Machine Learning

Seyoung Yun

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

1. AI? ML?
2. Machine Learning
3. Deep Learning

Artificial Intelligent (AI)

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

Hair Salon

Quiz

Please describe any differences between

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

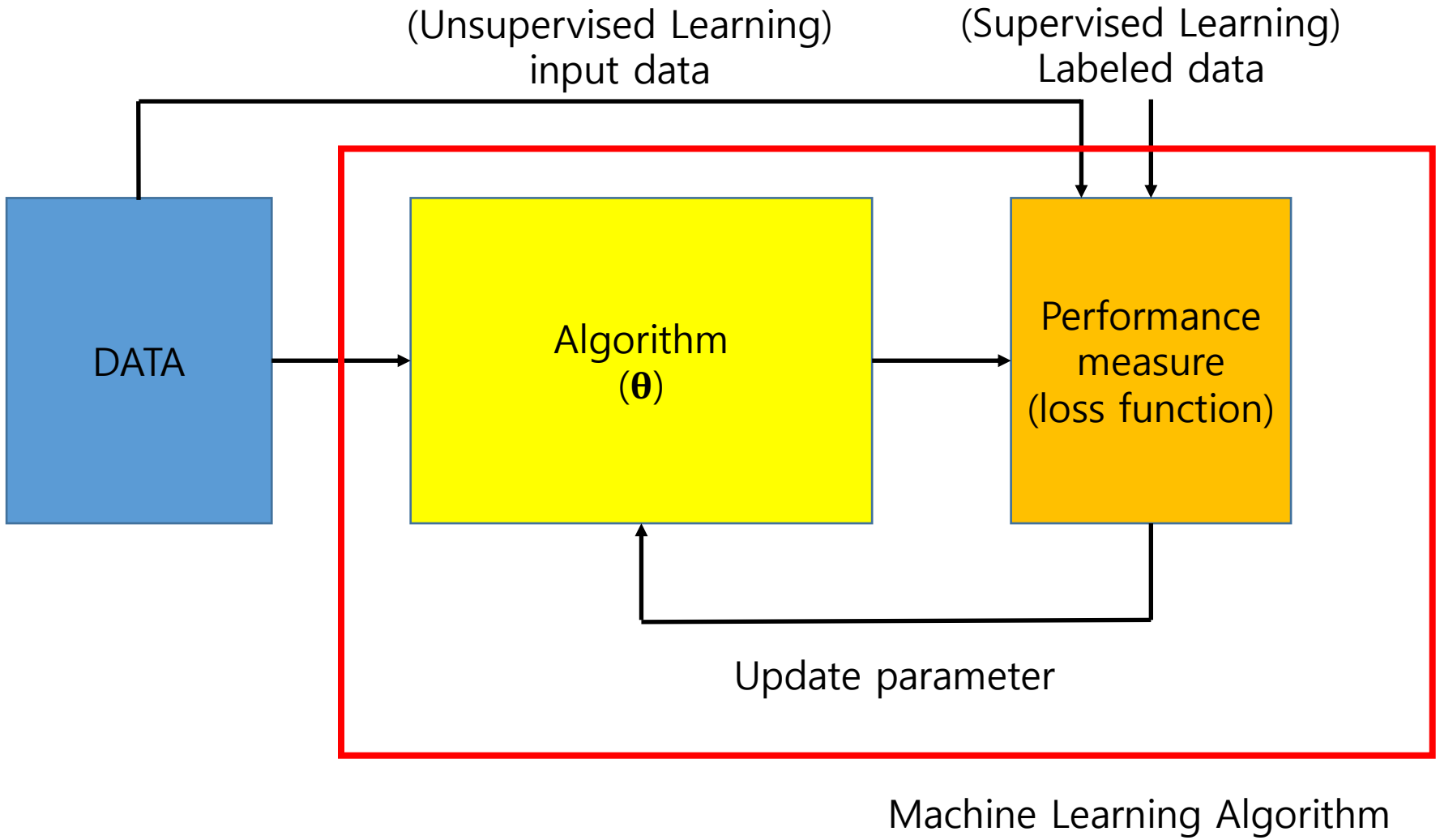
Answer

- Artificial Intelligent (AI)
  - Intelligence demonstrated by machines (planning, communication, object detection, voice recognition, QA,…)
  - General AI: like human-being
  - Narrow AI: task specific AI (e.g.,: object detection algorithms cannot communicate with human)
- Machine Learning (ML)
  - A subset of AI.
  - Algorithms and statistical models that learn from data
- Deep Learning (DL)
  - A subset of ML
  - Inspired from neural networks, also known as Artificial Neural Network (ANN)
  - Deep structure

목차

1. AI? ML?
2. Machine Learning
3. Deep Learning

Machine Learing



Supervised Learning

ML: training parameters (edge weight, bias) from data

- Optimize the objective function
- We can design the objective function in many different ways based on our task

Regression vs Classification (according to output type)

- Regression: continuous output (e.g. house price prediction)
- Classification: discrete class values (e.g., classify cat and dog)

