KUBIG 24-S
여름방학 BASIC STUDY SESSION

NLP SESSION
WEEK1

KUBIG
DATA SCIENCE & AI

CONTENTS

KUBIG
DATA SCIENCE & AI

Session 중에는 모두 캠을 켜주시기 바랍니다 :)

01 NLP SESSION 소개

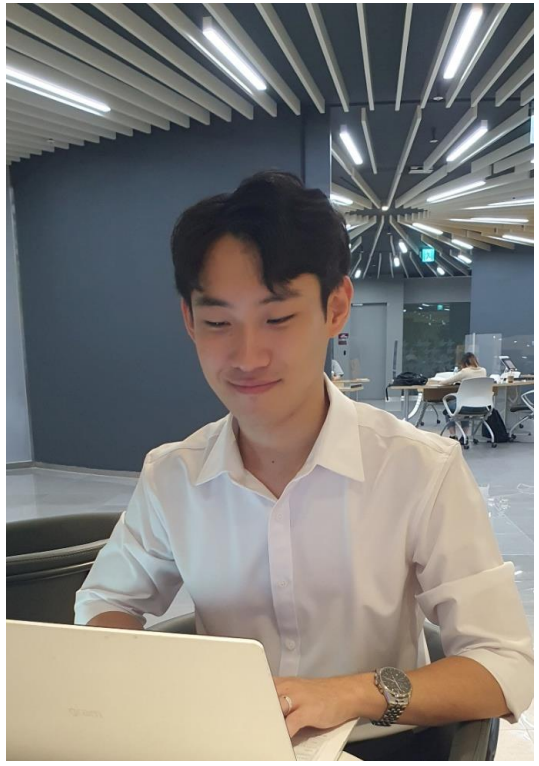
방학동안 무엇을 하나요?

KUBIG
DATA SCIENCE & AI

1-1. 분반장 소개

KUBIG DATA SCIENCE & AI

한 학기동안 잘 부탁 드립니다 ! !



분반장 18기 정해원

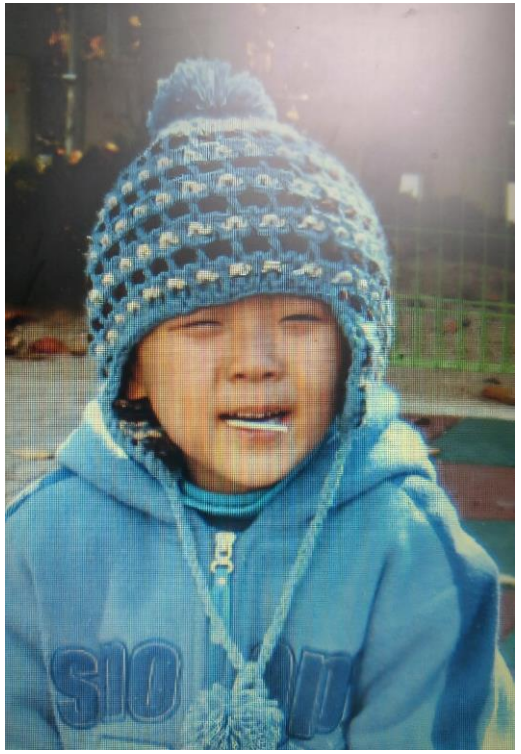
- 분반장 18기 정해원
- 전 홍보팀장 출신
- DL 초심자에서 분반장까지 순항 중
- 최근 우리말 겨루기를 시청하며 언어 능력 습득 중

분반장 17기 김연규

관심 분야: 데이터 분석 -> 인공지능

프로그래밍과 더 친해지려고 노력 중

스터디와 친목 둘 다 환영



분반장 17기 김연규

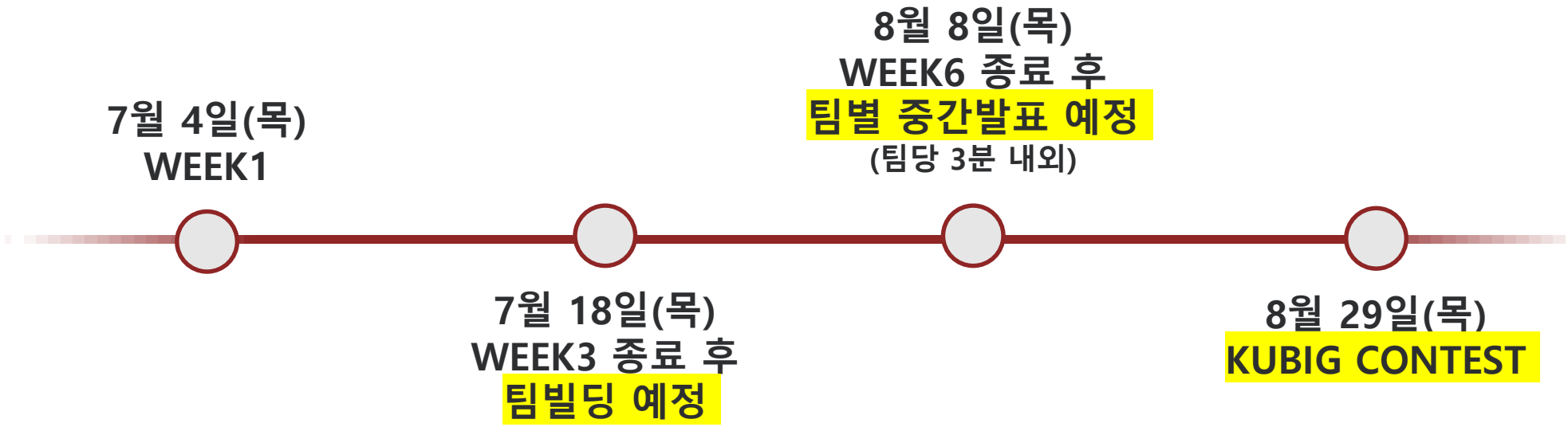
1-2. 커리큘럼

KUBIG DATA SCIENCE & AI

주차	복습과제	학습내용	예습과제
1주차	- 해당 주차에 배운 내용에 대한 코드 실습 과제 부여 - ex) week1 복습과제: deep learning reminder(pytorch basic) - session 시작 초반부에 우수 코드 선정자가 5분 가량 코드 구현 과정 발표(별도 발표자료 없이 코드를 화면공유하여 발표) - 마감기한: 수요일 오후 6시	OT, DL Reminder	- 다음 주차에 배울 내용에 대한 코드 실습 과제 부여 - ex) week1 예습과제: 텍스트 전처리 pipeline 구현 코드 - week1 예습과제라 함은, week1 session이 끝나고 부여되는, week2 내용에 대한 예습과제를 의미합니다! - 마감기한: 수요일 오후 6시
2주차		텍스트 전처리, 워드 임베딩 (Word2Vec, GloVe)	
3주차		RNN, LSTM, GRU, ELMo	
4주차		Attention, Transformer	
5주차		BERT	
6주차		GPT	
7주차		LLM 기초, HuggingFace	

1-2. 커리큘럼

KUBIG
DATA SCIENCE & AI

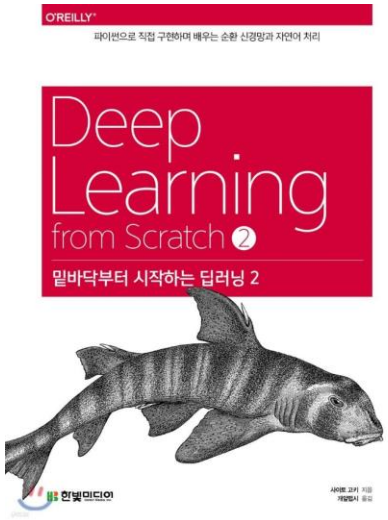


1-2. 커리큘럼

KUBIG
DATA SCIENCE & AI

매주 목 19:00~21:00 총 2시간 진행

1-3. 참고 교재



밑바닥부터 시작하는 딥러닝2

딥러닝을 이용한 자연어 처리 입문
위키독스에서 무료 이용 가능(부분 유료)

고려대학교 DSBA 연구실
유튜브 채널에서 paper review 영상 참고

02 자기 소개

친해지길 바래!

