광주 ICT 이노베이션스퀘어

인공지능 융복합교육

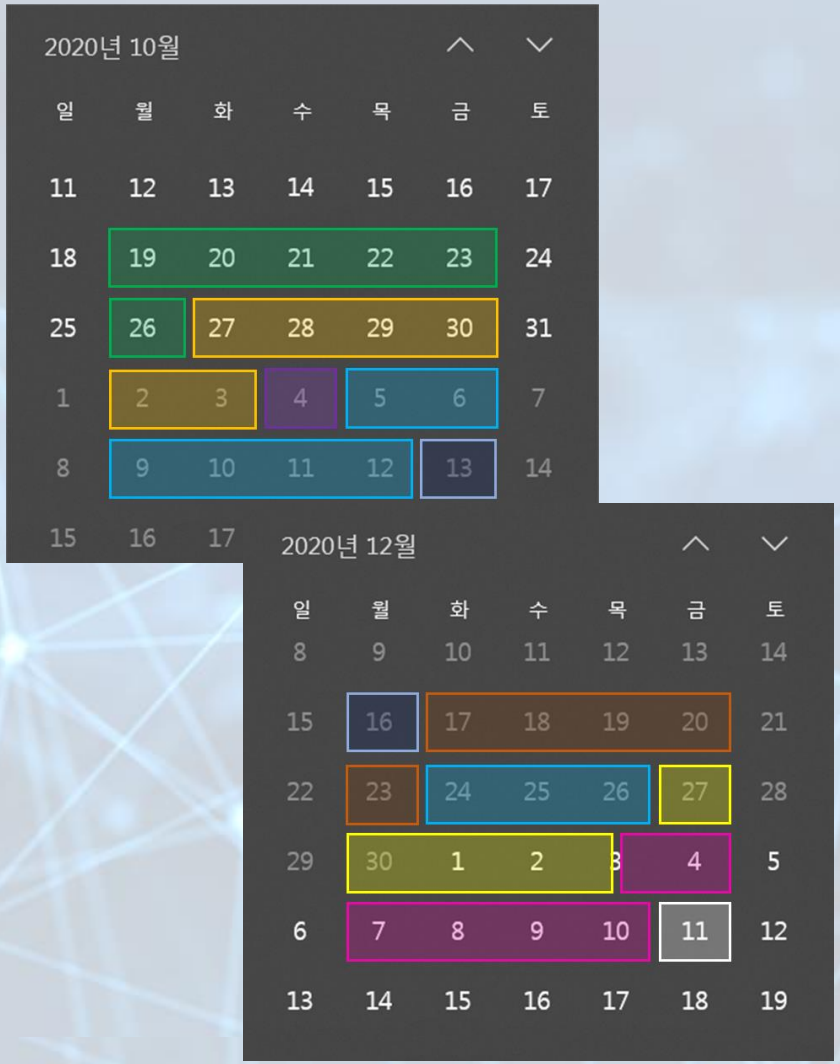
Final Project

- 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
repositivator@gmail.com

Course Overview

* 아래 커리큘럼의 세부 사항은 변동될 수 있습니다.
* 진도 상황에 따라 1~2일 정도 차이가 발생할 수 있습니다.



파이썬 프로그래밍 기초
파이썬 정형 데이터 분석 (데이터 탐색 / 데이터 전처리 / 데이터 시각화)

파이썬을 활용한 데이터 수집 & 웹 스크레이핑 (+ 자동화 프로그램 개발)
파이썬 기반 텍스트 데이터 분석

Python 기반 기초 통계분석 (빈도분석 / 기술통계 / 교차검정 / 평균차이검정)

1차 세미 프로젝트 (데이터 수집 / 전처리 / 통계분석 / 시각화 + 팀별 발표)

SQL 기초 프로그래밍 (Data Modeling / SQL CRUD / Adv. Techniques)

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

2차 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

파이널 프로젝트 (데이터 수집 / 탐색 & 전처리 / 시각화 + ML & DL model tuning)

파이썬을 활용한 Easy-AI & Automation-tools 프로그래밍 실습

* 매 정시마다 메인 회의실에서 출석 체크 진행합니다.