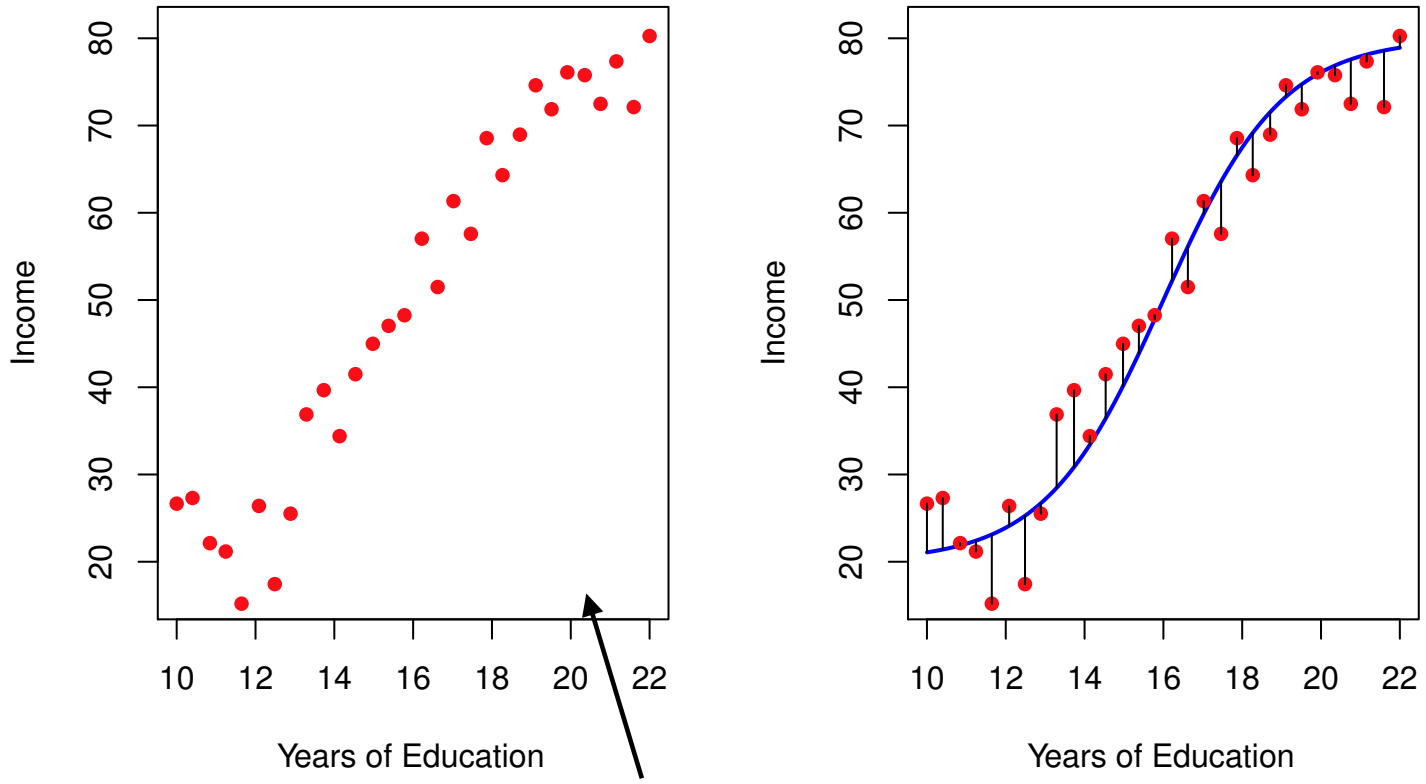
Objective functions

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

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

regularization

Regression



Training data

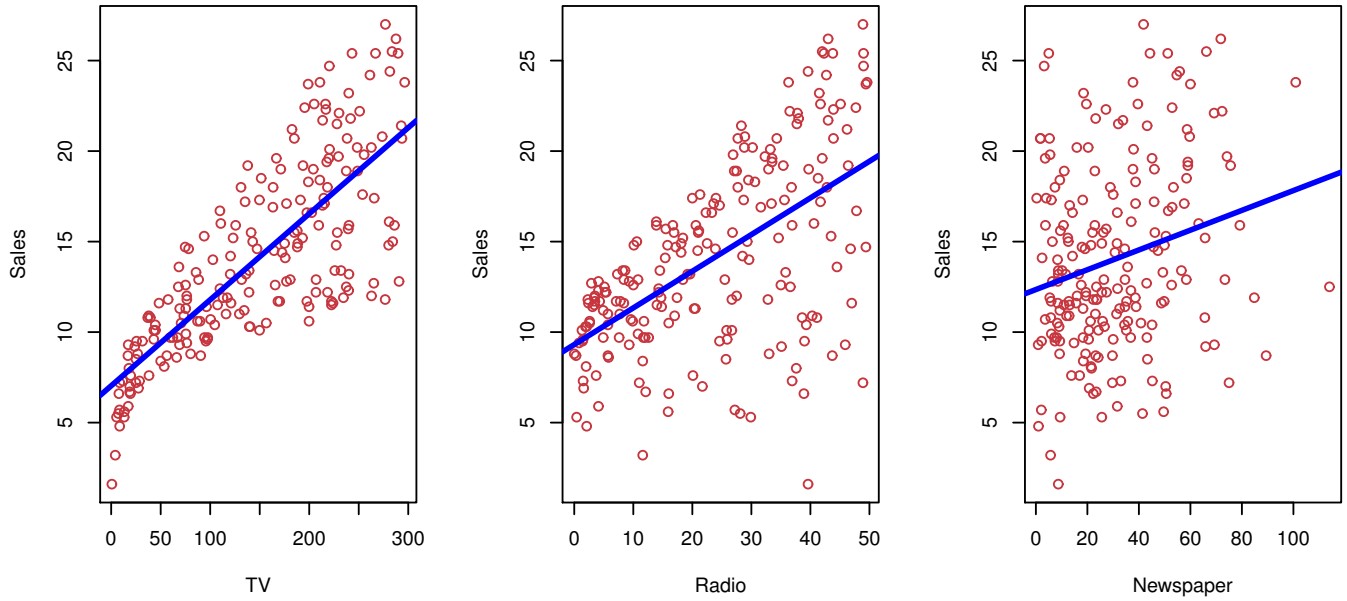
- Try to find a function that explains the input and output pairs in the training set
- Q: which functino? ?

Parametric Statistics

Parametric Statistics

- Assume that data is generated from a class of function defined with a set of parameters
- Learn the parameters from data
- e.g.: linear regression

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



Parameter Learning

Define loss function

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

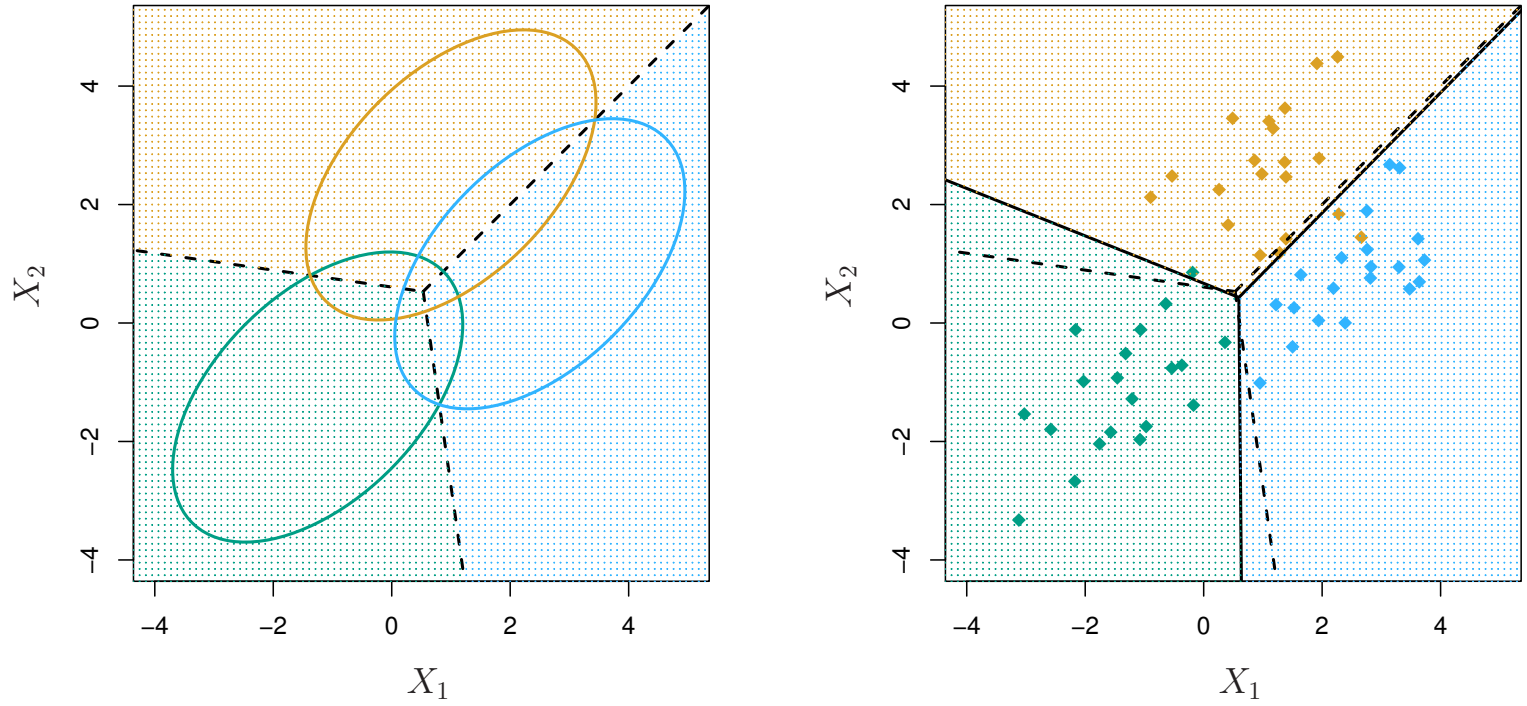
$$\begin{aligned}\text{RSS} =& e_1^2 + e_2^2 + \cdots + e_n^2 \\ =& \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2\end{aligned}$$