KUBIG
DATA SCIENCE & AI

2. 자기 소개

KUBIG
DATA SCIENCE & AI

한 학기 동안 함께 하실 분들 !

강동헌	강지윤	심승현
이동주	이소희	이승준
정종락	최지우	하진우
강민정	김채원	박준희
윤시호	이세은	이유진

여러분을 소개해주세요!

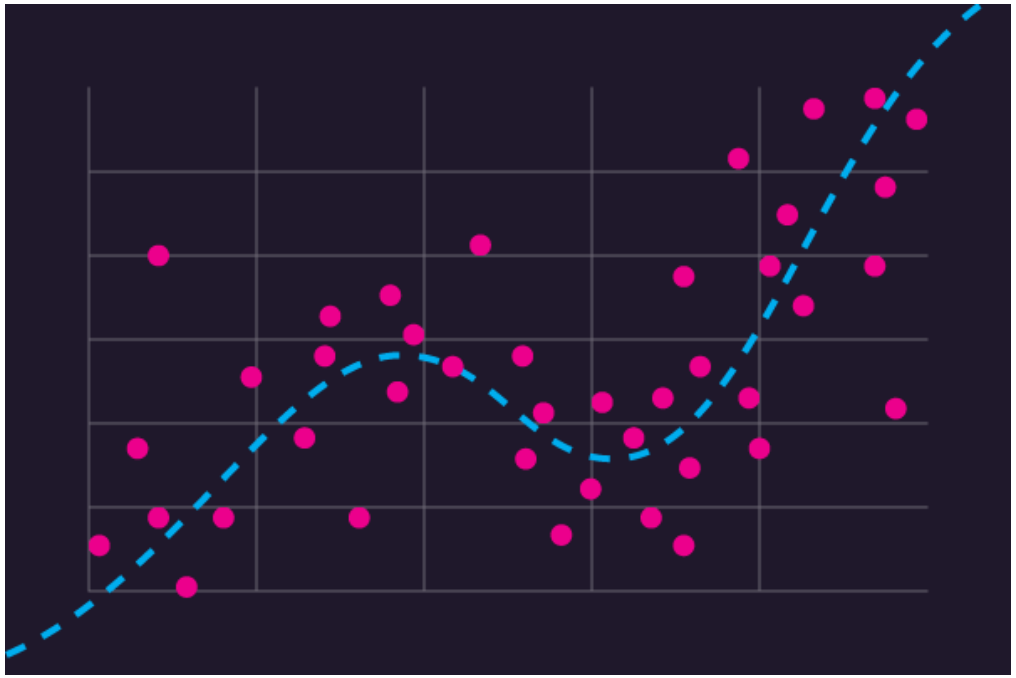
기수, 이름, 학과, 나이, 사는 곳, 취미

NLP 경험, NLP 선택 이유, 원하는 방향성

03 Theoretical Background of Deep Learning

KUBIG
DATA SCIENCE & AI

3-0. Estimate the True Function



What is the True Function?

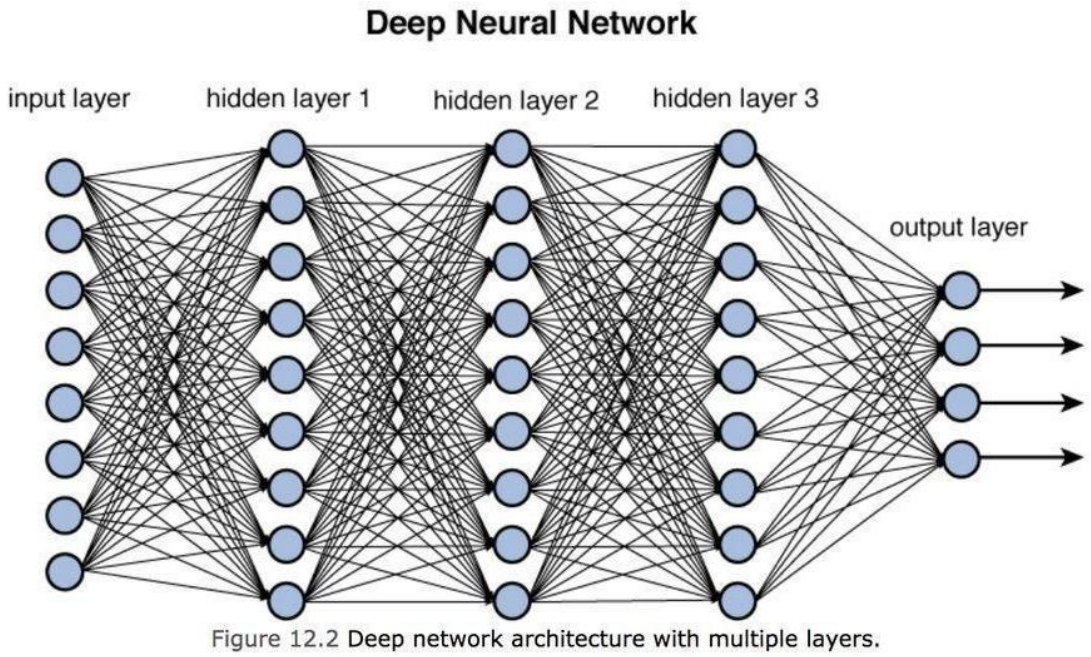
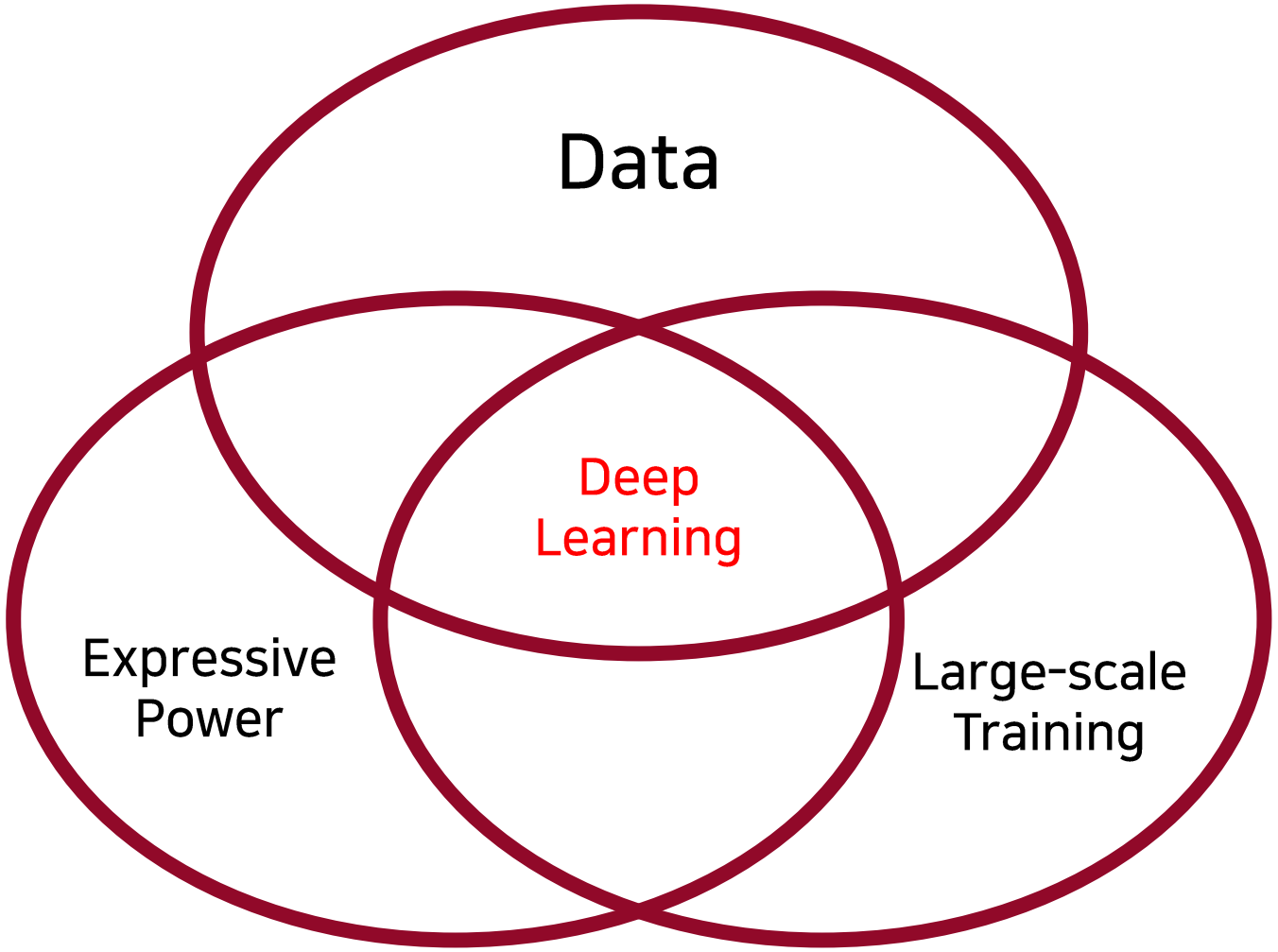


