



Artificial Intelligence Tutorial

박 경 규

강사소개



- 소속 : IT부문 SW개발본부
- 이름 : 박경규
- 강의경력: 딥러닝, 파이썬, 에너지솔루션
- 깃허브 주소 : <https://github.com/kgpark88>
- 개발경력

PMS(Project Management System) 개발

ADD플랫폼(AI Data Discovery Platform) 개발

BEMS(Building Energy Management System) 개발

- 관심분야 : Self Development, Book, Overseas Travel, Investment

목 차

1. 코드(CODE)
2. 인공지능 개요
3. 데이터 분석과 시각화
4. 머신러닝 핵심 알고리즘
5. 스타트 딥러닝
6. Further Study

학습 목표

AI 분야에서 반드시 알아야 하는 핵심 개념을 단계적으로 이해합니다.

데이터를 수집하고 분석이 가능한 형태로 정리하는 기술을 습득합니다.

머신러닝 알고리즘과 딥러닝 심층신경망 원리를 파악하고 활용합니다.

인공지능 기술을 직접 실무에 활용하며 AI 전문가로 성장합니다.

자료

실습 파일 <https://github.com/kgpark88/ai-summary>

파이썬 스터디
참고자료 <https://github.com/kgpark88/python>

딥러닝 스터디
참고자료 <https://github.com/kgpark88/deeplearning>

주관식

다음 유튜브 신조어의 뜻을 아는 대로 쓰시오.

① 불소

⑦ 설참

② 실매

⑧ 닉차

③ 톡디

⑨ 전공

④ 반신

⑩ 임구

⑤ 반박

⑪ 짙테

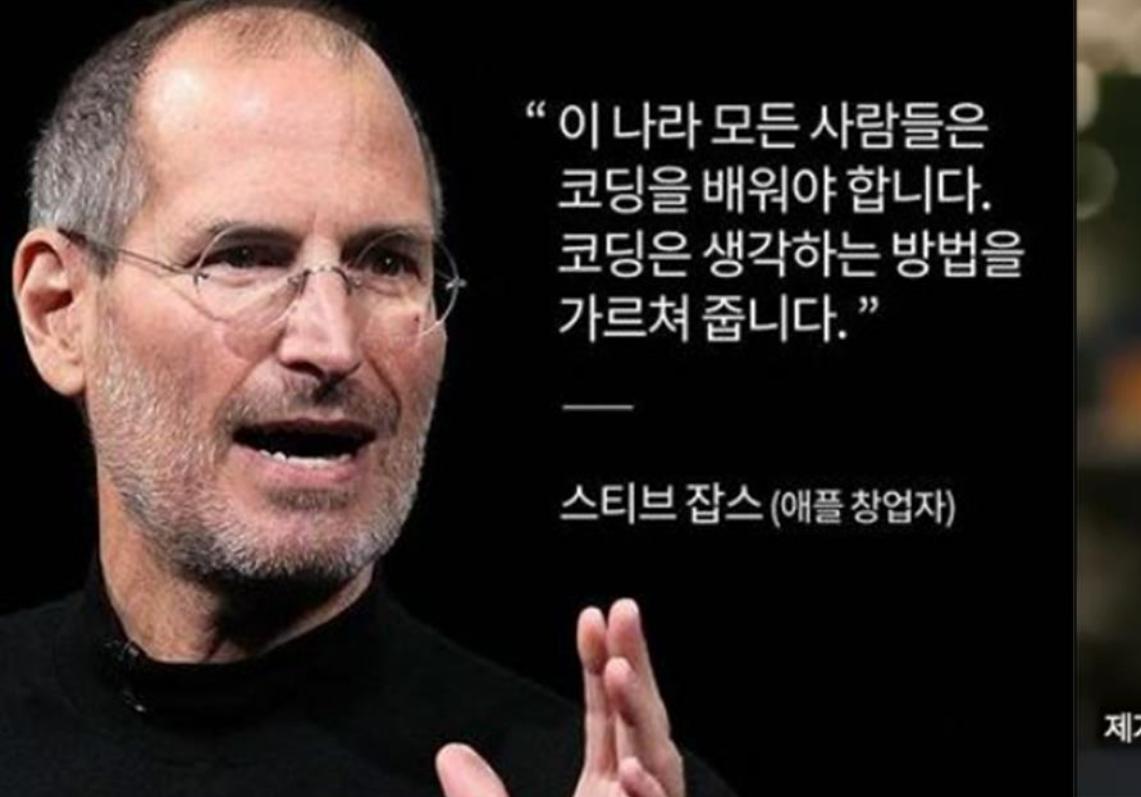
⑥ 윰차

⑫ 구취

1. 코드(CODE)

컴퓨터 코드(computer code) 또는 **프로그램 코드**(program code)는 컴퓨터가 실행하는 컴퓨터 프로그램을 구성하는 명령어들의 모임이다. 컴퓨터 하드웨어에서 실행되는 소프트웨어의 두 요소 가운데 하나이며, 다른 하나는 데이터이다.

출처 : <https://ko.wikipedia.org/>



“이 나라 모든 사람들은
코딩을 배워야 합니다.
코딩은 생각하는 방법을
가르쳐 줍니다.”

스티브 잡스 (애플 창업자)

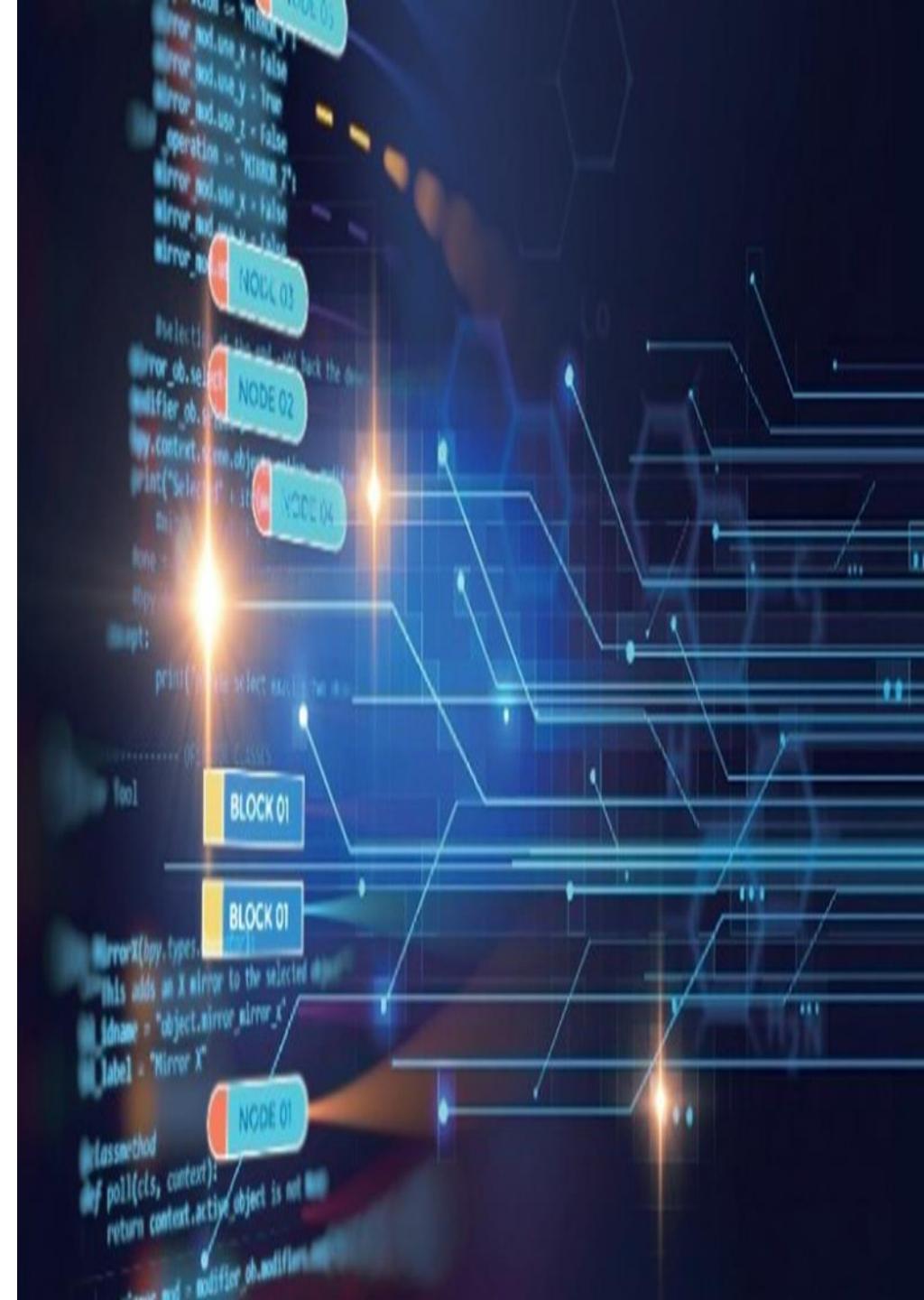


차세대 프로그래머는 미래의 마법사입니다. 다른 사람과 비교했을 때 마치 마법 능력이 있는 것처럼 보여질 거에요.



프로그래미란?

- 컴퓨터 프로그램은 특정 문제를 해결하기 위해 고안된 **특정 작업을 수행하기 위한 일련의 명령문의 집합체**
- 스마트폰, 태블릿 등에서는 ‘앱’이라는 용어를 사용
- 소프트웨어는 하드웨어의 반대 개념으로서의 의미이지만, 일반적으로는 프로그램과 같은 의미
- 프로그래밍은 주어진 문제를 해결하기 위해 컴퓨터 프로그램을 만들고 실행하는 전 과정
절차 : 문제분석 → 입출력설계 → 알고리즘설계
→ 코딩 → 프로그램실행



코딩이란?

- 코딩[Coding]은 컴퓨터 프로그램 언어로 프로그램을 작성하는 것 입니다.
좁은 의미의 프로그래밍입니다.
- 영어를 배우는 것 만큼 코딩을 배우는 게 중요합니다.
- 프로그램 언어 규칙에 따라 글을 쓰는 것만으로 컴퓨터에게 원하는 일을 시킬 수 있습니다.
- 코딩교육을 통해 소프트웨어시대의 경쟁력을 갖출 수 있게 될 것입니다.

코딩은 문제해결을 위한
절차적 사고와 논리적 사고력을
키워 줍니다.



코딩을 배워야 하는 이유

- 소프트웨어의 발전, 특히 소프트웨어를 만드는 소프트웨어 개발툴의 발전으로
모든 사람이 코딩을 할 수 있는 시대가 되었습니다.
- 코딩은 프로그래머 직종에 한정돼 있지 않으며,
데이터 분석, 과학, 의학, 엔지니어링, 영업, 농작물 재배 등에도 코딩 능력이 필요합니다.
- 전 산업이 소프트웨어를 이용해서 자동화 지능화로 혁신하고 있으며,
모든 소프트웨어는 코딩으로 만들어집니다.



You Can Create Anything You Want



Instant Scalability



Good Income

파이썬(Python)

<https://www.python.org/>

The screenshot shows the Python.org homepage. At the top, there's a navigation bar with links for Python, PSF, Docs, PyPI, Jobs, and Community. Below the header is the Python logo and a search bar with buttons for 'Donate', 'Search', 'GO', and 'Socialize'. A navigation menu below the header includes links for About, Downloads, Documentation, Community, Success Stories, News, and Events. On the left, there's a code snippet in a terminal window:

```
# Python 3: Fibonacci series up to n
>>> def fib(n):
    >>>     a, b = 0, 1
    >>>     while a < n:
    >>>         print(a, end=' ')
    >>>         a, b = b, a+b
    >>>     print()
    >>> fib(1000)
0 1 1 2 3 5 8 13 21 34 55 89 144 233 377 610 987
```

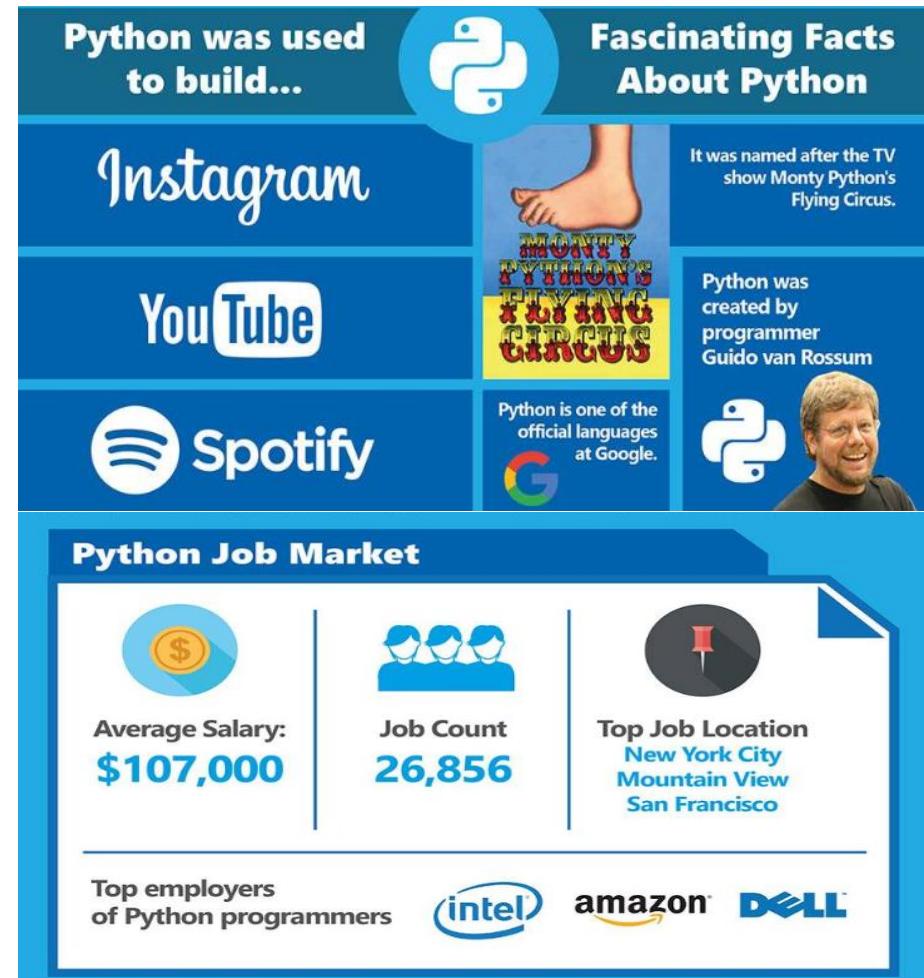
To the right of the code, there's a section titled "Functions Defined" with the following text:

The core of extensible programming is defining functions. Python allows mandatory and optional arguments, keyword arguments, and even arbitrary argument lists. [More about defining functions in Python 3](#)

At the bottom of the page, there's a footer with a call to action: "Python is a programming language that lets you work quickly and integrate systems more effectively. [» Learn More](#)".

파이썬이 유명한 이유

- ✓ Easy to Learn and Use
- ✓ Mature and Supportive Python Community
- ✓ Support from Renowned Corporate Sponsors
- ✓ Hundreds of Python Libraries and Frameworks
- ✓ Versatility, Efficiency, Reliability, and Speed
- ✓ Big data, Machine Learning and Cloud Computing
- ✓ First-choice Language
- ✓ The Flexibility of Python Language
- ✓ Use of python in academics
- ✓ Automation



파이썬 주요 라이브러리(패키지)



NumPy

행렬과 다차원 배열을 쉽게 처리 할 수 있게 해주는 라이브러리

pandas

데이터를 처리하고 분석하는 데 효과적인 라이브러리



데이터를 차트나 플롯(Plot)으로 그려주는 시각화 라이브러리



matplotlib을 기반으로 다양한 색상 테마와 통계용 차트 등의 기능을 추가한 시각화 라이브러리



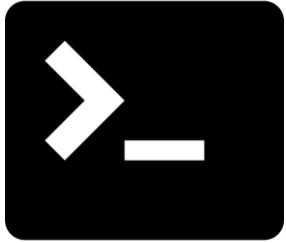
교육 및 실무를 위한 머신러닝 라이브러리



구글에서 만든 오픈소스 딥러닝 라이브러리

개발환경

Terminal

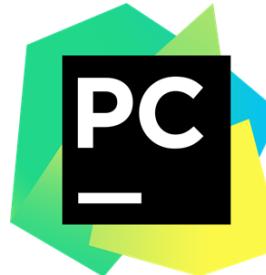


Code Editor



Visual Studio Code

IDE



PyCharm

Notebook



Jupyter
Notebook

WEB



<https://colab.research.google.com/>



Sublime Text



Spyder



JupyterLab



<https://repl.it/>



<https://glot.io/>

개발환경 - Chrome

컴퓨터에 Chrome이 설치되어 있지 않은 경우, [Chrome을 다운로드하여 설치](#)하고 기본 웹브라우저를 Chrome로 설정하세요.

기본 웹브라우저로 Chrome 설정

Windows 10



1. 컴퓨터에서 시작 메뉴 를 클릭합니다.
2. 설정 을 클릭합니다.
3. 기본 앱을 엽니다.
 - 기존 버전: 시스템 > 기본 앱을 클릭합니다.
 - 크리에이터스 업데이트: 앱 > 기본 앱
4. 하단의 '웹 브라우저'에서 현재 브라우저를 클릭합니다. 일반적으로 Microsoft Edge입니다.
5. '앱 선택' 창에서 **Chrome**을 클릭합니다.

나중에 손쉽게 Chrome을 열려면 작업 표시줄에 단축키를 추가하세요.

1. 컴퓨터에서 Chrome을 엽니다.
2. 하단의 Windows 작업 표시줄에서 Chrome을 마우스 오른쪽 버튼으로 클릭합니다.
3. 작업 표시줄에 고정을 클릭합니다.

출처 : <https://bit.ly/30DvgKY>

개발환경 - 코랩(Colab)

개발툴 설치없이 구글클라우드에서 데이터분석과 AI 모델을 개발할 수 있는 환경으로 딥러닝에 필요한 GPU를 사용할 수 있습니다.

<https://colab.research.google.com>

구글 계정 필요

The screenshot shows the Google Colab interface. At the top, there's a blue header bar with the URL and an orange bar that says "구글 계정 필요". Below the header is the Colaboratory menu bar with "런타임" (Runtime) selected. A dropdown menu for "런타임" is open, showing various options like "모두 실행" (Run All), "이전 셀 실행" (Run Previous Cell), and "선택항목 실행" (Run Selected Cells). In the middle of the screen, a modal window titled "Colaboratory의 즈오 기능을 가다하게 알아보세요" (Discover Colaboratory's cool features) is displayed. It has a large blue arrow pointing to the right. On the right side of the screen, there's a small image of a yellow TensorFlow logo with the text "Coding TensorFlow".

Colaboratory에 오신 것을 환영합니다

파일 수정 보기 삽입 런타임 도구 도움말

+ 코드 + 텍스트 드라이브

목차 코드 스니펫

Colaboratory 소개

시작하기

추가 리소스

머신러닝 예제: Seedbank

섹션

모두 실행 Ctrl+F9

이전 셀 실행 Ctrl+F8

초점이 맞춰진 셀 실행 Ctrl+Enter

선택항목 실행 Ctrl+Shift+Enter

이후 셀 실행 Ctrl+F10

실행 중단 Ctrl+M I

런타임 다시 시작 Ctrl+M R

다시 시작 및 모두 실행

런타임 초기화

런타임 유형 변경

세션 관리

런타임 로그 보기

Google Colab

Colaboratory의 즈오 기능을 가다하게 알아보세요

노트 설정

하드웨어 가속기 GPU

Colab를 최대한 활용하려면 필요하지 않은 경우 GPU를 사용하지 않는 것이 좋습니다.
[자세히 알아보기](#)

이 노트를 저장할 때 코드 셀 출력 생략

취소 저장

Coding TensorFlow



고성능GPU(Graphics Processing Unit)

개발환경 - Jupyter Notebook 단축기

- Esc : 커맨드 모드로 진입
- Enter : 선택 셀의 코드 입력 모드로 돌아가기
- Shift + Enter : 셀 실행, 아래 셀 선택
- Alt + Enter : 셀 실행, 아래에 셀 추가
- Ctrl + Enter : 셀 실행
- Tab : 함수 자동완성
- Shift + Tab : 함수 설명 보기
- D + D : 선택 셀 삭제
- M : Markdown으로 변경
- Y : Code로 변경
- A : 위에 셀 추가
- B : 아래에 셀 추가
- Shift + Down : 현재 셀과 아래 셀을 같이 선택
- Shift + Up : 현재 셀과 위의 셀을 같이 선택
- Shift + M : 선택 셀과 아래 셀과 합치기
- Ctrl + S : 파일 저장

파이썬 기초

■ 변수 할당(Variable Assignment)

```
x = 2
```

```
y = 3
```

```
z = x + y
```

```
x = 'hello'
```

```
x = "hello"
```

```
x
```

```
[Out] 'hello'
```

Single Quotation
작은 따옴표

Double Quotation
쌍 따옴표

■ 출력

```
print(x)
```

```
[Out] 'hello'
```

■ 리스트(List)

```
[1, 2, 3]
```

```
['a', 'b', 'c']
```

```
my_list = [1, 2, 'apple', True]
```

```
my_list.append(100)
```

```
my_list[0]
```

```
my_list[:-1]
```

```
my_list[-1]
```

Bracket
대괄호

■ 딕셔너리(Dictionary)

```
d = {'key1': 'item1', 'key2': 'item2'}
```

```
d['key1']
```

```
[Out] 'item1'
```

Brace
중괄호

파이썬 실습



https://github.com/kgpark88/ai-summary/blob/main/01_Python.ipynb

kgpark88 / ai-summary

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main / ai-summary / 01_Python.ipynb Go to file ...

kgpark88 Colaboratory를 통해 생성됨 Latest commit 59fc46e 6 minutes ago History

1 contributor

2413 lines (2413 sloc) | 49.5 KB

Open in Colab

파이썬 기본

- 데이터 타입
 - 숫자(Number)
 - 문자열(String)

2. AI 개요

인공지능 또는 AI는 인간의 학습능력, 추론능력, 지각능력,
그 외에 인공적으로 구현한 컴퓨터 프로그램 또는 이를 포함한 컴퓨터 시스템이다.

출처 : <https://ko.wikipedia.org/>

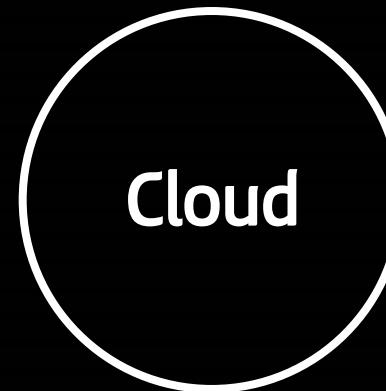
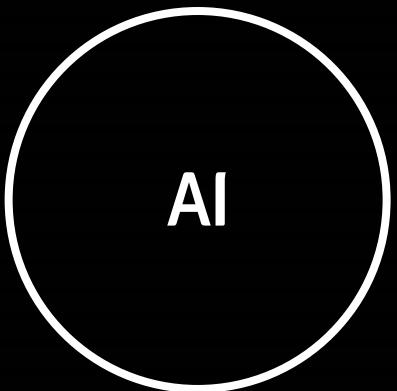
Tech and Change

증기기관
내연기관

전기
에너지

컴퓨터
인터넷

Tech and Change



iamai





인공지능 활용사례 - 이미지 분류

이미지넷(ImageNet) 제공 이미지 데이터
1,000여 카테고리로 분류된 100만 개의 이미지

airplane



automobile



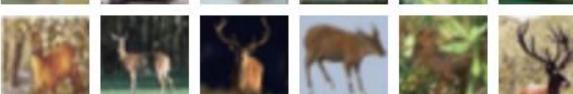
bird



cat



deer



dog



frog



horse



ship

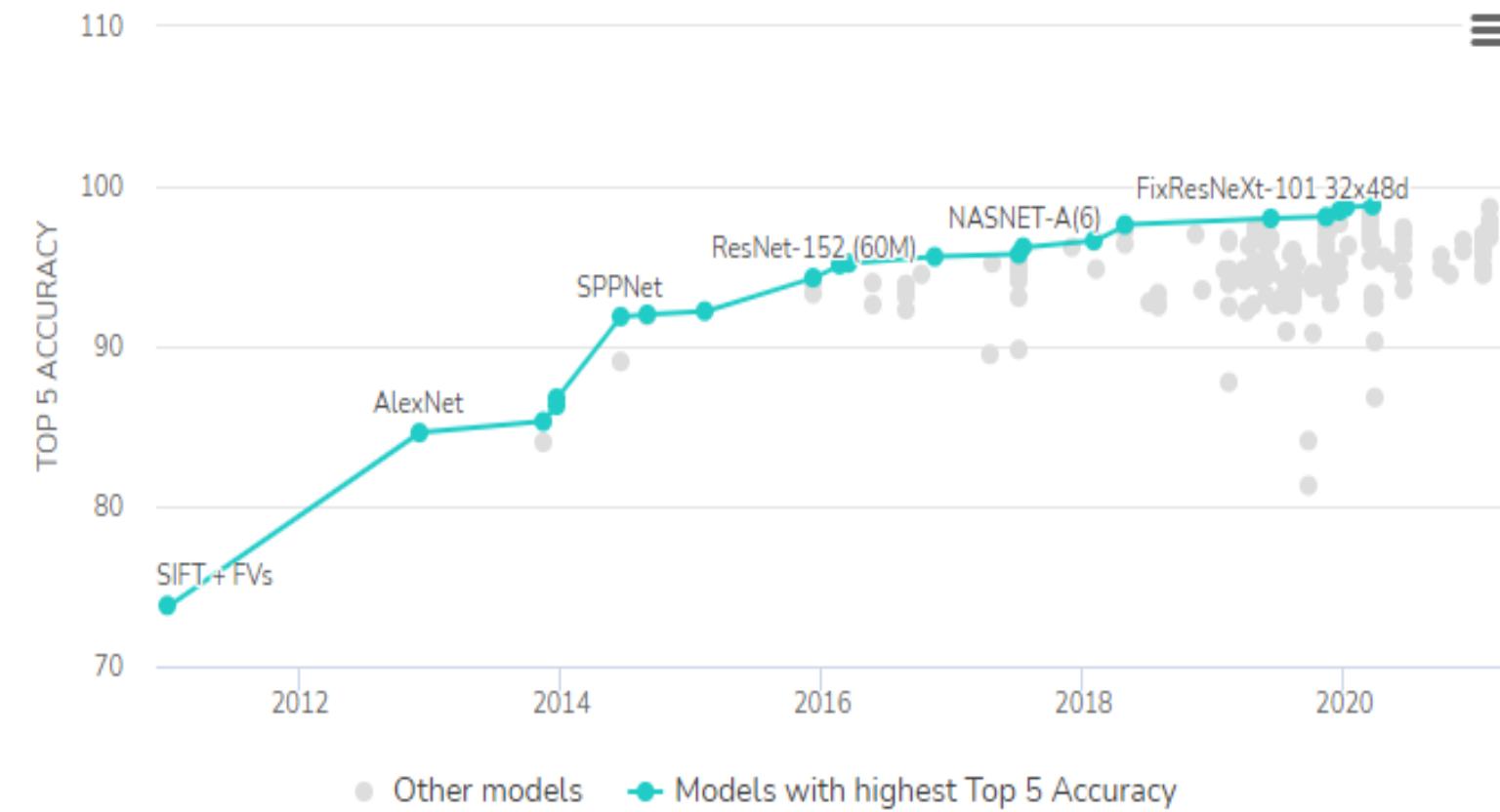


truck

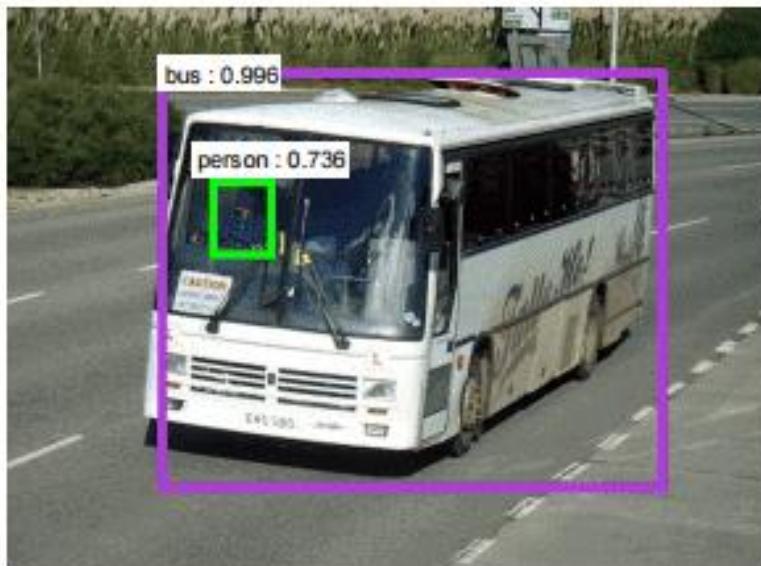
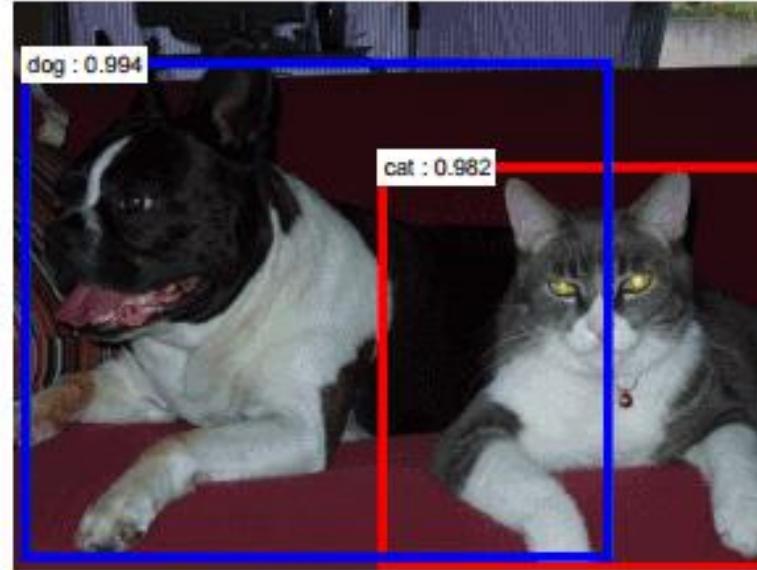
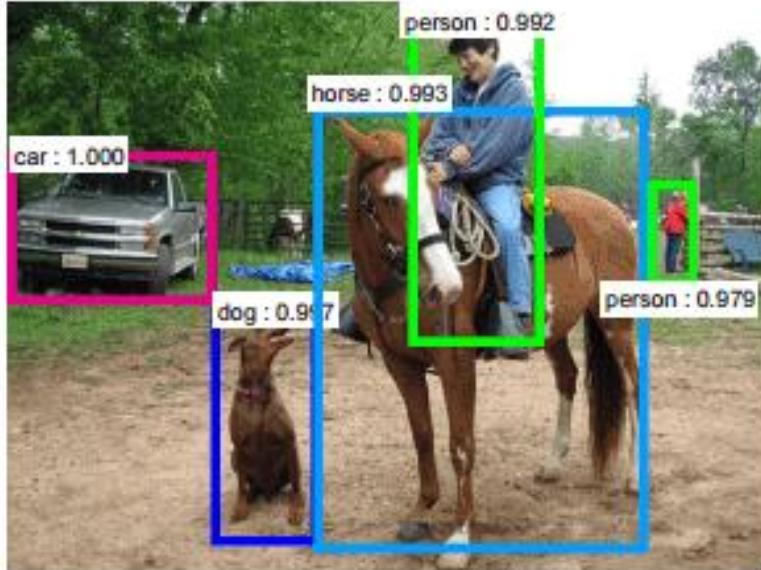