- How to find parameters?
  - Optimize RSS

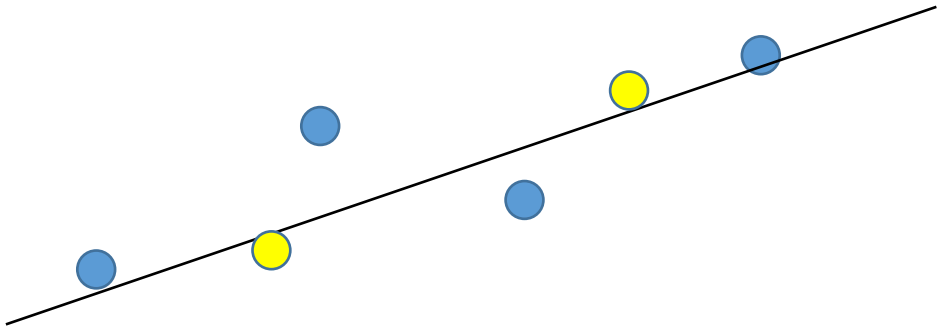
Parametric Learning

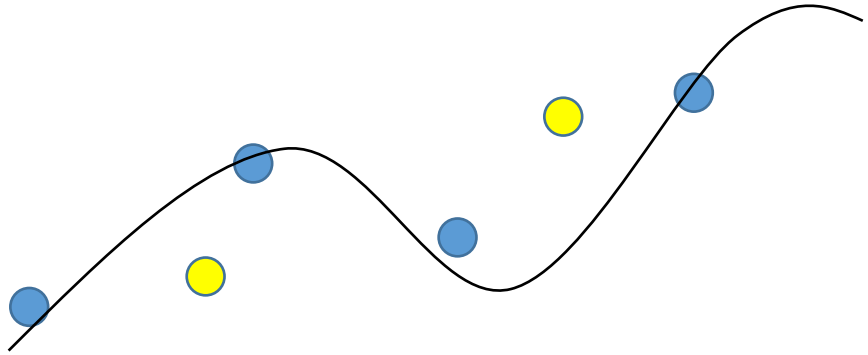
Parametric Statistics

- e.g.: Linear Discriminant Analysis (LDA) for classification
  - Linear decision boundary



Model Selection



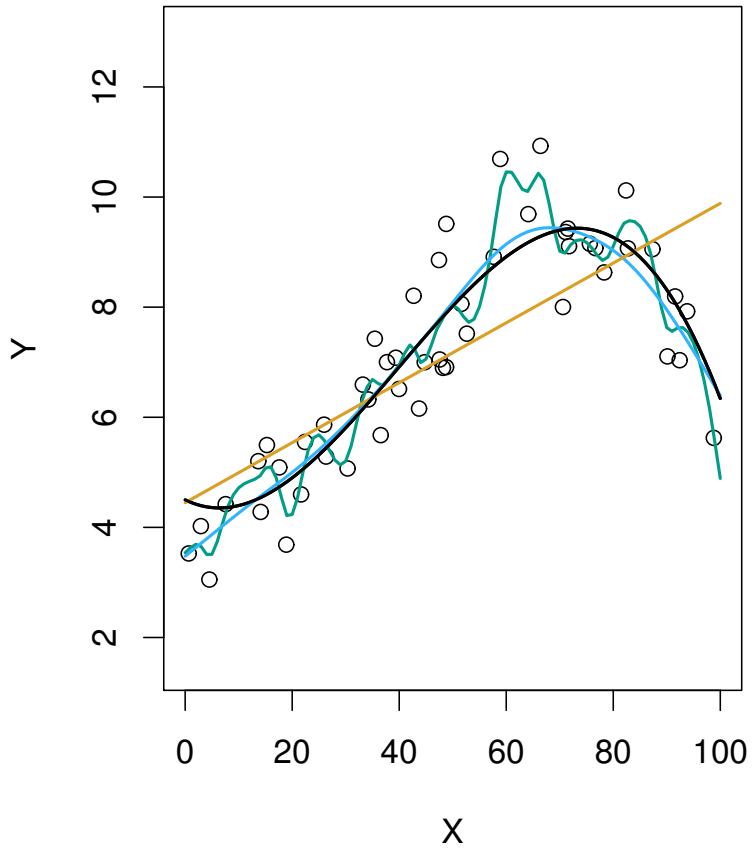


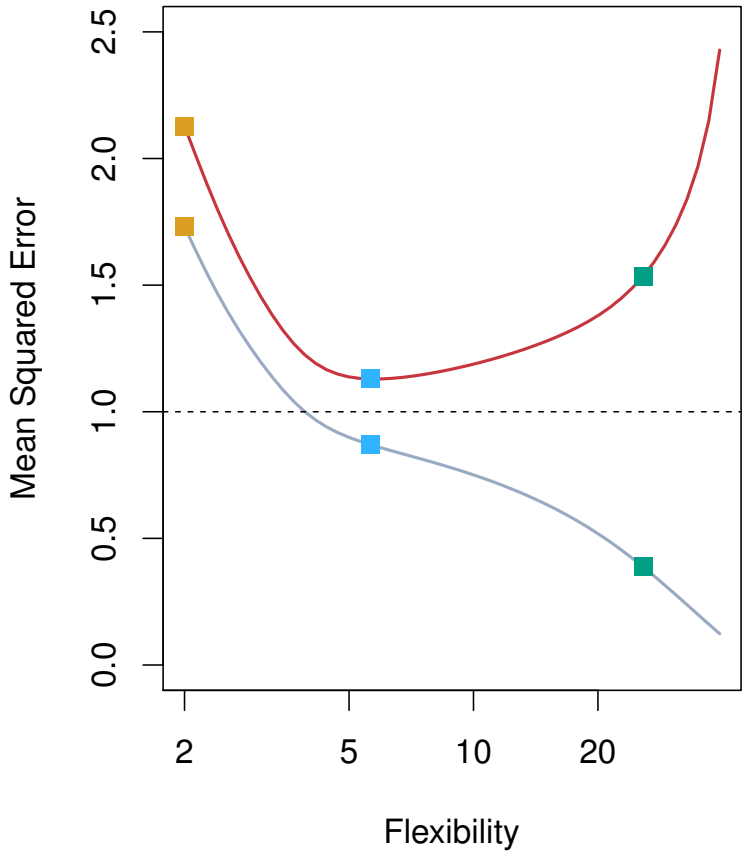
How to select the function class

- Bias vs variance

Bias-Variance Tradeoff

High flexible algorithms (complex, many parameters) can explain the training Data very well, but the algorithms could be very bad for the test data..