Figure 12.2 Deep network architecture with multiple layers.

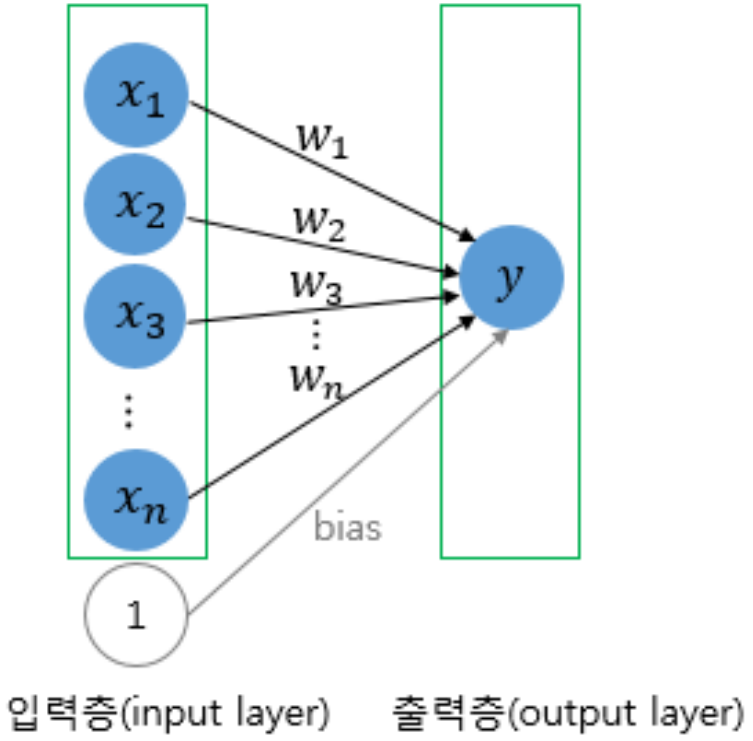
3-0. Principles of Deep Learning



출처: https://wikidocs.net/24958

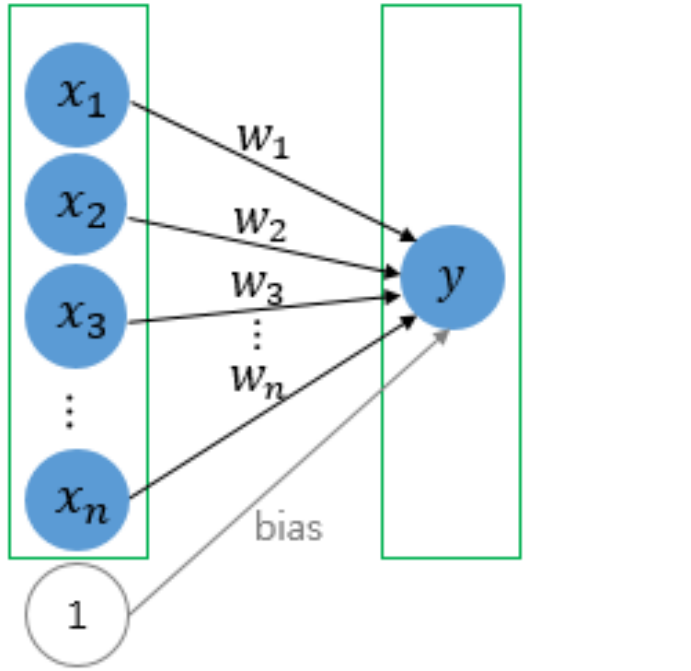
3-1. Single-Layer Perceptron

single-layer perceptron



출처: https://wikidocs.net/24958

single-layer perceptron



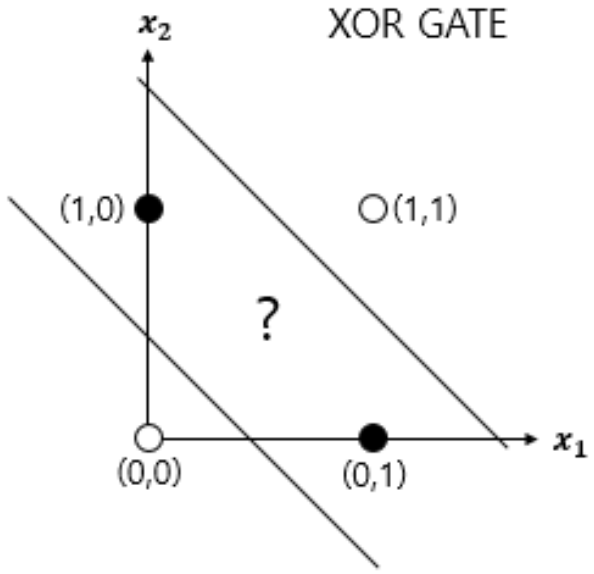
Q) How many parameters?

Q) What Dimension?

출처: https://wikidocs.net/24958

XOR problem

Can you separate B/W by a single line *(in 2-dimension space)*?