Leaderboard

Dataset



인공지능 적용사례 - 객체 탐지(Object Detection)



출처 : <https://sigmoidal.io/dl-computer-vision-beyond-classification>

인공지능 활용사례 - 이미지 생성(Style Transfer)

ORIGINAL PHOTO



REWORKED PHOTO



ORIGINAL PHOTO



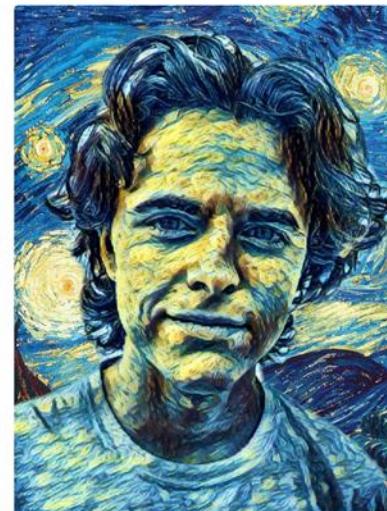
REWORKED PHOTO



ORIGINAL PHOTO



REWORKED PHOTO



ORIGINAL PHOTO



REWORKED PHOTO



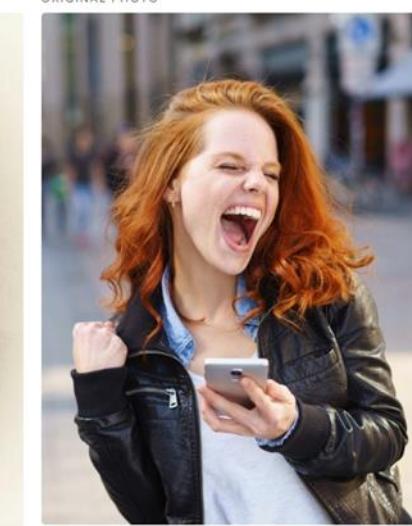
ORIGINAL PHOTO



REWORKED PHOTO



ORIGINAL PHOTO



REWORKED PHOTO



인공지능 활용사례 - 이미지 생성(GAN: generative adversarial network)



Original



Change Hair Color



Change Eye Color



Change Hair Style

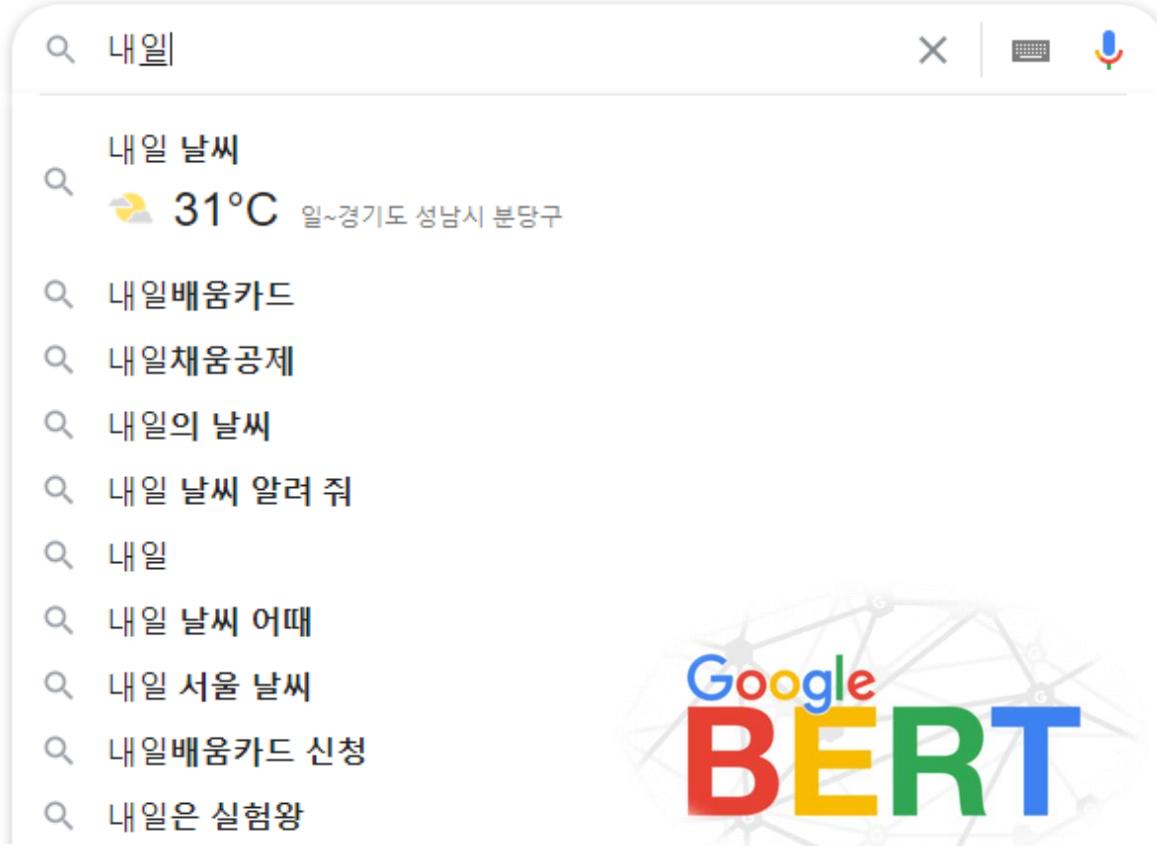


Open Mouth



Add Assesories

인공지능 활용사례 - 자연어 처리



인공지능 활용사례 - Improving our world with AI



인공지능(Artificial Intelligent)



인공 지능

인간의 지적능력(추론, 인지)을 구현하는 모든 기술

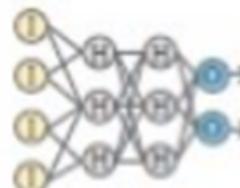


머신 러닝

알고리즘으로 데이터를 분석, 학습하여 판단이나 예측을 하는 기술

선형회귀
로지스틱회귀
K-최근접 이웃
결정트리
랜덤포레스트
서포트 벡터 머신

클러스터링 차원축소



딥러닝

인공신경망 알고리즘을 활용하는 머신러닝 기술

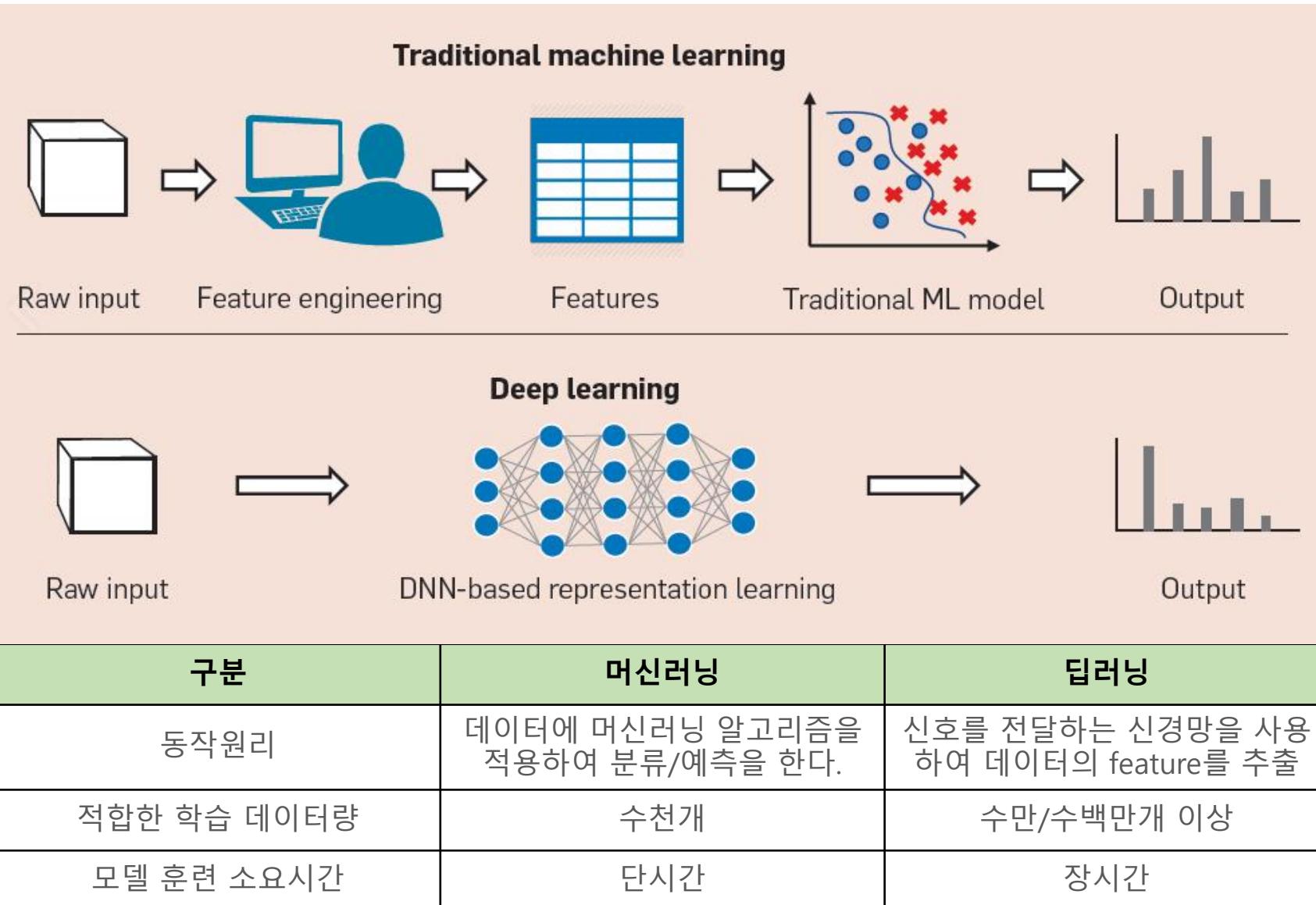
심층신경망
(DNN)

합성곱 신경망
(CNN)

순환 신경망
(RNN)

강화 학습

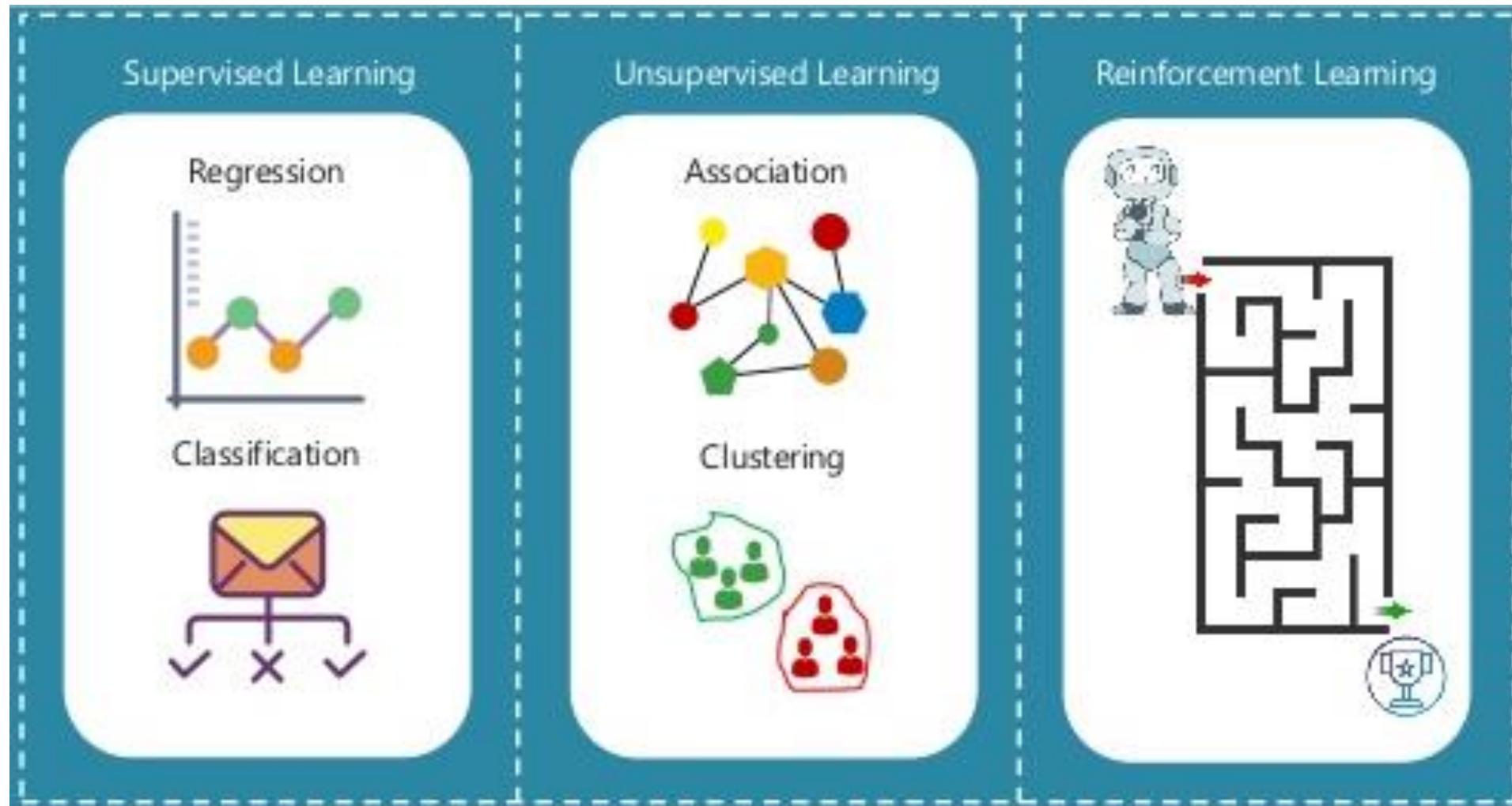
머신러닝 VS 딥러닝



머신러닝에서는 데이터로부터 속성(Feature)을 찾아내는 역할을 컴퓨터(Machine)가 담당

딥러닝에서는 신경망으로 데이터/이미지를 ‘있는 그대로’ 학습하며, 데이터에 포함된 중요한 속성을 컴퓨터가 스스로 학습

머신러닝/딥러닝 학습 방법



정답지(Label)로 학습
분류(Classification)
예측(Regression)

정답지(Label) 없이 학습
군집(Clustering)
차원 축소

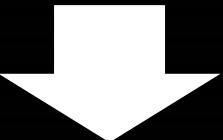
시뮬레이션 반복 학습
성능 강화 등에 사용
마르코프 결정 과정(Markov Decision Process)

AI 시대의 경쟁력

문제의 본질을 파악하는 능력과 데이터를 만드는 능력이 중요

인공지능을 활용하여 기존의 나의 일을 효율화 하는 것이 실력

AI를 활용하여 기존의 일을 효율적으로 바꾸는 일을 주도하는 것이 경쟁력



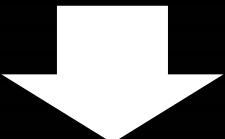
Let's start my game!

3. 데이터 분석과 시각화

빠르게 쌓여가는 방대한 데이터로 빅데이터(Big Data) 시대

빅데이터 저장, 처리, 분석에 필요한 컴퓨팅 자원을 저렴한 비용으로 사용 가능

컴퓨팅 파워의 대중화는 최적의 학습 환경과 연구 인프라를 제공



데이터 자체가 가장 중요한 자원이며, 데이터를 수집하고 분석이 가능한 형태로 정리하는 것이 중요

관점의 변화가 필요하다.

데이터 활용

데이터 분석을 통한 인사이트 발굴 및 의사결정에 활용



기존 데이터를 사용해서 새로운 가치가
더해진 데이터(Value Added DATA)를 만들어 낸다. ('예측 확률 값')

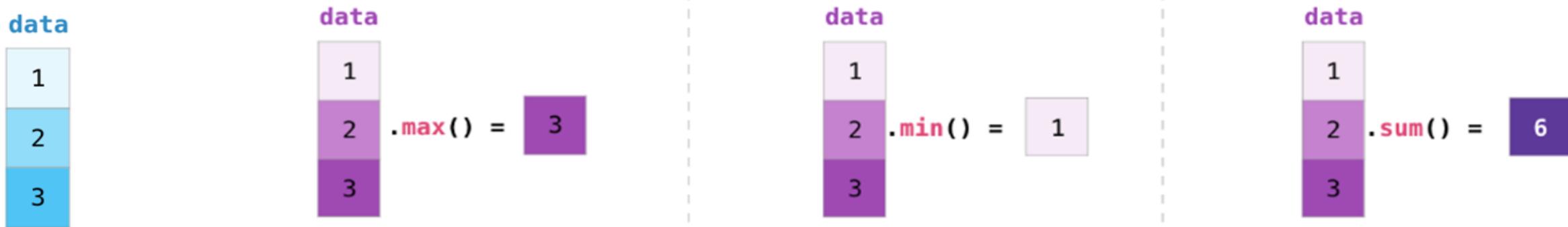


새로운 가치 데이터를 바탕으로 비즈니스적으로 의미 있는 변화를 만든다.
기존 업무와 프로세스를 변경하여 매출 증대, 비용 절감, 효율 증대 등의 변화를 만든다.

넘파이(Numpy)

NumPy(Numerical Python)는 데이터 분석, 수학/과학연산을 위한 파이썬 기본 패키지로 고성능의 다차원 배열 객체와 다양한 객체에 대해 고속 연산을 가능하게 합니다.

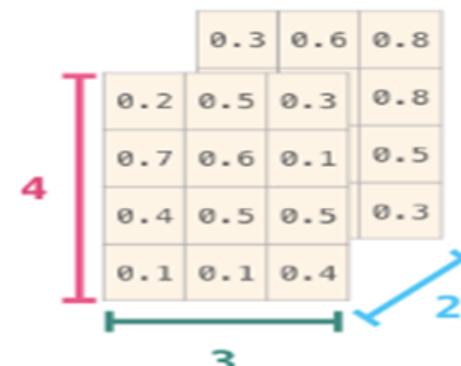
```
data = np.array([1,2,3])
```



```
np.array([[1,2],[3,4],  
         [[5,6],[7,8]]])
```

| | | |
|---|---|---|
| 1 | 2 | 5 |
| 3 | 4 | 6 |
| | | 8 |

```
np.random.random((4,3,2))
```



넘파이(Numpy)



https://github.com/kgpark88/ai-summary/blob/main/02_Numpy.ipynb

■ Numpy 라이브러리 임포트

```
import numpy as np
```

■ Numpy Array 생성

```
my_list = [1, 2, 3]  
np.array(my_list)  
[Out] array([1, 2, 3])
```

```
my_matrix = [[1, 2, 3], [4, 5, 6], [7, 8, 9]]  
np.array(my_matrix)  
[Out] array([[1, 2, 3],  
           [4, 5, 6],  
           [7, 8, 9]])
```

데이터 분석 필수 라이브러리 판다스(Pandas)

판다스(Pandas)는 데이터 처리와 분석을 위해 널리 사용되는 파이썬 라이브러리

데이터사이언티스트에게 필요한 기본적이면서도 아주 중요한 도구

행과 열로 이루어진 데이터 객체를 만들어 다룰 수 있음

데이터를 수집하고 정리하는 데 최적화 된 도구



판다스는 시리즈(Series)와 데이터프레임(DataFrame)이라는 구조화된 데이터 형식을 제공

시리즈(Series) : 1차원 배열

데이터프레임(DataFrame) : 2차원 배열

판다스 시리즈(Series)

1차원의 배열의 값(values)과 각 값에 대응하는 인덱스(index)를 부여할 수 있는 데이터 구조

```
import pandas as pd
```

```
sr = pd.Series([20000, 18000, 5000])  
print(sr)
```

| index | values |
|--------------|--------|
| 피자 | 20000 |
| 치킨 | 18000 |
| 맥주 | 5000 |
| dtype: int64 | |

```
sr = pd.Series([20000, 18000, 5000], index = ['피자', '치킨', '맥주'])  
print(sr.index)  
print(sr.values)  
print(sr)
```

```
sr = pd.Series({'피자': 20000, '치킨': 18000, '맥주': 5000})  
print(sr)
```

판다스 데이터프레임(DataFrame)

데이터프레임은 행과 열을 가지는 자료구조로 인덱스(index), 열(columns), 값(values)으로 구성

열(columns)

| | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService |
|----------------|------------|--------|---------------|---------|------------|--------|--------------|------------------|-----------------|
| 인덱스 (index) | | | | | | | | | |
| 0 | 7590-VHVEG | Female | 0 | Yes | No | 1 | No | No phone service | DSL |
| 1 | 5575-GNVDE | Male | 0 | No | No | 34 | Yes | No | DSL |
| 2 | 3668-QPYBK | Male | 0 | No | No | 2 | Yes | No | DSL |
| 3 | 7795-CFOCW | Male | 0 | No | No | 45 | No | No phone service | DSL |
| 4 | 9237-HQITU | Female | 0 | No | No | 2 | Yes | No | Fiber optic |

값(values)

판다스 실습

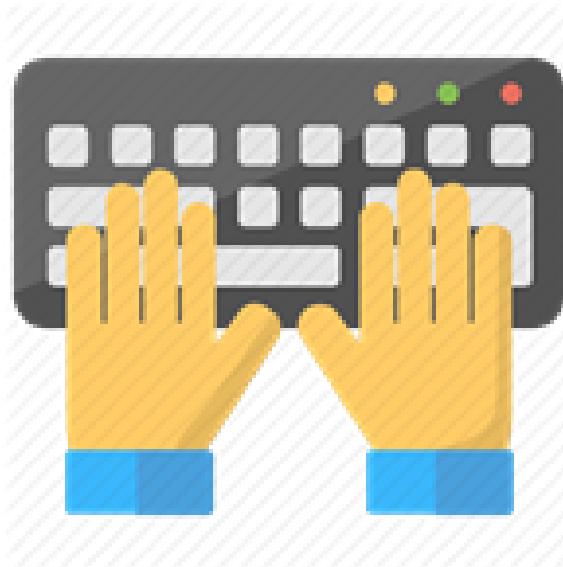


<https://github.com/kgpark88/ai-summary>

- [03_01_Pandas_Series.ipynb](#)
- [03_02_Pandas_DataFrame.ipynb](#)
- [03_03_Pandas_MissingData.ipynb](#)
- [03_04_Pandas_Groupby.ipynb](#)
- [03_05_Pandas_DataInputOutput.ipynb](#)
- [03_06_Pandas_Operation.ipynb](#)

판다스 Exercise

<https://bit.ly/3bnwEHT>





Question 1 – Define Python Pandas.

Question 2 – What Are The Different Types Of Data Structures In Pandas?

Question 6 – What Are The Most Important Features Of The Pandas Library?

**Question 8 – What are the different ways of creating DataFrame in pandas?
Explain with examples.**

Question 9 – Explain Categorical Data In Pandas?

Question 14 – How Can You Iterate Over Dataframe In Pandas?

Question 17 – What Is Groupby Function In Pandas?