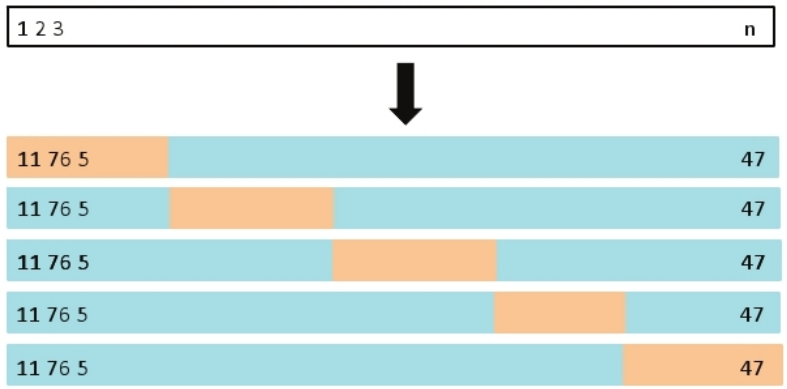
How to Select Model

Training data vs. Test data (cross-validation)

To learn parameters

Test the trained model

(optional) k-fold cross validation: (Deep Learning researchers do not use this)

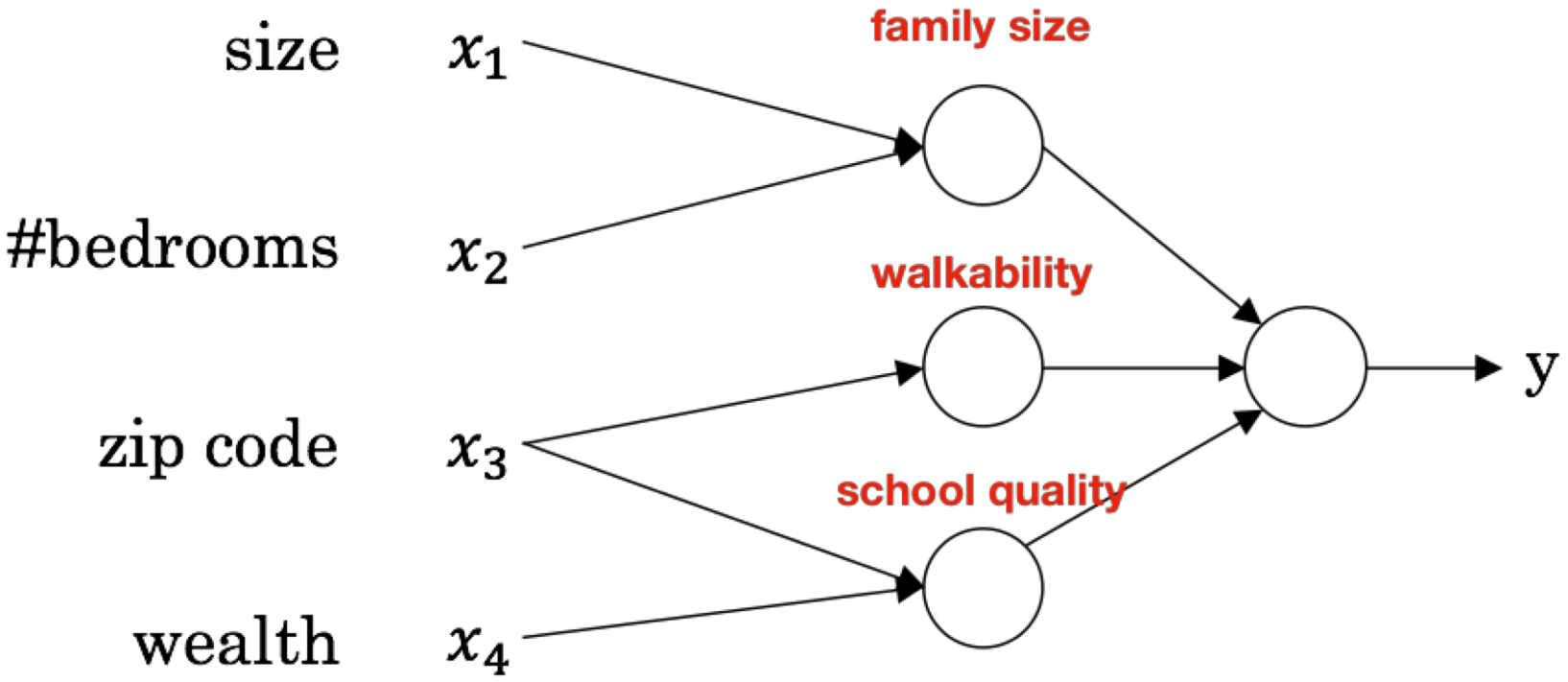


목차

1. AI? ML?
2. Machine Learning
3. Deep Learning

Example) House price prediction

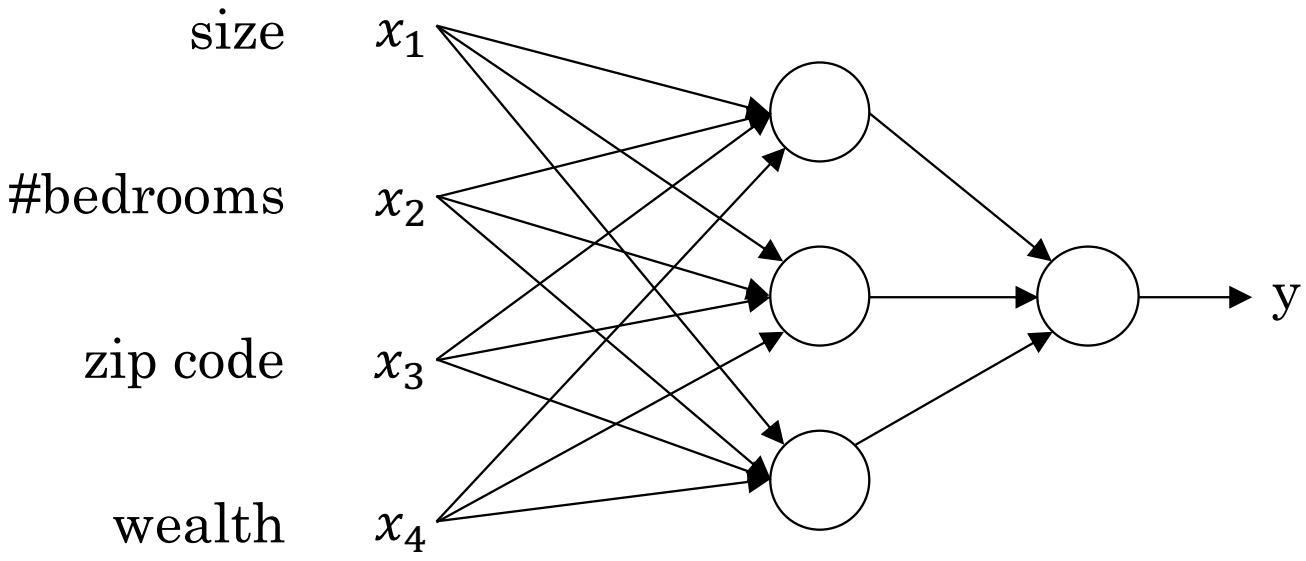
Let's predict house prices from house information



Rule-based system: we need experts to formulate the rules

Example) House price prediction

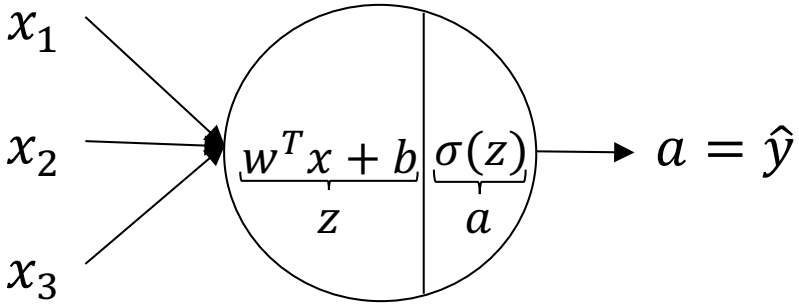
Neural Net?

