

[2022 기업멤버십 SW캠프]

클라우드 활용 빅데이터서비스 개발자 부트캠프

Semi-Project for Part 1~2

- Data collection / exploration / visualization
- Data preprocessing & Feature engineering
- Train & test traditional ML algorithms
- Train & test deep-learning models
- Compare various models & deliver the result

Daeyeon Jo
repositorator@gmail.com

본 교안 및 실습자료는 저작권법에 의거하여 본 교육 外 배포/게시/공개를 금합니다.

Course Overview

2023년 3월						
일	월	화	수	목	금	토
26	27	28	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	1
2	3	4				

2023년 4월						
일	월	화	수	목	금	토
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	1	2	3	4	5	6
7	8	9	10	11	12	13

- * 아래 커리큘럼의 세부 사항은 변동될 수 있습니다.
- * 상황에 따라 특정 파트가 아예 제외될 수 있습니다.
- * 진도 상황에 따라 2~3일 정도 차이가 발생할 수 있습니다.

머신러닝 핵심 이론 & 주요 알고리즘 이론
파이썬 기반 머신러닝 알고리즘 실습 (Scikit-learn)
+ 데이터 분석 관련 직무 & 학습 리소스 소개

1차 세미 프로젝트 (Feature engineering & applying ML algorithms)

딥러닝 핵심 이론 & 인공신경망 최적화 이론
파이썬 기반 딥러닝 알고리즘 실습 (Tensorflow & Keras)
+ 분야별 머신러닝 & 딥러닝 활용 사례 소개 + 각종 자동화 도구 실습

2차 세미 프로젝트 (데이터 수집 / 탐색 / 전처리 / 시각화 + ML&DL model tuning)

머신러닝 & 딥러닝 모델 활용을 위한 웹 프로그래밍
- Django Basic / Intermediate / Advanced
- 10 Steps to create a landing page
- ML & DL Models for NLP web services

Final-Project (ML/DL Model serving via webpage)

Final-Project 최종 발표

수업 관련 공지사항

*** 정형데이터 활용 권장**

*** 데이터 수집 & 전처리 / Model & Metric 선택 모두 자유입니다.**(배운 내용의 복습에 Focus!)

*** Part 1~2 에서 배운 지식들을 최대한 모두 활용하는데 초점을 맞춰주세요.**(웹 크롤링 필수 X)

- 3/31(금)~4/3(월) : **문제 정의 & 데이터 수집 / 데이터 탐색 & 시각화 / 데이터 전처리**

- 4/4(화)~4/6(목) : **ML & DL 모델 적용 / 모델 튜닝 & 성능 비교 / 최종 모델 선택**

- 4/7(금) : **모델 개선 & 발표 준비 / 최종 발표**

- **4/7 금요일 16:50** : 팀별 발표 및 질의응답 (20분 내외/팀, **최대 25분**)

: **4/7 (금) 16:30 까지** 발표 자료 & Jupyter notebook(+원본 데이터) 제출 @ 슬랙 DM

*** 1차 Semi-Project 발표자는 발표 X & 도움이 필요할 경우 슬랙 채널에서 호출**

수업 관련 공지사항

- 팀별 소통채널 개설
- 팀장 선정 & 강사 DM

1팀 : 김현수, 송재원, 이동민, 이화정, 조현민

2팀 : 김주환, 박은영, 신주용, 이도원, 허우영

자습1 : 김효경, 이재용, 이지원, 최민정

자습2 : 김찬욱, 이영재, 진광환, 최소운

* 수강생 인원 변동에 따라 팀 구성에 변동이 있을 수 있습니다. (기존 팀에서 옮겨갈 수도 있습니다.)