데이터 분석 실습 - 타이타닉 데이터셋



https://github.com/kgpark88/ai-summary/blob/main/04_DataAnalysis_Titanic.ipynb

■ Seaborn 라이브러리 임포트

```
import seaborn as sns
```

■ 파일에서 데이터를 로드

```
df = sns.load_dataset('titanic')
```

■ 데이터 확인

```
df.head()
```

```
df.head(20)
```

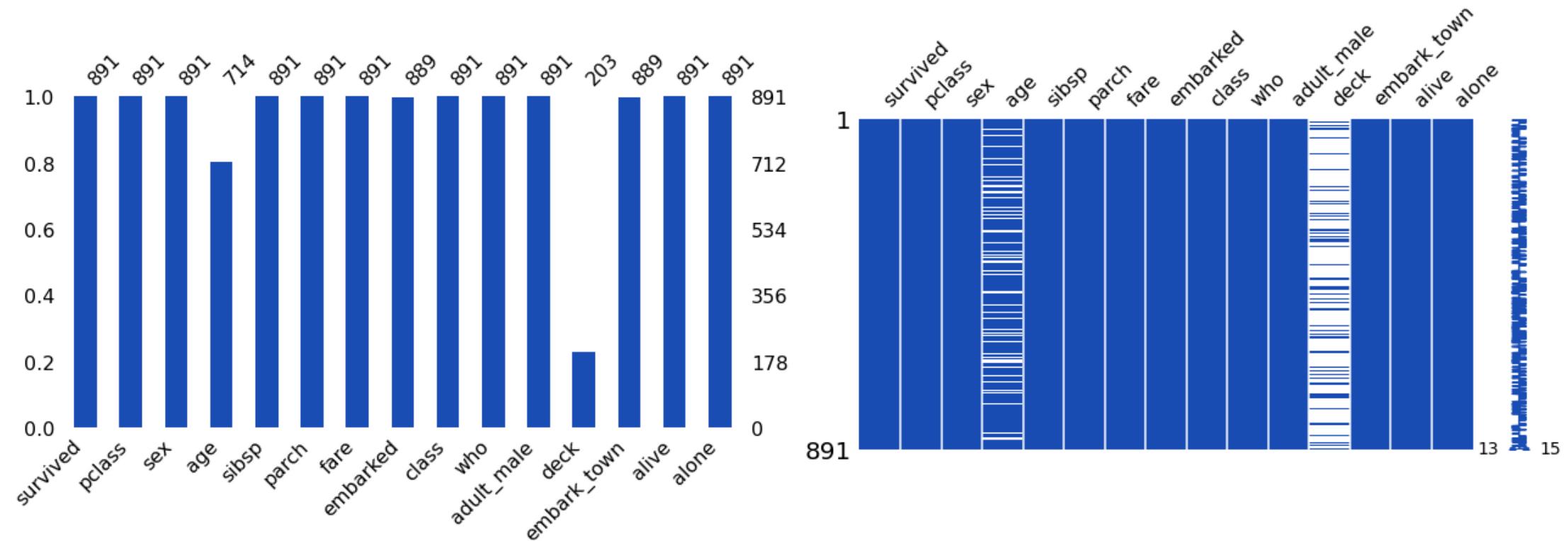
```
df.tail()
```

| | survived | pclass | sex | age | sibsp | parch | fare | embarked | class | who | adult_male |
|---|----------|--------|--------|------|-------|-------|---------|----------|-------|-------|------------|
| 0 | 0 | 3 | male | 22.0 | 1 | 0 | 7.2500 | S | Third | man | True |
| 1 | 1 | 1 | female | 38.0 | 1 | 0 | 71.2833 | C | First | woman | False |
| 2 | 1 | 3 | female | 26.0 | 0 | 0 | 7.9250 | S | Third | woman | False |
| 3 | 1 | 1 | female | 35.0 | 1 | 0 | 53.1000 | S | First | woman | False |
| 4 | 0 | 3 | male | 35.0 | 0 | 0 | 8.0500 | S | Third | man | True |

데이터 분석 실습 - 타이타닉 데이터셋

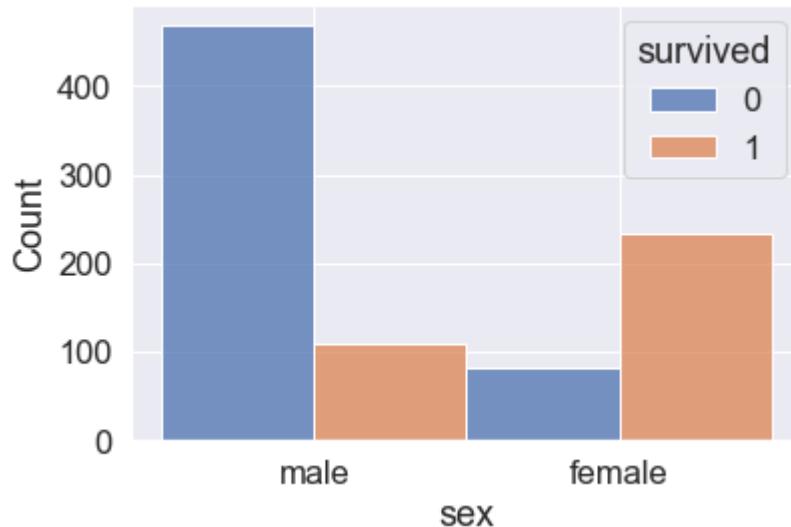
■ 결측값 확인

```
!pip install missingno  
import missingno as msno  
msno.bar(df, figsize=(10, 5), color=(0.1, 0.3, 0.7))  
msno.matrix(df, figsize=(10, 5), color=(0.1, 0.3, 0.7))
```

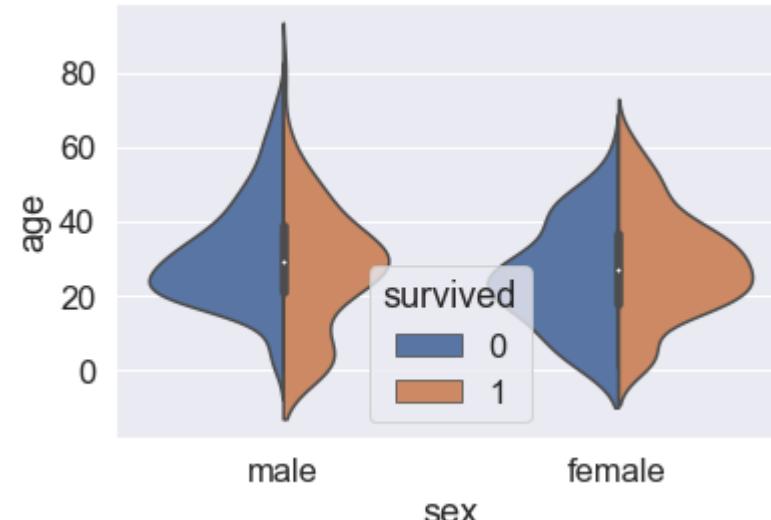


데이터 분석 실습 - 타이타닉 데이터셋

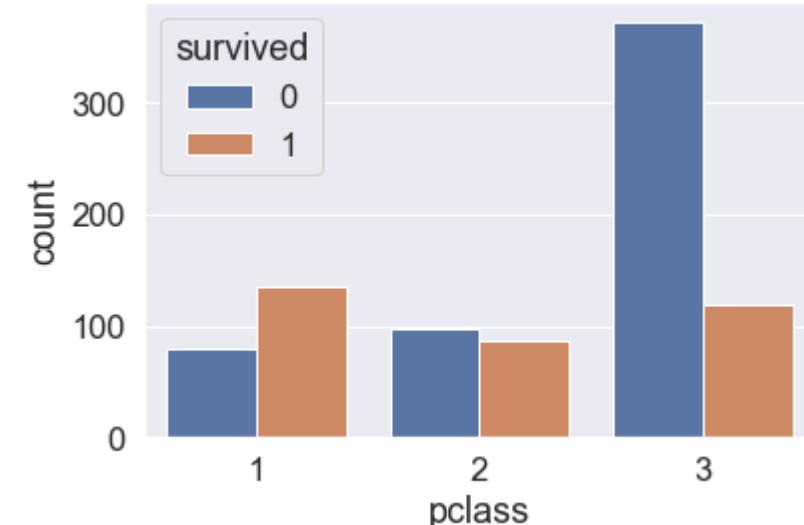
■ 성별(sex)에 따른 생존율 분포



■ 승객 나이와 생존 여부와의 관계



■ 객실등급과 생존율



```
sns.histplot(x='sex', hue='survived', multiple='dodge', data=df)
```

```
sns.violinplot(x='sex', y='age', hue='survived', data=df, split=True)
```

```
sns.countplot(x='pclass', hue='survived', data=df)
```

데이터 분석 실습 - 통신사 이탈고객 데이터셋



https://github.com/kgpark88/ai-summary/blob/main/05_DataAnalysis_Telecom.ipynb

■ Pandas 라이브러리 임포트

```
import pandas as pd
```

■ 파일에서 데이터 로드

```
df = pd.read_csv('churn_data.csv')
```

■ 데이터 확인

```
df.head()
```

```
df.head(20)
```

```
df.tail()
```

| | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines |
|---|------------|--------|---------------|---------|------------|--------|--------------|------------------|
| 0 | 7590-VHVEG | Female | 0 | Yes | No | 1 | No | No phone service |
| 1 | 5575-GNVDE | Male | 0 | No | No | 34 | Yes | No |

데이터 분석 실습 - 통신사 이탈고객 데이터셋

■ 데이터구조 파악

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   customerID      7043 non-null    object  
 1   gender          7043 non-null    object  
 2   SeniorCitizen   7043 non-null    int64  
 3   Partner         7043 non-null    object  
 4   Dependents     7043 non-null    object  
 5   tenure          7043 non-null    int64  
 ....., 
 18  MonthlyCharges 7043 non-null    float64 
 19  TotalCharges   7043 non-null    object  
 20  Churn          7043 non-null    object  
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

■ 데이터 타입 확인

df.dtypes

```
customerID        object
gender            object
SeniorCitizen    int64
Partner           object
Dependents        object
tenure            int64
PhoneService      object
MultipleLines     object
InternetService   object
OnlineSecurity    object
OnlineBackup      object
DeviceProtection  object
TechSupport       object
StreamingTV       object
StreamingMovies   object
Contract          object
PaperlessBilling  object
PaymentMethod     object
MonthlyCharges    float64
TotalCharges      object
Churn             object
dtype: object
```

■ Null 데이터 확인

df.isnull().sum()

```
customerID      0
gender          0
SeniorCitizen   0
Partner         0
Dependents      0
tenure          0
PhoneService    0
MultipleLines   0
InternetService 0
OnlineSecurity  0
OnlineBackup    0
DeviceProtection 0
TechSupport     0
StreamingTV     0
StreamingMovies 0
Contract        0
PaperlessBilling 0
PaymentMethod   0
MonthlyCharges  0
TotalCharges    0
Churn          0
dtype: int64
```

데이터 분석 실습 - 통신사 이탈고객 데이터셋

■ 통계 정보

df.describe()

| | SeniorCitizen | tenure | MonthlyCharges |
|--------------|---------------|-------------|----------------|
| count | 7043.000000 | 7043.000000 | 7043.000000 |
| mean | 0.162147 | 32.371149 | 64.761692 |
| std | 0.368612 | 24.559481 | 30.090047 |
| min | 0.000000 | 0.000000 | 18.250000 |
| 25% | 0.000000 | 9.000000 | 35.500000 |
| 50% | 0.000000 | 29.000000 | 70.350000 |
| 75% | 0.000000 | 55.000000 | 89.850000 |
| max | 1.000000 | 72.000000 | 118.750000 |

■ 데이터 상관관계 분석

df.corr()

| | SeniorCitizen | tenure | MonthlyCharges |
|----------------|---------------|----------|----------------|
| SeniorCitizen | 1.000000 | 0.016567 | 0.220173 |
| tenure | 0.016567 | 1.000000 | 0.247900 |
| MonthlyCharges | 0.220173 | 0.247900 | 1.000000 |

데이터 분석 /전처리 실습 - 통신사 이탈고객 데이터셋

■ 데이터 전처리

입력 데이터에서 제외: drop()

Null 데이터 처리: dropna(), fillna()

누락데이터 처리: replace()

데이터타입 변환 : astype()

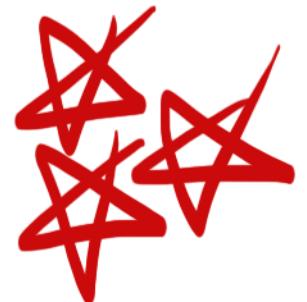
특성 추출 (feature engineering) : df['new_feature'] = df['f_1']/df['f_2']

```
df.drop('customerID', axis=1, inplace=True)
```

```
df['TotalCharges'].replace([' '], [0], inplace=True)
```

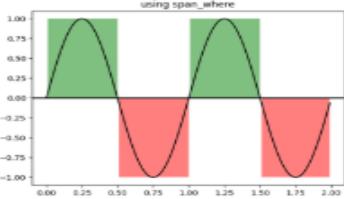
```
df['TotalCharges'] = df['TotalCharges'].astype(float)
```

```
df['Churn'].replace(['Yes', 'No'], [1, 0], inplace=True)
```

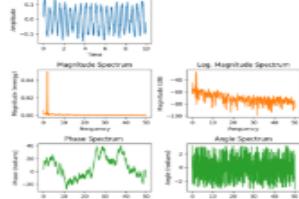


데이터 시각화 - 맷플롯립(Matplotlib)

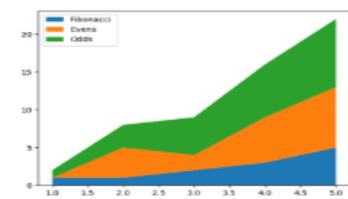
데이터를 차트나 플롯(Plot)으로 표시할 때 가장 많이 사용되는 데이터 시각화 라이브러리



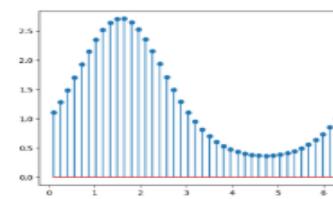
Using span_where



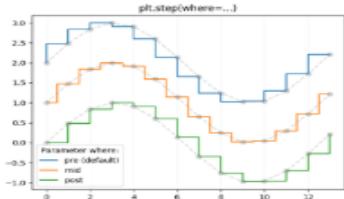
Spectrum Representations



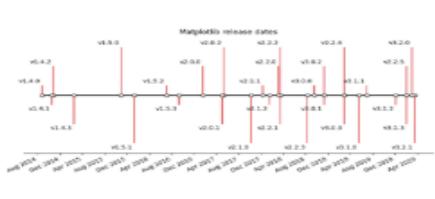
Stackplot Demo



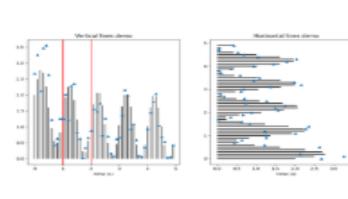
Stem Plot



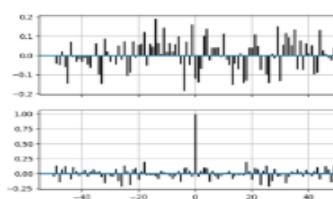
Step Demo



Creating a timeline with lines, dates, and text



hlines and vlines



Cross- and Auto-Correlation Demo

데이터 시각화 - 맷플롯립(Matplotlib)

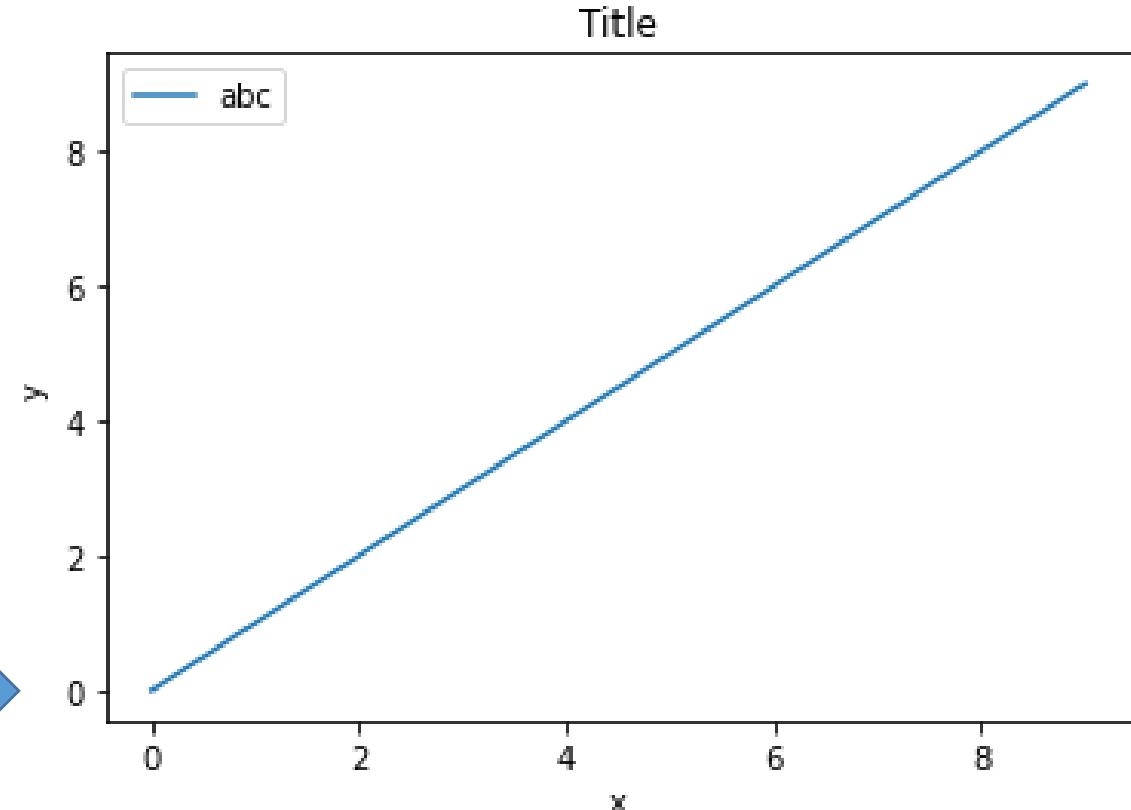
■ 라이브러리 임포트

```
import matplotlib.pyplot as plt  
%matplotlib inline
```

■ Matplotlib 사용법(예시)

```
x = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]  
y = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
```

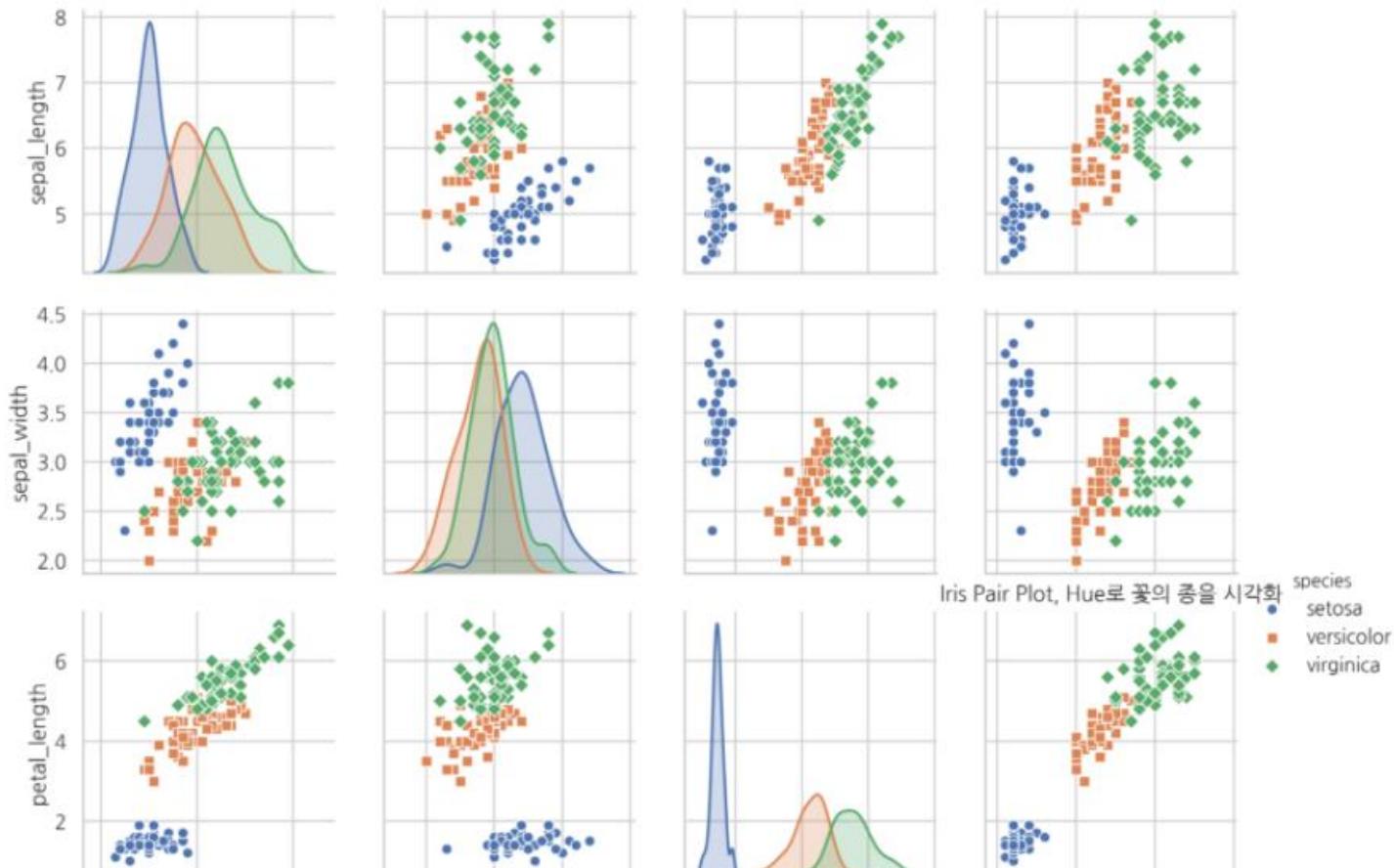
```
plt.plot(x, y)  
plt.title('Title')  
plt.xlabel('x')  
plt.ylabel('y')  
plt.legend(['abc'])  
plt.show()
```



데이터 시각화 - 씨본(Seaborn)

Matplotlib을 기반으로 다양한 색상 테마와 통계용 차트 등의 기능을 추가한 시각화 라이브러리

```
sns.pairplot(iris, hue="species", markers=["o", "s", "D"])
plt.title("Iris Pair Plot, Hue로 꽃의 종을 시각화")
plt.show()
```



데이터 시각화 - 씨본(Seaborn)

■ 패키지 설치

```
!pip install seaborn
```

■ 라이브러리 임포트

```
import seaborn as sns
```

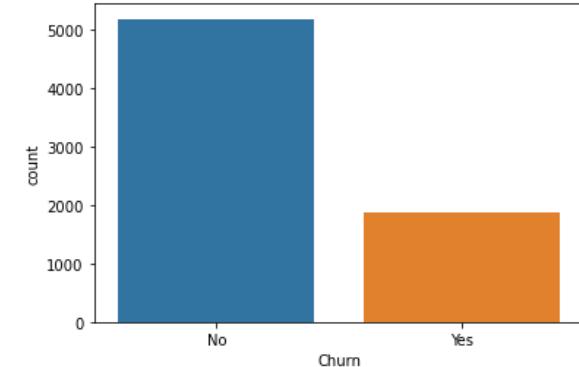
■ 상관관계 히트맵

```
sns.heatmap(df.corr(), annot=True)
```



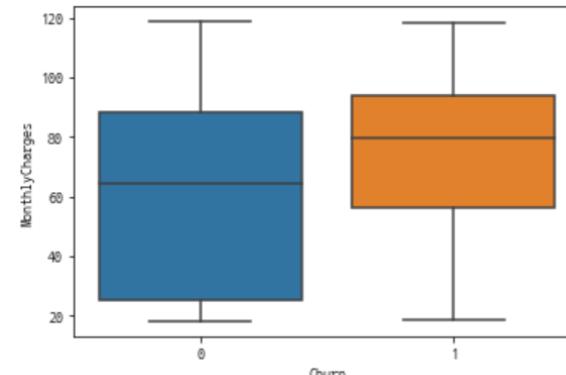
■ 카운트 플롯

```
sns.countplot(x='Churn', data=df)
```



■ 박스 플롯

```
sns.boxplot(x='Churn', y='MonthlyCharges', data=df)
```



•알쓸新JOB• 데이터 과학자

통계적 사고관을 갖추고
데이터 과학 기초를 이해하면
변화하는 세상을 살아가는데
분명 도움이 될 겁니다.



데이터 과학자 권재명

4. 머신러닝 핵심 알고리즘

데이터

| | total_bill | tip | sex | smoker | day | time | size |
|---|------------|------|--------|--------|-----|--------|------|
| 0 | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| 1 | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| 2 | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| 3 | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |
| 4 | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |



머신러닝

머신러닝은 컴퓨터 알고리즘이 데이터를 학습하여 입력과 출력간의 관계를 찾는 과정입니다.
학습할 때 정답 레이블을 있는지 여부에 따라 지도학습과 비지도학습으로 구분을 합니다.

■ 지도학습(supervised learning)

- 학습시 정답을 알려 주면서 진행하는 학습으로, 학습시 데이터와 레이블(정답)이 함께 제공됩니다.
- **레이블(Label)** = 정답, 실제값, 타깃, 클래스, y
- 예측된 값 = 예측값, 분류값, \hat{y} (y hat)
- 데이터마다 레이블을 달기 위해 많은 시간을 투자해야 합니다.
- 지도학습 **모델**에는 **분류모델**(이진분류, 다중분류)과 **회귀모델**(주가예측 등)이 있습니다.

■ 비지도학습(unsupervised learning)

- 레이블(정답) 없이 진행되는 학습으로, 데이터 자체에서 패턴을 찾아내야 할 때 사용합니다.
- 비지도학습의 대표적인 예는 군집화(clustering)와 차원축소가 있습니다.

지도학습

지도 학습은 정답이 있는 데이터를 활용해 데이터를 학습시키는 것입니다. 입력 값(X data)이 주어지면 입력값에 대한 Label(Y data)를 주어 학습시키며 분류모델과 회귀모델이 있습니다.

■ 모델

- 데이터들의 패턴을 대표할 수 있는 함수, 예) $f(x) = ax + b$
- 함수의 입력은 독립변수이고 출력은 종속변수로, 독립변수들에 의해 출력값이 정해집니다.

■ 분류 모델(Classification)

- 레이블의 값들이 이산적으로 나눠질 수 있는 문제에 사용합니다.
- 예) 스팸 메일 분류, 품종 분류

Classification



■ 회귀 모델(Regression)

- 레이블의 값들이 연속적인 문제에 사용합니다.
- 예) 날씨 예측, 주가 예측

Regression



지도학습 데이터셋 구조

각 열(column)을 특징/속성(feature) 이라고 합니다.
데이터 컬럼(column)중에 하나를 선택해서 레이블로 사용합니다.

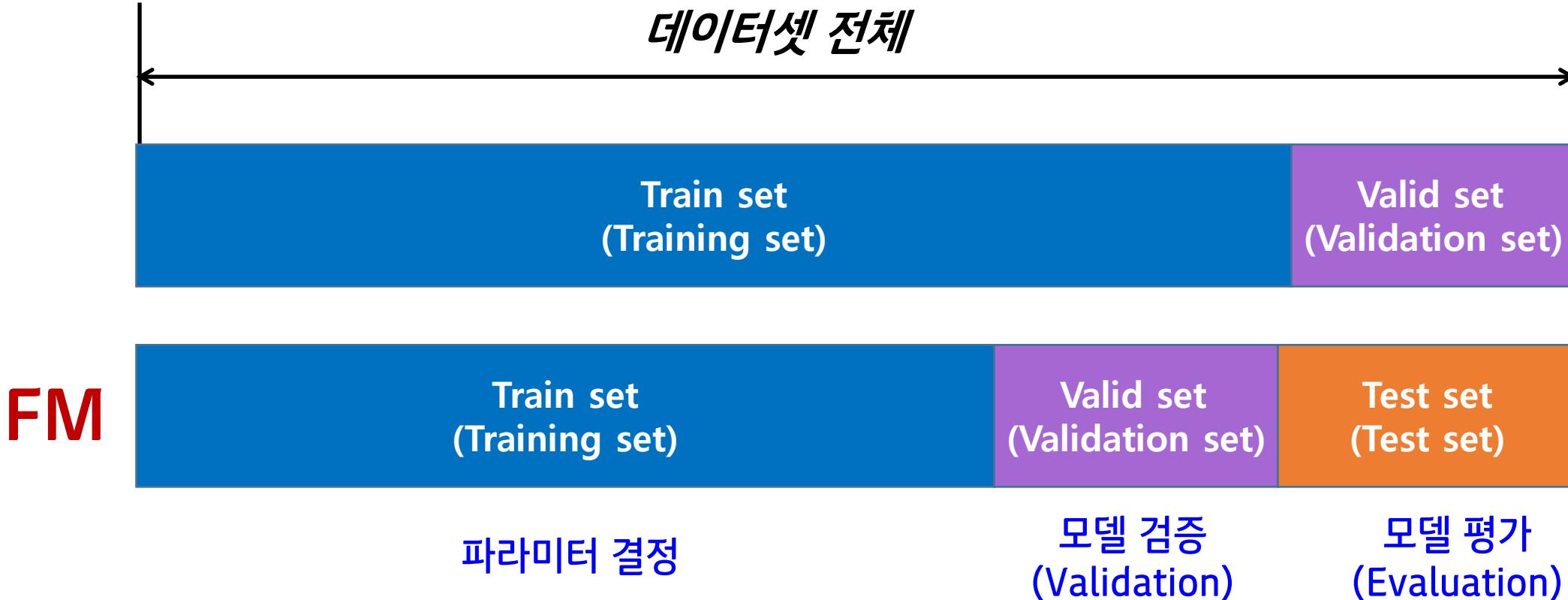
| | total_bill | tip | sex | smoker | day | time | size |
|---|------------|------|--------|--------|-----|--------|------|
| 0 | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| 1 | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| 2 | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| 3 | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |
| 4 | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |