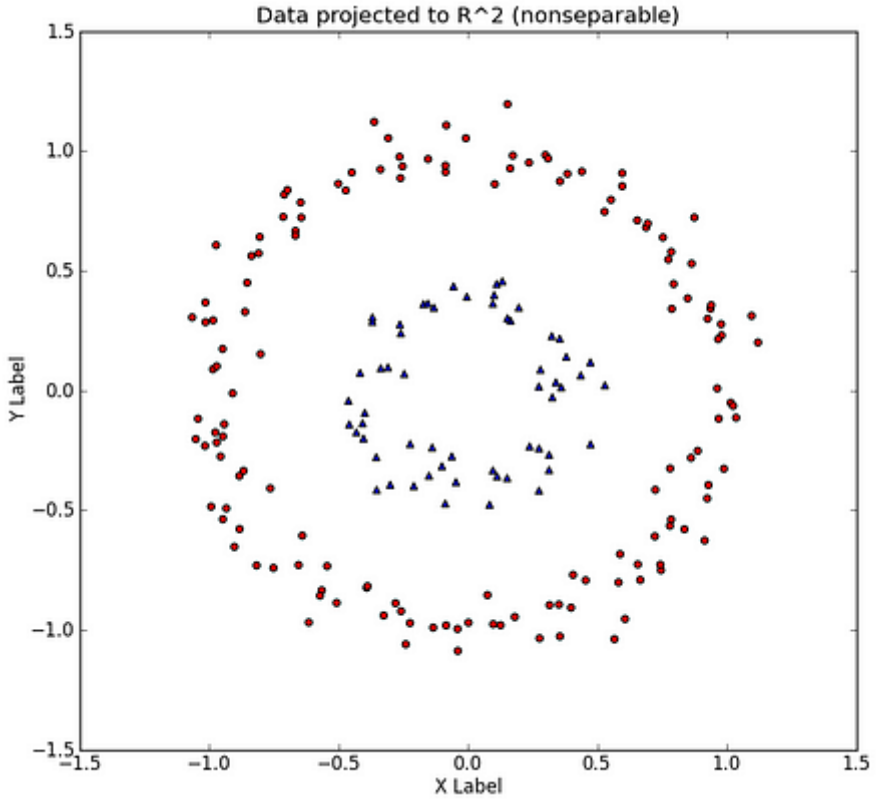


x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

출처: https://wikidocs.net/24958

Linearly Non-Separable

(in 2-dimension space)

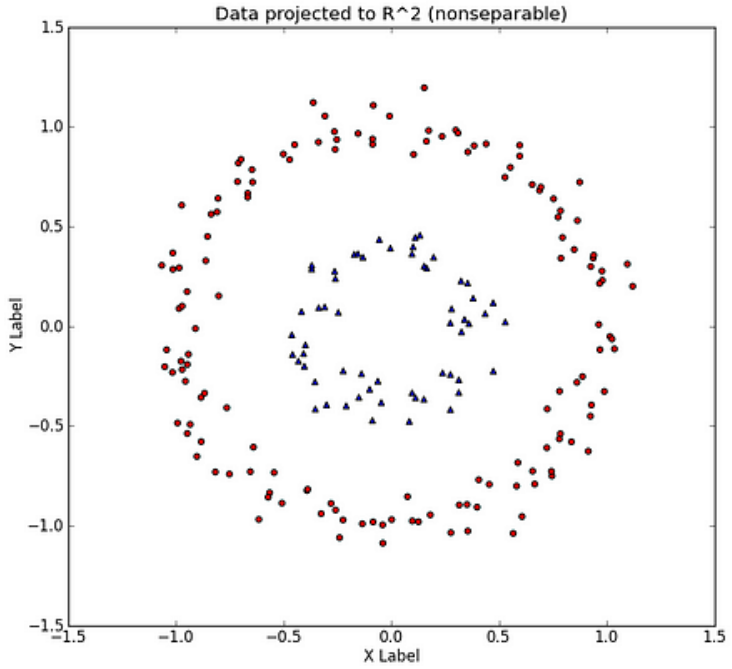


출처: https://vivek-yadav.medium.com/

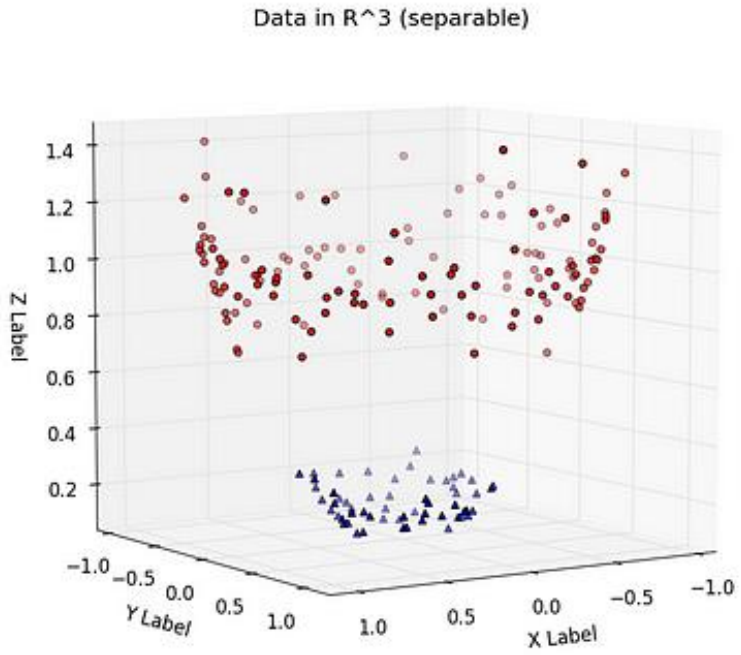
KUBIG DATA SCIENCE & AI

Linearly Non-Separable

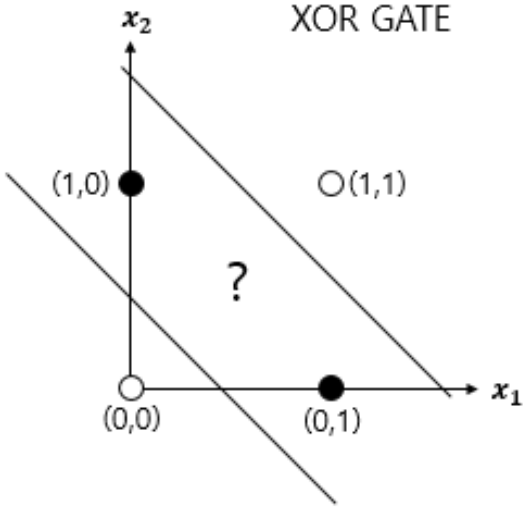
(in 2-dimension space)



