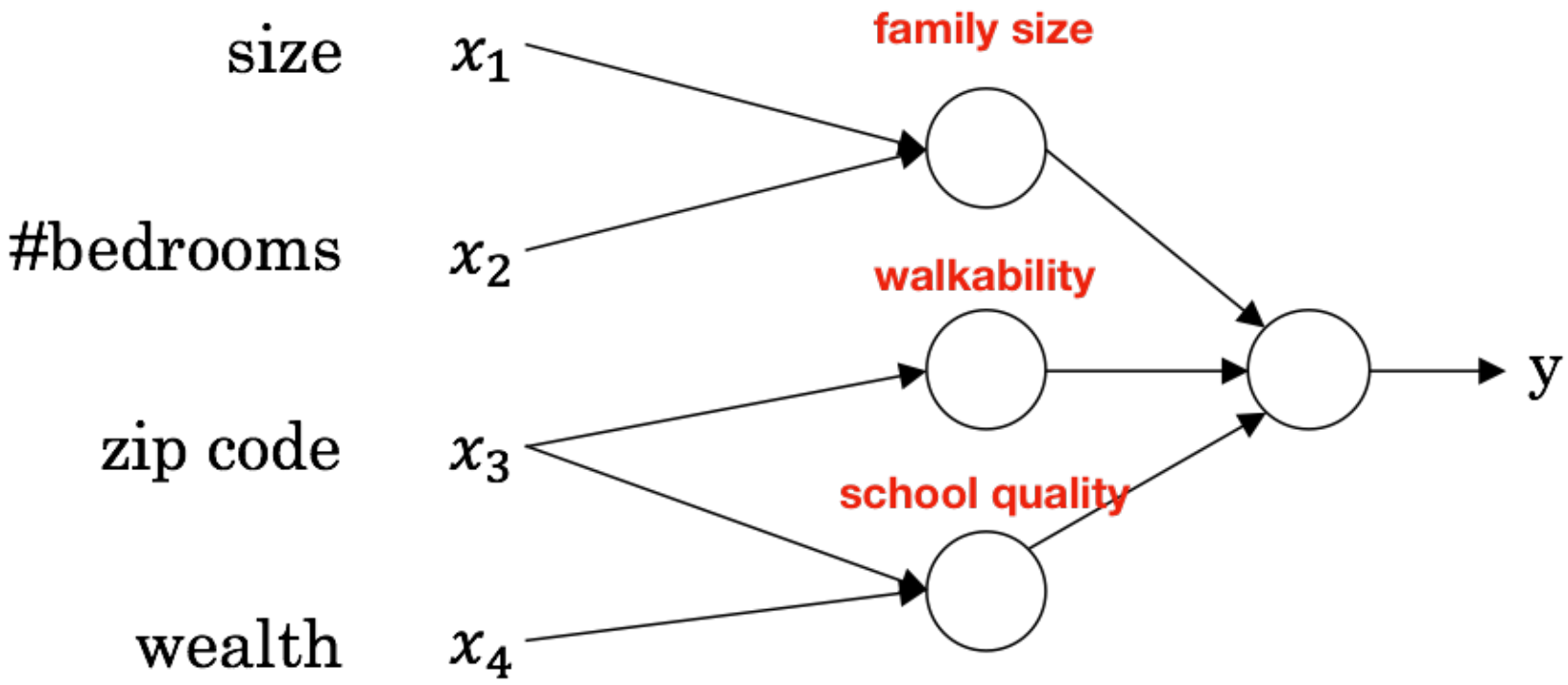
Q2. Overfitting은 무엇인가요?

Q3. 모델은 어떻게 선택 할 수 있을까요?

목차

예제) 집값 예측

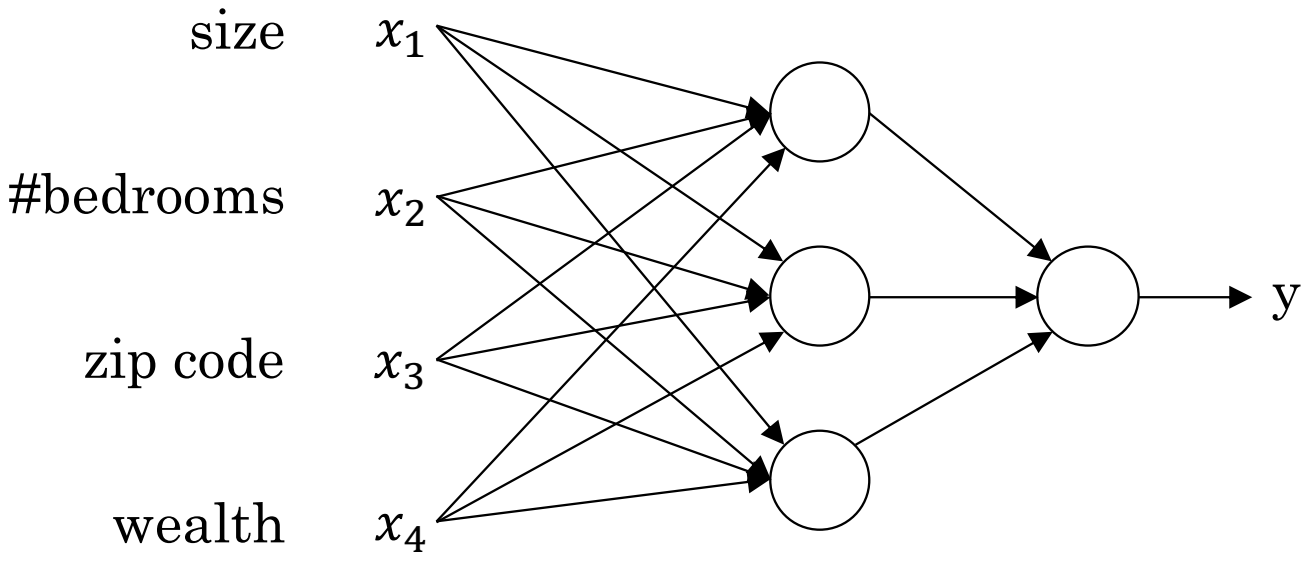
주어진 정보를 바탕으로 집값을 예측하자



Rule-based system: 전문가가 필요하다

예제) 집값 예측

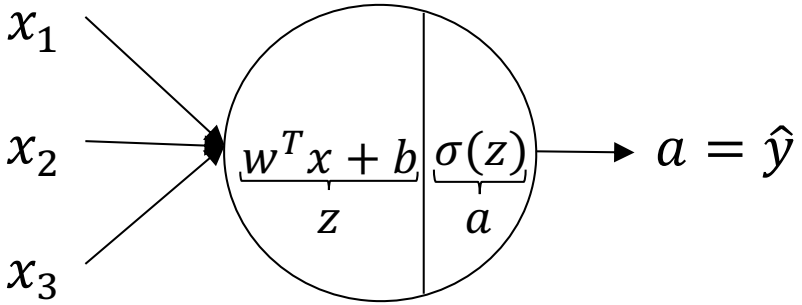
Neural Net?