각 행(row)을
예제(Example)
데이터라고
합니다

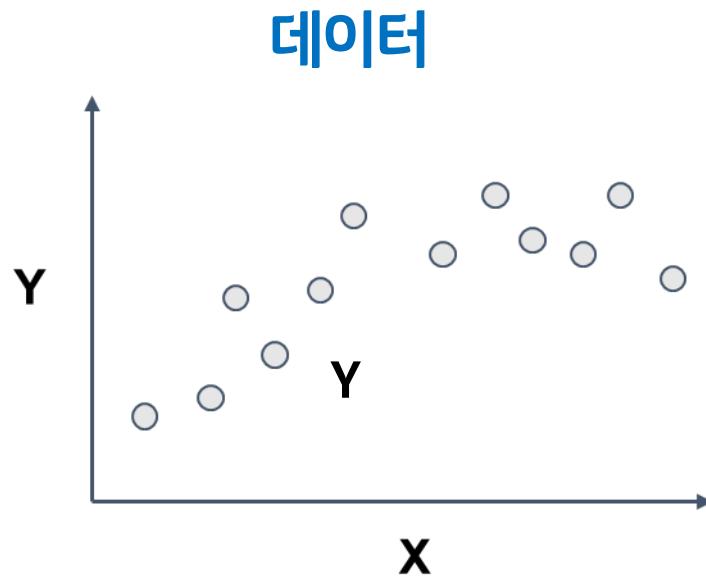
예측 모델(회귀 모델)
팁의 크기를 예측

분류 모델
손님의 성별을 예측

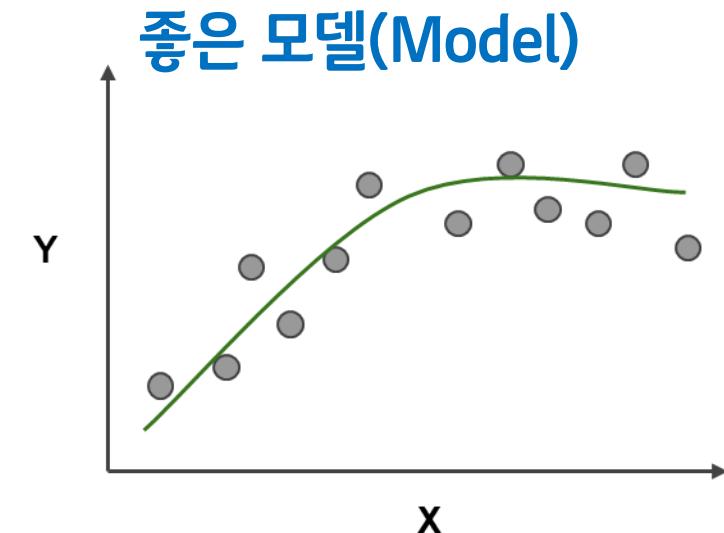
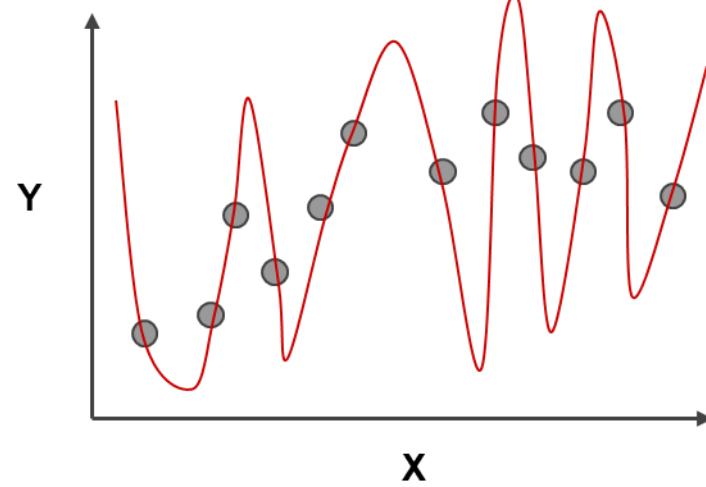
데이터셋 분리



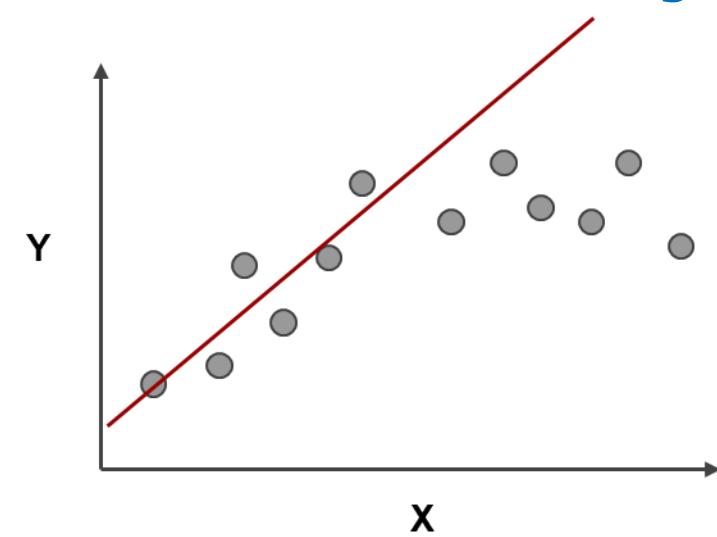
모델 선택



과적합(Overfitting)

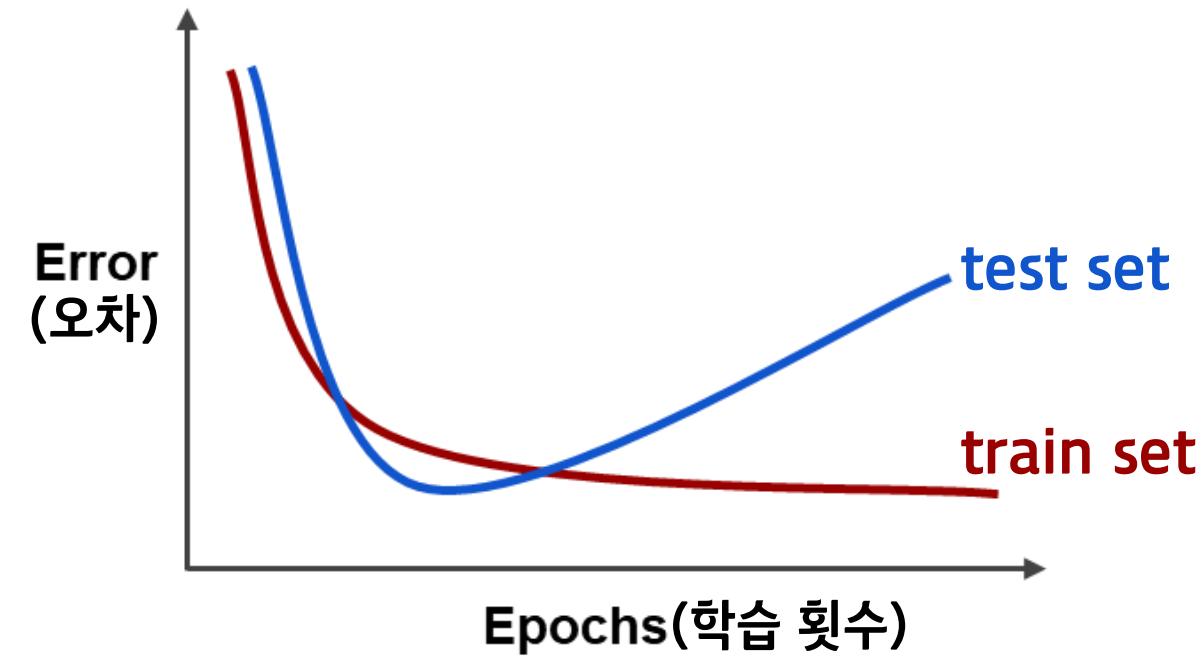
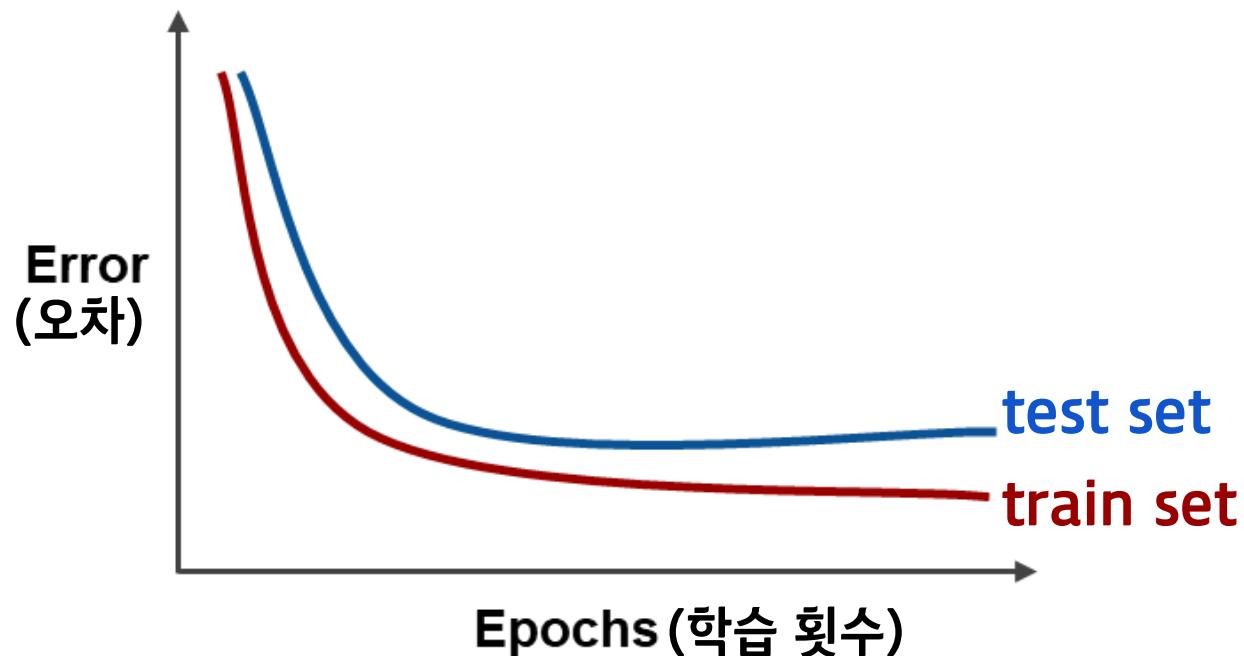


과소적합(Underfitting)



모델 성능 측정

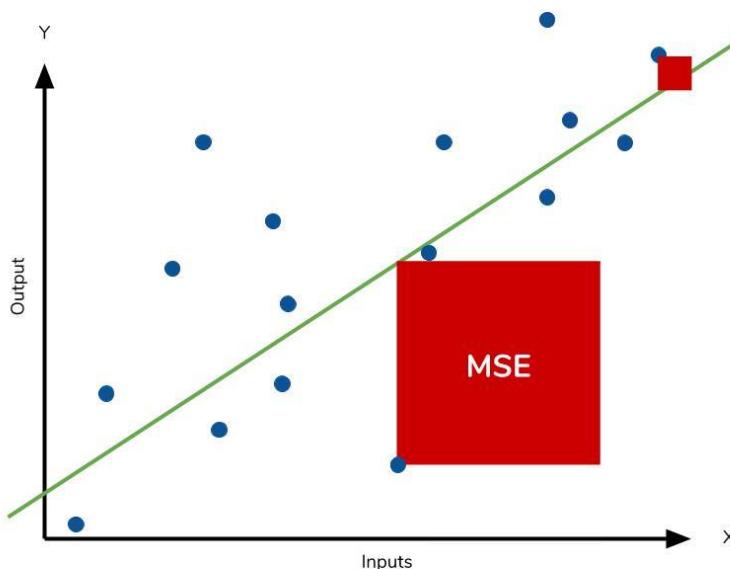
모델학습을 진행하면서 손실(Loss, Error, 오차)을 지속적으로 측정합니다.



회귀모델 손실함수(Loss Function)

회귀모델(Regression)에서는 주로 평균제곱오차(MSE)를 손실함수로 사용합니다.

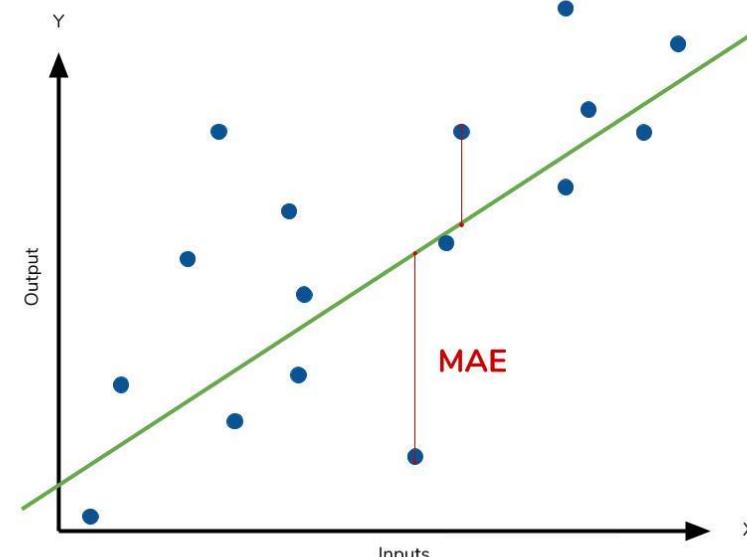
■ MSE(Mean Squared Error)



$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

레이블 값
(실제값)
 y_i
(모델이 예측한 값)
 \hat{y}_i

■ MAE(Mean Absolute Error)



$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

분류모델 손실함수(Loss Function)

이진분류는 Binary Cross Entropy 를 다중분류는 Categorical Cross Entropy 손실함수를 사용합니다.

■ 이진분류(Binary Classification)

0



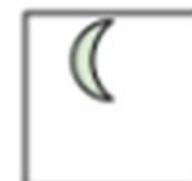
1



■ 다중분류(Multi-Class Classification)

$C = 3$

Samples



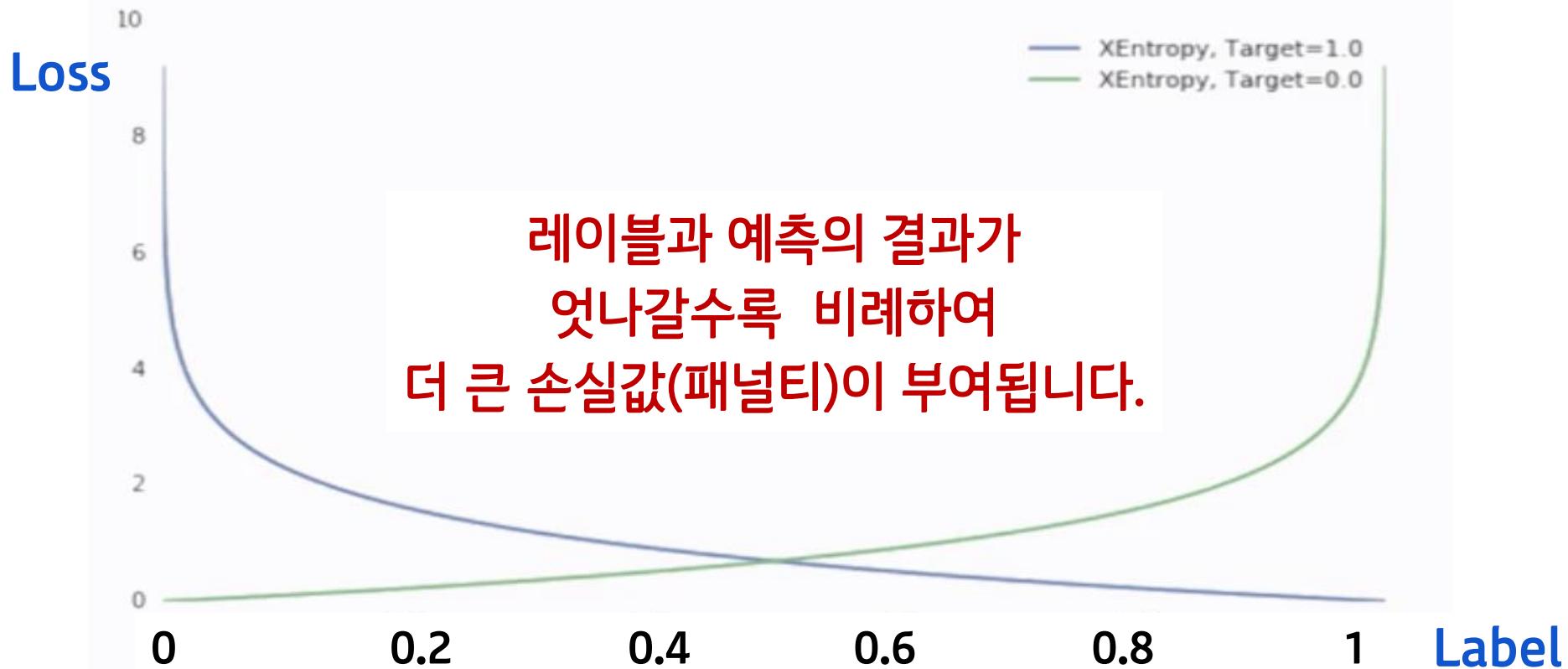
Labels (t)

[0 0 1]

[1 0 0]

[0 1 0]

분류모델 손실함수(Loss Function)



$$\frac{-1}{N} \times \sum_{i=1}^N y_i \times \log(\hat{y}_i) + (1 - y_i) \times \log(1 - \hat{y}_i)$$

참고자료 : <https://youtu.be/Jt5BS71uVfl>

<http://kocw-n.xcache.kincdn.com/data/document/2017/kumoh/kojaepil0302/8.pdf>

오차 행렬(Confusion Matrix)

분류모델 성능평가에 사용하며 대략적인 성능확인과 모델의 성능을 오차행렬을 기반으로 수치로 표현할 수 있습니다.

| | | 실제값 | |
|-----|----------|----------------|----------------|
| | | Positive | Negative |
| 예측값 | Positive | True Positive | False Positive |
| | Negative | False Negative | True Negative |

Positive = True = 1

Negative = False = 0

TP(True Positive) : Positive(값 1)로 예측했는데, 실제값 역시 Positive(값 1)

TN(True Negative) : Negative(값 0)으로 예측는데, 실제값 역시 Negative(값 0)

FP(False Positive) : Positive(값 1)로 예측 했는데, 실제값은 Negative(값 0)

FN(False Negative) : Negative(값 0)으로 예측했는데, 실제 값은 Positive(값 1)

모델성능 평가지표

■ 정밀도(Precision)

모델이 True(Positive) 라고 분류한 것 중에서 실제 True(Positive) 인 것의 비율
날씨 예측 모델이 맑다로 예측했는데, 실제 날씨가 맑았는지는 나타낸 지표입니다.

$$(Precision) = \frac{TP}{TP + FP}$$

■ 재현율/회수율(Recall)

실제 True인 것 중에서 모델이 True라고 예측한 것의 비율
실제 날씨가 맑은 날 중에서 모델이 맑다고 예측한 비율을 나타낸 지표입니다.

$$(Recall) = \frac{TP}{TP + FN}$$

■ 정확도(Accuracy)

가장 직관적으로 모델의 성능을 나타낼 수 있는 평가 지표
혼동행렬상에서는 대각선(TP)을 전체 셀로 나눈 값에 해당합니다.
한달 동안에 맑은 날이 28일이고 비가 오는 날이 이틀인 경우, 비가 오는 것을
예측하는 성능은 매우 낮을 수 밖에 없으므로 이를 보완할 지표가 필요합니다.

$$(Accuracy) = \frac{TP + TN}{TP + FN + FP + TN}$$

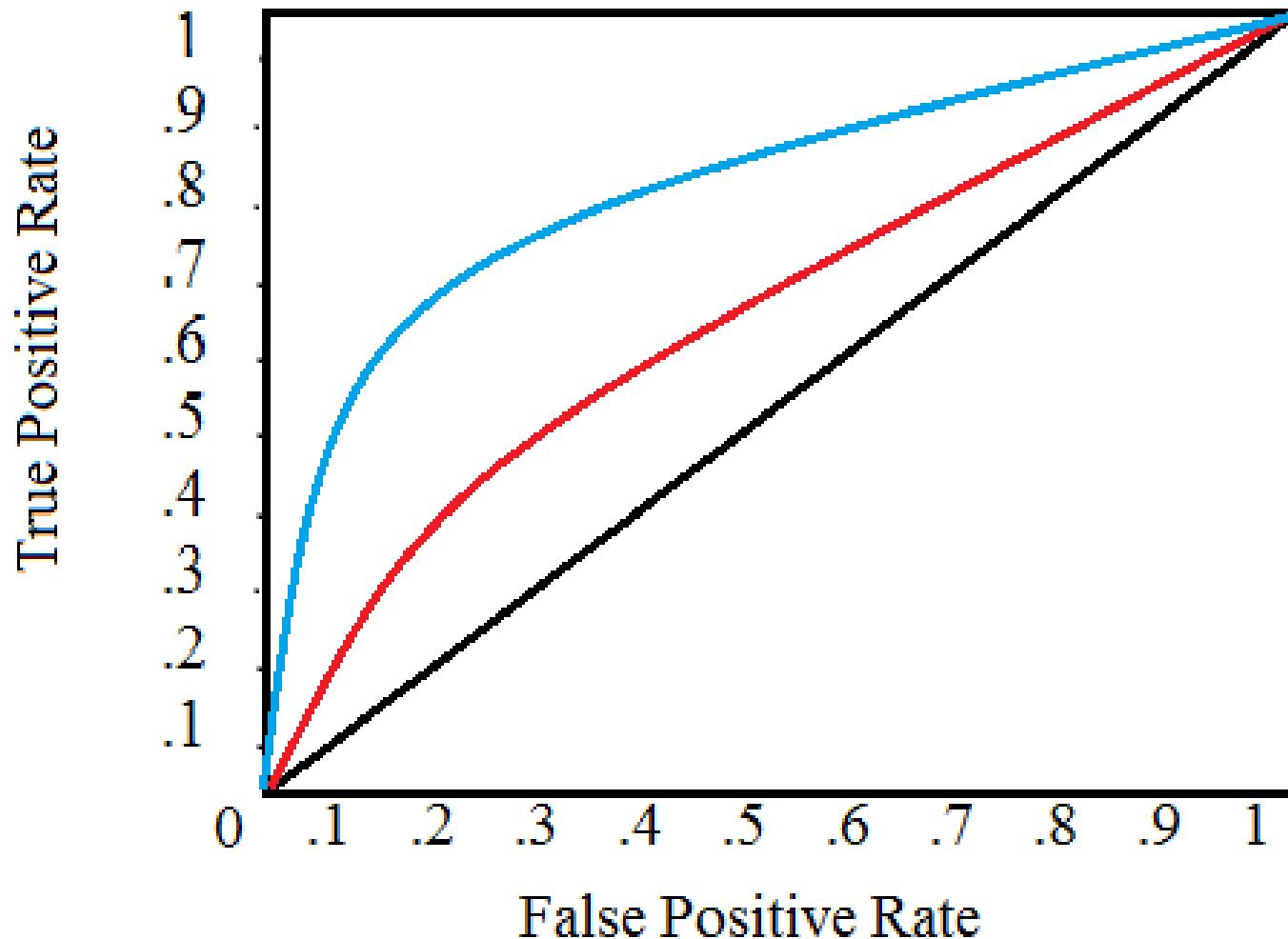
■ F1 점수(F1 score)

정밀도와 재현율의 조화평균입니다.

$$(F1-score) = 2 \times \frac{\frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

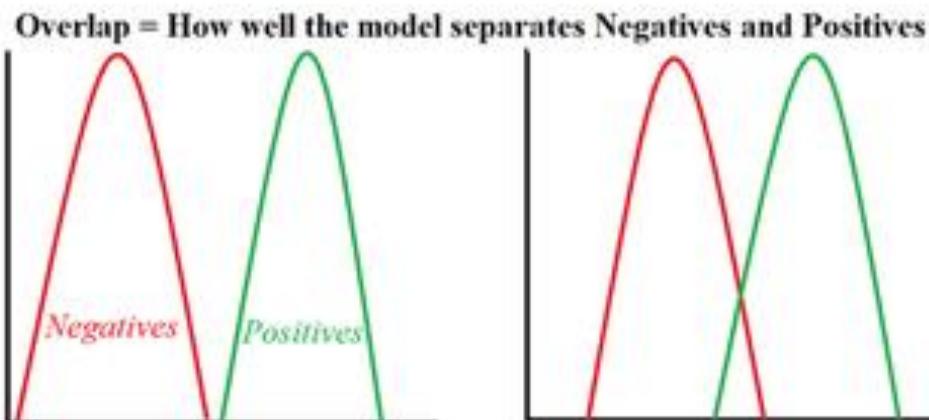
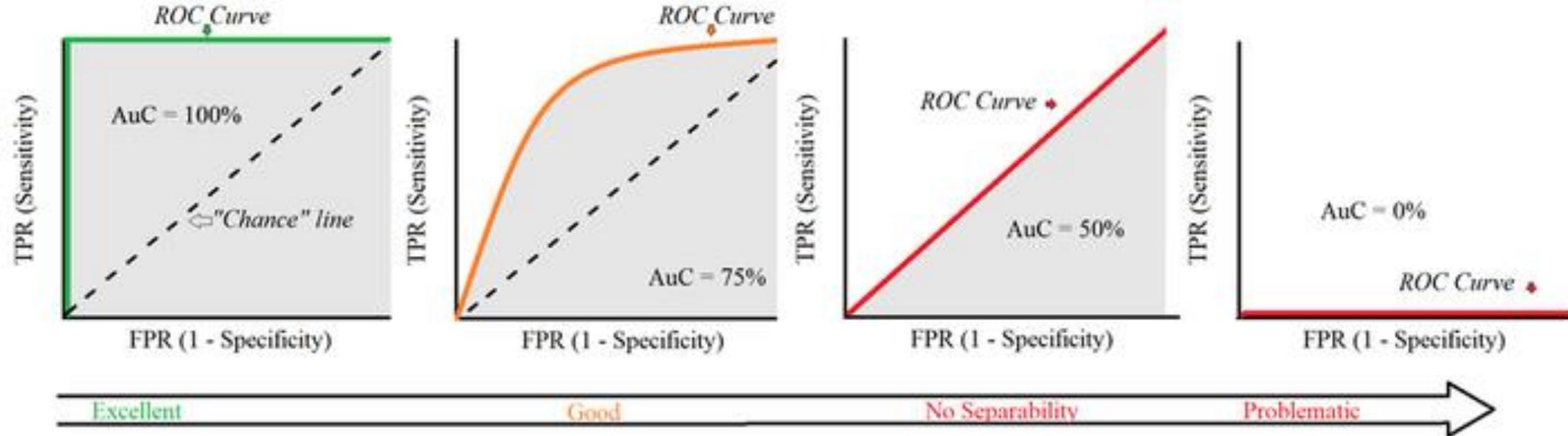
ROC 곡선(Receiver Operating Characteristic)

거짓 양성 비율(False Positive Rate)에 대한 진짜 양성 비율(True Positive Rate, 재현율)의 곡선입니다.
ROC 곡선을 통해 모델의 성능을 평가하거나 최적의 분류기준(threshold)을 찾을 수 있습니다.

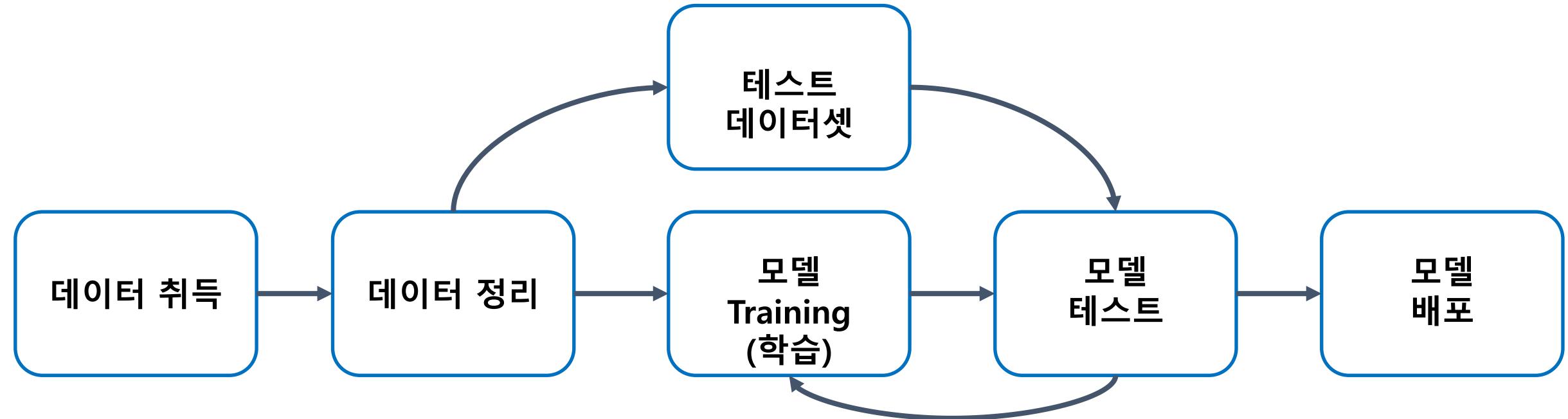


AUC(Area under the Curve)

ROC 곡선 아래의 면적(AUC)을 측정하면 모델의 성능을 평가하거나 최적의 분류기준(threshold)을 찾을 수 있습니다.



머신러닝 프로세스



사이킷런(Scikit-learn)

가장 인기있는 머신러닝 패키지이며, 많은 머신러닝 알고리즘이 내장되어 있습니다.