Various data collection – etc (Datasets / Data repository)

Awesome Public Datasets @ <https://github.com/awesomedata/awesome-public-datasets>

Google AI Datasets @ <https://ai.google/tools/datasets>

Google Dataset Search @ <https://toolbox.google.com/datasetsearch>

Kaggle competition datasets @ <https://www.kaggle.com/datasets>

(ex. Google Play Store Apps data @ <http://j.mp/2PDhbKR>)

<https://data-on.co.kr> – 데이터온 (대한민국의 모든 데이터를 한 곳에서, 누구나 쉽게 찾고 활용하는 데이터플랫폼)

<http://www.aihub.or.kr> – AI 오픈이노베이션 허브 (한국어 음성 & 대화, 한국인 안면, 법률/특허/헬스케어/관광/농업/이미지 데이터)

<https://openapi.kftc.or.kr> & <https://developers.kftc.or.kr/dev> – 금융결제원 오픈API 통합포털 (오픈뱅킹 & 금융인증)

<https://golmok.seoul.go.kr> – 서울시 우리마을가게 상권분석 서비스

<http://data.seoul.go.kr> – 서울 열린 데이터 광장

<https://www.dataquest.io/blog/free-datasets-for-projects> – 19 Places to Find Free Data Sets for Data Science Projects

<http://dataportals.org> – A Comprehensive List of Open Data Portals from Around the World

<https://www.kdnuggets.com/datasets/index.html> – Datasets for Data Mining/Science

* Public APIs (Github) @ <https://bit.ly/3a5ReOI>

* 각종 데이터분석 관련 공모전/대회/프로젝트사례 모음 @ <http://j.mp/2MPDfON>

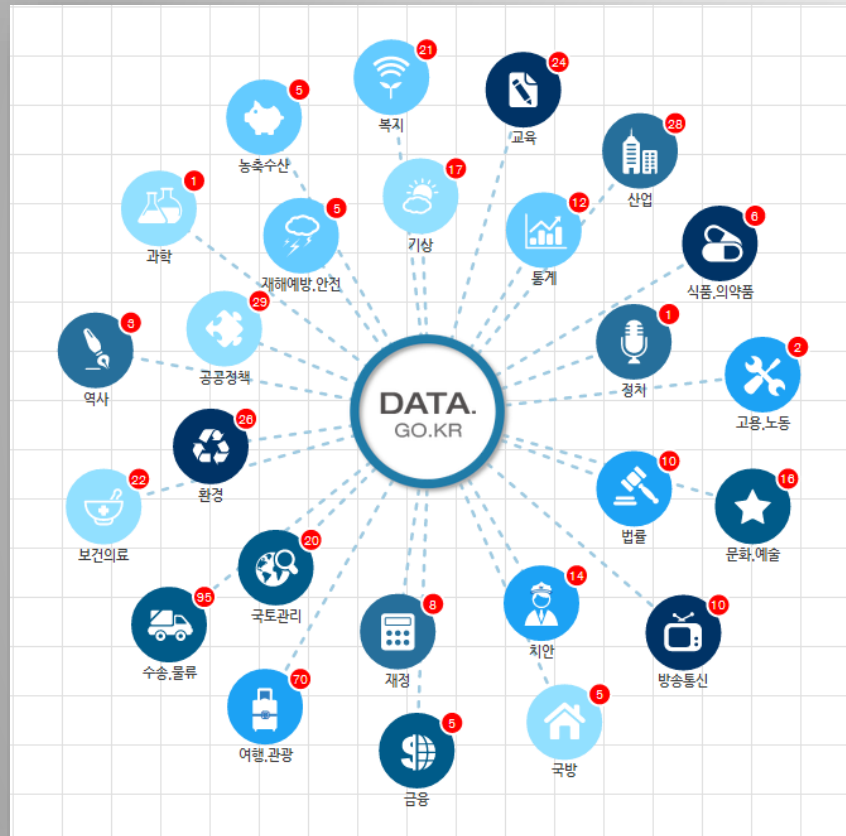
* 2020 문화체육관광 빅데이터플랫폼 데이터 설명회 (한국문화정보원, 국립중앙도서관, 국민체육진흥공단 등) <https://j.mp/30KFuJH>

* KLUE: Korean Language Understanding Evaluation (한국어 NLP 데이터셋 & pretrained models) @ <https://j.mp/3woY9ly> / <https://j.mp/3f7p0mU> / <https://j.mp/3gxEdhn>

본 교안 및 실습자료는 저작권법에 의거하여 본 교육 외 배포/게시/공개를 금합니다.