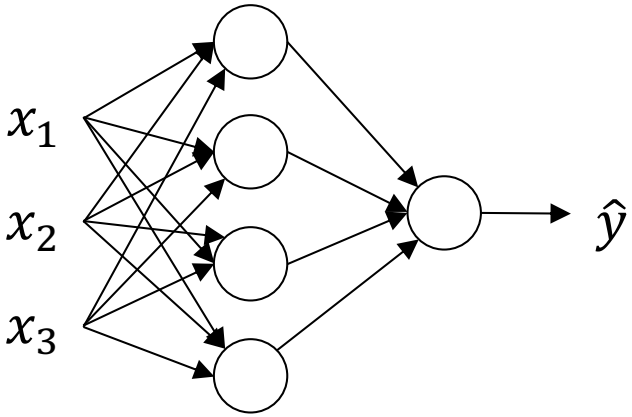


Machine Learning and Neural Net

- Without experts, prior knowledge,
- Artificial Neural Net automatically generates the function mapping input to output from data (Machine Learning)

Line and Circle





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

$$a = \sigma(z)$$

Line: edge, arrow,..

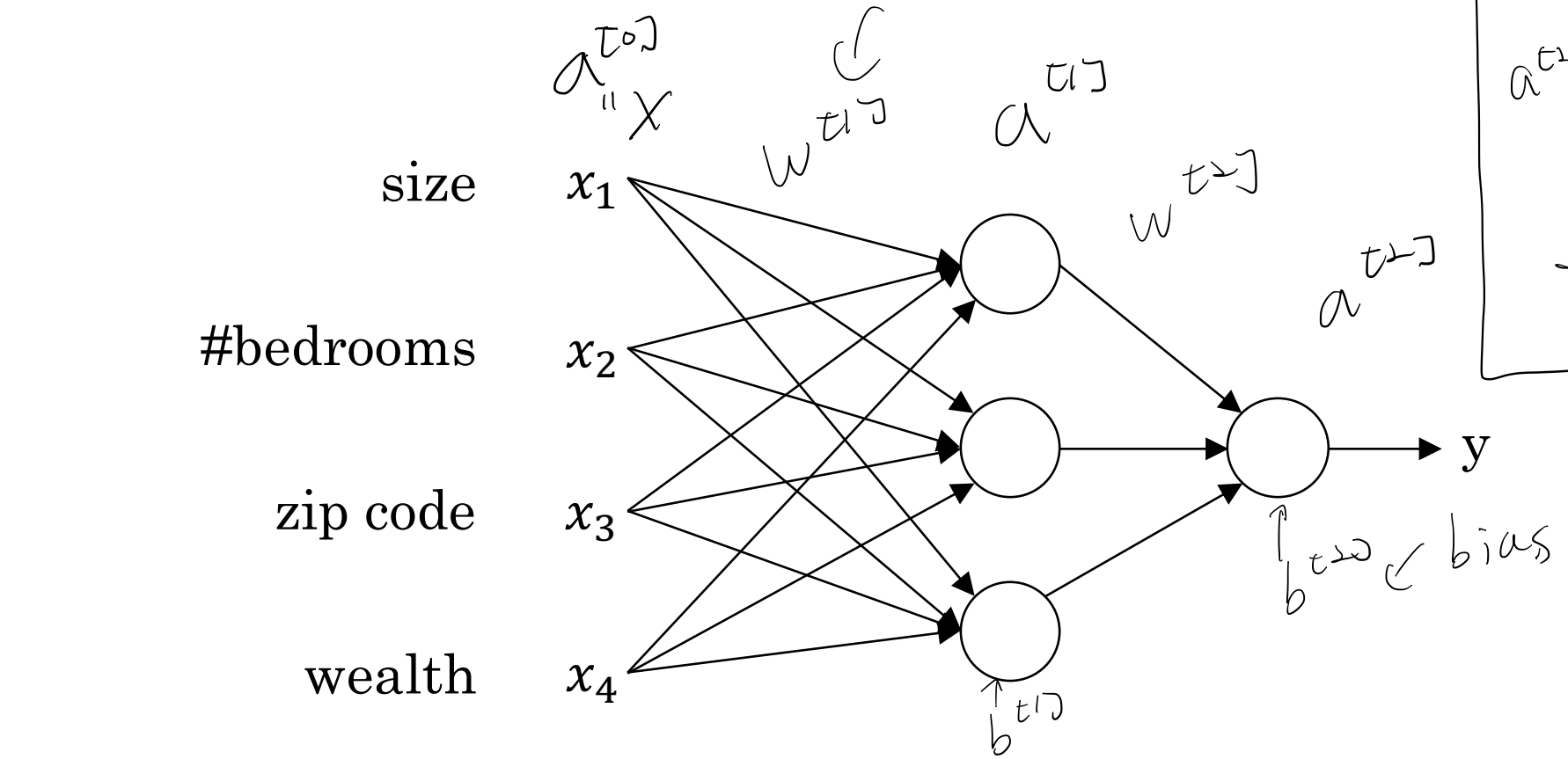
- Each line has a parameter and multiplies incoming value and the parameter value

Circle: Neuron

- Add all incoming values and pass through a non-linear activation function

Quiz

Please describe the neural network function.