High-dimensional Space

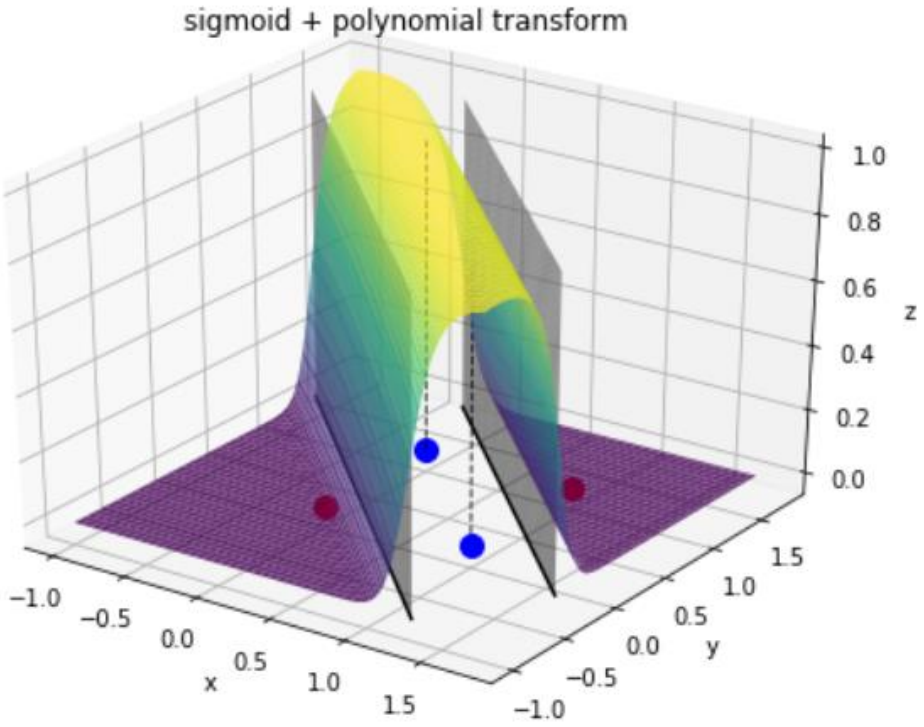


출처: https://vivek-yadav.medium.com/



x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

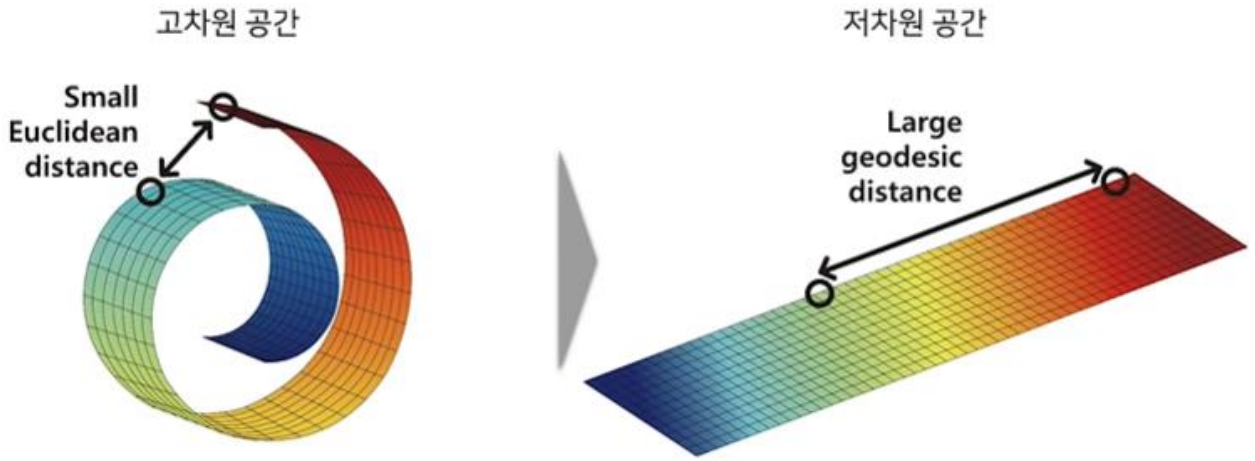
High-dimensional Space



출처: https://datascience.stackexchange.com/questions/11589/creating-neural-net-for-xor-function

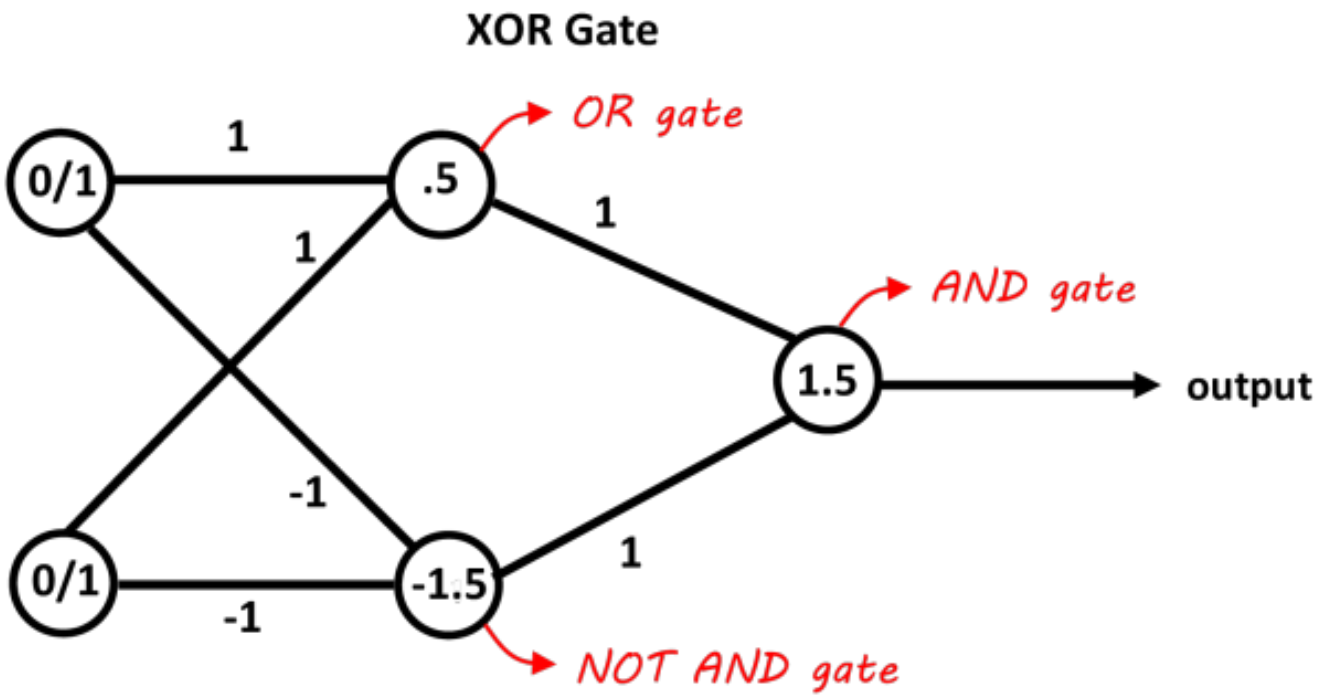
KUBIG DATA SCIENCE & AI

Manifold Hypothesis



- Hypothesis: High-dimensional data tend to lie in the vicinity of a low-dimensional manifold
- We can map data to a high-dimensional space through a smooth representation if the manifold hypothesis holds

출처: https://velog.io/@sheoyonj/NN-Network-3

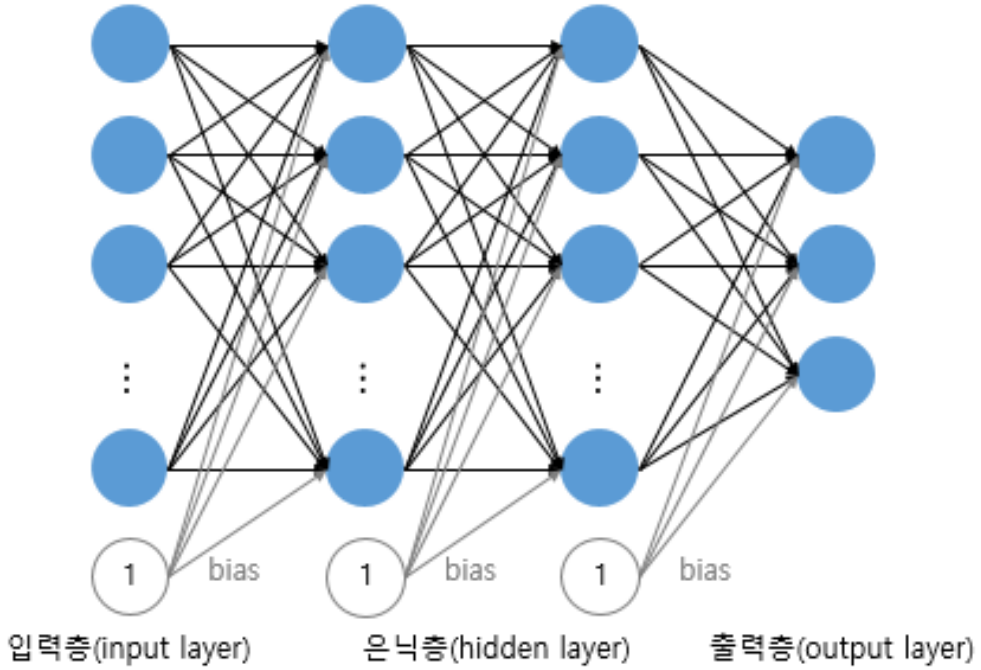


출처: https://datascience.stackexchange.com/questions/11589/creating-neural-net-for-xor-function

KUBIG
DATA SCIENCE & AI

Multi-Layer Perceptron

(More Layers, much more parameters)



출처: https://wikidocs.net/24958

KUBIG DATA SCIENCE & AI

Universal Approximation Theorem

Why deep learning is powerful

Universal approximation theorem — Let $C(X, \mathbb{R}^m)$ denote the set of continuous functions from a subset X of a Euclidean $\mathbb{R}^n$ space to a Euclidean space $\mathbb{R}^m$. Let $\sigma \in C(\mathbb{R}, \mathbb{R})$. Note that $(\sigma \circ x)_i = \sigma(x_i)$, so $\sigma \circ x$ denotes σ applied to each component of x.