1. Available resources for data collection

Various data collection – Public data & Open data (APIs & files)



- 공공 데이터 포털 : <https://www.data.go.kr>
- 국가 통계 포털 : <http://kosis.kr>
- MDIS (MicroData Integrated Service) : <https://mdis.kostat.go.kr>

* 오픈 API를 통한 공공데이터 수집 (서울열린데이터 광장) @ <http://j.mp/2AWRA5g>

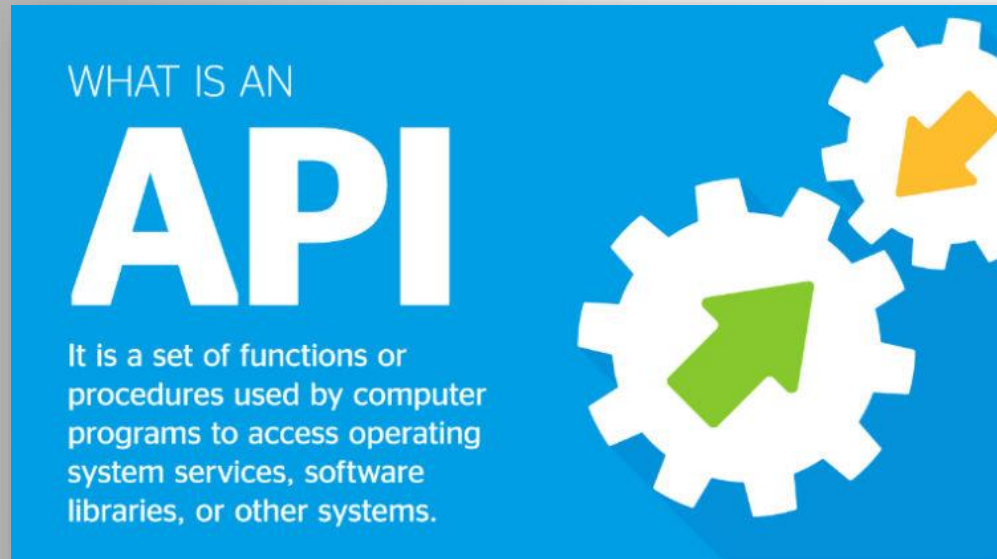
* 공공데이터 OpenAPI 활용을 쉽게 해주는 라이브러리 PublicDataReader @ <https://bit.ly/3aOWoPV>

* 코로나 확진자 동향 데이터 수집 및 시각화 (공공데이터포털 코로나19 감염 현황 OpenAPI 활용) @ <https://j.mp/3bHa8dC>

본 교안 및 실습자료는 저작권법에 의거하여 본 교육 외 배포/게시/공개를 금합니다.

1. Available resources for data collection

Various data collection – Unowned data



Use APIs & Web scraper

- APIs (Twitter, Facebook, Instagram, etc)
- Bots (Web crawler, Web scraper)

* Public APIs (Github) @ <https://bit.ly/3a5ReOI>
* 대법원, 야놀자 정보 크롤링 한 여기어때 창업주 '무죄' @ <https://bit.ly/37H0Wq2> / robots.txt 10분 안에 끝내는 총정리 가이드 @ <https://bit.ly/3b7NOfi>
* Listly (크롬 확장프로그램 for 웹크롤링) @ <https://j.mp/2LSb8kh> / 네이버 크롤링 라이브러리 Kocrawl (날씨/미세먼지/지도/맛집/맞춤법) @ <https://j.mp/2CbdRA8>
* Web Scraping Tool & Web Data Extractor : ScrapeStorm @ <http://j.mp/2Y4porj> / Octoparse @ <https://j.mp/3o5i23q> / Automatio @ <https://automatio.co>

본 교안 및 실습자료는 저작권법에 의거하여 본 교육 외 배포/게시/공개를 금합니다.