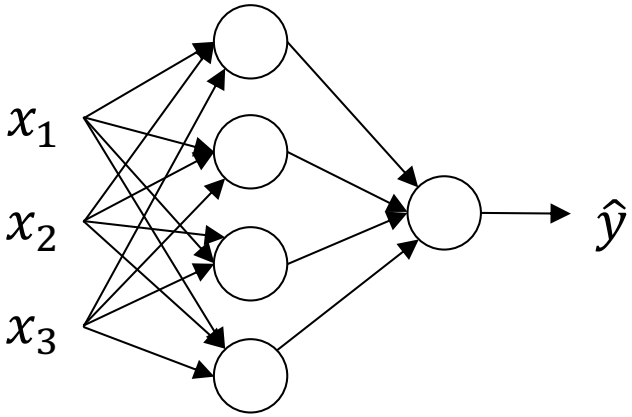


Machine Learning and Neural Net

- 어떠한 사전 지식 없이, 주어진 데이터를 기반으로
- 계층적인 추론 구조(Neural Net)의 연산식을 획득(Machine Learning)

원과 선의 의미





$$z = w^T x + b$$

$$a = \sigma(z)$$

선: edge, arrow,..

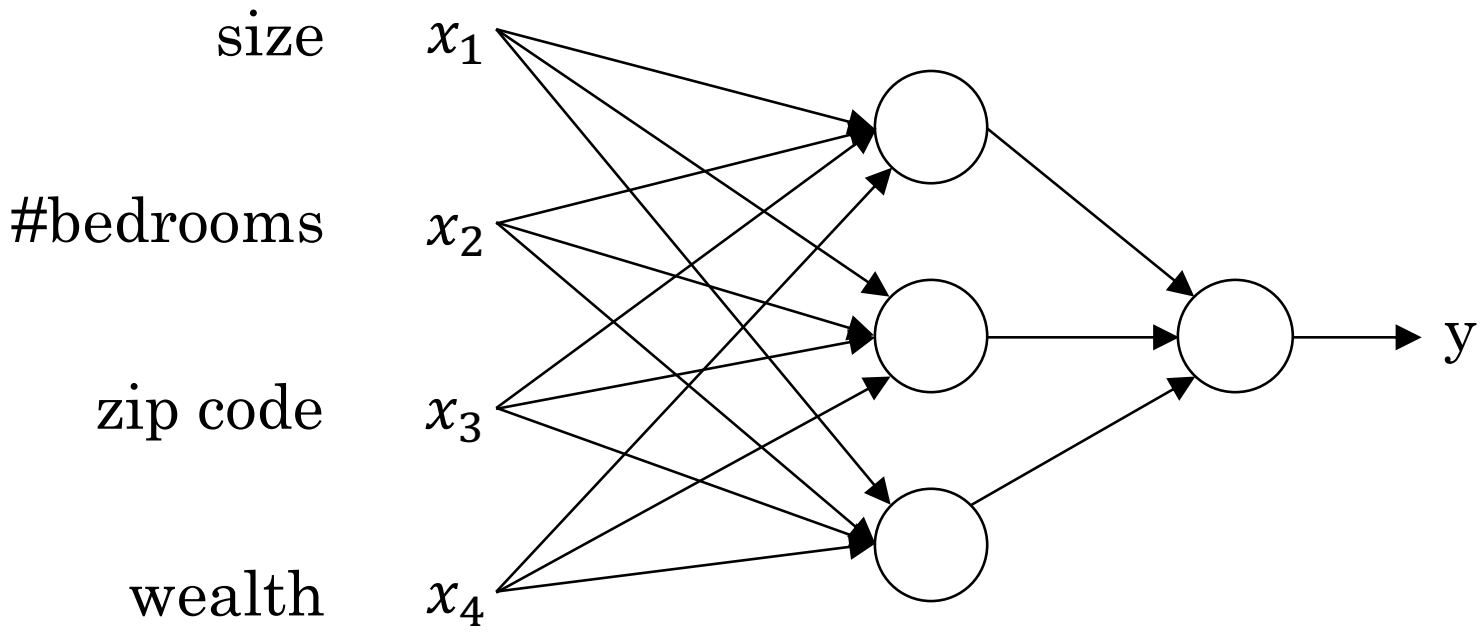
- 각 선은 각각의 값을 가지고 있으며 화살표 방향으로 흘러오는 정보에 해당 값을 곱함

원:Neuron

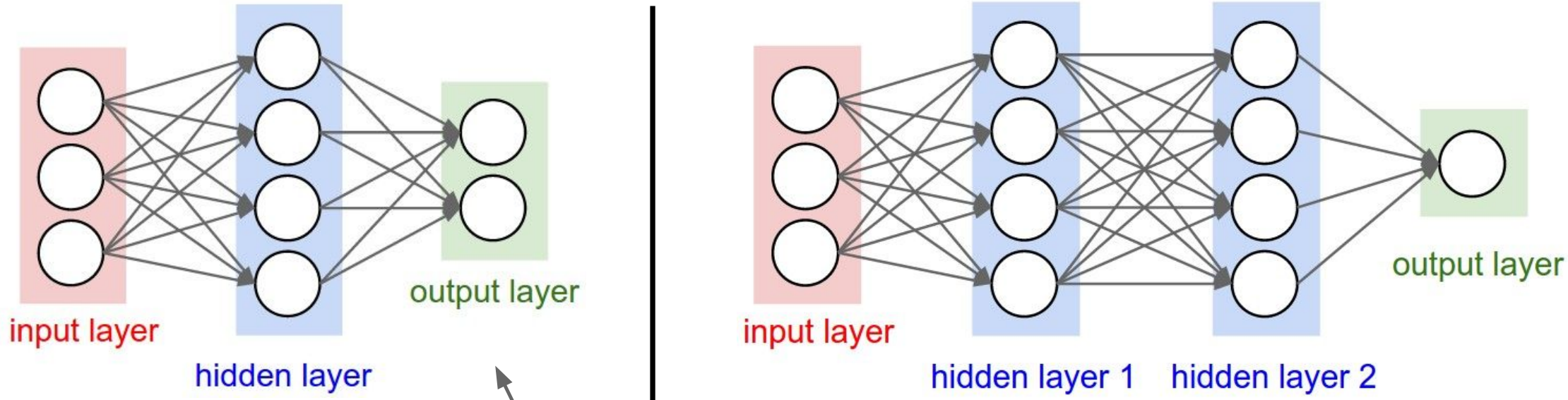
- 흘러 들어오는 정보를 모두 더한 후 non-linear activation function에 통과시킴

연습문제

다음 neural network의 함수식을 표현하시오.