Then σ is not polynomial if and only if for every $n \in \mathbb{N}$, $m \in \mathbb{N}$, compact $K \subseteq \mathbb{R}^n$, $f \in C(K, \mathbb{R}^m), \varepsilon > 0$ there exist $k \in \mathbb{N}$, $A \in \mathbb{R}^{k \times n}$, $b \in \mathbb{R}^k$, $C \in \mathbb{R}^{m \times k}$ such that

$$\sup_{x \in K} \|f(x) - g(x)\| < \varepsilon$$

where $g(x) = C \cdot (\sigma \circ (A \cdot x + b))$

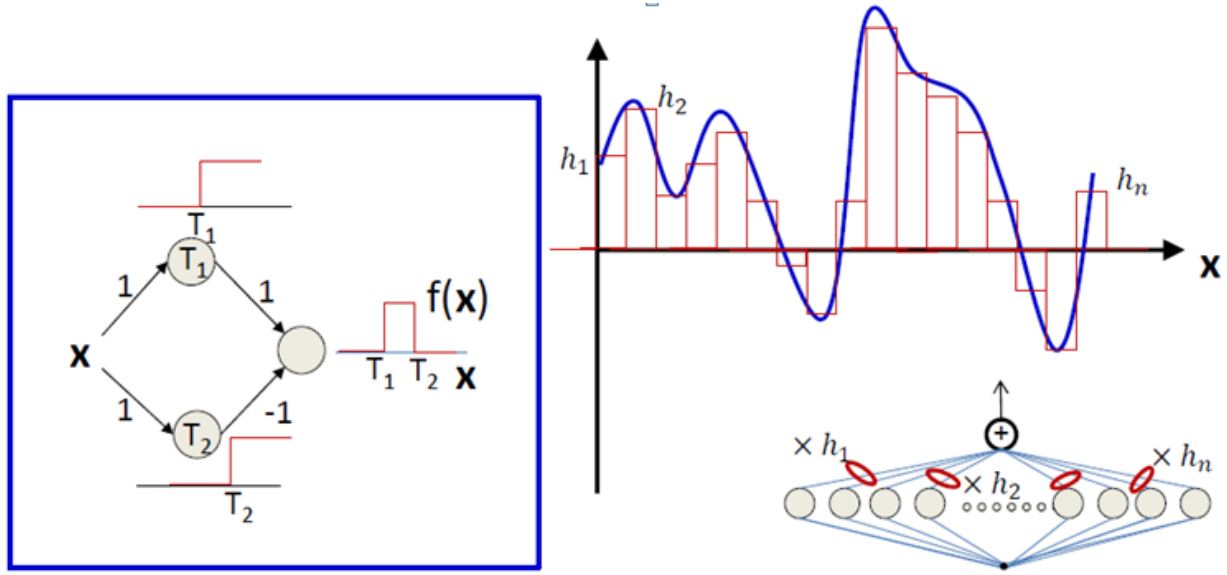
(Universal Approximation Theorem) For a given arbitrary continuous function on a bounded domain and an error bound, there always exists a one-hidden-layer neural network that can approximate the given continuous function within the error bound.

출처: https://en.wikipedia.org/wiki/Universal_approximation_theorem

KUBIG DATA SCIENCE & AI

Universal Approximation Theorem

Why deep learning is powerful



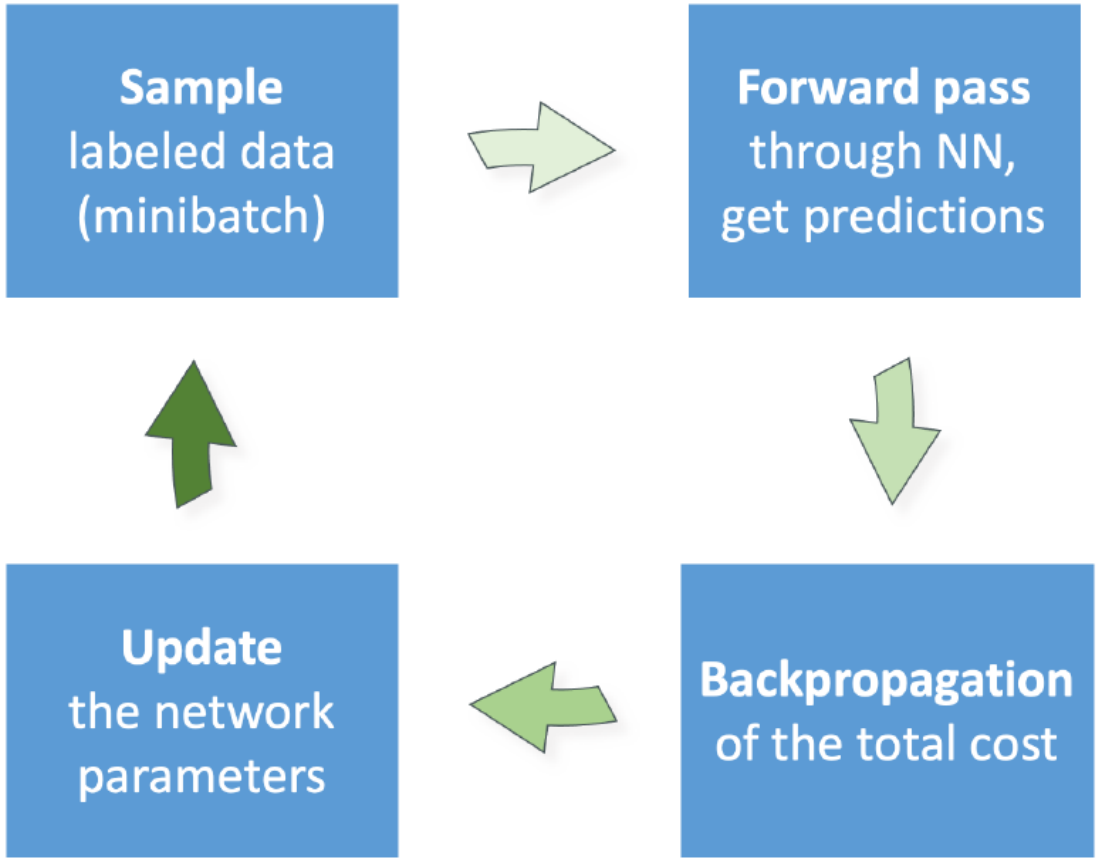
Increase in # of hidden units -> increase in # of parameters
-> increase in # of regions -> increase in # of patterns a function can represent