2. Possible pathways for data preprocessing

- + Check & adjust data for handling **Missing data & Outlier**
- + Select important columns (or just use all columns & improve your model later)
- + Change characters to numbers (binary-num, categorical-num, one-hot vector, etc.)
- + (If applicable & useful) **Select features** with Tree-based models
- + (If applicable & useful) **Modify the scale of features** with **Scaler & Encoder** (fit on training data & use Pipeline)
- + (If applicable & useful) **Reduce dimension** with PCA
- + (If applicable & useful) Try **other traditional ML Models** for enhancing the result (+ **DL/NN**)

서울시 범죄현황 통계자료 분석 및 시각화

2. 서울시 범죄현황 통계자료 분석 및 시각화 Last Checkpoint: an hour ago (unsaved changes)

File Edit View Insert Cell Kernel Help Trusted

서울시 범죄현황 통계자료 분석 및 시각화

```
In [1]: 1 import numpy as np
2 import pandas as pd
3 import seaborn as sns
4
5 import matplotlib.pyplot as plt
6 from matplotlib import font_manager, rc
```

1. 데이터 입력 및 데이터 전처리

```
In [2]: 1 df = pd.read_excel('관서별 5대범죄 발생 및 검거.xlsx', encoding='utf-8')
2 df.head()
```

	관서명	소계(발생)	소계(검거)	살인(발생)	살인(검거)	강도(발생)	강도(검거)	강간(발생)	강간(검거)	절도(발생)	절도(검거)	폭력
0	계	126481	82688	163	156	276	257	5449	5869	55387	21842	652
1	중부서	2860	1716	2	2	3	2	185	65	1395	477	135
2	중로서	2472	1589	3	3	6	5	115	98	1878	413	127
3	남대문서	2894	1226	1	0	6	4	65	46	1153	382	869
4	서대문서	4829	2579	2	2	5	4	154	124	1812	738	285

Scikit-learn practices & Appendix

(Appendix 0) Missing data visualization (with missingno).zip

(Appendix 1) Tuning HyperParams with Hyperopt (+ LightGBM).zip

(Appendix 2) Auto-ScikitLearn (with Breast cancer data).zip

(Appendix 3) Auto-feature-engineering with FeatureTools.zip

(Appendix 4) PCA for BreastCancer & Cifar10 (딥러닝 학습 후 추가 학습).zip

(Appendix 5) IQR 기반 Outlier 탐지 & SMOTE 기반 Oversampling.zip

(Appendix 6) Census Income Dataset Classification (EDA, 결측치, FeatureEng., ...)

1. (Cheat Sheet) Scikit_Learn.zip

2. Hands On MachineLearning with ScikitLearn and TensorFlow.zip

3. Model saving & loading (Scikit-learn) + Model stacking.zip

4. Pipeline for feature-transformer (StandardScaler & OneHotEncoder).zip

파이썬을 활용한 기초 통계분석

2. 빈도 분석 & 기술통계량 분석

File Edit View Insert Cell Kernel Widgets Help

3. Outlier의 탐지 및 제거와 전후 분포 비교

```
In [218]: 1 df.boxplot(column='amount') # 위 아래의 작은 점들은 결측치 선이 상/하한선, 그 밖의 경우는
```

<matplotlib.axes._subplots.AxesSubplot at 0x1b1170839b8>

* Should binary features be one-hot encoded? @ <https://j.mp/39aFGpf>

* Top 6 Python libs for Visualization: Matplotlib/Seaborn/Plotly/Bokeh/Altair/Folium (장단점) @ <https://j.mp/30772sU>