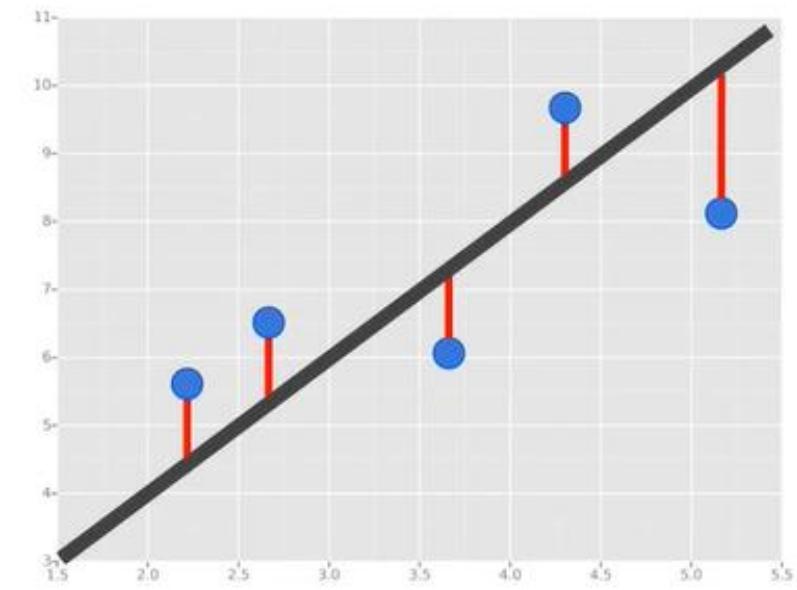
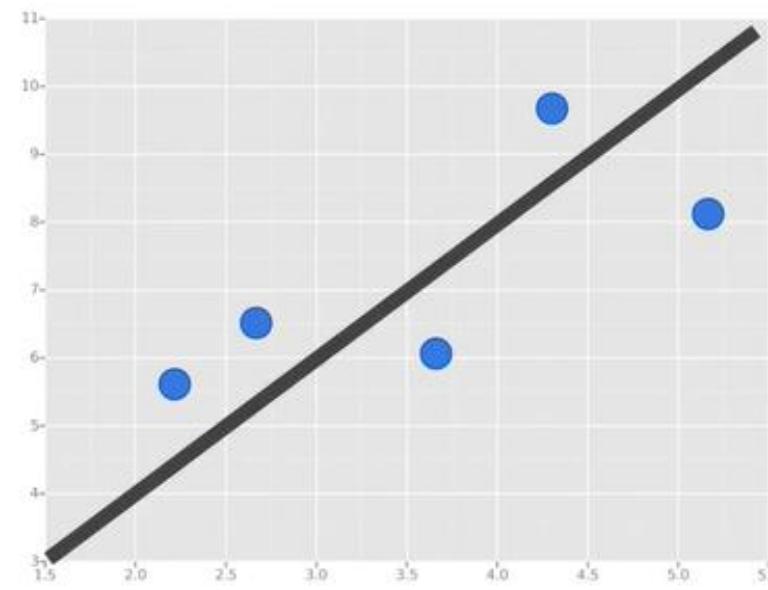
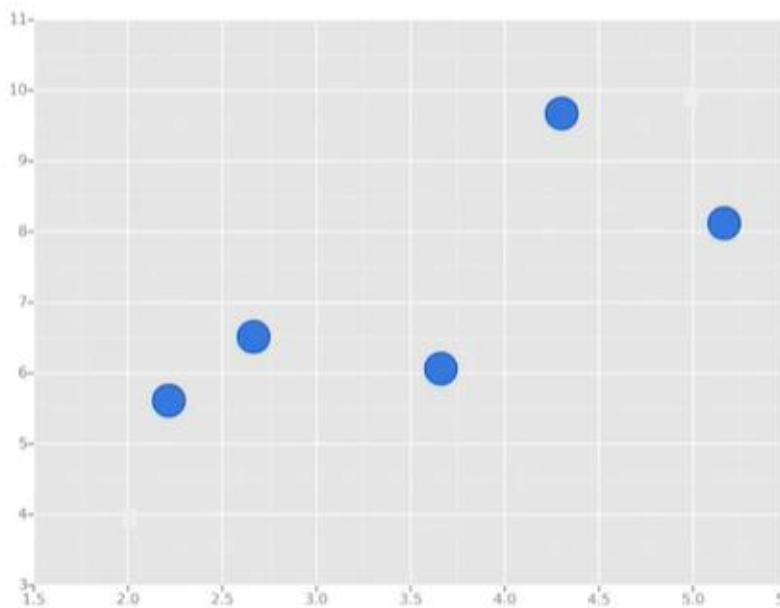
The screenshot shows the official website for scikit-learn (<https://scikit-learn.org>). The top navigation bar includes links for 'Install', 'User Guide', and 'API'. Below the header, there's a large 'scikit-learn' logo and the tagline 'Machine Learning in Python'. A blue banner highlights the following features:

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

The main content area is divided into three sections:

- Classification**: Describes identifying which category an object belongs to. It lists applications like spam detection and image recognition, and algorithms like SVM, nearest neighbors, random forest, and more. It includes a grid of small plots showing decision boundaries for various classifiers.
- Regression**: Describes predicting a continuous-valued attribute associated with an object. It lists applications like drug response and stock prices, and algorithms like SVR, nearest neighbors, random forest, and more. It includes a line plot titled "Boosted Decision Tree Regression" showing target values versus data points for different numbers of estimators.
- Clustering**: Describes automatic grouping of similar objects into sets. It lists applications like customer segmentation and grouping experiment outcomes, and algorithms like k-Means, spectral clustering, mean-shift, and more. It includes a scatter plot titled "K-means clustering on the digits dataset (PCA-reduced data)" showing data points grouped into four clusters with their respective centroids marked by white crosses.

선형 회귀(Linear Regression)



선형 회귀(Linear Regression)



06_01_LinearRegression.ipynb

```
from sklearn.model_selection import train_test_split  
from sklearn.linear_model import LinearRegression  
  
X = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10 ]).reshape(-1,1)  
y = np.array([13, 25, 34, 47, 59, 62, 79, 88, 90, 100])  
X_train, X_test, y_train, y_test = train_test_split(X, y,  
                                                    test_size=0.3, random_state=42)  
  
model = LinearRegression()  
model.fit(X_train, y_train)  
predictions = model.predict(X_test)
```

분류(Classification)



08_Classification.ipynb

■ setosa



■ versicolor



■ virginica



```
from sklearn import datasets  
from sklearn.model_selection import train_test_split  
iris = datasets.load_iris()
```

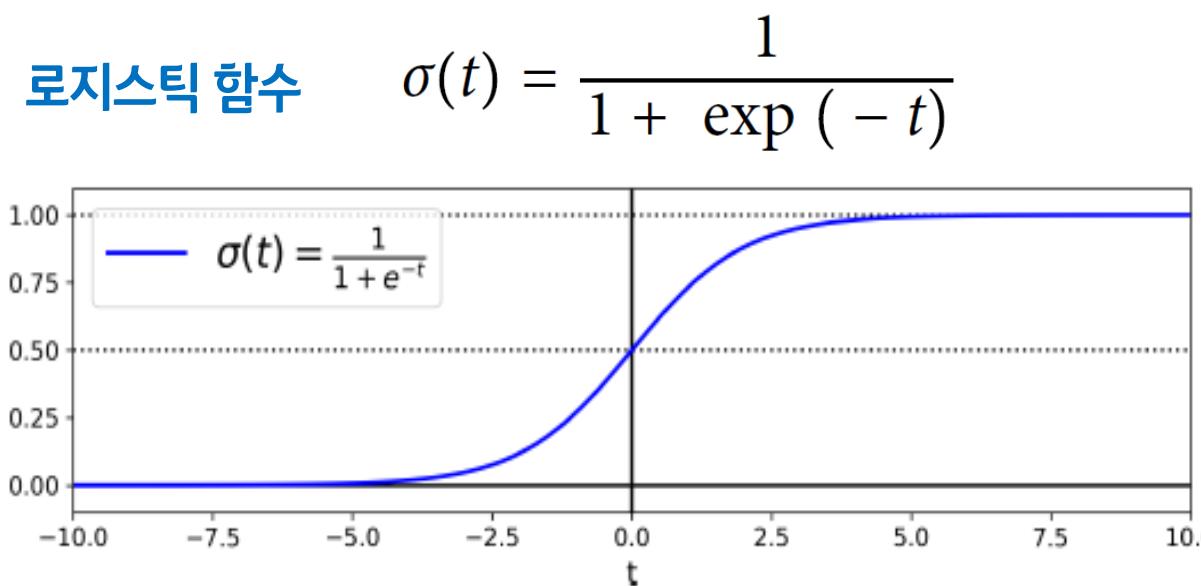
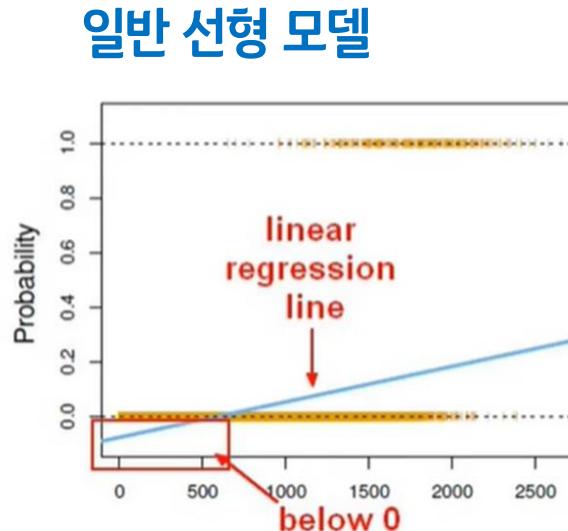
Train Test 데이터셋 분할

```
from sklearn.model_selection import train_test_split  
  
X_train, X_test, y_train, y_test = train_test_split(  
    iris['data'],  
    iris['target'],  
    test_size=0.3,  
    shuffle=True,  
    stratify=iris.target,  
    random_state=42)
```



로지스틱 회귀(Logistic Regression)

- 이진 분류 규칙은 0과 1의 두 클래스를 갖는 것으로, 일반 선형 회귀 모델을 이진분류에 사용할 수 없습니다.
- 대신 선형 회귀를 로지스틱 회귀 곡선으로 변환 할 수 있으며, 로지스틱 회귀 곡선은 0과 1 사이에서만 이동할 수 있으므로 분류에 사용할 수 있습니다.
- 로지스틱 회귀는 선형 회귀처럼 바로 결과를 출력하지 않고 로지스틱(logistic)을 출력합니다.
- 로지스틱 회귀는 샘플이 특정 데이터에 속할 확률을 추정(이진분류)하는 데 사용됩니다.
- 추정 확률이 50%가 넘으면 모델은 그 샘플이 해당 클래스에 속한다고 예측합니다.



로지스틱 확률모델

$$\hat{p} = h_{\theta}(x) = \sigma(x^T \theta)$$
$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5 \\ 1 & \text{if } \hat{p} \geq 0.5 \end{cases}$$

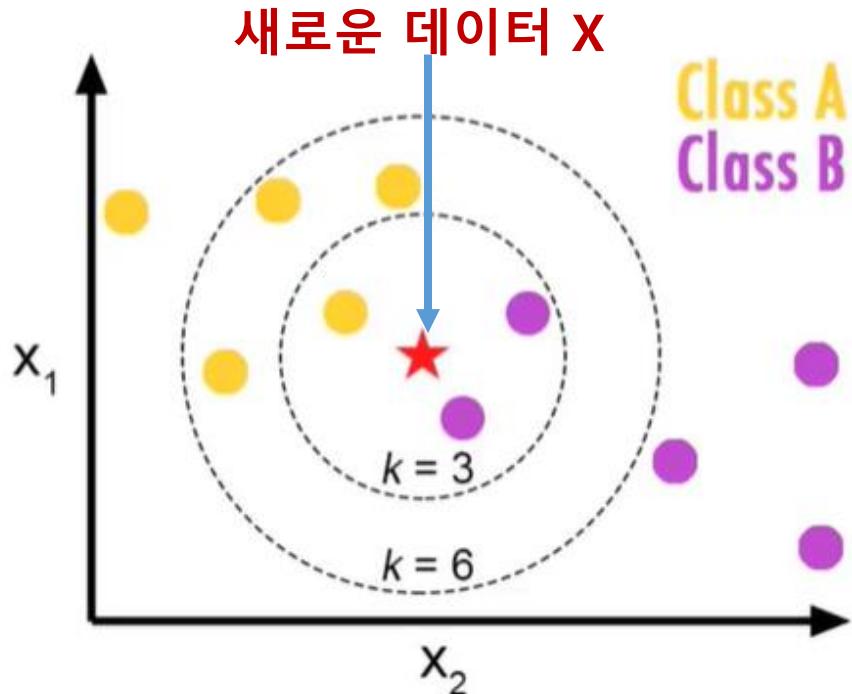
로지스틱 회귀(Logistic Regression)

```
from sklearn.linear_model import LogisticRegression
# 모델 학습
lr = LogisticRegression()
lr.fit(X_train, y_train)
# 예측
pred = lr.predict(X_test)
print('예측값: ', pred[:10])

# 모델 성능 평가
accuracy = accuracy_score(y_test, pred)
print(f'Mean accuracy score: {accuracy:.4f}')
# 확률값
prob = lr.predict_proba(X_test)
print("Probability: ", prob[0])
```

KNN(K-Nearest Neighbor)

- KNN은 새로운 데이터가 주어졌을 때 기존 데이터 가운데 가장 가까운 k개 이웃의 정보로 새로운 데이터를 예측하는 방법론입니다. 아래 그림처럼 검은색 점의 범주 정보는 주변 이웃들을 가지고 추론해낼 수 있습니다.
- 만약 k값이 3이면 Class B, k가 6이면 Class A로 분류(classification)하는 것입니다.
- 만약, 회귀(regression) 문제라면 이웃들 종속변수(y)의 평균이 예측값이 됩니다.
- 알고리즘이 간단하며 큰 데이터셋과 고차원 데이터에 적합하지 않은 단점이 있습니다.



| K | 이웃(Neighbor) | 예측값 |
|---|--------------|---------|
| 3 | | Class B |
| 7 | | Class A |

KNN(K-Nearest Neighbor)

```
from sklearn.linear_model import KNeighborsClassifier
```

모델 학습

```
knn = KNeighborsClassifier(n_neighbors=7)  
knn.fit(X_train, y_train)
```

예측

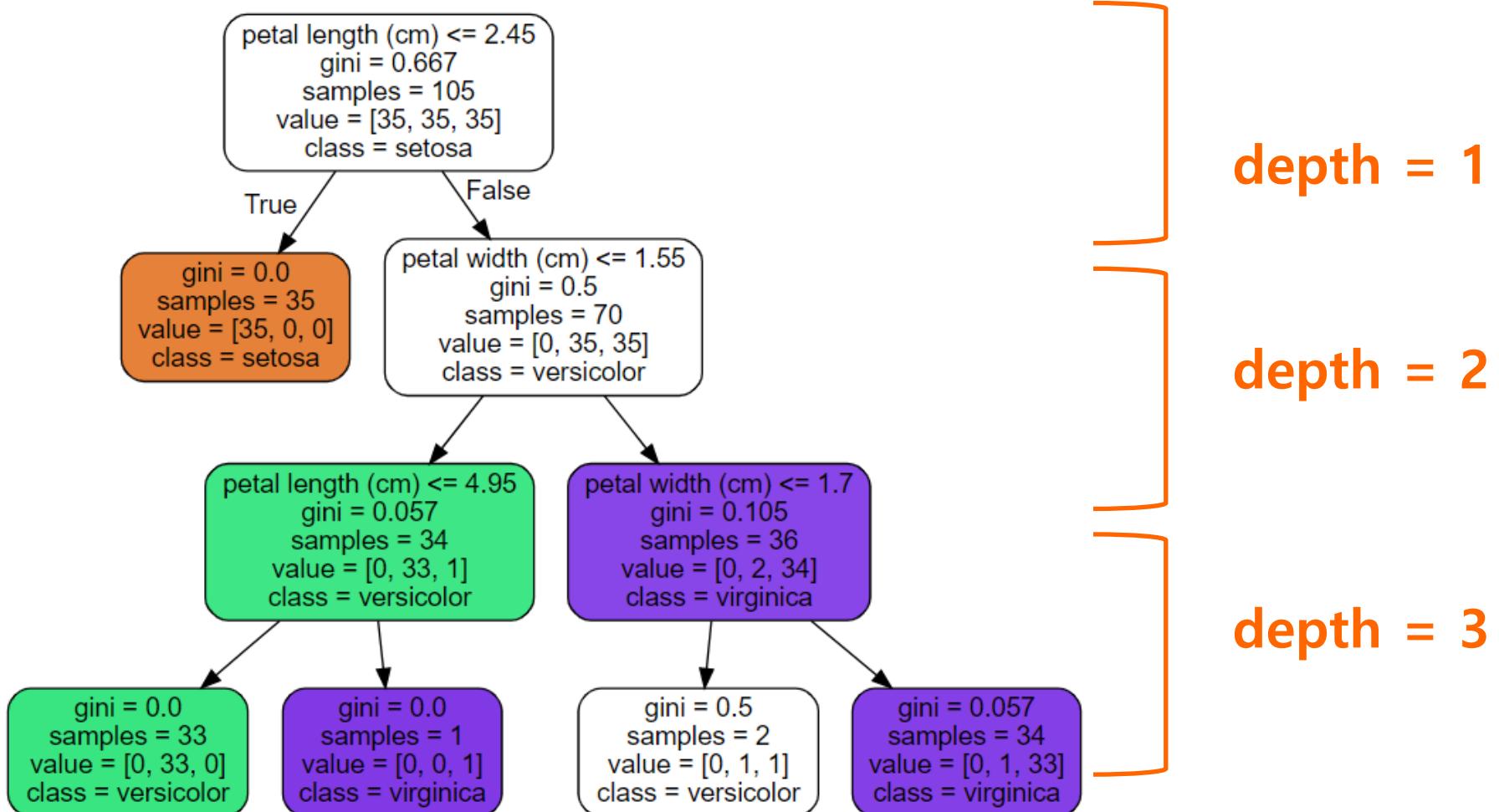
```
pred = knn.predict(X_test)  
print ('예측값: ', pred[:10])
```

모델 성능 평가

```
accuracy = accuracy_score(y_test, pred)  
print(f'Mean accuracy score: {accuracy:.4}')
```

의사결정트리(Decision Tree)

의사결정트리 모델은 트리(Tree) 알고리즘을 사용합니다. 트리의 각 분기점(node)에 데이터셋의 Feature를 하나씩 위치시키고, 각 분기점(node)에서 임의의 조건식으로 가지를 나무면서 데이터를 구분합니다.



depth = 1

depth = 2

depth = 3

의사결정트리(Decision Tree)

```
from sklearn.linear_model import DecisionTreeClassifier

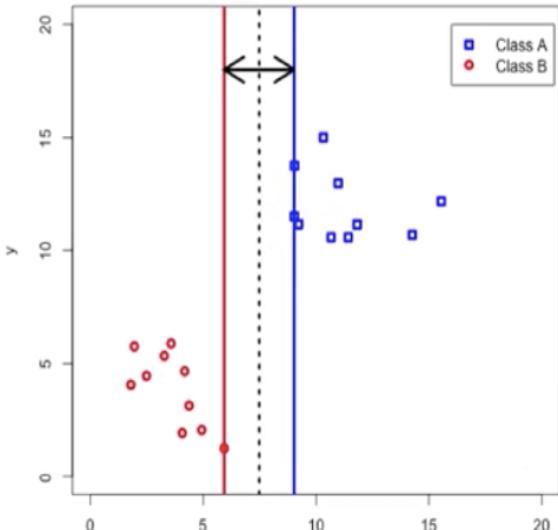
# 모델 학습
dtc = DecisionTreeClassifier(max_depth=3, random_state=42)
dtc.fit(X_train, y_train)
# 예측
pred = dtc.predict(X_test)
print('예측값: ', pred[:10])
# 모델 성능 평가
accuracy = accuracy_score(y_test, pred)
print(f'Mean accuracy score: {accuracy:.4f}')
# 확률값
prob = dtc.predict_proba(X_test)
print("Probability: ", prob[0])
```

서포트 벡터 머신(SVM)

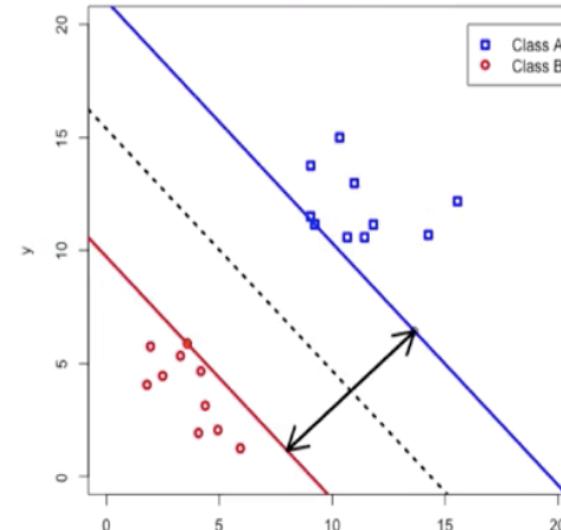
- 서포트 벡터 머신은 선형/비선형 분류, 회귀, 이상치 탐색에도 사용할 수 있는 다목적 머신러닝 모델입니다.
- SVM은 복잡한 분류 모델에 잘 들어 맞으며 작거나 중간 크기의 데이터셋에 적합합니다.

SVMs maximize the margin between two classes

Small margin

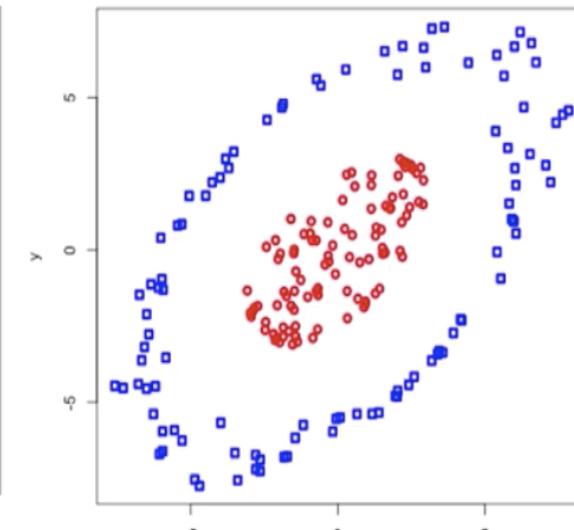


Large margin



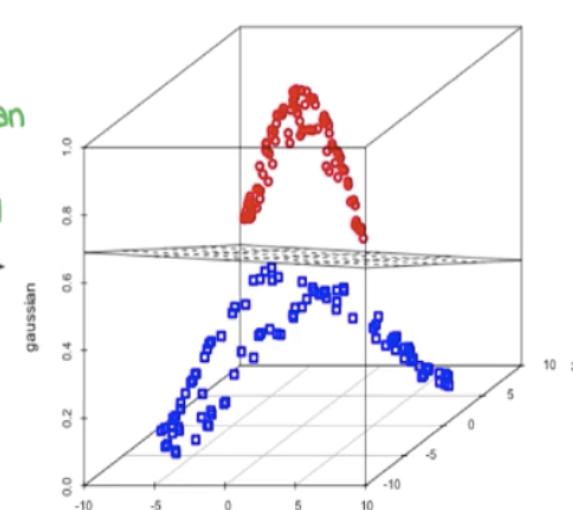
Kernels transform the input space into a more usable feature space

Class A
Class B



Gaussian
RBF
Kernel

Class A
Class B



■ 서포트 벡터 머신(SVM)

```
from sklearn.svm import SVC
```

```
# 모델 학습
```

```
svc = SVC(kernel='rbf')  
svc.fit(X_train, y_train)
```

```
# 예측
```

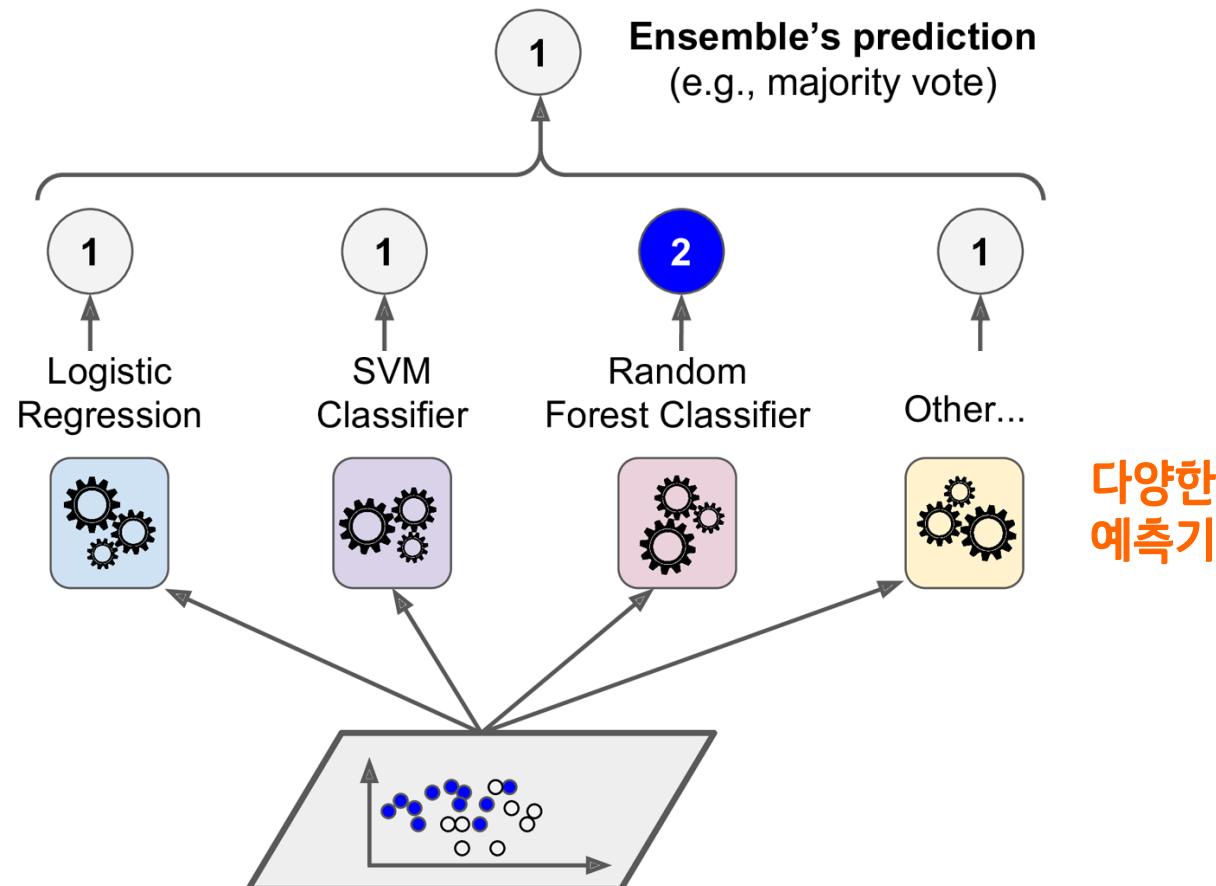
```
pred = svc)  
print ('예측값: ', pred[:10])
```

```
# 모델 성능
```

```
acc = accuracy_score(y_test, pred)  
print("Accuracy: {:.4f}".format(acc))
```

앙상블 학습(Ensemble Learning)

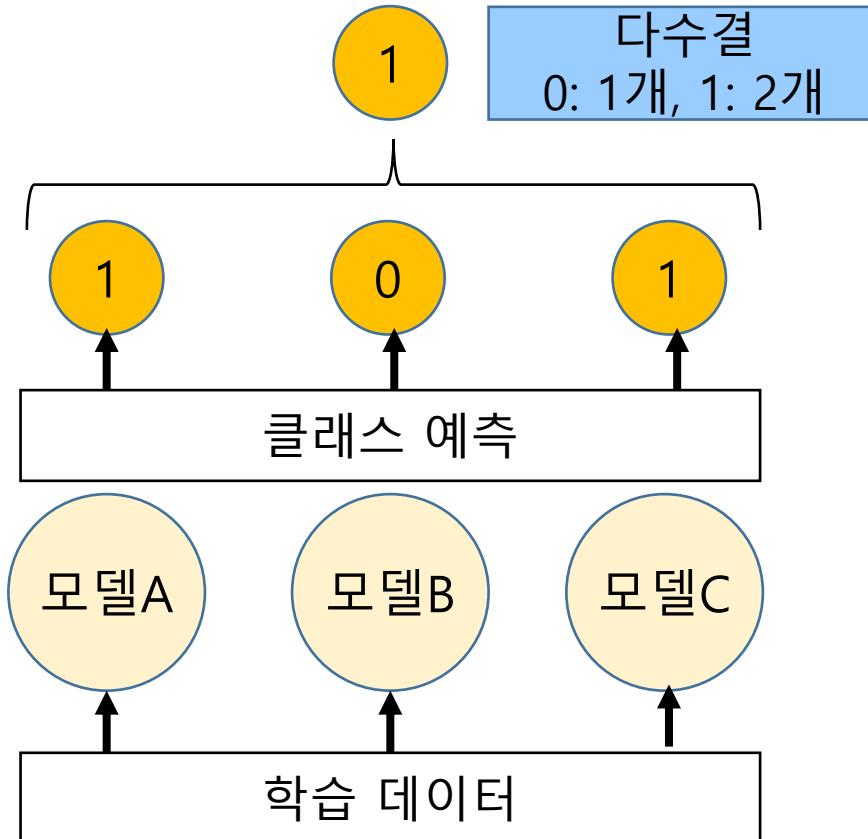
- 일련의 예측기(분류, 회귀)로부터 예측을 수집하면, 가장 좋은 모델 1개보다 더 좋은 예측을 얻을 수 있을 것입니다.
- 일련의 예측기를 앙상블이라 부르고 이를 앙상블 학습(Ensemble Learning)이라고 합니다.
- 가장 인기 있는 앙상블 방법에는 배깅, 부스팅이 있습니다.