출처: https://medium.com/analytics-vidhya/

04 Factors of Deep Learning

KUBIG
DATA SCIENCE & AI

4-1. Deep Learning Training Cycle

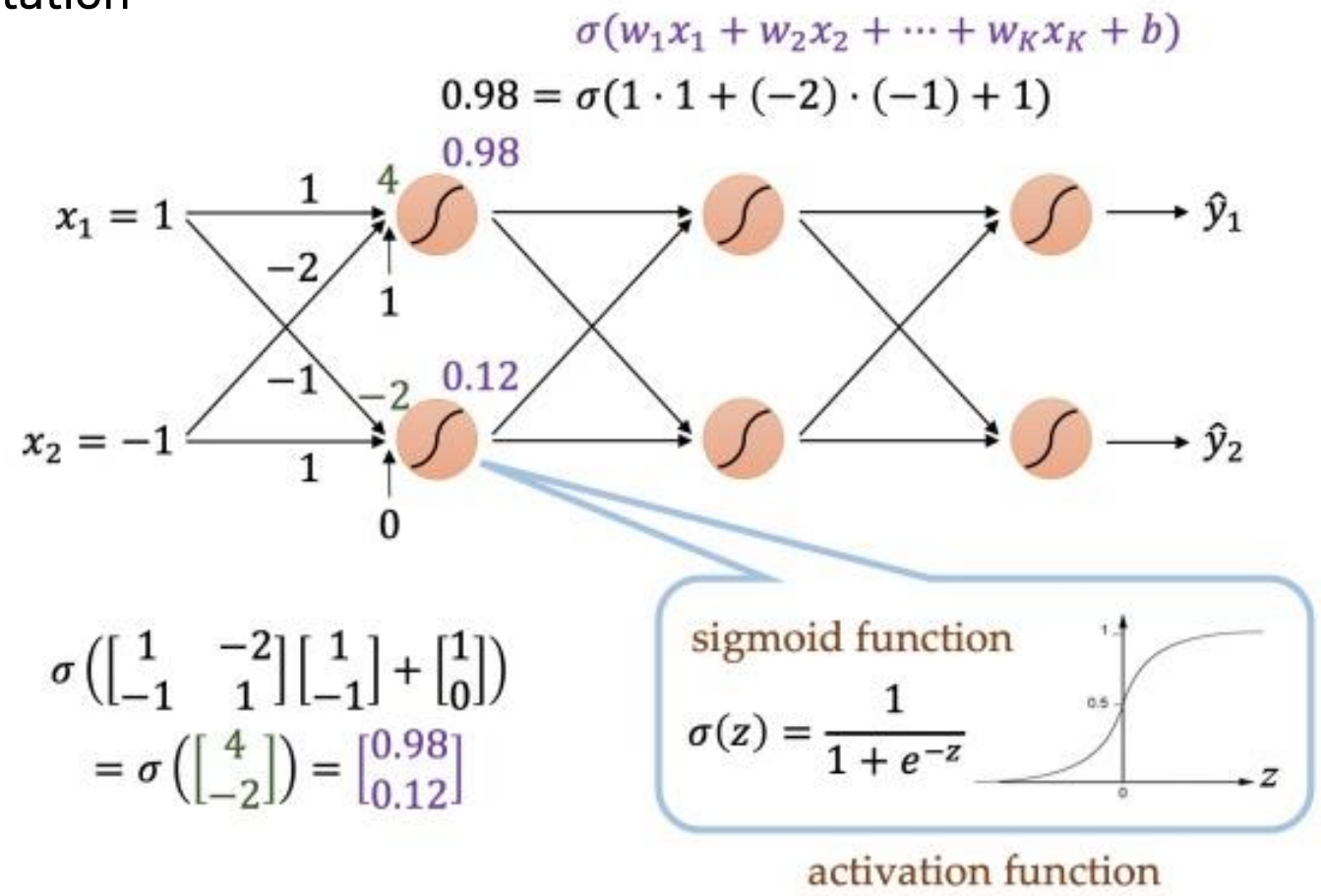


출처: 딥러닝과수학 Seungsang Oh

4-2. Forward Pass

KUBIG DATA SCIENCE & AI

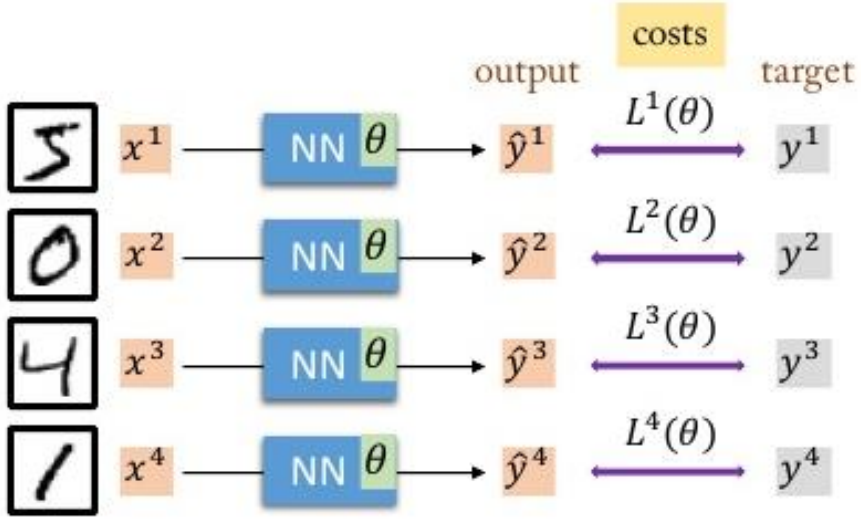
Forward Pass computation



출처: 딥러닝과수학 Seungsang Oh

KUBIG DATA SCIENCE & AI

Total Cost



Total Cost

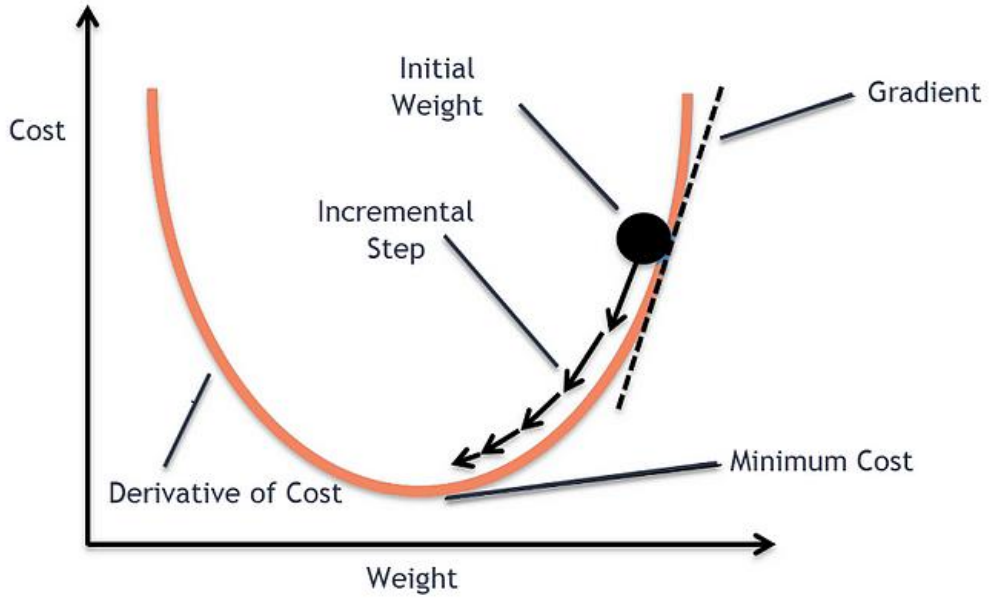
$$C(\theta) = \sum_{n=1}^{N} L^n(\theta)$$