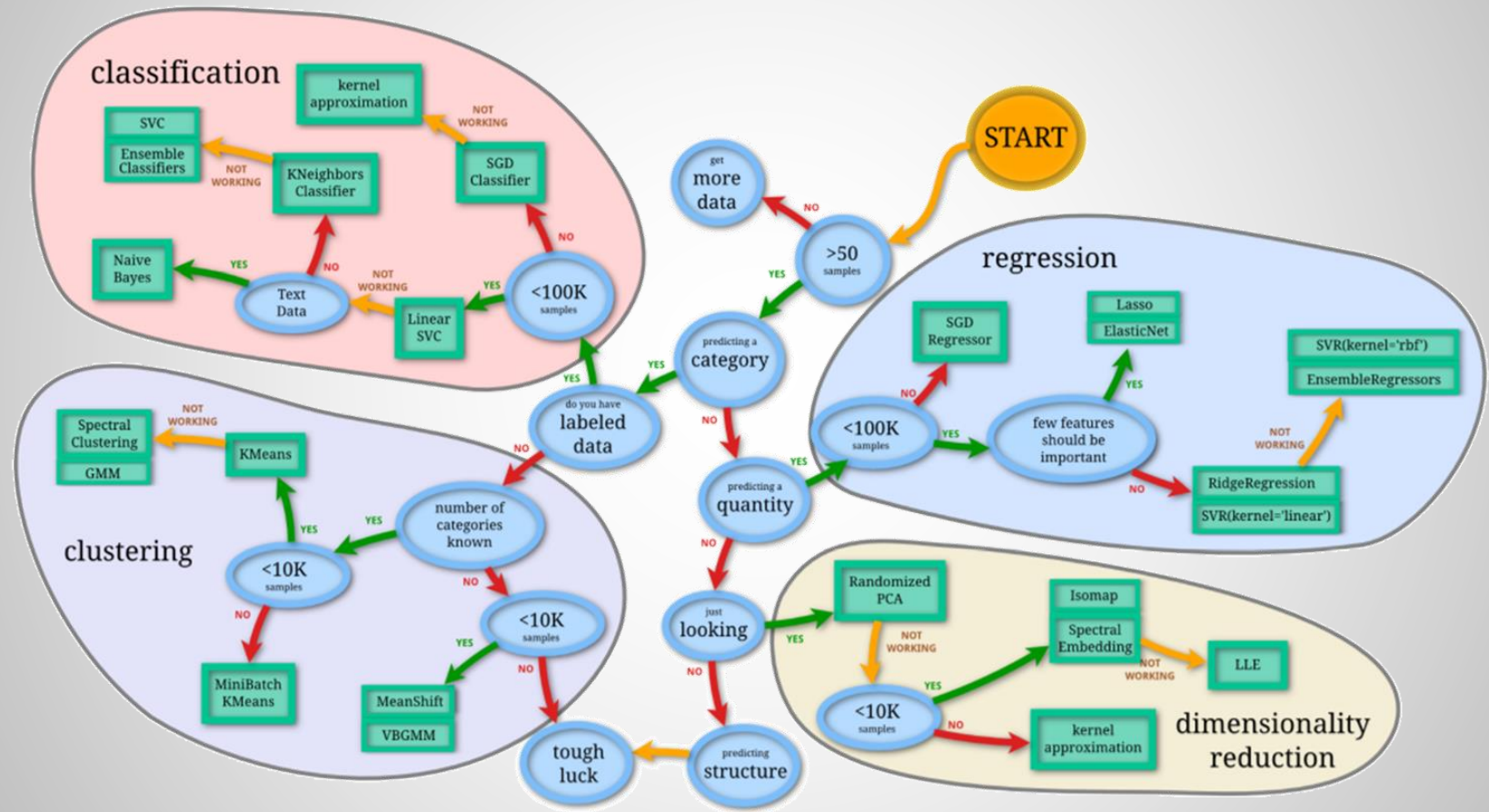
* 딥러닝에서 클래스 불균형을 다루는 방법 (Weight balancing & Focal loss / Over & under sampling / SMOTE) @ <http://j.mp/2qrkTuM> & <https://j.mp/3kR6Sxl>

본 교안 및 실습자료는 저작권법에 의거하여 본 교육 외 배포/게시/공개를 금합니다.