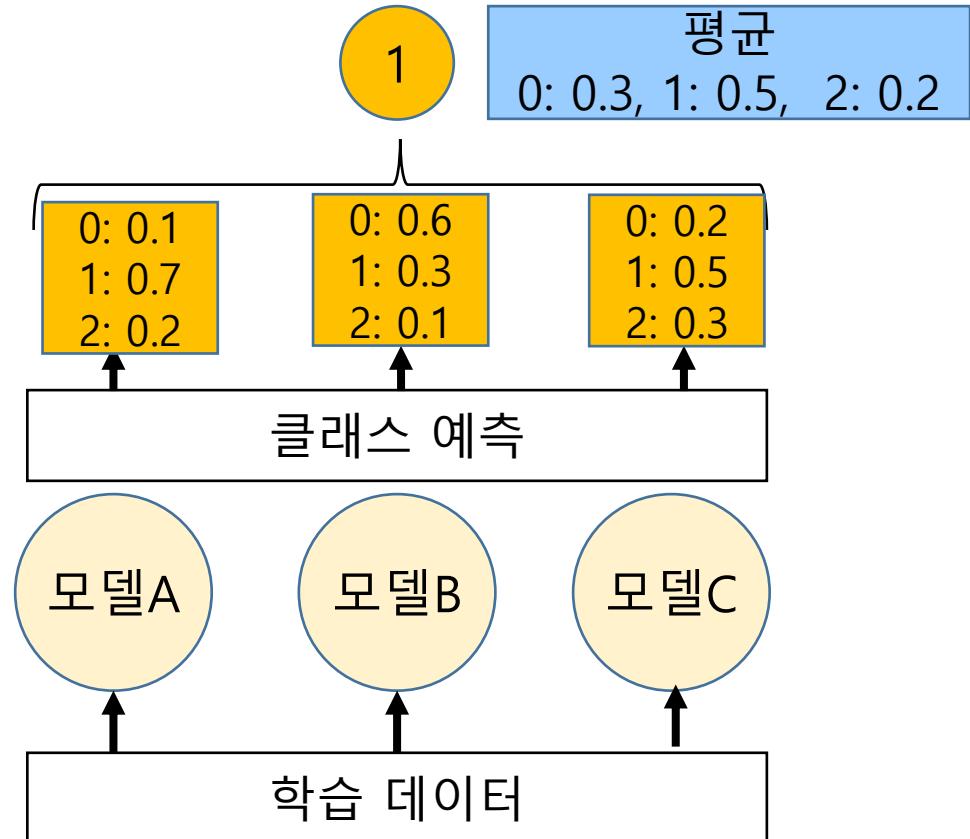
앙상블 모델 - 보팅(Voting)

- 보팅(Voting)은 여러 개의 모델이 예측한 값을 결합하여 최종 예측값을 결정하는 앙상블 방법입니다.
- 하드 보팅(hard voting)은 모델이 예측한 값 중에서 다수결로 최종 분류 클래스를 정합니다.
- 소프트 보팅(soft voting)은 각 분류 클래스별 예측 확률을 평균하여 최종 분류 클래스를 정합니다.

■ 하드 보팅(hard voting)



■ 소프트 보팅(soft voting)



앙상블 모델 - 보팅(Voting)

```
from sklearn.ensemble import VotingClassifier
```

```
# 모델 학습
```

```
hvc = VotingClassifier(estimators=[('KNN', knn), ('DT', dtc),
        ('SVM', svc)], voting='hard')
```

```
hvc.fit(X_train, y_train)
```

```
# 예측
```

```
pred = hvc.predict(X_test)
```

```
print('예측값:', pred[:10])
```

```
# 모델 성능 평가
```

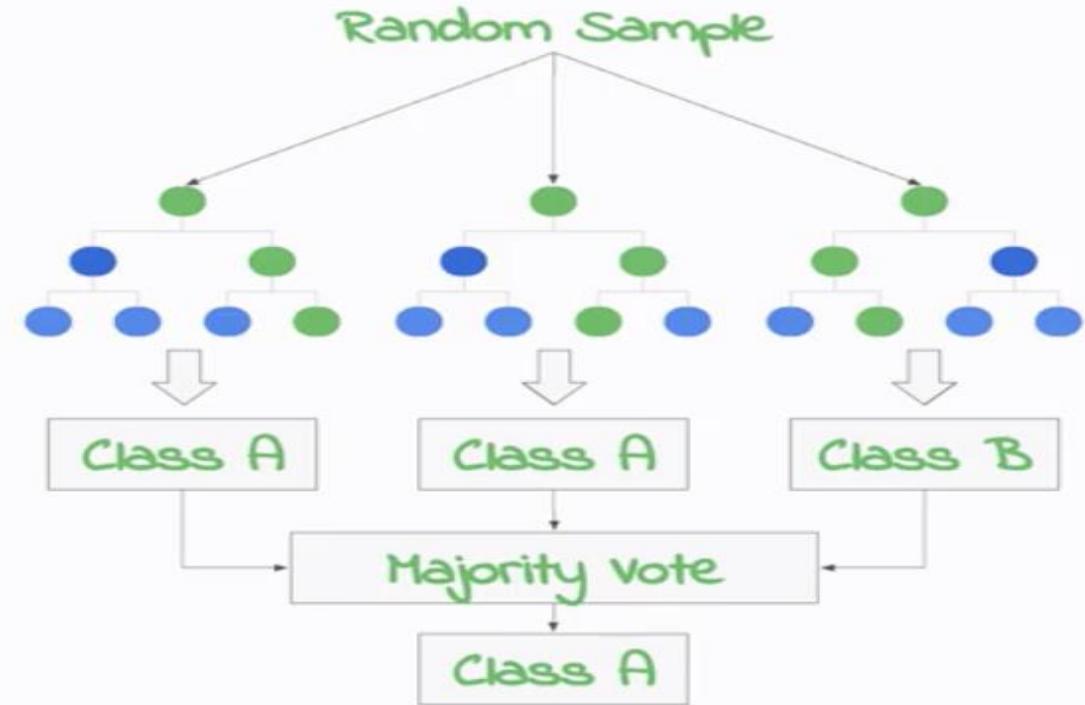
```
acc = accuracy_score(y_test, pred)
```

```
print("Accuracy: {:.4f}".format(acc))
```

앙상블 모델 - 배깅(Bagging)

- 다양한 분류기를 만드는 각기 다른 훈련 알고리즘을 사용하는 것과, 같은 알고리즘을 사용하고, 훈련 세트의 서브셋을 무작위로 구성하여 각기 다르게 학습시키는 방법이 있습니다.
- 훈련세트에서 중복을 허용하여 샘플링 하는 방식을 **bootstrap aggregating**, 배깅(**bagging**)라고 합니다.
- 통계학에서 중복을 허용한 리샘플링을 부트스트래핑(bootstrapping)이라고 합니다.
- 중복을 허용하지 않고 샘플링 하는 방식은 페이스팅(pasting)이라고 합니다.
- 랜덤 포레스트(Random Forest)는 일반적으로 배깅(또는 페이스팅)을 적용한 의사결정트리의 앙상블입니다.

Random forest: Strong learner from many weak learners



앙상블 모델 - 랜덤 포레스트(Random Forest)



```
from sklearn.ensemble import RandomForestClassifier
```

```
# 모델 학습
```

```
rfc = RandomForestClassifier(n_estimators=50, max_depth=3,  
                            random_state=20)
```

```
rfc.fit(X_train, y_train)
```

```
# 예측
```

```
pred = rfc.predict(X_test)  
print('예측값:', pred[:10])
```

```
# 모델 성능 평가
```

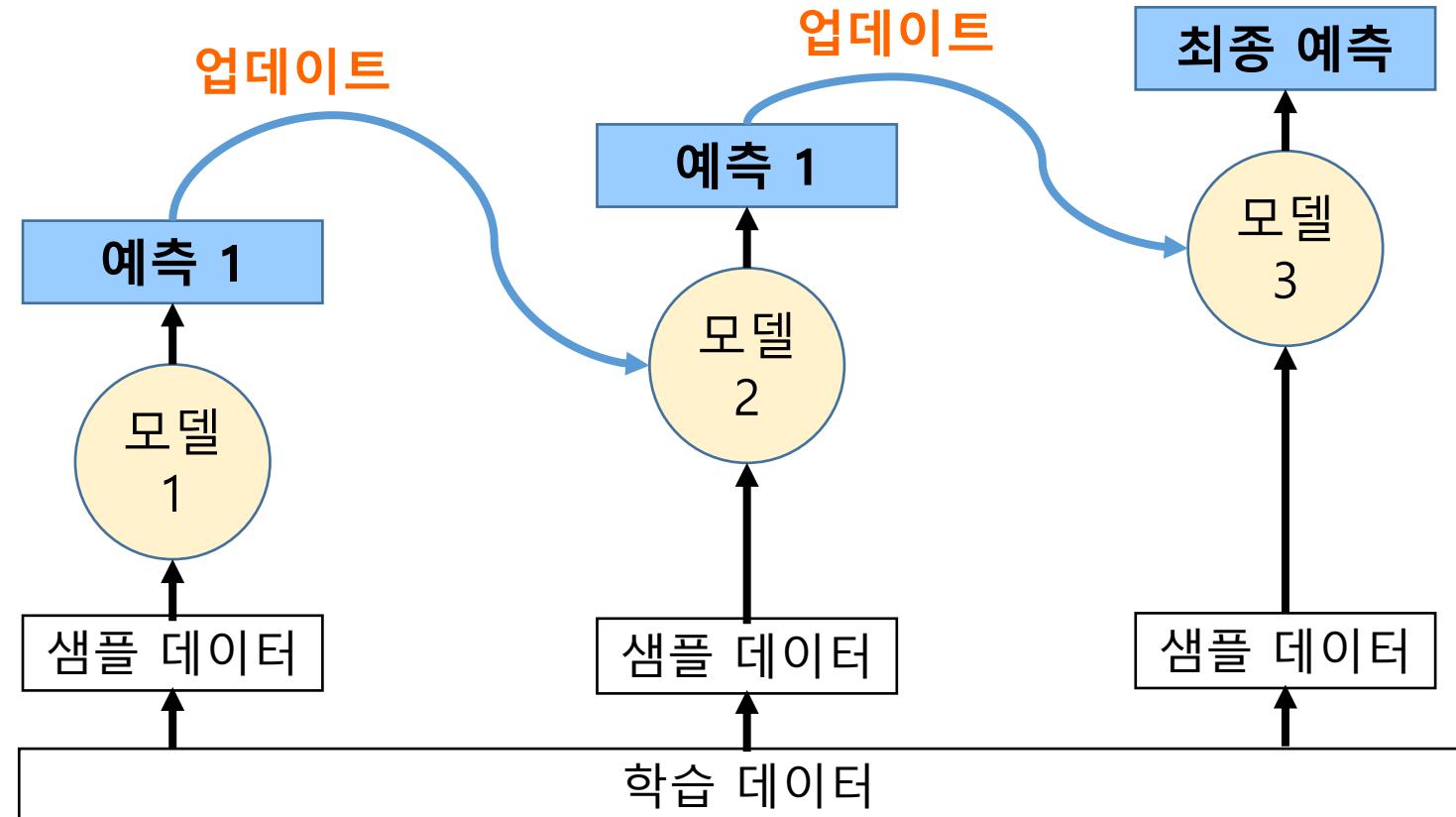
```
accuracy = accuracy_score(y_test, pred)  
print(f'Mean accuracy score: {accuracy:.4f}')
```

shift+tab키 : 함수 설명 보기

```
Init signature:  
RandomForestClassifier(  
    n_estimators=100,  
    *,  
    criterion='gini',  
    max_depth=None,  
    min_samples_split=2,  
    min_samples_leaf=1,  
    min_weight_fraction_leaf=0.0,  
    max_features='auto',
```

앙상블 모델 - 부스팅(Boosting)

- 부스팅(Boosting)은 여러 개의 모델을 순차적으로 학습합니다.
- 잘못 예측한 데이터에 대한 예측 오차를 줄일 수 있는 방향으로 모델을 계속 업데이트 합니다.
- XGBoost 모델은 Kaggle(<https://www.kaggle.com/>) 경진대회에서 많이 사용되고 있는 알고리즘의 하나입니다.



앙상블 모델 - 부스팅(XGBoost)



```
!pip install xgboost
```

```
from xgboost import XGBClassifier
```

```
# 모델 학습
```

```
xgbc = XGBClassifier(n_estimators=50, max_depth=3,  
                      random_state=42)
```

```
xgbc.fit(X_train, y_train)
```

```
# 예측
```

```
pred = xgbc.predict(X_test)
```

```
print('예측값: ', pred[:10])
```

```
# 모델 성능 평가
```

```
acc = accuracy_score(y_test, pred)
```

```
print(f'Mean accuracy score: {accuracy:.4}')
```

머신러닝 모델구현 실습



데이터파일 : churn_data.csv

| customerID | gender | SeniorCitizen | TotalCharges | Churn |
|------------|--------|---------------|--------------|-------|
| 7590-VHVEG | Female | 0 | 29.85 | No |
| 5575-GNVDE | Male | 0 | 1889.5 | No |
| 3668-QPYBK | Male | 0 | 108.15 | Yes |
| 7795-CFOCW | Male | 0 | 1840.75 | No |
| 9237-HQITU | Female | 0 | 151.65 | Yes |
| 9305-CDSKC | Female | 0 | 820.5 | Yes |
| 1452-KIOVK | Male | 0 | 1949.4 | No |
| 6713-OKOMC | Female | 0 | 301.9 | No |

- customerID: 고객ID
- gender: 고객 성별
- SeniorCitizen: 고객이 노약자인가 아닌가
- Partner: 고객에게 파트너가 있는지 여부(결혼 여부)
- Dependents: 고객의 부양 가족 여부
- tenure: 고객이 회사에 머물렀던 개월 수
- PhoneService: 고객에게 전화 서비스가 있는지 여부
- MultipleLines: 고객이 여러 회선을 사용하는지 여부
- InternetService: 고객의 인터넷 서비스 제공업체
- OnlineSecurity: 고객의 온라인 보안 여부
- OnlineBackup: 고객이 온라인 백업을 했는지 여부
- DeviceProtection: 고객에게 기기 보호 기능이 있는지 여부
- TechSupport: 고객이 기술 지원을 받았는지 여부
- StreamingTV: 고객이 스트리밍TV를 가지고 있는지 여부
- StreamingMovies: 고객이 영화를 스트리밍하는지 여부
- Contract: 고객의 계약기간
- PaperlessBilling: 고객의 종이 없는 청구서 수신 여부(모바일 청구서)
- PaymentMethod: 고객의 결제 수단
- MonthlyCharges: 매월 고객에게 청구되는 금액
- TotalCharges: 고객에게 청구된 총 금액
- Churn: 고객 이탈 여부 (label, y)

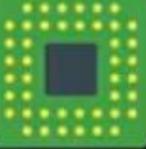
머신러닝 모델구현 실습



09-ML-Exercise.ipynb



1. TotalCharges 컬럼의 공백값을 문자 '0'으로 변경하고 수치형 데이터 타입으로 변환하세요.
2. 고객이탈여부 데이터를 변수 y에 할당하고 나머지 데이터를 변수 X에 할당하세요.
3. X, y 데이터셋을 70%:30% 비율로 훈련데이터셋과 검증데이터셋으로 분할하세요
4. 랜덤 포레스트 모델로 이탈고객 예측분류기를 만들고 모델성능을 출력하세요.
5. XGBoost 모델로 이탈고객 예측분류기를 모델성능을 측정하세요.



제프리 힌튼



훈련 데이터

출처 : <https://youtu.be/C2sqt9pG6K0>



5. 스타트 딥러닝

인공신경망(Artificial Neural Network, ANN)

심층신경망(Deep Neural Network, DNN)

가중치(Weight)

활성화 함수(Activation Function)

손실함수(Loss Function)

딥러닝 학습

순전파(Forward Propagation)

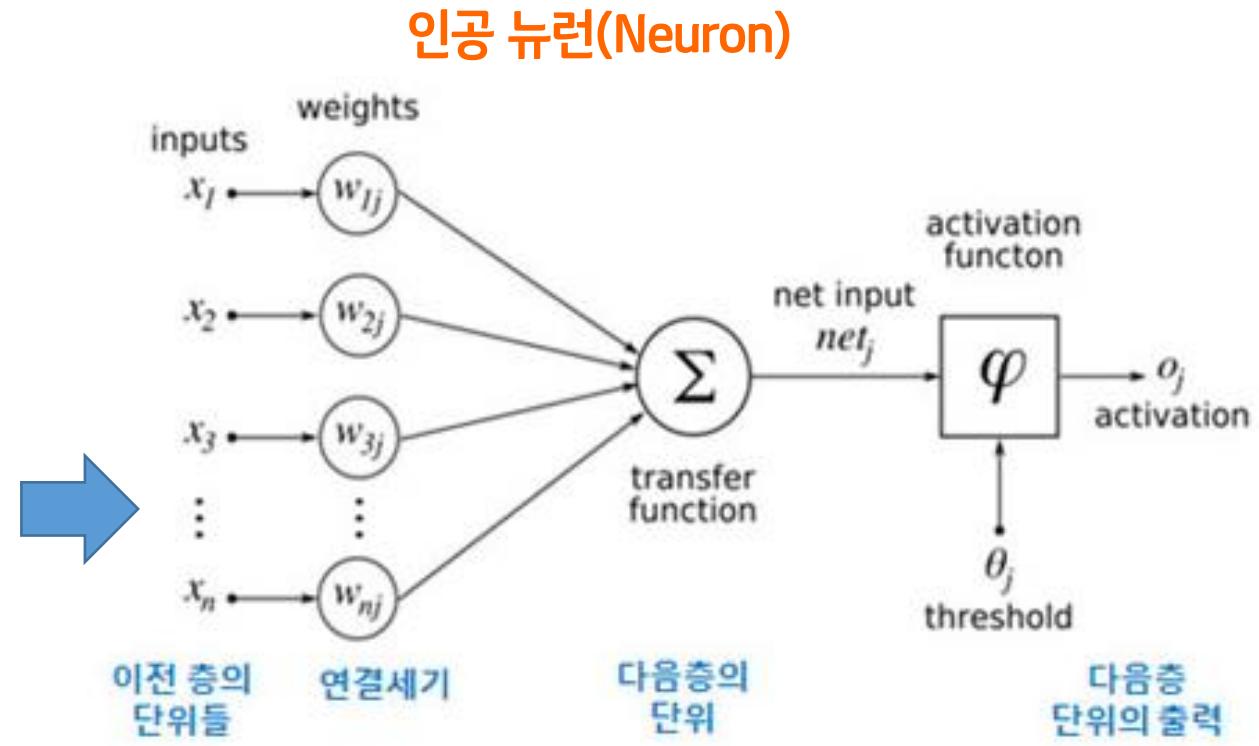
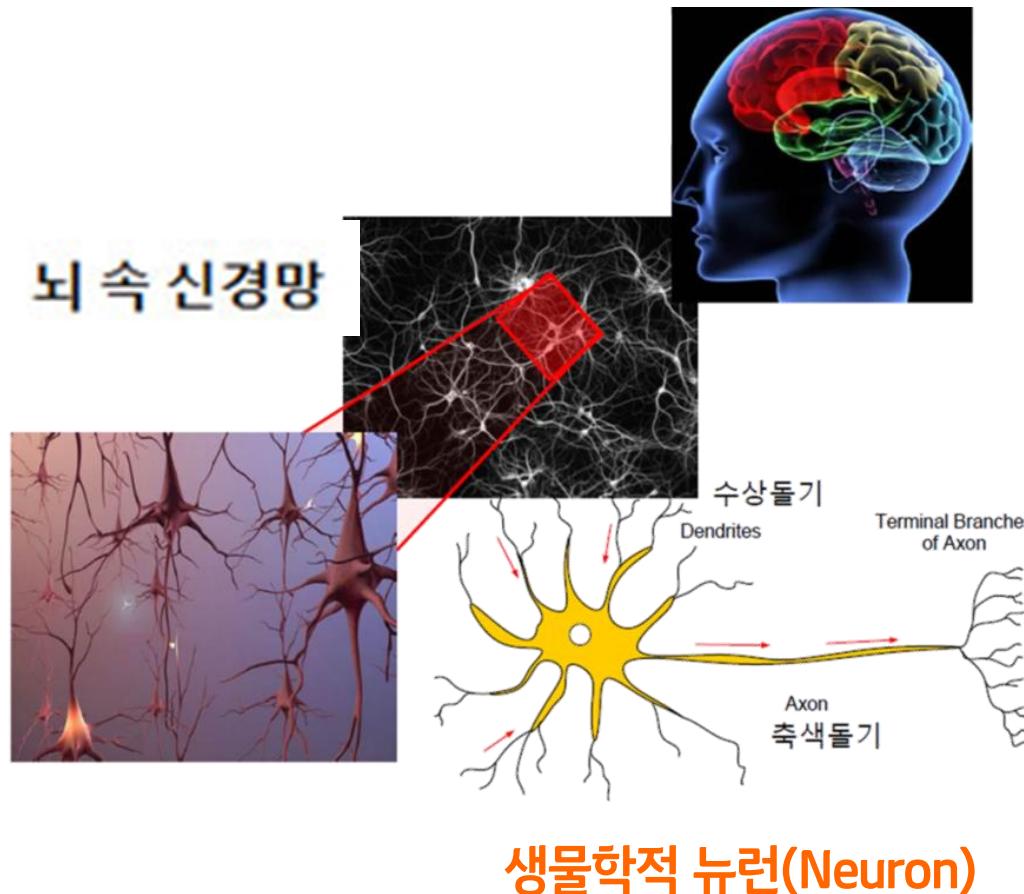
오차역전파>Error Back Propagation)

경사하강법(Gradient Descent)

옵티마이저(Optimization Algorithm)

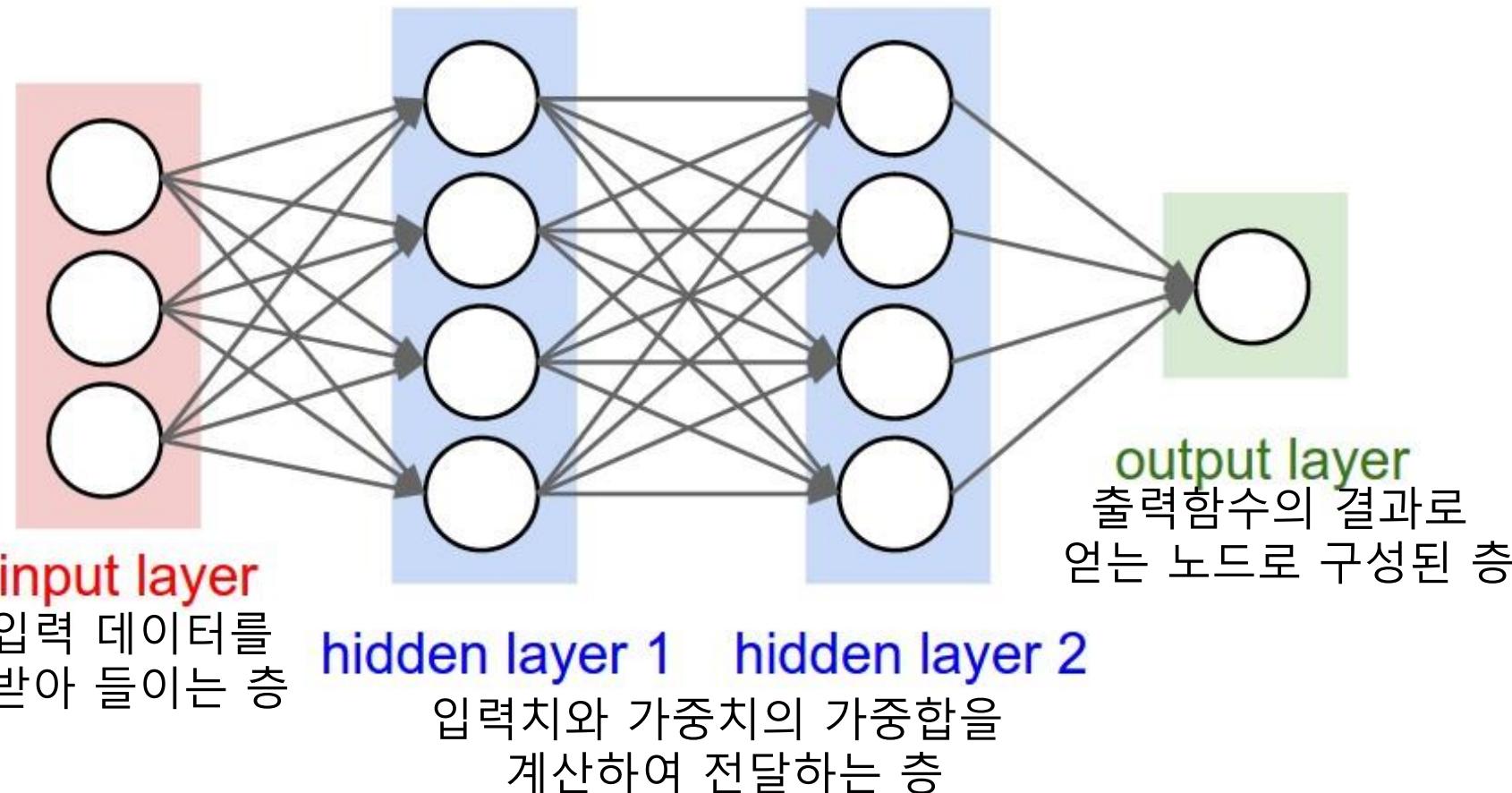
인공신경망(Artificial Neural Network, ANN)

- 뇌신경은 수많은 신경세포(뉴런, neuron)들이 연결되어 정보를 처리하고 전달합니다.
- 인공신경망은 뇌 신경계의 정보처리 구조를 모방하여 만든 계산 알고리즘입니다.
- 뇌 신경계와 같이 수많은 계산 함수를 연결하여 복잡한 정보를 처리하는 네트워크 구조입니다.

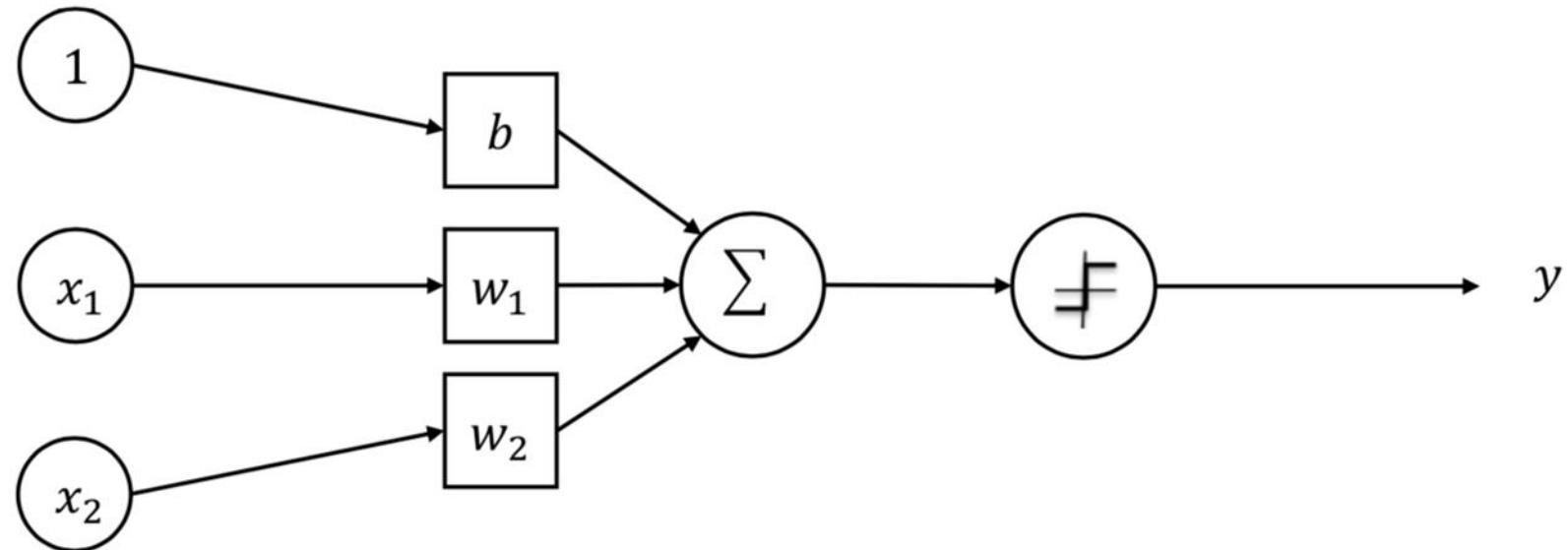


심층신경망(Deep Neural Network, DNN)

- 딥러닝은 여러 층(layer)을 가진 인공신경망(Artificial Neural Network, ANN)을 사용하여 학습을 수행하는 것입니다.
- 심층신경망은 입력층과 출력층사이에 다수의 은닉층(hidden layer)을 포함하는 인공신경망입니다.
- 머신러닝에서는 비선형 분류를 위해 여러 trick을 사용하지만, DNN은 다수의 은닉층으로 비선형 분류가 가능해집니다.