$$a^{[1]} = \sigma^{[1]}(W^{[1]}x + b^{[1]})$$

activation function

$$a^{[2]} = \sigma^{[2]}(W^{[2]}a^{[1]} + b^{[2]})$$

$$y = a^{[2]}$$

$a^{[0]} = x$

$W^{[1]}$ weight matrix

$a^{[1]}$

$W^{[2]}$

$a^{[2]}$

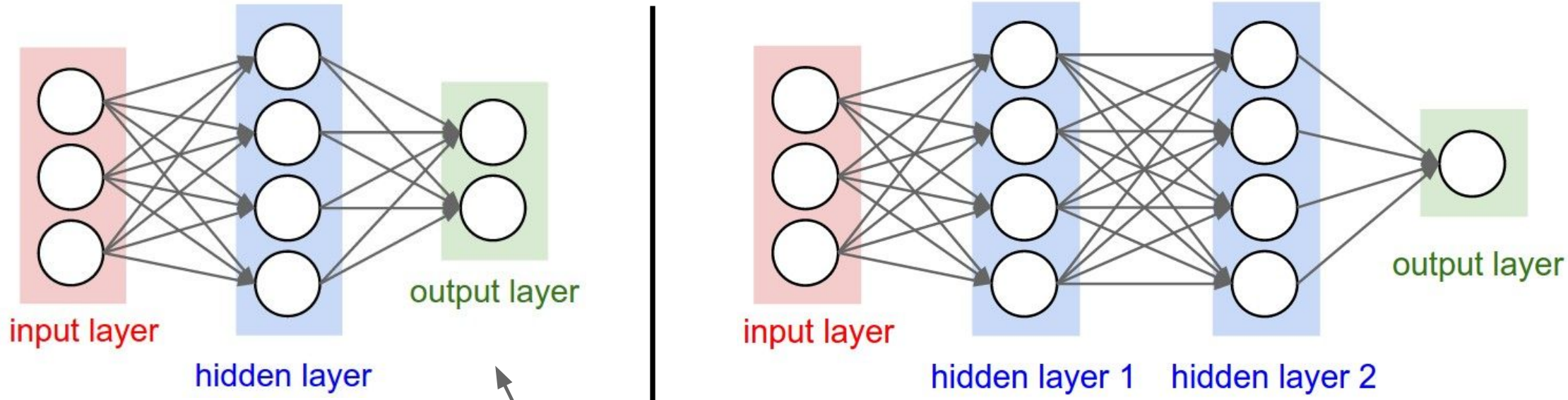
$b^{[2]}$ bias

$b^{[1]}$

$$W^{[1]} = \begin{bmatrix} w_{1,1}^{[1]} \\ \vdots \end{bmatrix}_{1 \le i \le 3,\ 1 \le j \le 4}$$

out, in

Neural Networks



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

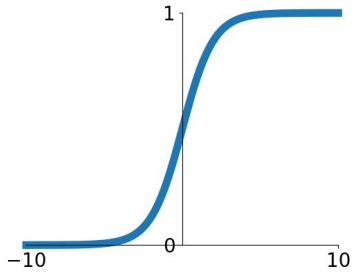
"3-layer Neural Net", or
"2-hidden-layer Neural Net"

"Fully-connected" layers

Activation 함수

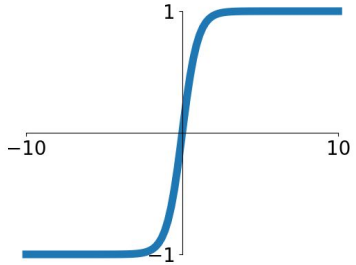
Sigmoid

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



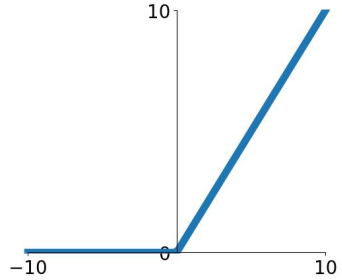
tanh

$$\tanh(x)$$



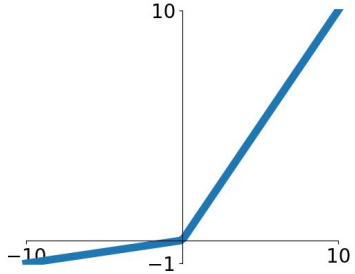
ReLU

$$\max(0, x)$$



Leaky ReLU

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

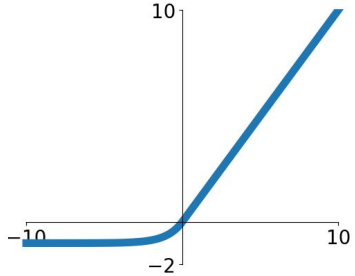


Maxout

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

ELU

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



Li et al, CS231 Lecture note

Why should we use non-linear activation functions?

When $\sigma(x) = x$

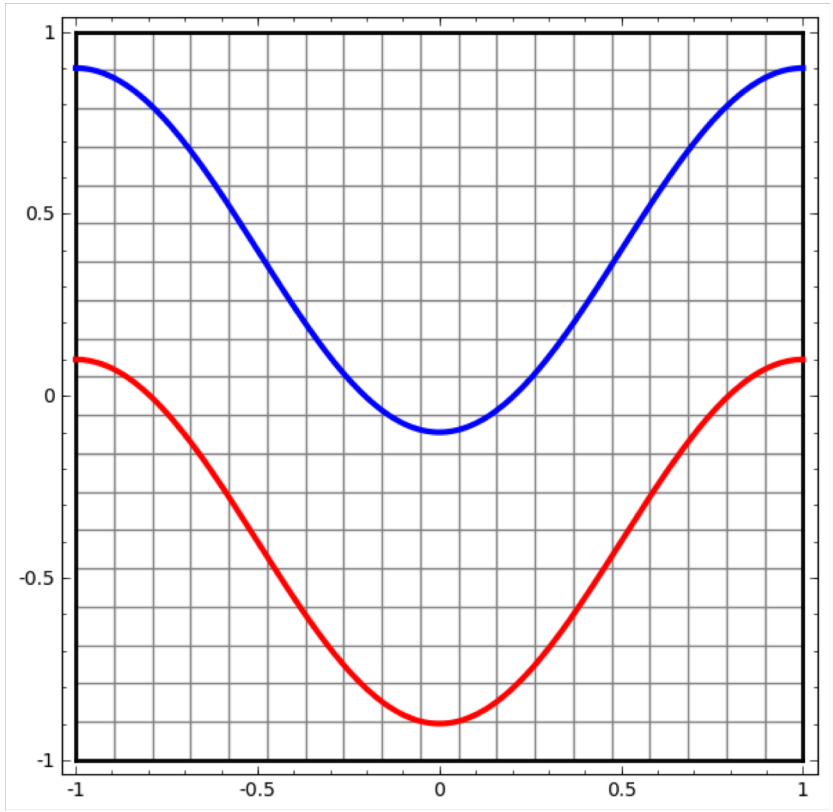
$$a^{[2]} = \sigma^{[2]}(w^{[2]} a^{[1]} + b^{[2]})$$
$$= \sigma^{[2]}(w^{[2]} \sigma^{[1]}(w^{[1]} a^{[0]} + b^{[1]}) + b^{[2]})$$
$$= w^{[2]} w^{[1]} a^{[0]} + w^{[2]} b^{[1]} + b^{[2]}$$

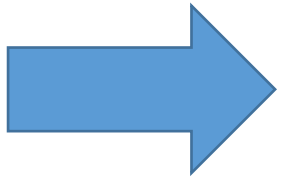
for any $w^{[1]}$, $w^{[2]}$, $b^{[1]}$ and $b^{[2]}$, we can make the same function