Neural Networks

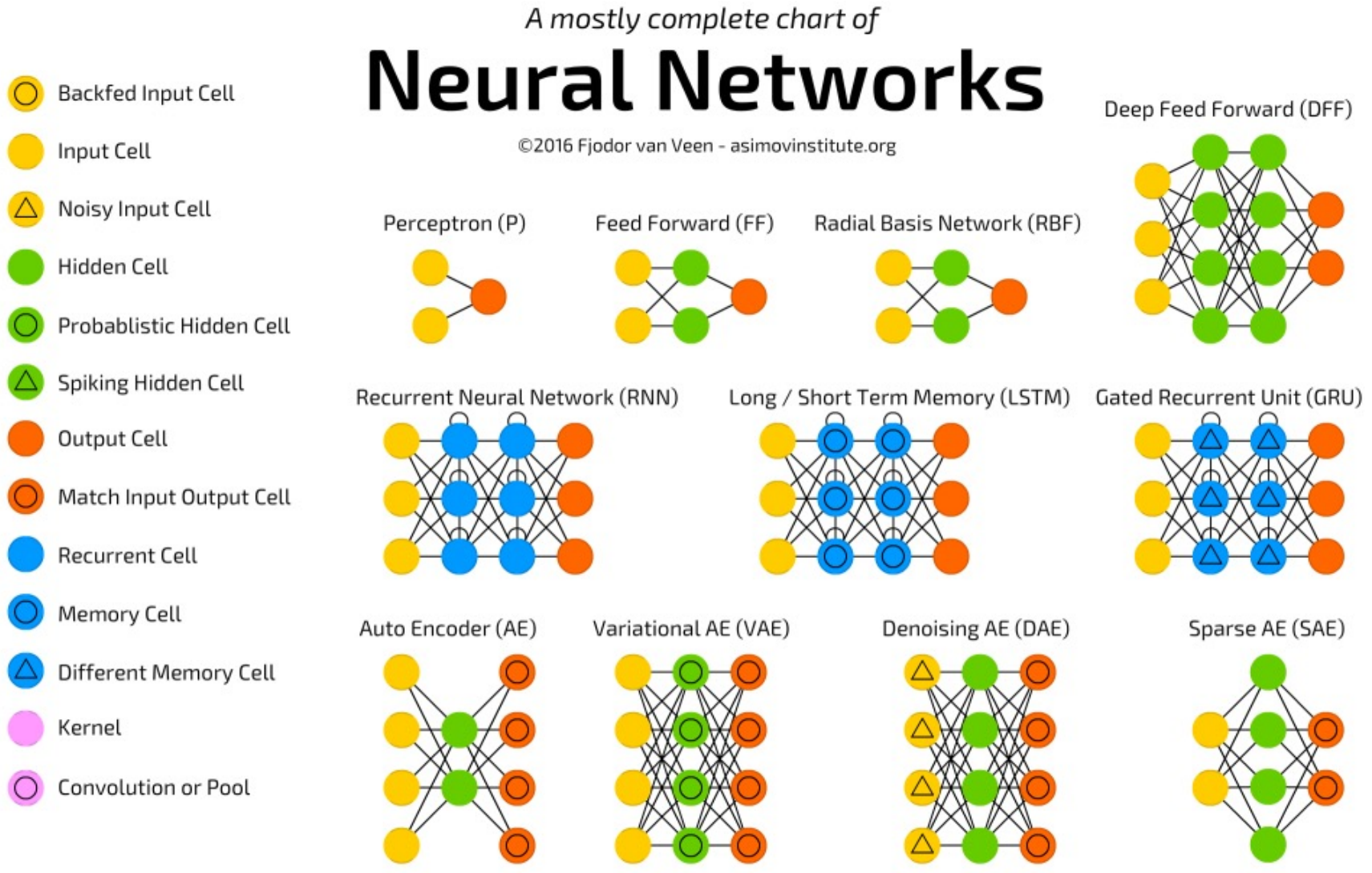


"2-layer Neural Net", or
"1-hidden-layer Neural Net"

"3-layer Neural Net", or
"2-hidden-layer Neural Net"

"Fully-connected" layers

Neural Networks

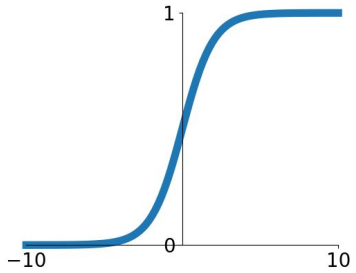


https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/neural_networks.html

Activation 함수

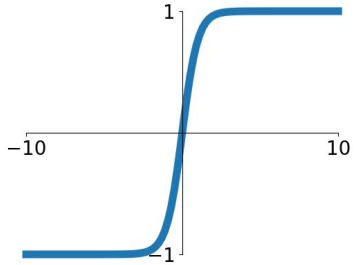
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



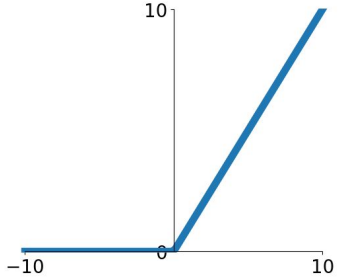
tanh

$$\tanh(x)$$



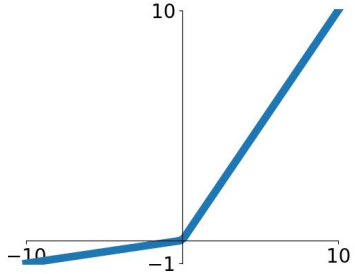
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