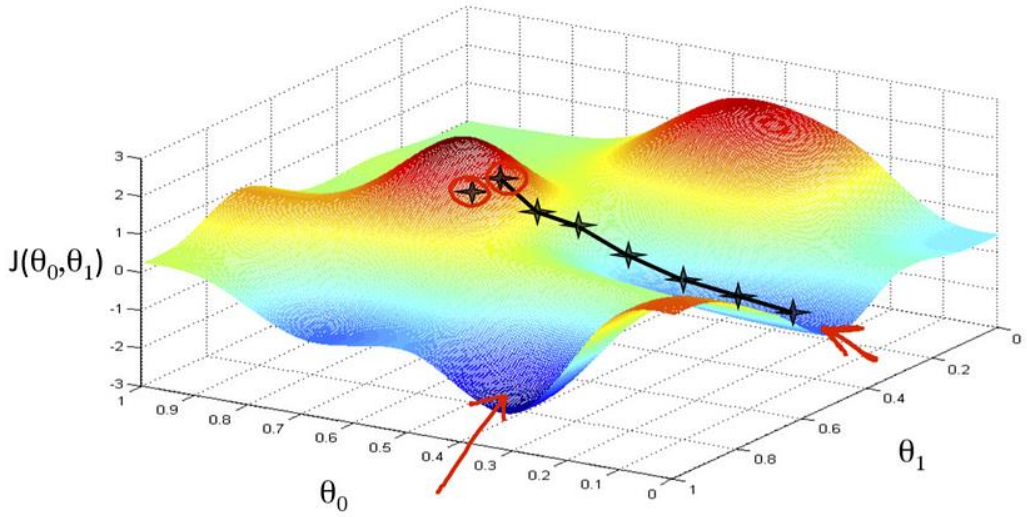
Total cost를 minimize하는 network parameter 찾기

출처: 딥러닝과수학 Seungsang Oh

KUBIG DATA SCIENCE & AI

Gradient Descent

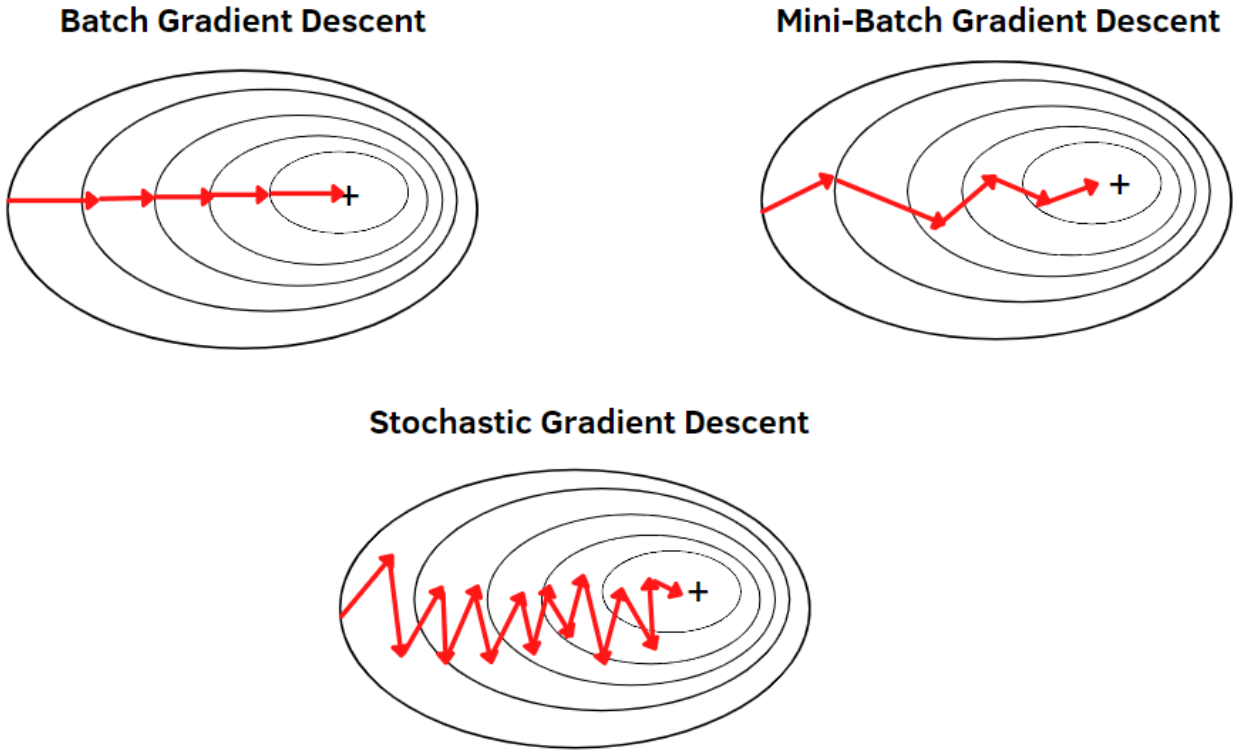




출처: https://towardsdatascience.com/an-intuitive-explanation-of-gradient-descent-83adf68c9c33
https://www.analyticsvidhya.com/blog/2022/07/gradient-descent-and-its-types/

KUBIG DATA SCIENCE & AI

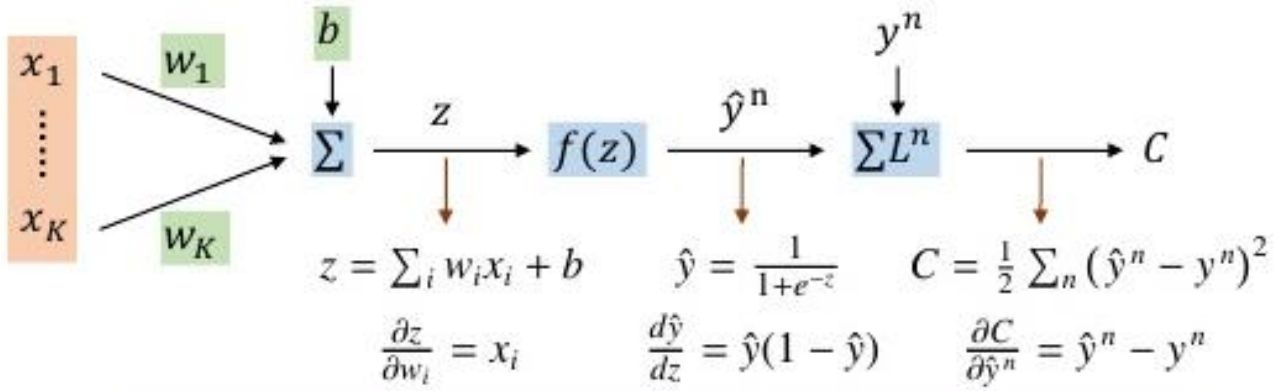
Gradient Descent



출처: https://towardsdatascience.com/an-intuitive-explanation-of-gradient-descent-83adf68c9c33
https://www.analyticsvidhya.com/blog/2022/07/gradient-descent-and-its-types/

4-5. Backpropagation

$$w_i(t+1) = w_i(t) - \eta \frac{\partial C}{\partial w_i}$$



$$\frac{\partial C}{\partial w_i} = \sum_{n=1}^{N} \frac{\partial z^n}{\partial w_i} \frac{d\hat{y}^n}{dz^n} \frac{\partial C}{\partial \hat{y}^n} = \sum_{n=1}^{N} x_i^n \hat{y}^n (1-\hat{y}^n)(\hat{y}^n - y^n)$$

Chain rule

출처: 딥러닝과수학 Seungsang Oh

05 과제 설명

1주차 복습과제, 예습과제

KUBIG
DATA SCIENCE & AI

4-1. 예습과제, 복습과제

KUBIG
DATA SCIENCE & AI

1주차 복습과제1	1주차 복습과제2	1주차 예습과제1	1주차 예습과제2
Pytorch 기본 익히기 code	내용 복습 / 추론	텍스트 전처리 + 워드클라우드 code	Word2vec code

코드과제의 파일형식은 ipynb로, KUBIG 24-2 **Github** repo에 업로드 될 예정입니다!
Colab 환경에서 제작된 과제들이므로 google colab에서 실행하시는 것을 권장드립니다.

수고하셨습니다