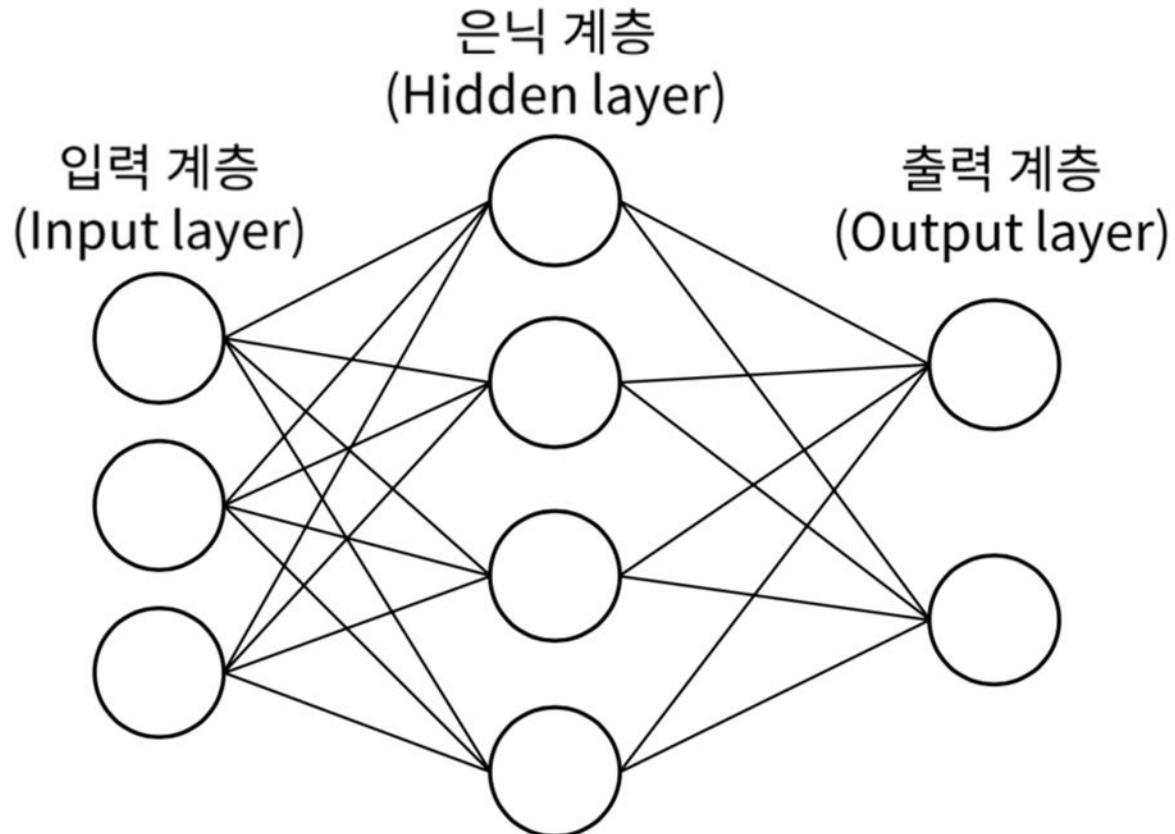
뉴런 Neuron



| 입력 (Input node) | 가중치와 편향 (Weights and bias) | 활성 함수 (Activation function) | 출력 (Output node) |
|--------------------|-------------------------------|--------------------------------|---------------------|
|--------------------|-------------------------------|--------------------------------|---------------------|

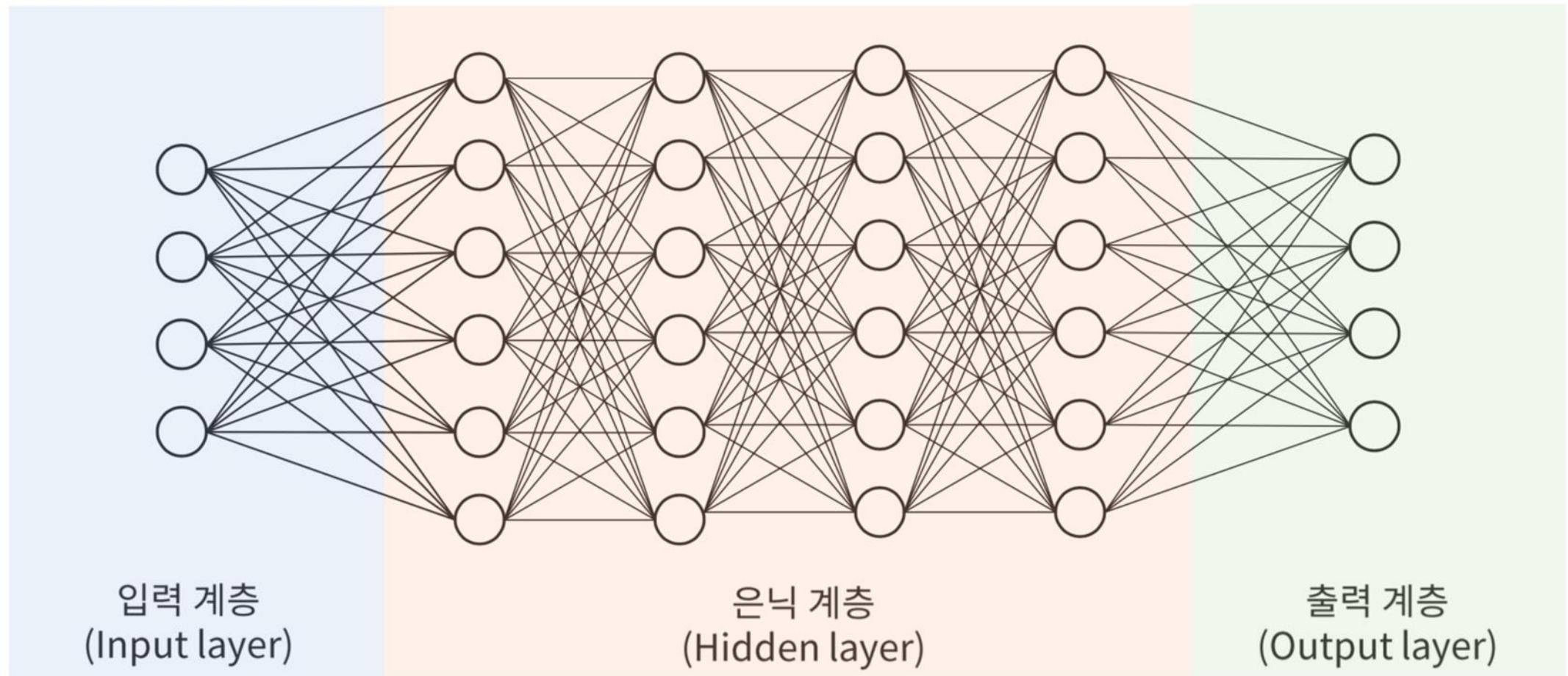
신경망은 **뉴런을 기본 단위**로 하며, 이를 조합하여 복잡한 구조를 이룬다.

얕은 신경망 Shallow Neural Network



가장 단순하고 얕은(은닉 계층이 1개인) 신경망 구조를 얕은 신경망이라고 한다.

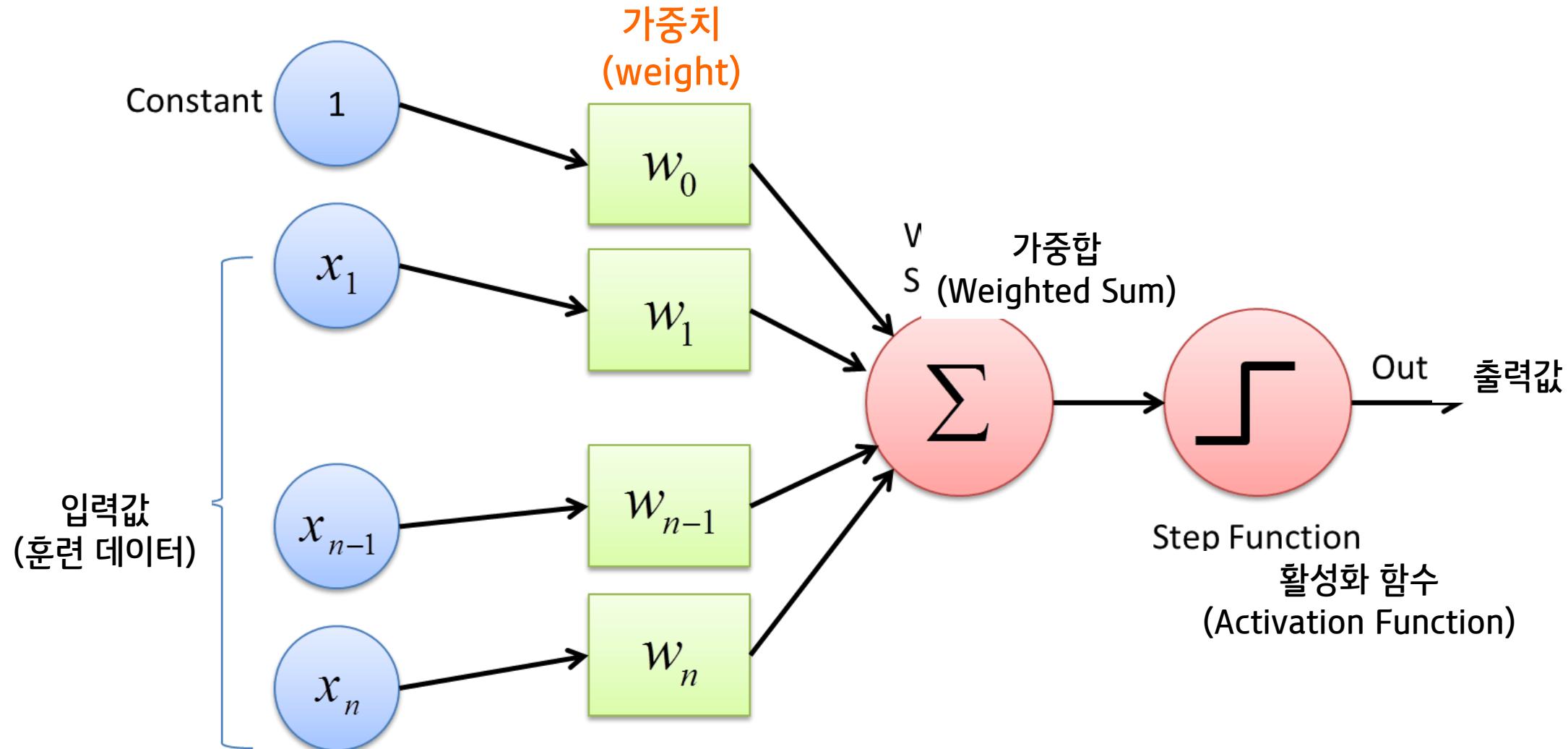
심층 신경망 Deep Neural Network (DNN)



- 얕은 신경망보다 은닉 계층이 많은 신경망을 DNN이라고 부른다.

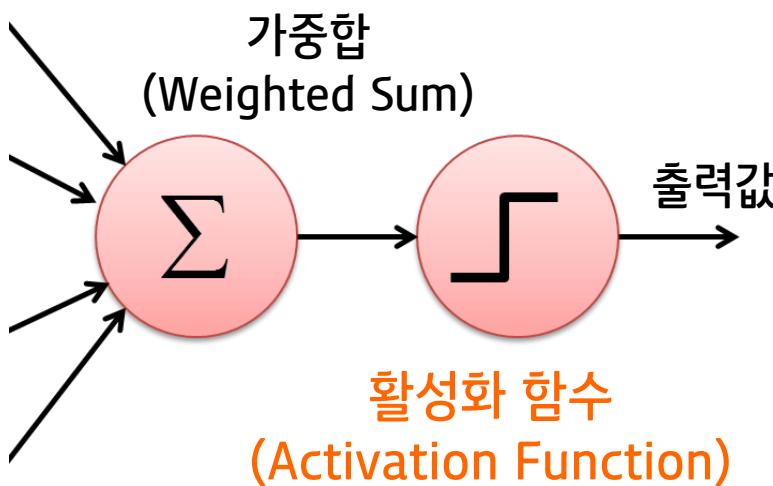
가중치(Weight)

가중치는 입력값이 연산결과에 미치는 영향도를 조절하는 요소입니다.



활성화 함수(Aactivation function)

입력값들의 수학적 선형결합을 다양한 형태의 비선형(또는 선형) 결합으로 변환하는 역할을 합니다.



| | | |
|--|--|---|
| Binary step | | $f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$ |
| Logistic (a.k.a. Sigmoid or Soft step) | | $f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}^{[1]}$ |
| TanH | | $f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$ |
| Rectified linear unit (ReLU) ^[12] | | $f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ x & \text{for } x > 0 \end{cases} = \max\{0, x\} = x\mathbf{1}_{x>0}$ |

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

참고 : <https://woonoo.tistory.com/209>

손실함수(Loss Function)

인공신경망 학습의 목적함수로 출력값(예측값)과 정답(실제값)의 차이를 계산합니다.

■ 회귀(regression)

평균제곱오차 : Mean Squared Error $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2$

평균절대오차 : mean absolute error $MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$

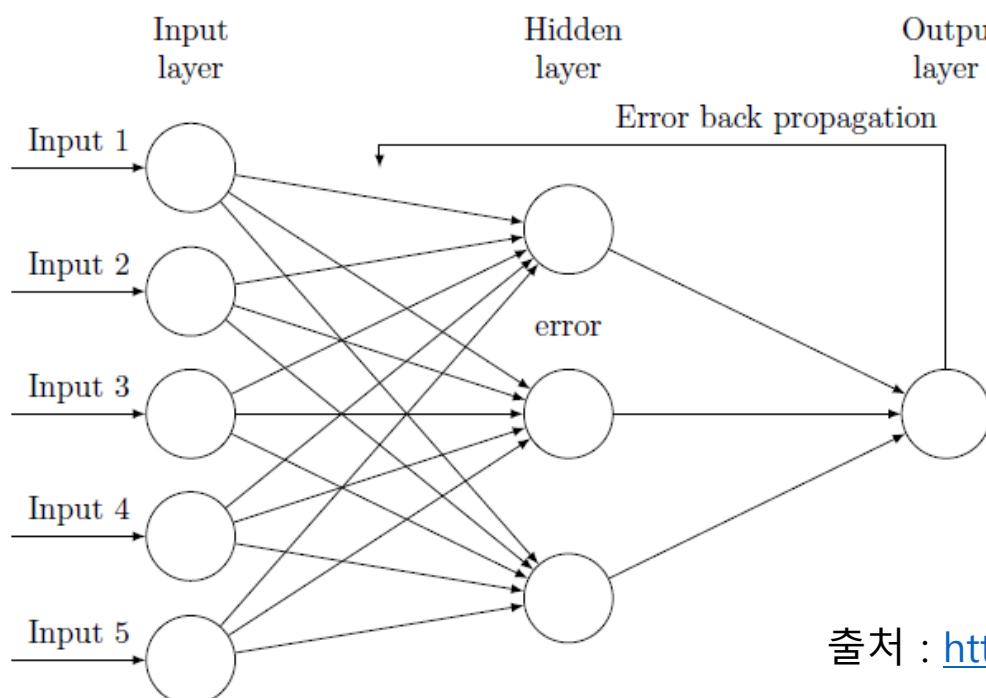
■ 분류(classification)

이진분류 : binary cross-entropy $L = -\frac{1}{N} \sum_{i=1}^N t_i \log(y_i) + (1 - t_i) \log(1 - y_i)$

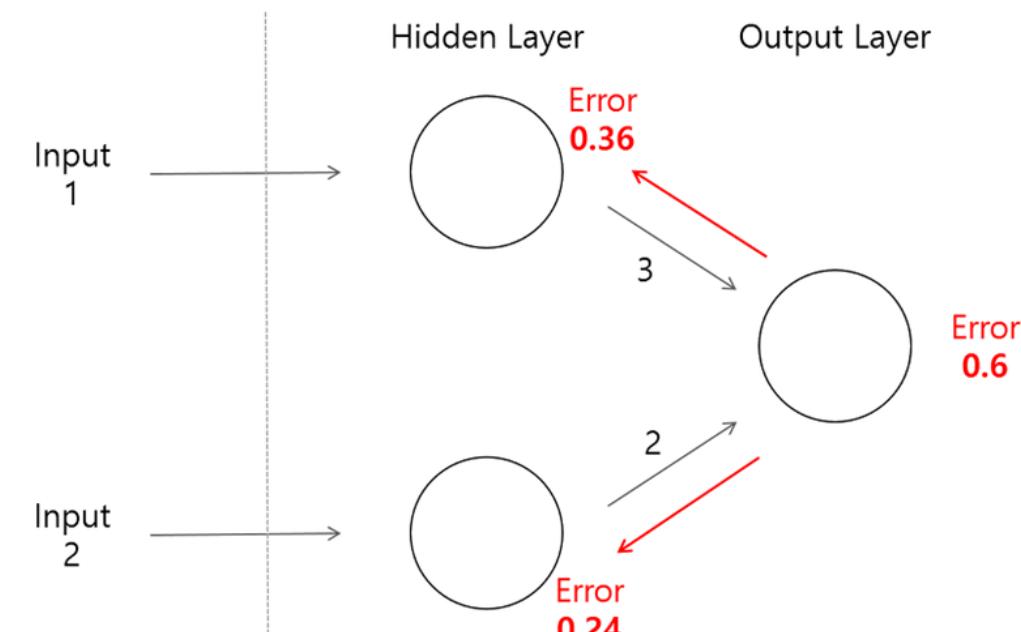
다중 분류: categorical cross-entropy $L = -\frac{1}{N} \sum_{j=1}^N \sum_{i=1}^C t_{ij} \log(y_{ij})$

딥러닝 학습

- 딥러닝 학습은 손실/에러(Loss/Error)를 최소화 하는 인공신경망의 가중치(weight)와 편향(bias)을 찾는 과정입니다.
- 딥러닝 학습은 순전파(Forward Propagation)와 오차역전파(Error Back Propagation)의 반복으로 진행이 됩니다.
- 순전파는 뉴럴 네트워크의 입력층부터 출력층까지 순서대로 변수들을 계산하고 저장하는 것입니다.
- 오차역전파는 결과값을 통해서 역으로 input 방향으로 오차(Error)를 다시 보내며 가중치를 재업데이트 하는 것으로, 결과에 영향을 많이 미친 노드(뉴런)에 더 많은 오차를 돌려 줍니다.

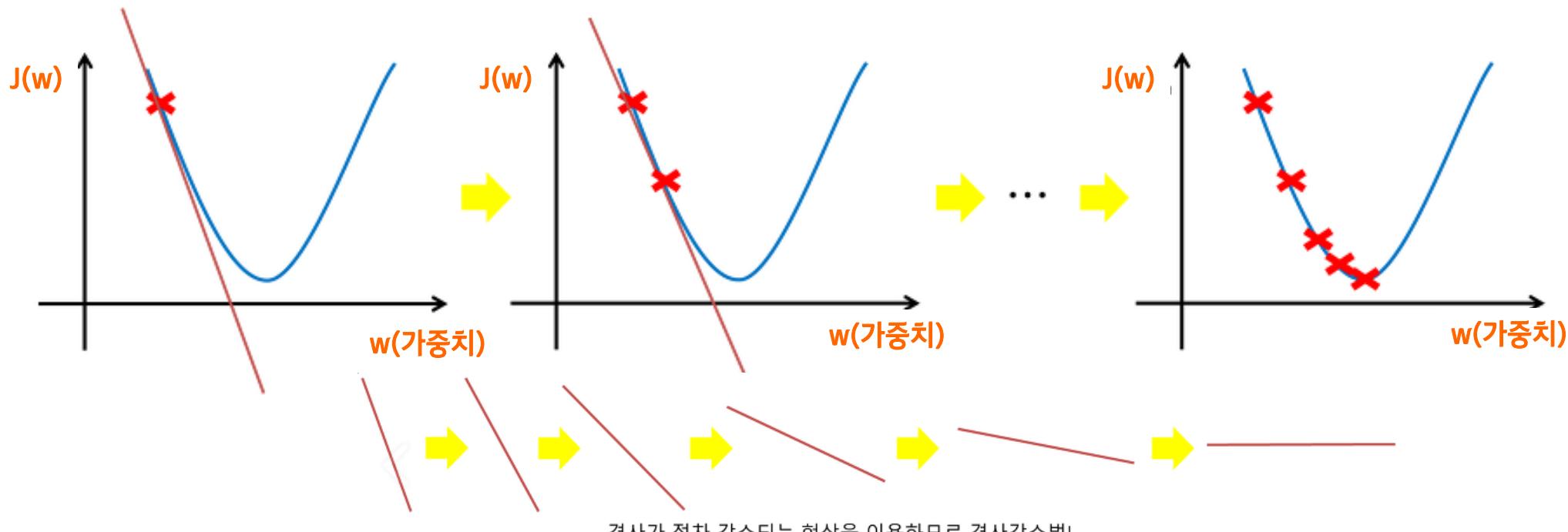


출처 : <https://sacko.tistory.com/19>



경사하강법(Gradient Descent)

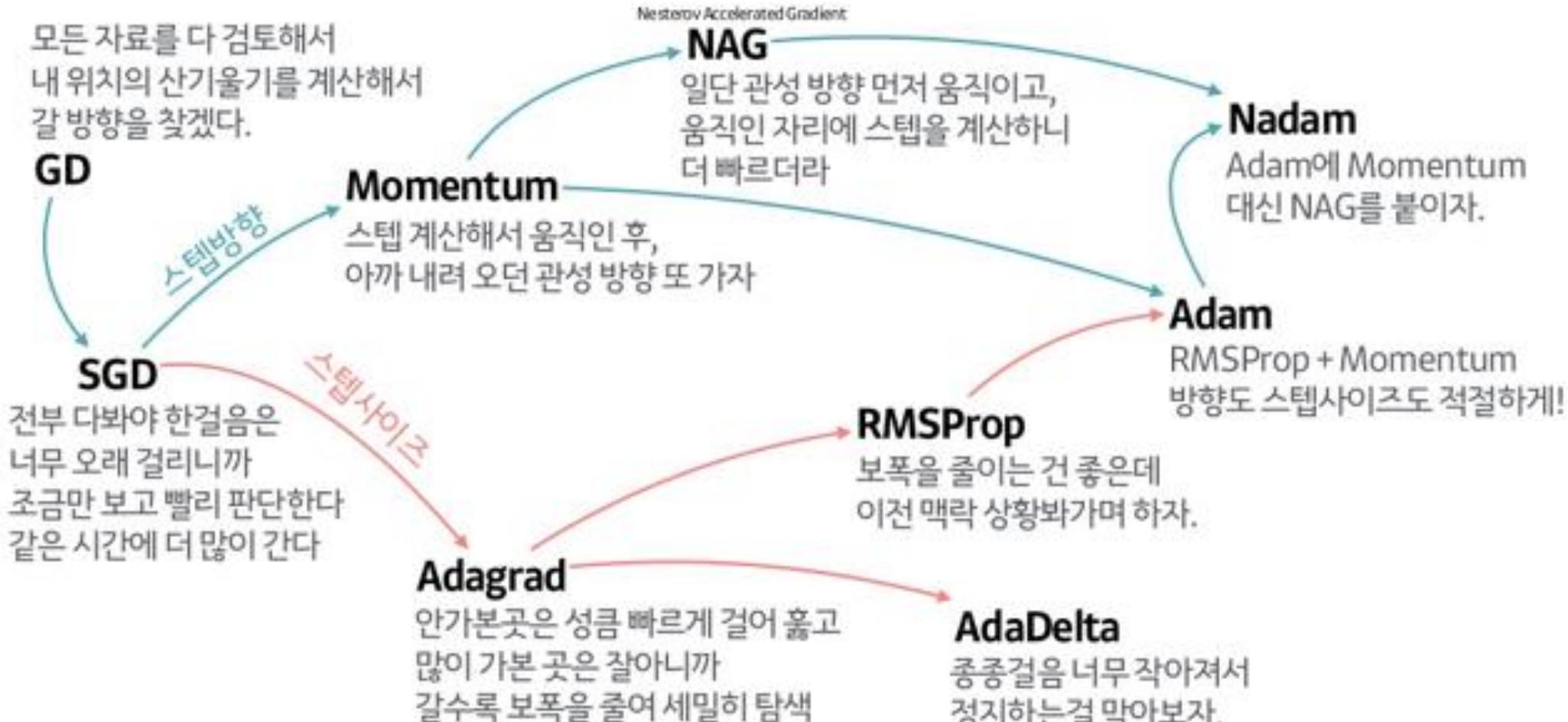
- 손실함수 $J(w)$ 는 가중치(w)의 함수로, 볼록함수 형태라면 미분으로 손실이 가장 작은 가중치를 찾을 수 있습니다.
- 하지만, 딥러닝에서는 손실함수가 복잡하고 계산량이 매우 크고, 미분이 0이 되는 값이 여러 개 존재하므로 미분만으로 최소값을 찾기 어려워 경사하강법(Gradient Descent)을 사용합니다.
- 경사하강법은 손실함수의 현 가중치에서 기울기를 구해서 손실(Loss)을 줄이는 방향으로 업데이트 해 나갑니다.



참고 : https://angeloyeo.github.io/2020/08/16/gradient_descent.html
https://angeloyeo.github.io/2020/08/16/gradient_descent.html

옵티마이저(Optimization Algorithm)

손실함수를 최소화하는 방향으로 가중치를 갱신하는 알고리즘입니다.



딥러닝 용어 정리

■ 딥러닝 학습방법

딥러닝 학습의 목표는 모델에 입력값을 넣었을 때의 출력값이 최대한 정답과 일치하게 하는 것입니다.

딥러닝 모델의 매개변수(weight, bias)를 무작위로 부여한 후,

반복학습(순전파-오차역전파)을 통해 모델의 출력값을 정답과 일치하도록 매개변수(weight, bias)를 조금씩 조정합니다.

■ 순전파(Forward Propagation)

딥러닝 모델에 값을 입력해서 출력을 얻는 과정입니다.

■ 오차역전파(Error Backpropagation)

실제값과 모델 결과값에서 오차를 구해서, 오차를 input 방향으로 보내서 가중치를 재업데이트 하는 과정입니다.

■ 손실함수(Loss Function)

손실함수는 신경망 학습의 목적으로(목적함수) 출력값과 정답의 차이를 계산합니다.

■ 과적합(overfitting)

생성된 모델이 학습 데이터와 지나치게 일치하여 새 데이터를 올바르게 예측하지 못하는 경우입니다.

딥러닝 용어 정리

■ 최적화(Optimization)

딥러닝 모델의 매개변수(weight, bias)를 조절해서 손실함수의 값을 최저로 만드는 과정으로 경사하강법(Gradient Descent)이 대표적입니다.

■ 경사하강법(gradient descent)

학습 데이터의 조건에 따라 모델의 매개변수를 기준으로 손실의 경사를 계산하여 손실을 최소화하는 기법입니다. 쉽게 설명하면, 경사하강법은 매개변수를 반복적으로 조정하면서 손실을 최소화하는 가중치와 편향의 가장 적절한 조합을 점진적으로 찾는 방식입니다.

■ 경사(gradients)

모든 독립 변수를 기준으로 한 편미분의 벡터입니다. 머신러닝에서 경사는 모델 함수의 편미분의 벡터입니다.

■ 일반화(generalization)

모델에서 학습에 사용된 데이터가 아닌 이전에 접하지 못한 새로운 데이터에 대해 올바른 예측을 수행하는 능력을 의미합니다.

딥러닝 용어 정리

■ 시퀀스 모델(sequence model)

입력에 순서 종속성이 있는 모델입니다

■ 밀집 연결층(Dense Layer)

완전 연결층(fully connected layer)이나 밀집 층(dense layer)라고 불리는 밀집 연결 층(densely connected layer)

■ 입력 레이어(input layer)

신경망의 첫 번째 레이어로서 입력 데이터를 수신합니다.