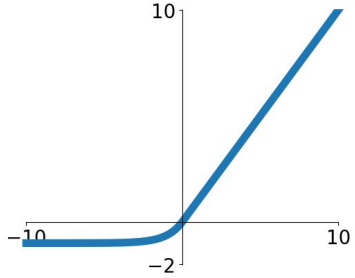


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

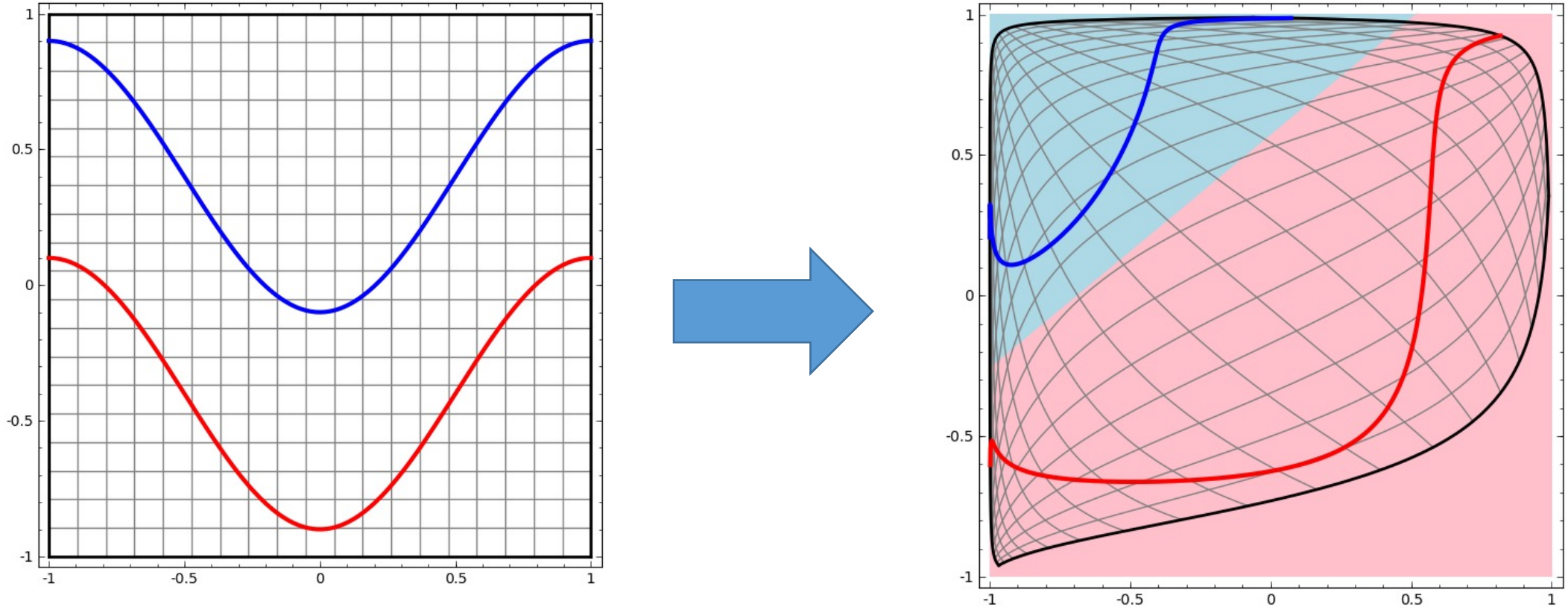
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Li et al, CS231 Lecture note

왜 Activation 함수는 비선형이어야 할까요?

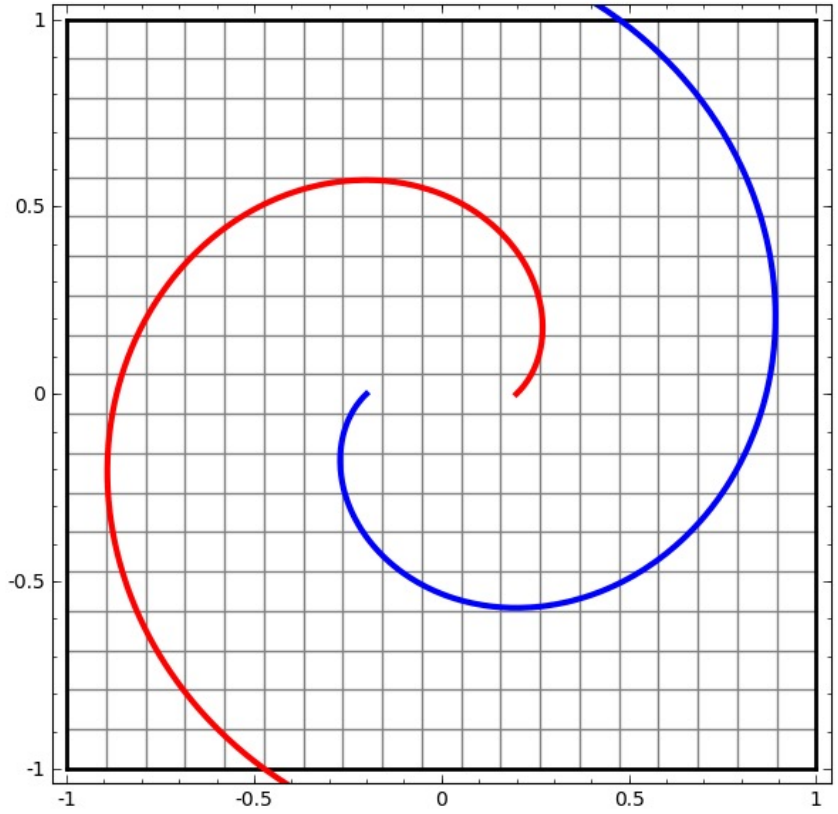
Multi-layered Neural Networks 의 깊이 (depth) 의미



http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

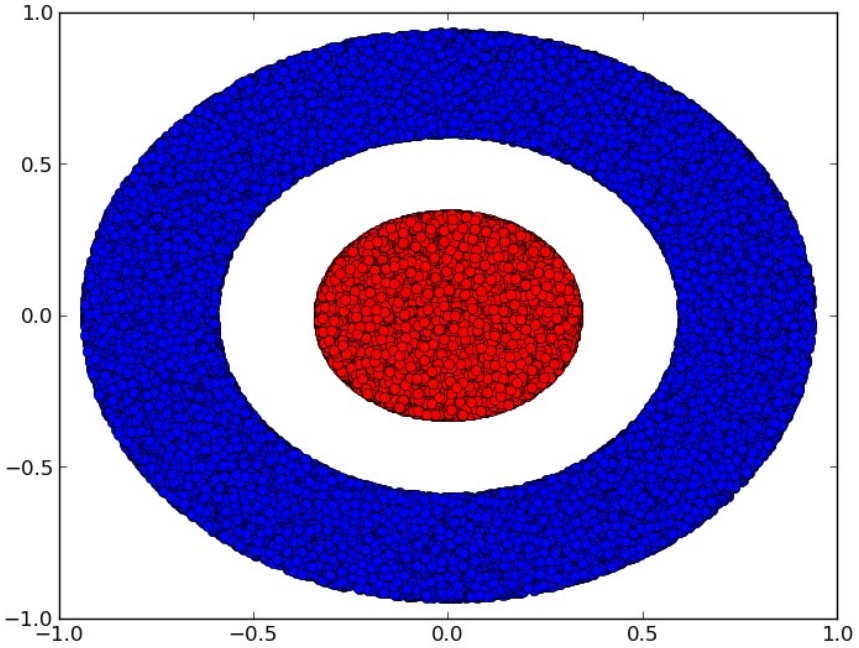
Multi-layered Neural Networks 의 깊이 (depth) 의미