3. Available traditional ML models (+ apply hyper-params optimization)

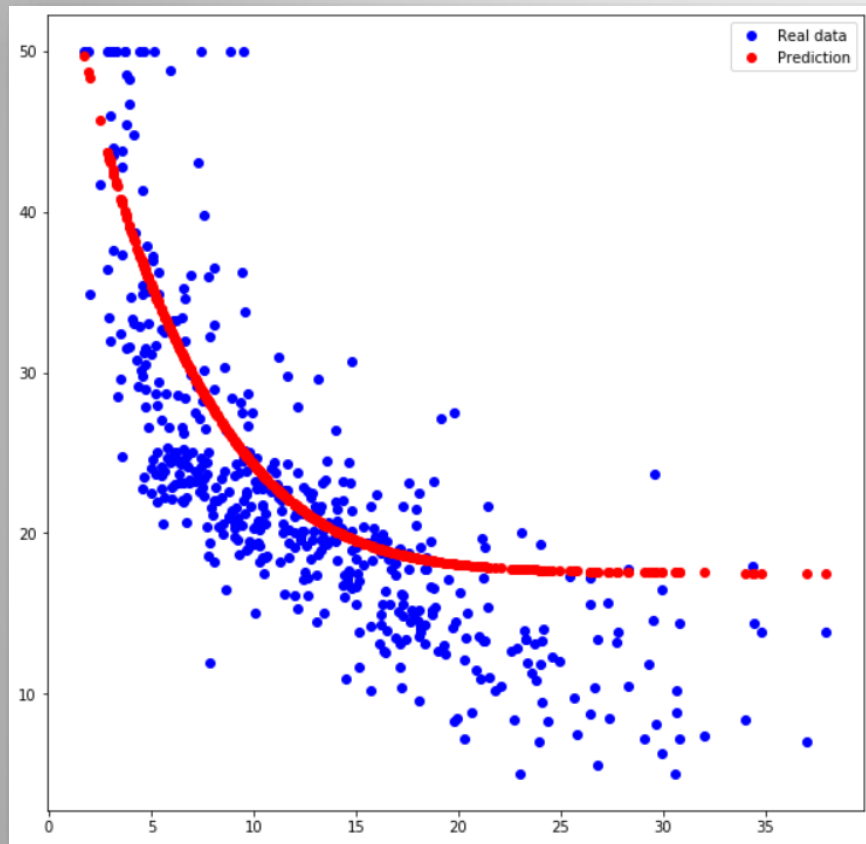
풀어내려는 문제의 종류와 데이터의 타입(형태, 수)에 따른 ML 알고리즘 선택 가이드

http://scikit-learn.org/stable/tutorial/machine_learning_map/ (각 알고리즘 별 예시 코드 有)



* An easy guide to choose the right Machine Learning algorithm for your task @ <http://j.mp/39eTmOC>
* 11가지 전통적인 시계열 예측 모델 with Python (영문, CheatSheet) @ <https://j.mp/39Hv9mt>
* TF 2.0 시계열 예측 공식튜토리얼 (CNN & LSTM) @ <https://j.mp/3dvFflq>
* PapersWithCode for 시계열 예측 @ <https://j.mp/3ulBDiR>

Neural-network modeling with TensorFlow & Keras



4-1. (UseThis) Classification with Keras (Titanic dataset).ipynb

4-2. (UseThis) Regression with Keras (Boston house price dataset).ipynb

Try other improvements,

- Other **activation functions** (tanh, relu)
- Other **optimizers** (Adam, Adagrad, RMSProp)
- Other **learning rates** (0.01, 0.0001)
- More **learning steps** (75000, 100000)
- More **layers & nodes** (64, 128, 256)

+ **Model stacking**

+ **AutoML** (Keras-tuner, Google AutoML Tables, FeatureTools 등)