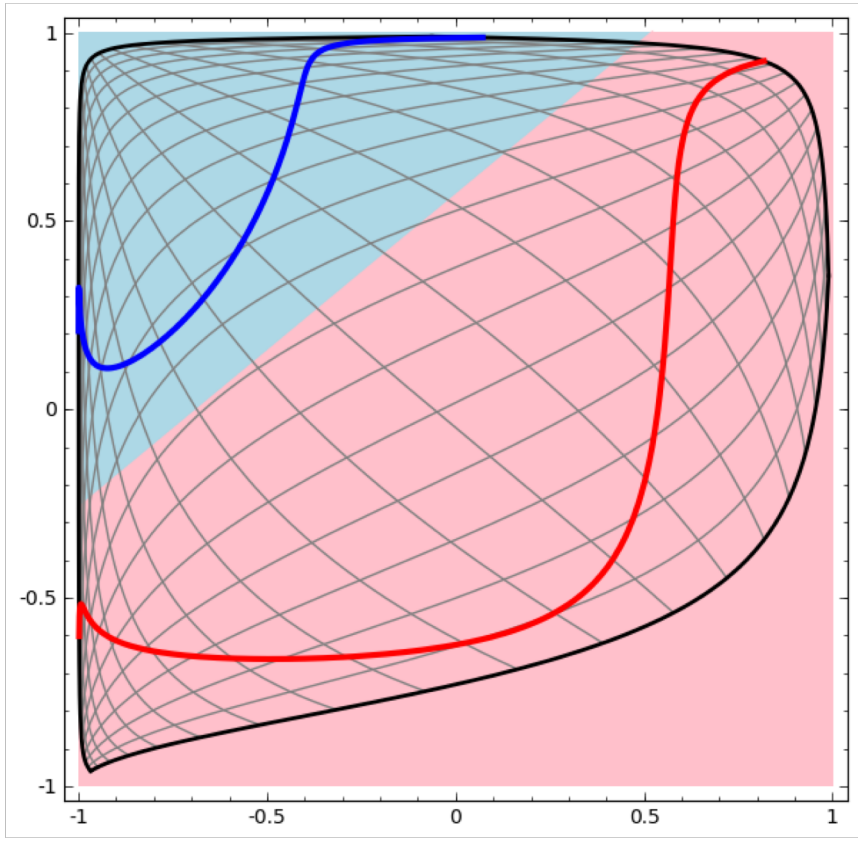
$$a^{[2]} = \underbrace{w}_{w^{[2]} w^{[1]}} a^{[0]} + \underbrace{b}_{w^{[2]} b^{[1]} + b^{[2]}}$$

So, $a^{[1]}$ is redundant.