다 층의 non-linear 구조를 통과하면서 구분이 용이한 상황으로 변환

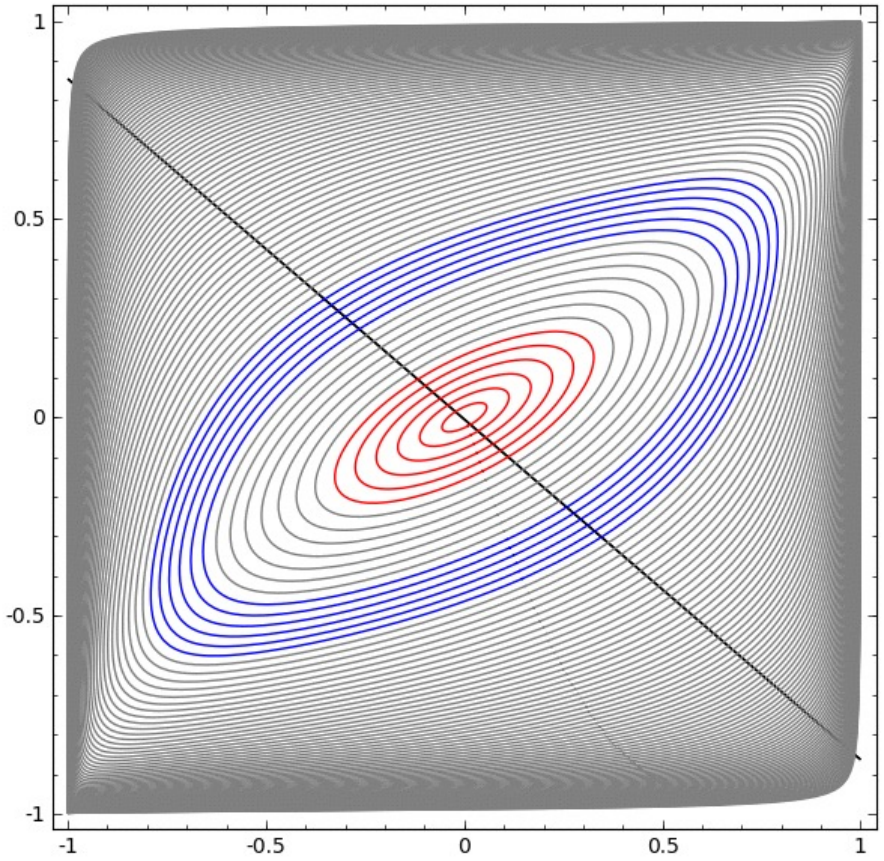


http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Multi-layered Neural Networks 의 넓이 (width) 의미

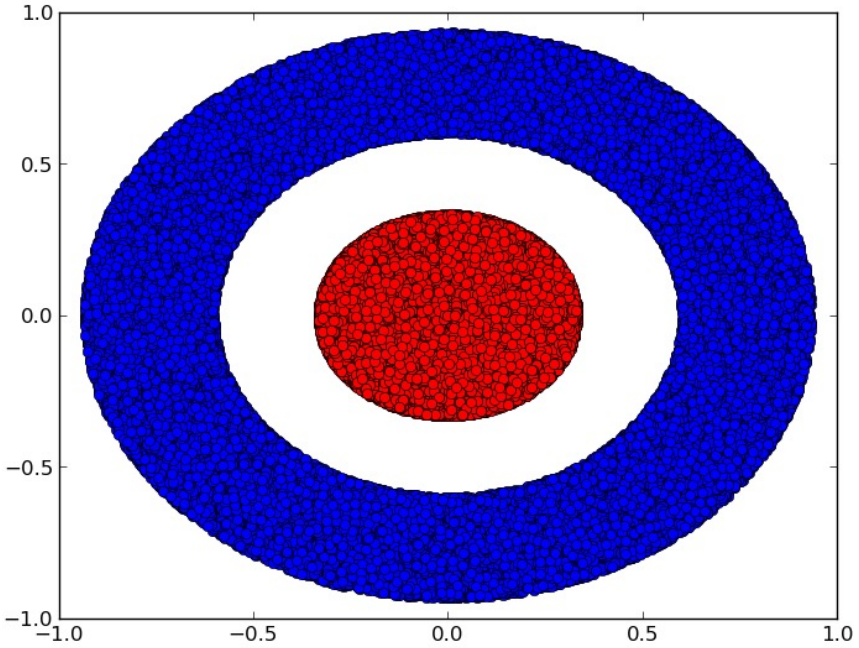


넓이가 2인 경우

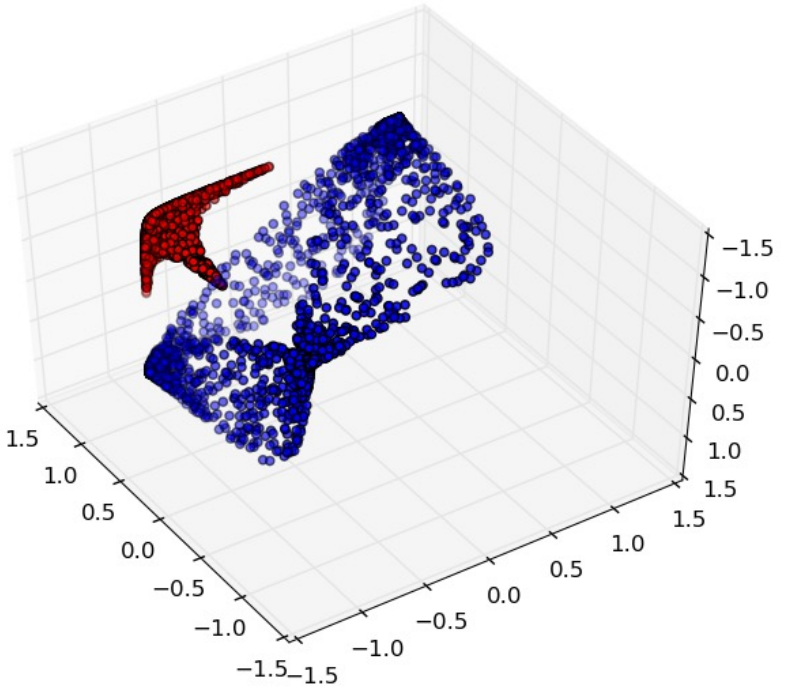


http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Multi-layered Neural Networks 의 넓이 (width) 의미