+ **Bayesian Hyperparams Optimization**

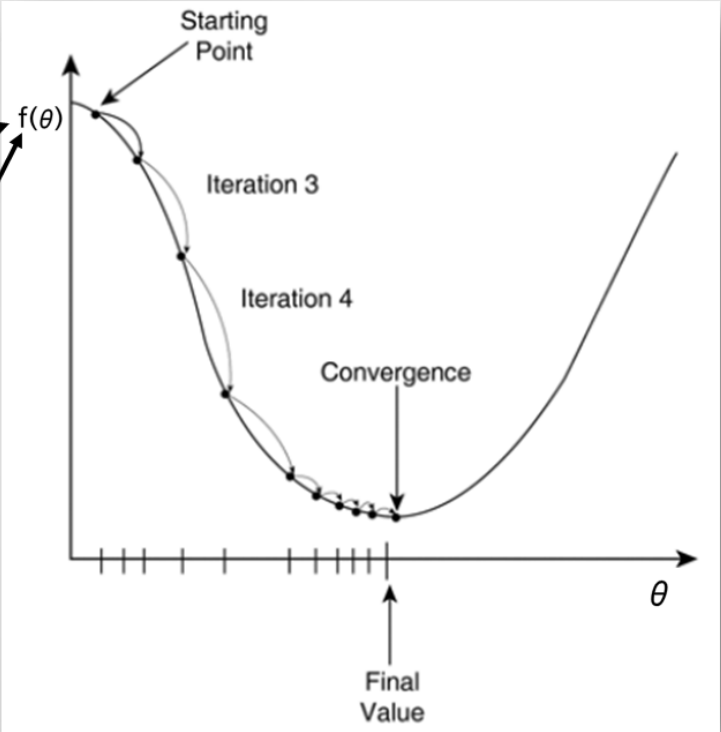
5. Test & compare models with appropriate metrics

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

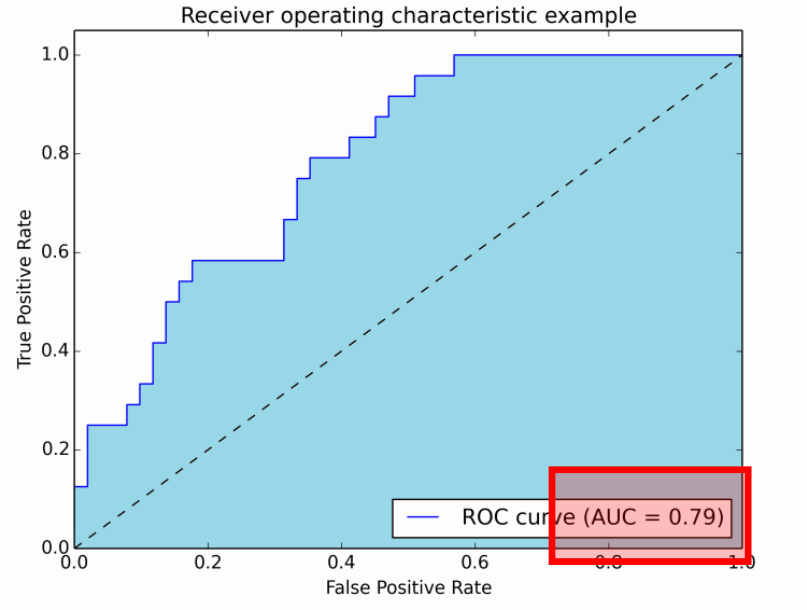
Mean squared error
for regression

$$J(\theta) = - \sum_i y^{(i)} \log(h_{\theta}(x^{(i)}))$$

Cross-entropy
for classification



AUC = Area Under the ROC Curve



- measures the **quality** of classifier.
- AUC = 0.5 : random classifier.
- AUC = 1 : **perfect** classifier.

* Recall, Precision, F1 and ROC/AUC @ <https://j.mp/30iZ9Ry> & <https://j.mp/2PhbWO1>
* Accuracy(정확도), Recall(재현율), Precision(정밀도), 그리고 F1-Score @ <https://j.mp/2AE4S9S>
* F-beta score (Recall vs Precision 가중치 부여) @ <https://j.mp/3pVfZ3t> & <https://j.mp/3760AVI>