■ 하든 레이어(hidden layer)

신경망에서 입력 레이어(특성)와 출력 레이어(예측) 사이에 위치하는 합성 레이어입니다.

신경망에 하나 이상의 하든 레이어가 포함될 수 있습니다.

■ 드롭아웃(dropout)

신경망의 과적합을 방지하는 방법으로 단일 경사 스텝이 일어날 때마다

특정 네트워크 레이어의 유닛을 고정된 개수만큼 무작위로 선택하여 삭제합니다.

■ 출력 레이어(output layer)

신경망의 '최종' 레이어입니다. 이 레이어에 답이 포함됩니다.

딥러닝 용어 정리



1 Epoch : 모든 데이터 셋을 한 번 학습

1 iteration : 1회 학습

minibatch : 데이터 셋을 batch size 크기로 쪼개서 학습

ex) 총 데이터가 100개, batch size가 100이면,

1 iteration = 10개 데이터에 대해서 학습

1 Epoch = $100/\text{batch size} = 10 \text{ iteration}$

TensorFlow

Google에서 만든, 딥러닝 프로그램을 쉽게 구현할 수 있도록 다양한 기능을 제공해주는 라이브러입니다.

<https://www.tensorflow.org/tutorials/>

The screenshot shows the TensorFlow website's 'TensorFlow 시작하기' (Getting Started) page. On the left, there is a sidebar with categories like 'ML 학습 및 사용', '연구 및 실험', '프로덕션 규모의 ML', etc. The main content area has a heading 'TensorFlow 시작하기' and a brief introduction. Below it, there is a section titled 'ML 학습 및 사용' with a sub-section about Keras API. A code snippet for a Keras model is displayed:

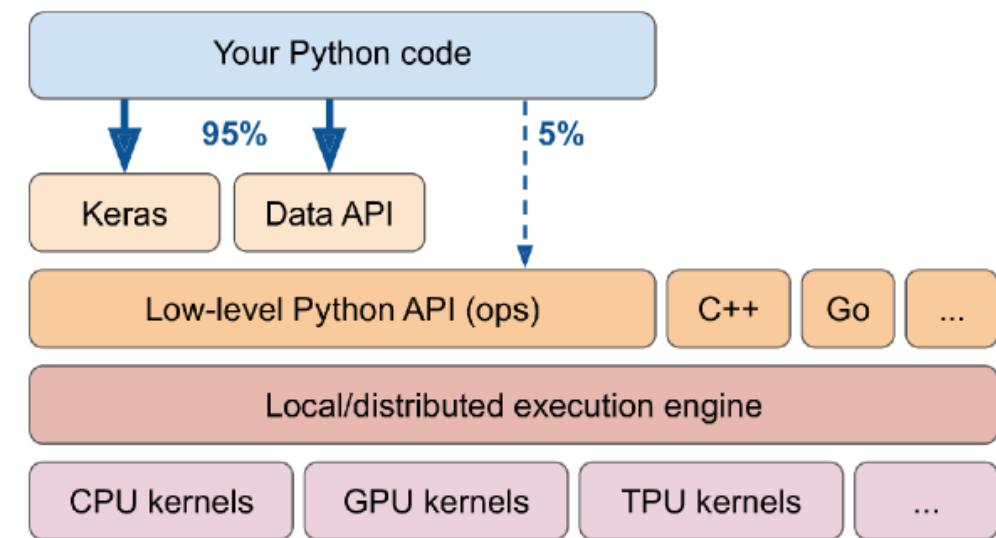
```
import tensorflow as tf
mnist = tf.keras.datasets.mnist

(x_train, y_train), (x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0

model = tf.keras.models.Sequential([
    tf.keras.layers.Flatten(input_shape=(28, 28)),
    tf.keras.layers.Dense(512, activation=tf.nn.relu),
    tf.keras.layers.Dropout(0.2),
    tf.keras.layers.Dense(10, activation=tf.nn.softmax)
])
model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

model.fit(x_train, y_train, epochs=5)
model.evaluate(x_test, y_test)
```

텐서플로 구조



TensorFlow

AI 모델 구현



10-DNN.ipynb

- ① 라이브러리 임포트(import)
- ② 데이터 가져오기>Loading the data)
- ③ 탐색적 데이터 분석(Exploratory Data Analysis)
- ④ 데이터 전처리(Data PreProcessing) : 데이터타입 변환, Null 데이터 처리, 누락데이터 처리, 카테고리 데이터, 더미특성 생성, 특성 추출 (feature engineering) 등
- ⑤ 훈련/테스트 데이터 분할(Train Test Split)
- ⑥ 데이터 정규화(Normalizing the Data)
- ⑦ 모델 개발(Creating the Model)
- ⑧ 모델 성능 평가(Evaluating Model Performance)

AI 모델 학습 데이터



데이터파일 : churn_data.csv

| customerID | gender | SeniorCitizen | TotalCharges | Churn |
|------------|--------|---------------|--------------|-------|
| 7590-VHVEG | Female | 0 | 29.85 | No |
| 5575-GNVDE | Male | 0 | 1889.5 | No |
| 3668-QPYBK | Male | 0 | 108.15 | Yes |
| 7795-CFOCW | Male | 0 | 1840.75 | No |
| 9237-HQITU | Female | 0 | 151.65 | Yes |
| 9305-CDSKC | Female | 0 | 820.5 | Yes |
| 1452-KIOVK | Male | 0 | 1949.4 | No |
| 6713-OKOMC | Female | 0 | 301.9 | No |

- customerID: 고객ID
- gender: 고객 성별
- SeniorCitizen: 고객이 노약자인가 아닌가
- Partner: 고객에게 파트너가 있는지 여부(결혼 여부)
- Dependents: 고객의 부양 가족 여부
- tenure: 고객이 회사에 머물렀던 개월 수
- PhoneService: 고객에게 전화 서비스가 있는지 여부
- MultipleLines: 고객이 여러 회선을 사용하는지 여부
- InternetService: 고객의 인터넷 서비스 제공업체
- OnlineSecurity: 고객의 온라인 보안 여부
- OnlineBackup: 고객이 온라인 백업을 했는지 여부
- DeviceProtection: 고객에게 기기 보호 기능이 있는지 여부
- TechSupport: 고객이 기술 지원을 받았는지 여부
- StreamingTV: 고객이 스트리밍TV를 가지고 있는지 여부
- StreamingMovies: 고객이 영화를 스트리밍하는지 여부
- Contract: 고객의 계약기간
- PaperlessBilling: 고객의 종이 없는 청구서 수신 여부(모바일 청구서)
- PaymentMethod: 고객의 결제 수단
- MonthlyCharges: 매월 고객에게 청구되는 금액
- TotalCharges: 고객에게 청구된 총 금액
- Churn: 고객 이탈 여부 (label, y)

심층신경망 구현

■ 라이브러리 임포트

```
import pandas as pd  
import seaborn as sns  
import matplotlib.pyplot as plt  
  
import tensorflow as tf  
from tensorflow.keras.models import Sequential  
from tensorflow.keras.layers import Dense, Activation, Dropout
```

■ 데이터 로드

```
df = pd.read_csv('churn_data.csv')
```

심층신경망 구현

■ 데이터 분석

```
df.info()  
df.isnull().sum()  
df.describe().transpose()  
df.corr()['MonthlyCharges'][:-1].sort_values().plot(kind='bar')  
sns.pairplot(df)
```

■ 데이터 전처리

```
df.drop('customerID', axis=1, inplace=True)  
df['TotalCharges'].replace([' '], ['0'], inplace=True)  
df['TotalCharges'] = df['TotalCharges'].astype(float)  
df['Churn'].replace(['Yes', 'No'], [1, 0], inplace=True)
```

심층신경망 구현

■ 데이터 전처리

```
cols = ['gender', 'Partner', 'Dependents', 'PhoneService',
'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup',
'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies',
'Contract', 'PaperlessBilling', 'PaymentMethod']
```

```
dummies = pd.get_dummies(df[cols], drop_first=True)
df = df.drop(cols, axis=1)
df = pd.concat([df, dummies], axis=1)
```

```
# df = pd.get_dummies(df)
# cols = list(df.select_dtypes('object').columns)
```

심층신경망 구현

■ 훈련/테스트 데이터 분할

```
from sklearn.model_selection import train_test_split  
X = df.drop('Churn', axis=1).values  
y = df['Churn'].values  
  
X_train, X_test, y_train, y_test = train_test_split(  
    X, y,  
    test_size=0.3,  
    random_state=42)  
  
X_train.shape  
[Out] (4930, 30)  
  
y_train.shape  
[Out] (4930,)
```



심층신경망 구현

■ 데이터 정규화/스케일링(Normalizing/Scaling)

```
from sklearn.preprocessing import MinMaxScaler
```

```
scaler = MinMaxScaler()  
scaler.fit(X_train)  
X_train = scaler.transform(X_train)  
X_test = scaler.transform(X_test)
```



심층신경망 구현

■ 모델 구성

```
model = Sequential()
```

input Layer

```
model.add(Dense(64, activation='relu', input_shape=(30,)))
```

hidden Layer

```
model.add(Dense(64, activation='relu'))
```

hidden Layer

```
model.add(Dense(32, activation='relu'))
```

output Layer

```
model.add(Dense(1, activation='sigmoid'))
```

심층신경망 구현

■ 모델 구성 - 과적합 방지

```
model = Sequential()  
model.add(Dense(128, activation='relu', input_shape=(30,)))  
model.add(Dropout(0.5))  
model.add(Dense(64, activation='relu'))  
model.add(Dropout(0.5))  
model.add(Dense(64, activation='relu'))  
model.add(Dropout(0.5))  
model.add(Dense(32, activation='relu'))  
model.add(Dropout(0.5))  
  
model.add(Dense(1, activation='sigmoid'))
```



심층신경망 구현

■ 모델 컴파일 - 이진 분류 모델

```
model.compile(optimizer='adam',  
              loss='binary_crossentropy',  
              metrics=['accuracy'])
```

■ 모델 컴파일 - 다중 분류 모델

```
model.compile(optimizer='adam',  
              loss='categorical_crossentropy',  
              metrics=['accuracy'])
```

■ 모델 컴파일 - 예측 모델

```
model.compile(optimizer='adam',  
              loss='mse')
```



심층신경망 구현

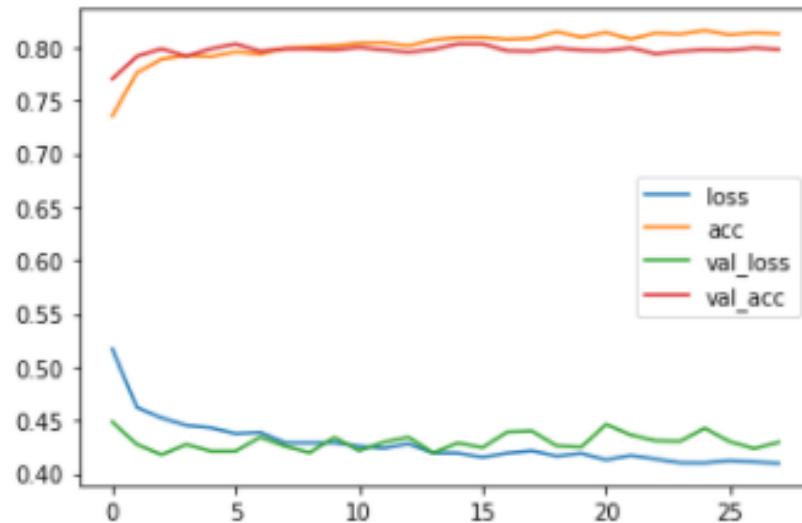
■ 모델 훈련

```
model.fit(X_train, y_train,  
          validation_data=(X_test, y_test), epochs=20, batch_size=10)
```



■ 모델 성능 평가

```
losses = pd.DataFrame(model.history.history)  
losses[['loss','val_loss']].plot()
```



심층신경망 구현 실습



11-DNN-Exercise.ipynb

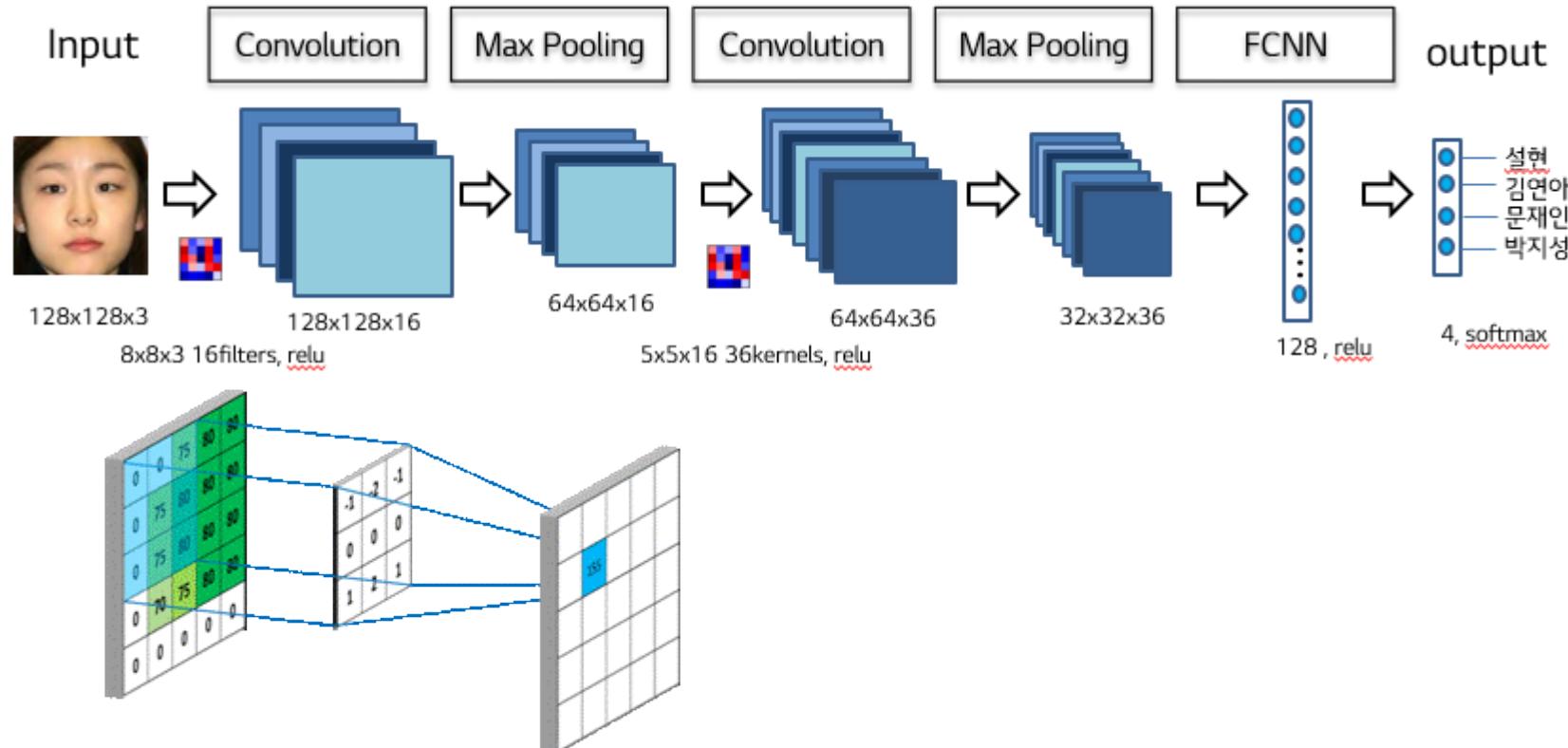


1. X_train, X_valid 값을 1,0 사이의 값으로 스케일링 하세요.
2. 딥러닝 심층신경망모델(DNN)로 통신사이탈고객을 예측하는 분류기를 만드세요.
검증정확도(val_acc)가 80% 이상이 나오도록 하이퍼 파라미터를 설정하세요.
그리고, 검증 정확도가 가장 높은 모델을 파일명 best_model.h5로 저장하세요.
3. 학습 정확도, 학습손실, 검증 정확도, 검증손실을 그래프로 표시하세요.

6. Further Study

합성곱 신경망(CNN, Convolutional Neural Network)

사람의 시각 피질 메커니즘에 영감을 받아 설계된 이미지, 영상등을 인식하는 신경망 모델
Convolution층에서는 각 filter가 입력 이미지의 픽셀 전체를 차례로 훑고 지나가며
linear combination을 진행하고 Feature Map을 구성합니다.

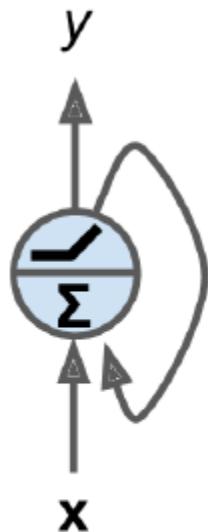


순환 신경망(RNN, Recurrent Neural Network)

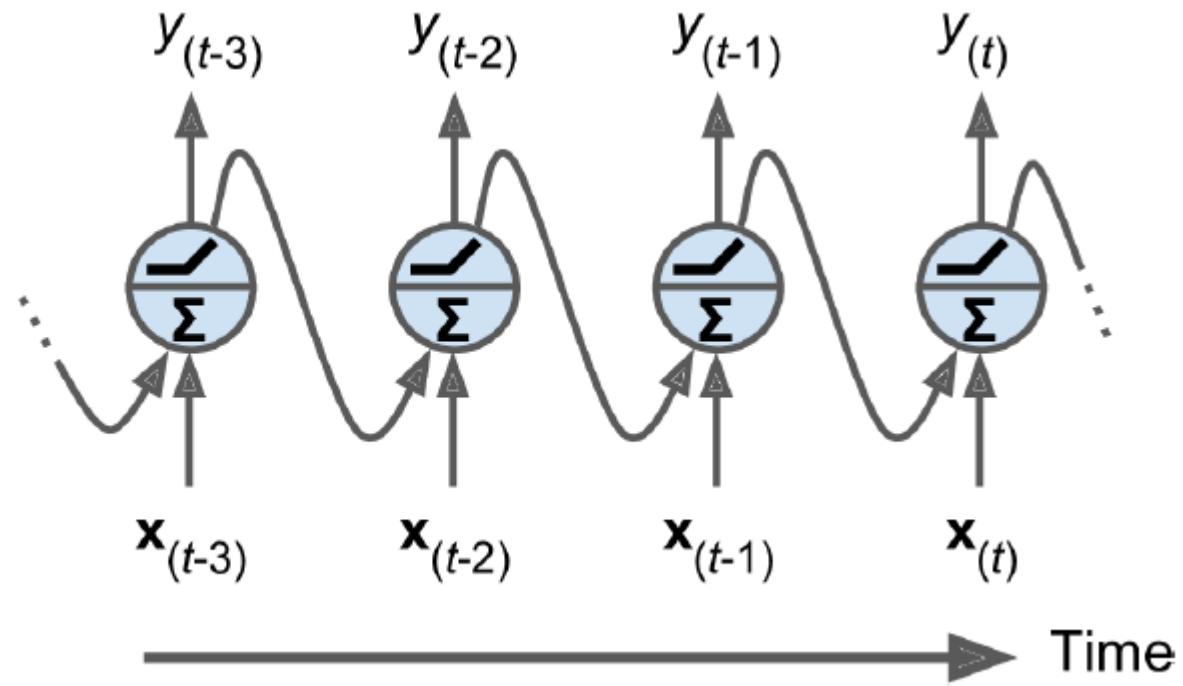
순환신경망은 고정 길이 입력이 아닌 임의 길이를 가진 시퀀스를 다룰 수 있습니다.

순환신경망은 시계열 데이터를 분석해서 미래값을 예측하고 문장, 오디오를 입력으로 받아 자동번역, 자연어처리에 유용합니다.

■ 순환뉴런

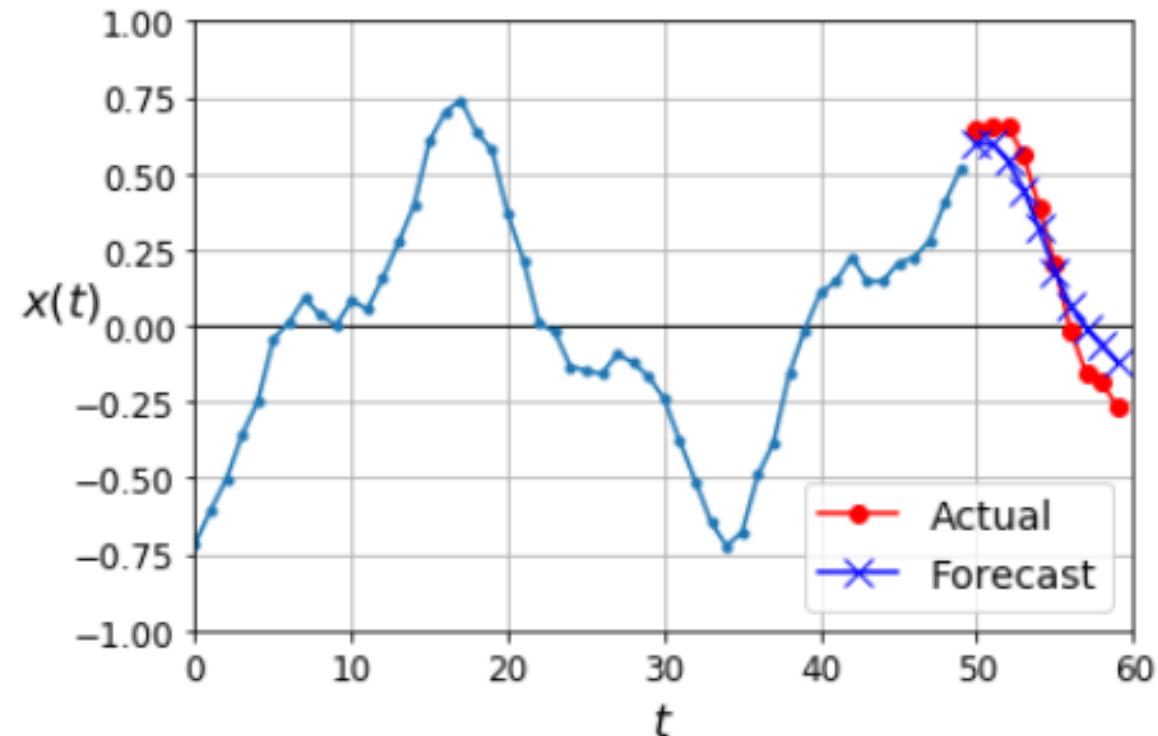
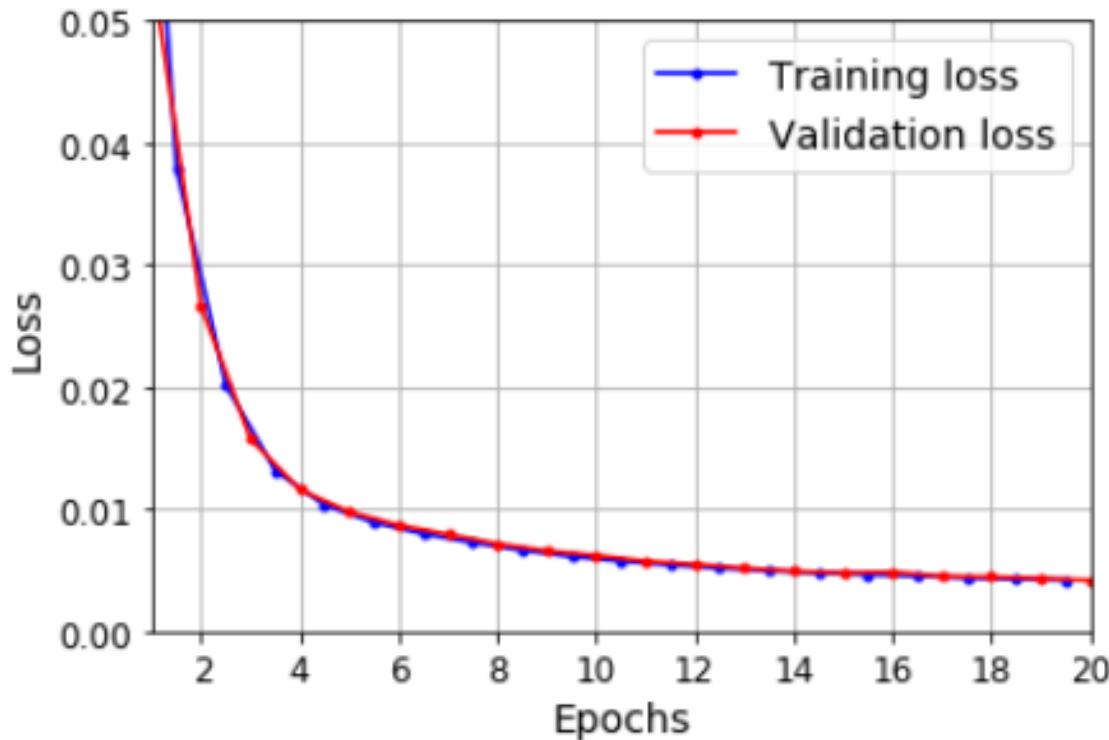
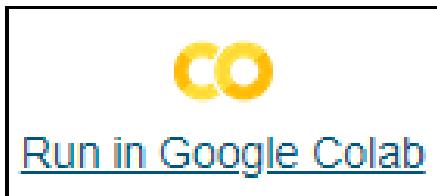


■ 순환뉴런을 타임 스텝으로 펼친 모습



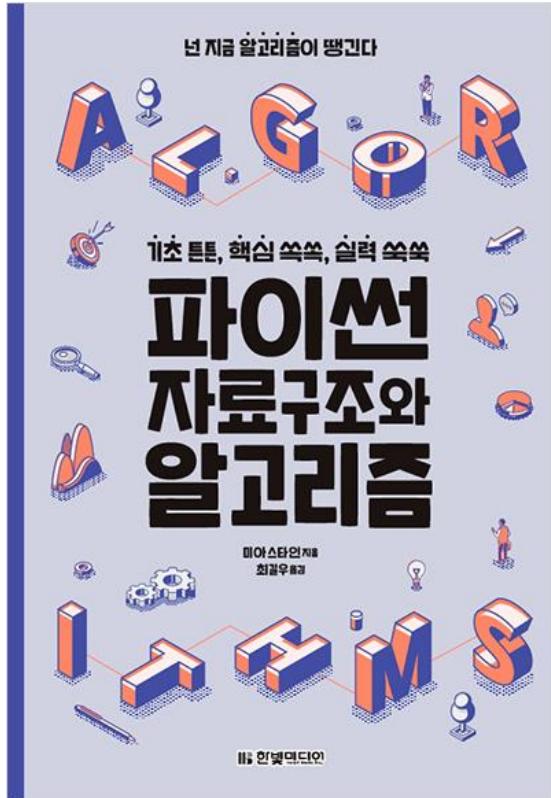
에너지 수요예측 모델

https://github.com/rickiepark/handson-ml2/blob/master/15_processing_sequences_using_rnns_and_cnns.ipynb



추천 도서

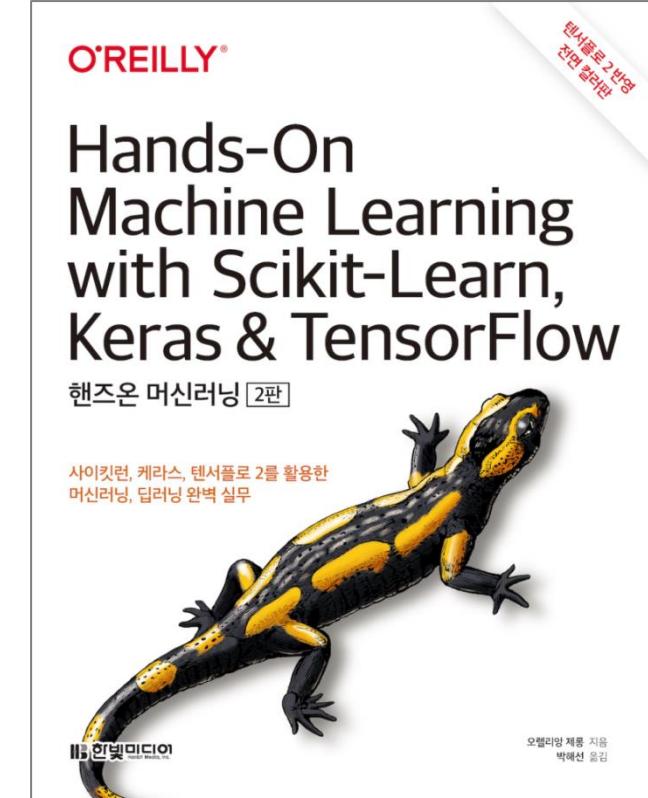
<https://bit.ly/38cErVX>



<https://bit.ly/3tbyWQf>



<https://bit.ly/36Ltr2X>



<https://tensorflow.blog/handson-ml2/>





수고하셨습니다. 감사합니다.