넓이가 3인 경우



http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

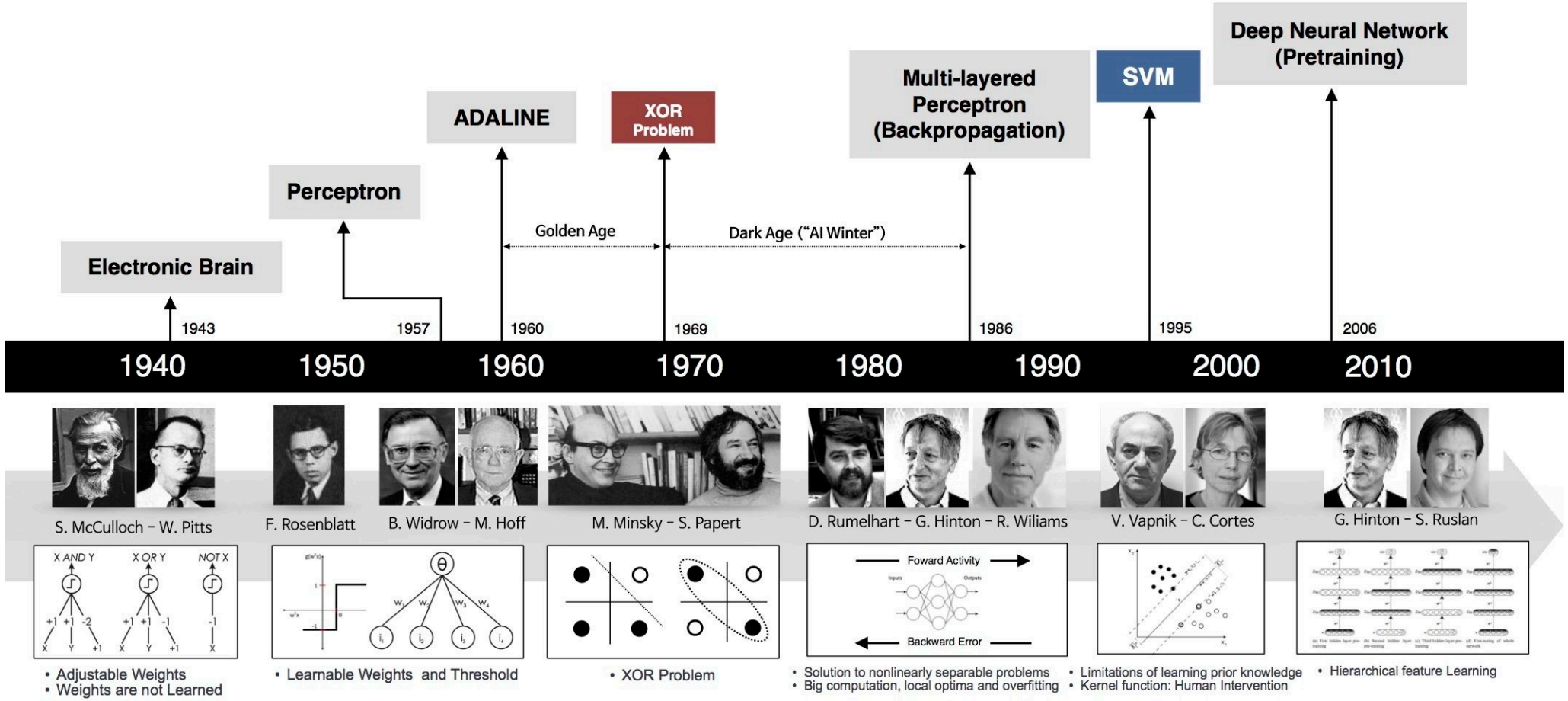
Quiz

Q1. Deep Learning에서 깊이와 넓이를 키우는 것이 무조건 도움이 되는 것인가요?

Q2. Deep Learning이 다른 ML에 비하여 무조건 좋은가요?

목차

Deep Learning 의 겨울과 봄

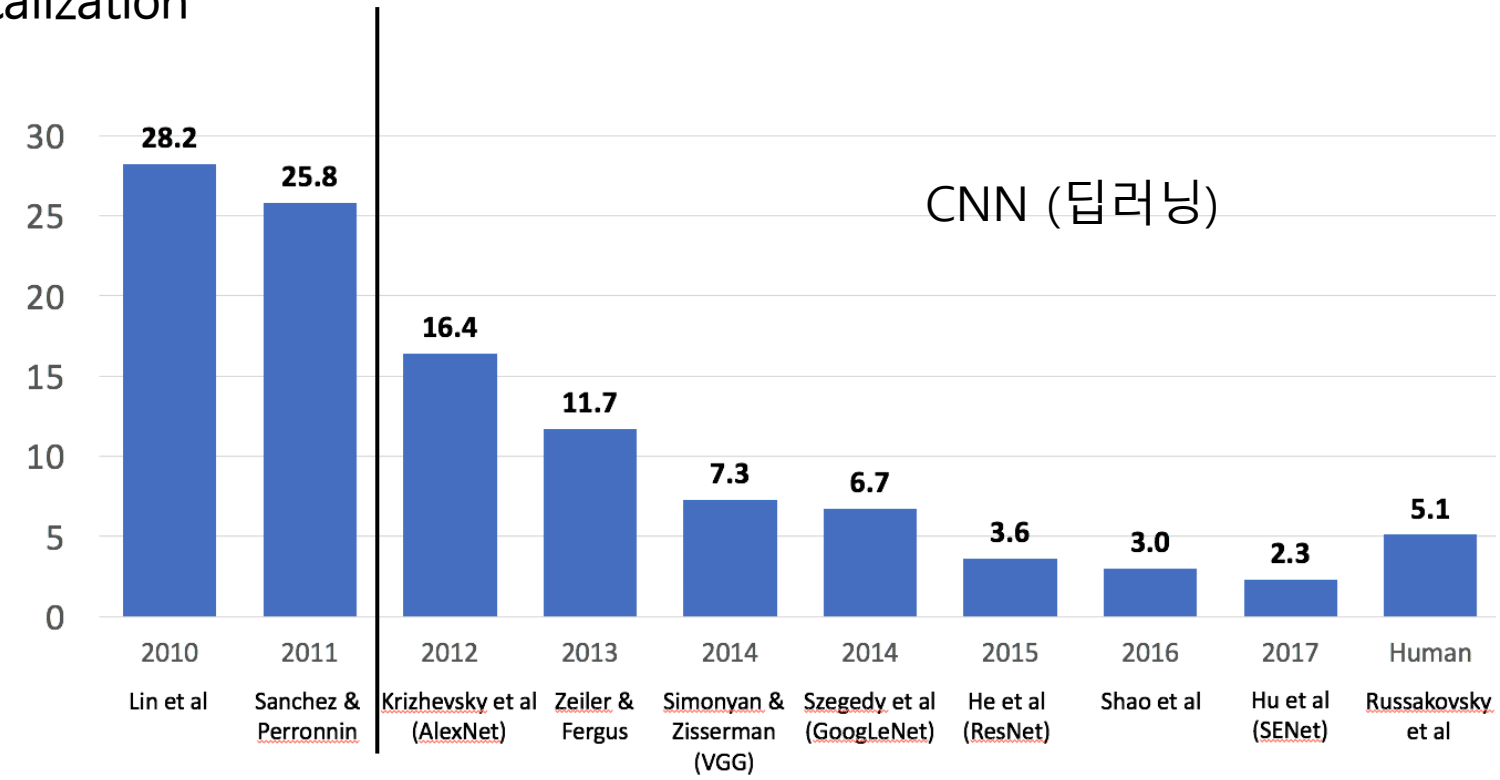


http://slides.com/beamandrew/deep-learning-101#/12

Deep Learning 은 왜 갑자기 관심을 얻었나?