6. Use various additional tools if needed

- 📁 (Appendix) 1. PyCaret을 활용한 low-code machine learning (Colab-only)
- 📁 (Appendix) 2. Keras-tuner를 활용한 Bayesian-HPO (Colab-only)
- 📁 (Appendix) 3. Build, Train, and Visualize CNN models (CNN Basic)
- 📁 (Appendix) 4. CNN Transfer-learning for Fashion MNIST & CIFAR10
- 📁 (Appendix) 5. TF2.0 Tensorboard & Keras for CNN MNIST (+ Colab GPU with GoogleDrive)
- 📁 (Appendix) 6. TF2.0 TF Lite & Quantization for CNN MNIST (+ 추가링크 for Android app)

- 📁 (Misc.) 1. cvlib을 활용한 편리한 얼굴 & 물체 검출
- 📁 (Misc.) 2. Clova Face Recognition & Papago NMT
- 📁 (Misc.) 3. Easy Speech-to-Text with Python (Google speech recognition API)
- 📁 (Misc.) 4. OS 매크로 도구 PyAutoGUI
- 📁 (Misc.) 5. Python을 활용한 자동 이메일 & 문자메시지 발송 (via twilio)
- 📁 (Misc.) 6. PyInstaller를 활용한 실행파일(.exe) 생성

```
필수 라이브러리 설치 (SpeechRecognition & PyAudio)

# pip install SpeechRecognition==3.8.1
# pip install PyAudio-0.2.11-cp37-cp37m-win_amd64.whl

• Windows OS의 경우 pip install PyAudio 명령어로 PyAudio가 설치되지 않을 수 있습니다.
  (PyAudio for Windows 링크의 단편에 있는 설명을 따라) https://pypi.org/project/PyAudio/에서 Python 버전 및 windows bit 값에 따른
  Wheel 파일을 다운로드한 다음, 해당 whl 파일을 바탕으로 pip 명령어를 입력해 직접 설치 진행이 가능합니다.
  (이 whl 파일의 경우 Python 3.7 버전, Windows 64bit 에 따른 설치 파일입니다.)
  • We need to install PyAudio library which used to receive audio input and output through the microphone and speaker. Basically, it helps to get
    our voice through the microphone.
```

```
# pip install opencv-python==4.1.1
# pip install cvlib==0.2.4
# pip show cvlib

Name: cvlib
Version: 0.2.4
Summary: A high level, easy to use, open source computer vision library for python
Home-page: https://github.com/arunponnusamy/cvlib
Author: Arun Ponnusamy
Author-email: hello@arunponnusamy.com
License: MIT
Location: c:\programdata\anaconda3\lib\site-packages
Requires: requests, pillow, imutils, imageio, numpy, progressbar
Required-by:

1
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)

import matplotlib.pyplot as plt
import numpy as np

import cv2
```

```
:pip install pyautogui==0.9.42

# # (한글 버전 윈도우에서) 'cp949' codec here decode 시
# pip install pygetwindow==0.0.1
# pip install pyautogui==0.9.42 # (or 0.9.39)

import pyautogui
import time

# 모니터 해상도 가져오기
width, height = pyautogui.size()
print('width={0}, height={1}'.format(width, height))

width=1920, height=1080

# 마우스 위치 가져오기
x, y = pyautogui.position()
print('x={0}, y={1}'.format(x, y))

x=1017, y=587
```

* 발표 시 포함할 사항 :

1. 프로젝트 주제 소개 (어떤 분석을 하였는가)
2. 데이터 소개/탐색/시각화 (출처, 형식, 분포 등)
3. 데이터 전처리 과정 (적용한 전처리 방법 & 이유)
4. 적용한 분석 기법 및 모델 소개
5. 분석 및 모델링 결과 (각종 지표 수치 제시)
(+ 가능 시 추가로 분석하면 좋을 과제 제시)

* 발표 시 제출할 사항 :

발표 자료
(PPT < PDF)

전체 코드 with 주석
(.ipynb, 단일 혹은 복수)

* 여러 파일 제출 시 파일 이름 내 넘버링 적용

전체 원본 데이터

* 1GB 넘는 데이터는 별도 링크로 첨부

[2022 기업멤버십 SW캠프]
클라우드 활용 빅데이터서비스 개발자 부트캠프

End of Document