Multi-layered Neural Networks and Depth



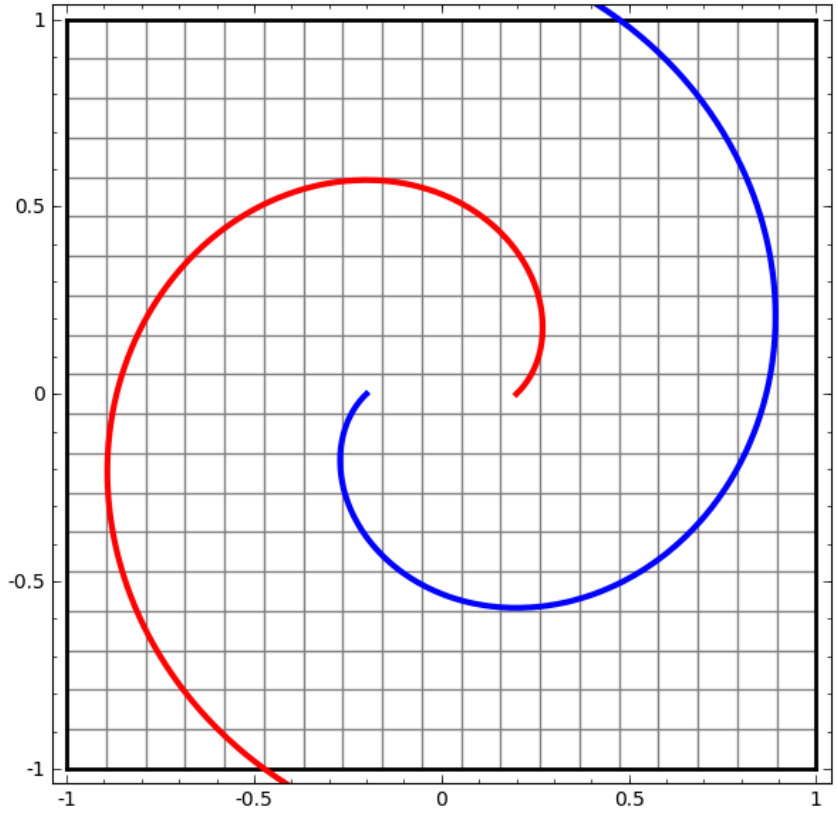




http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

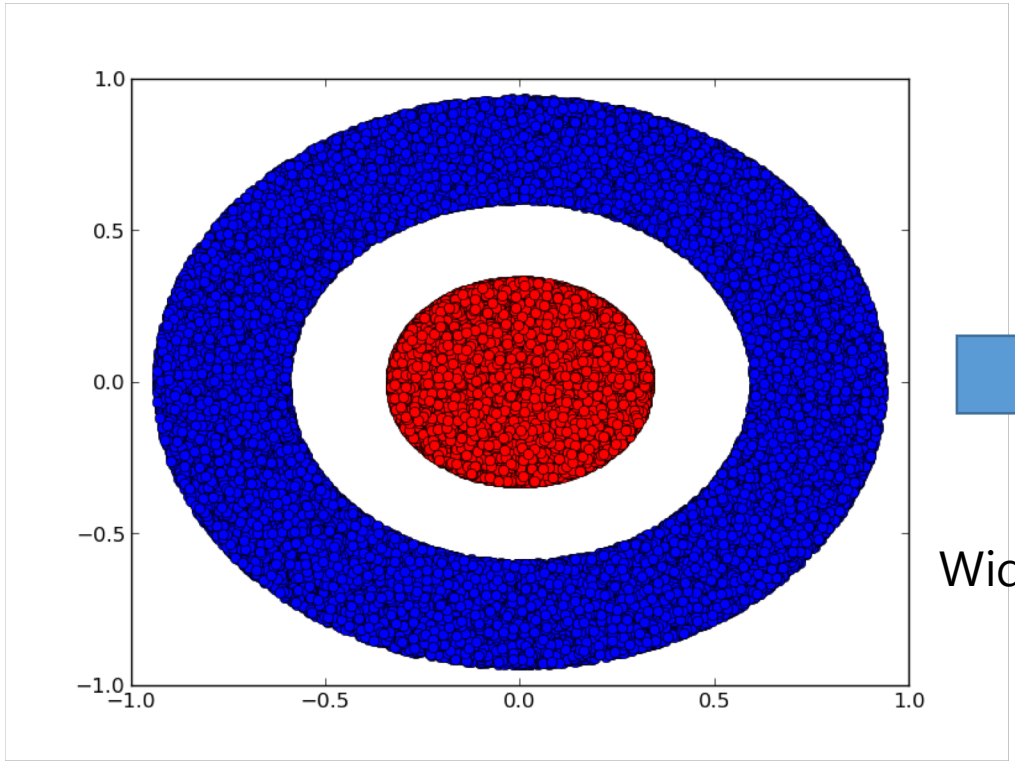
Multi-layered Neural Networks and Depth

다 층의 non-linear 구조를 통과하면서 구분이 용이한 상황으로 변환

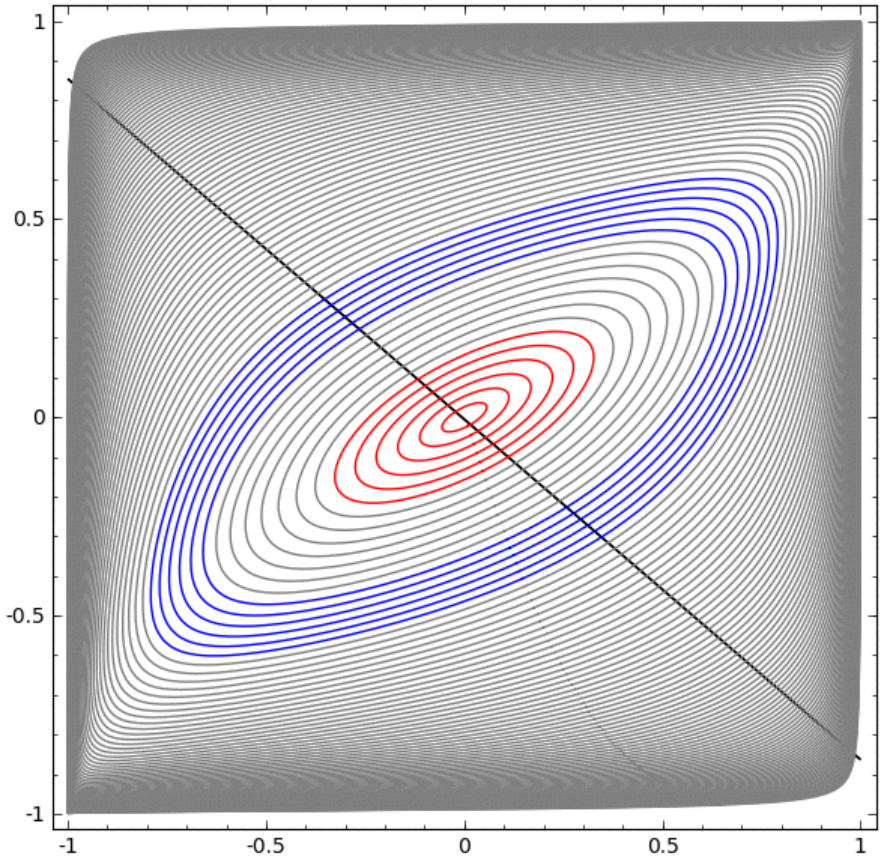


http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Multi-layered Neural Networks and width

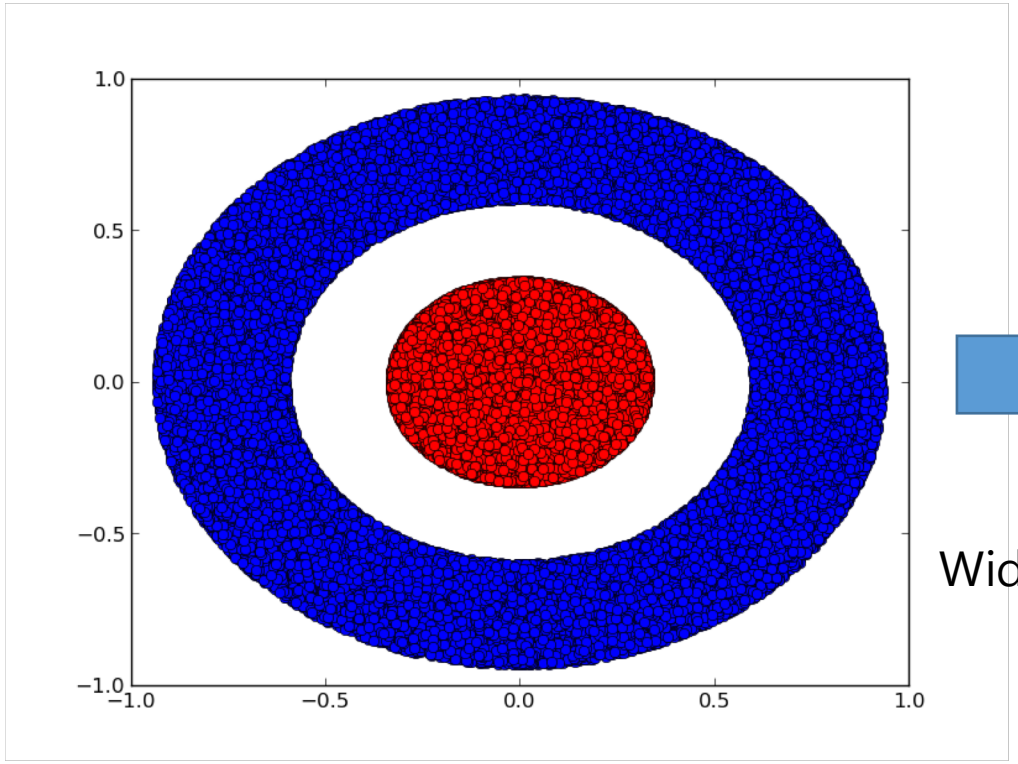


Width = 2

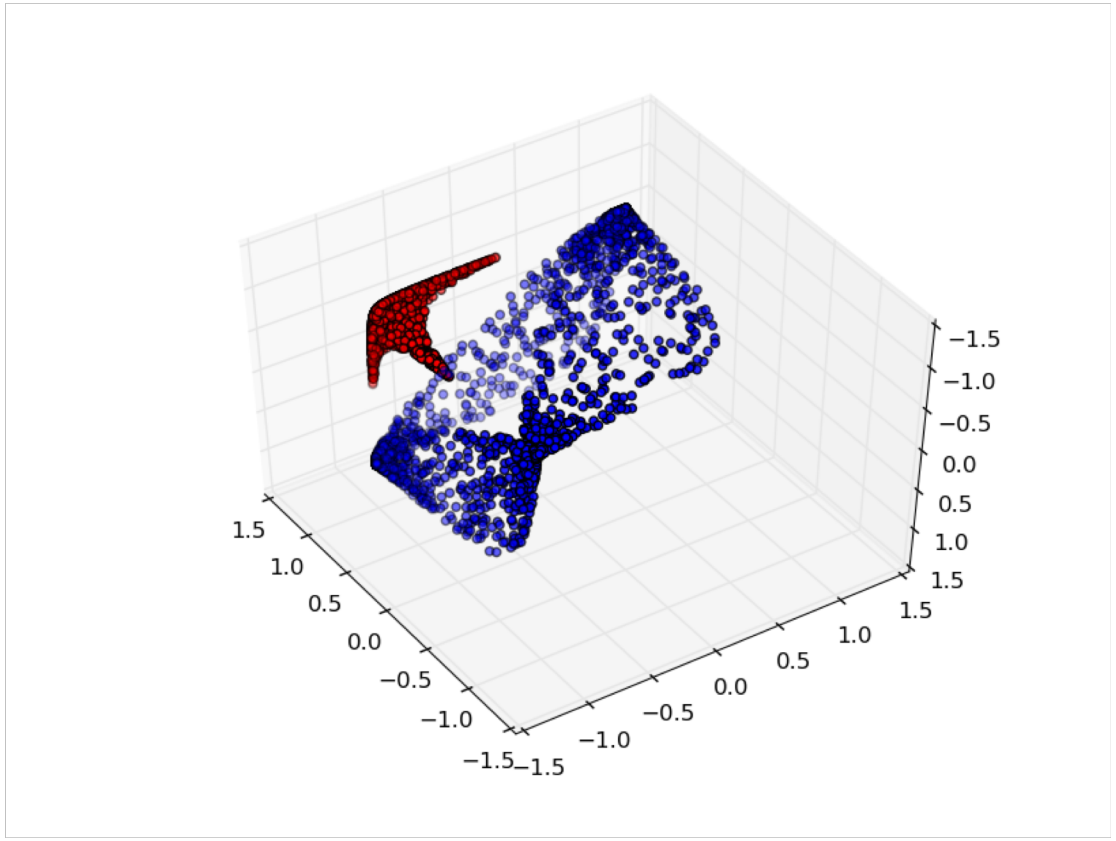


http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Multi-layered Neural Networks and width



Width = 3



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