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

- Image classification
- Single-object localization
- Object Detection



Ex) Image Classification Challenge: 1,000 object classes 1,431,167 images

Russakovsky et al. IJCV 2015
Li et al, CS231 Lecture note

진화과정

Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

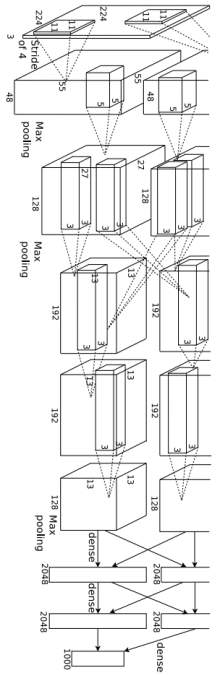
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

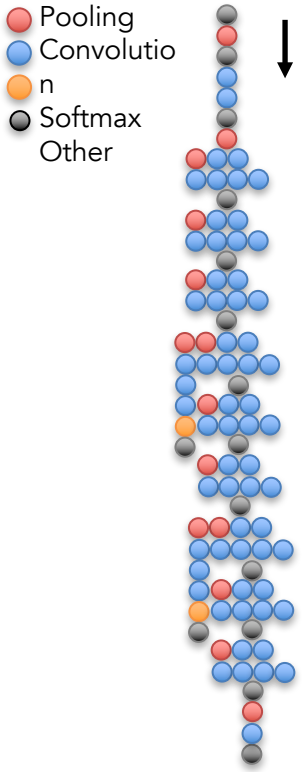
SuperVision



[Krizhevsky NIPS 2012]

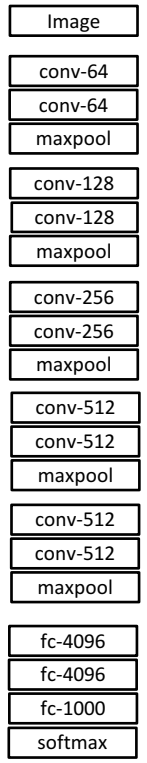
Year 2014

GoogLeNet



[Szegedy arxiv 2014]

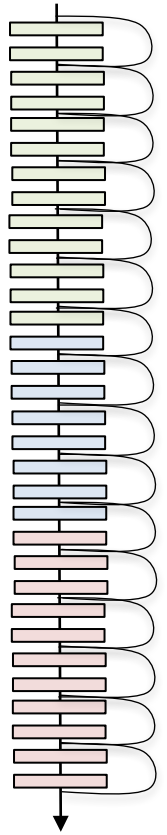
VGG



[Simonyan arxiv 2014]

Year 2015

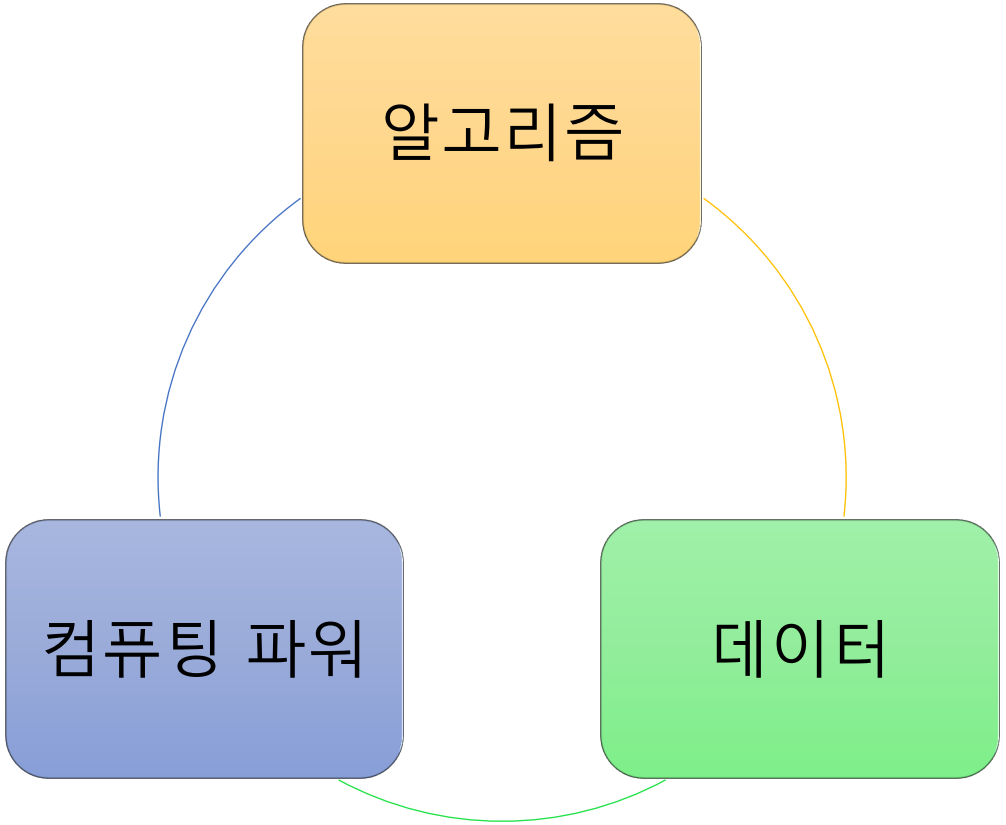
MSRA



[He ICCV 2015]

Li et al, CS231 Lecture note

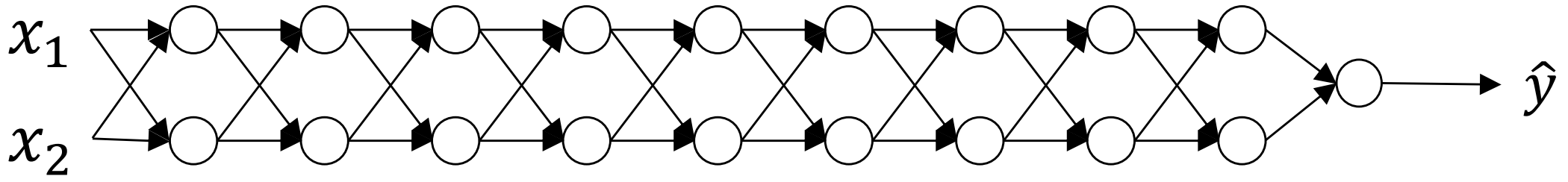
Deep Learning의 발전 3요소



목차

Backpropagation

매우 큰 Neural Network를 어떻게 학습시키나?



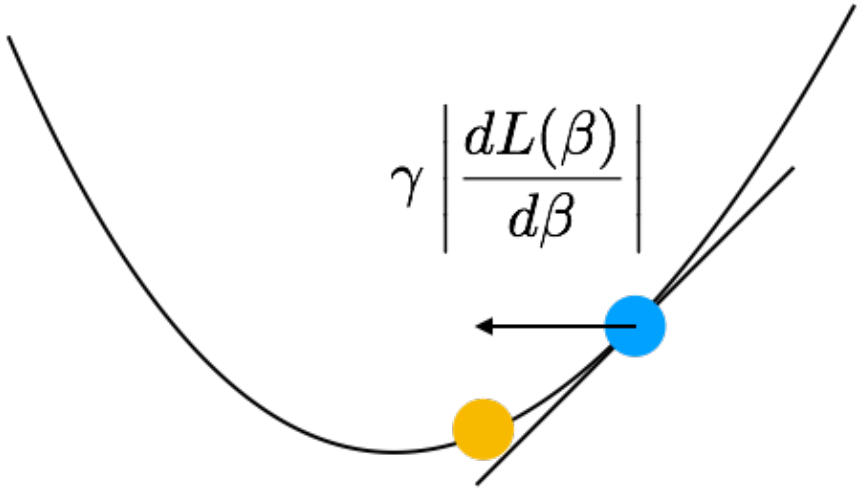
학습이 어려워서 NN이 널리 안쓰였었음.

NN의 혁신의 가장 큰 역할을 한 학습방법이 **Backpropagation** 임

- Backpropagation의 이해를 위해서는 gradient descent 알고리즘과 Computation graph에 대한 이해가 필요함
- Backpropagation은 Chain rule을 활용하는 방법

Gradient Descent

가장 간단한 최적화 방법 중 하나



Gradient: $\frac{dL(\beta)}{d\beta} = \lim_{h \to 0} \frac{L(\beta + h) - L(\beta)}{h}$

Gradient Descent: $\beta(t+1) = \beta(t) - \gamma(t) \nabla L(\beta(t))$

step size

여러가지 방법으로 step size를 각 축마다 다르게 줄 수 있음. 예, AdaGrad, RMSProp, Adam,..

Neural Net 에서의 Gradient 계산

Backpropagation: a simple example

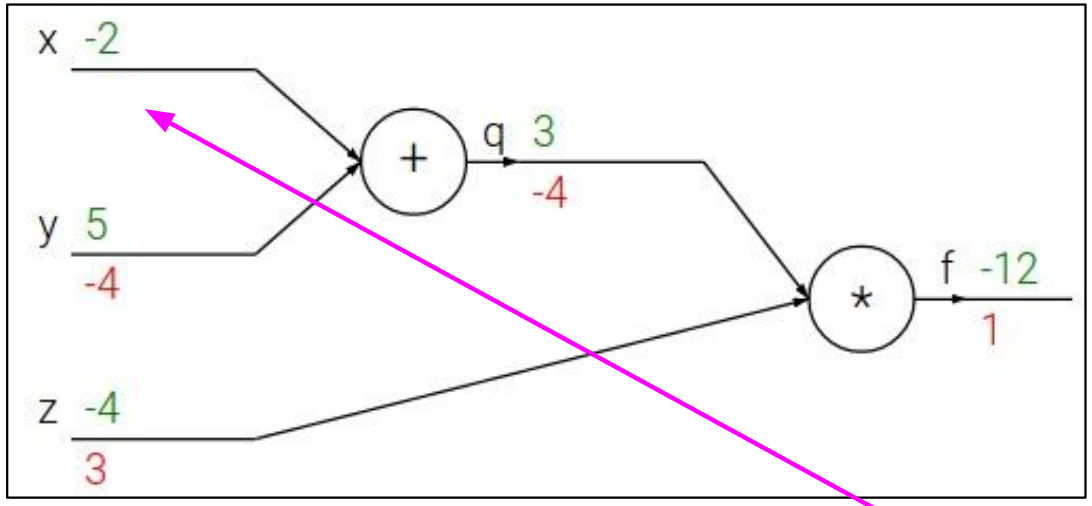
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain rule:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial x}$$

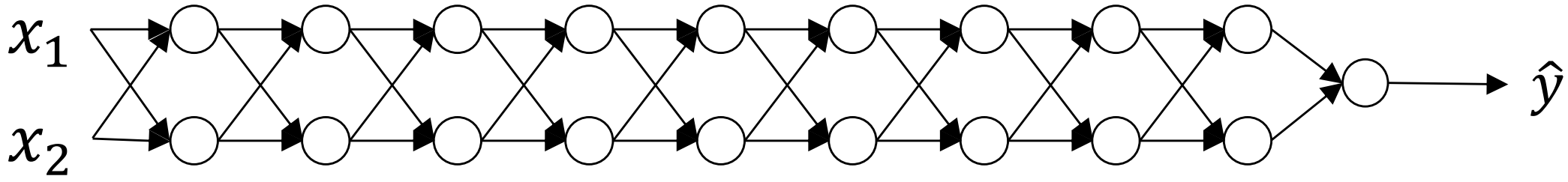
Upstream gradient

Local gradient

Vanishing Gradient/Exploding Gradient

Backpropagation식을 보면, 출력부터 Gradient가 순차적으로 곱해지게 된다.

=〉 Vanishing/Exploding Gradient 문제 발생



Layer를 증가시키면 오히려 성능이 떨어지는 문제가 많이 발생 〈= 학습의 어려움 때문

해결 방안: ReLu를 사용, 학습이 용이한 구조를 사용,….

Stochastic Gradient Descent

데이터가 굉장히 많은 상황에서 모든 데이터를 활용하여 Gradient 값을 계산하기 위해서는 너무 많은 시간이 필요함.

=〉 잘게 쪼개서 계산하자!

$$L(W) = \frac{1}{N}\sum_{i=1}^{N} L_i(x_i, y_i, W) + \lambda R(W)$$

$$\nabla_W L(W) = \frac{1}{N}\sum_{i=1}^{N} \nabla_W L_i(x_i, y_i, W) + \lambda \nabla_W R(W)$$

Full sum expensive when N is large!

Approximate sum using a **minibatch** of examples
32 / 64 / 128 common

Li et al, CS231 Lecture note

감사합니다!