수업 관련 공지사항

* 데이터 선정 / 전처리 프로세스 / 모델 & 지표 선택 모두 자유입니다. (배운 내용의 복습에 Focus!)

* Part 1~6 에서 배운 지식들을 최대한 모두 활용하는데 초점을 맞춰주세요. (크롤링 필수 X)

- 12/3(목)~12/4(금) : 문제 정의 / 데이터 선정 / 데이터 탐색 & 시각화
- 12/7(월)~12/9(수) : 데이터 전처리 / ML & DL 모델 적용 / 모델 튜닝 & 최종 모델 선택
- 12/10(목) : 발표 준비 / 최종 발표

- 12/10 (목) 3~4교시 : 팀별 발표 및 질의응답 (15~20분 내외/팀)
  : 목요일 2교시 종료 전까지 발표 자료 & Jupyter notebook 제출 : 슬랙 DM or repositivator@gmail.com

* 발표 시작 시간은 일정에 따라 변동될 수 있습니다 & 도움이 필요할 경우 슬랙 채널에서 호출

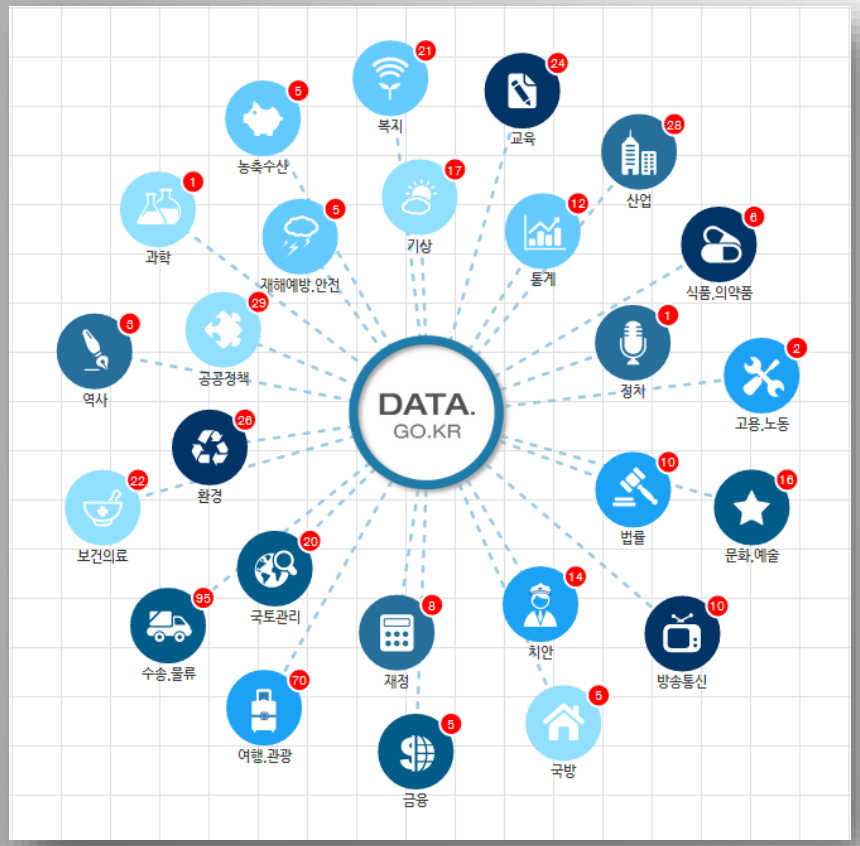
Various data collection – etc (Datasets / Data repository)

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
SKT BigData Hub @ https://www.bigdatahub.co.kr
Kaggle competition datasets @ https://www.kaggle.com/datasets
(ex. **Google Play Store Apps data** @ http://j.mp/2PDhbKR)

http://www.aihub.or.kr – **AI 오픈이노베이션 허브 (한국어 음성 & 대화, 한국인 안면, 법률/특허/헬스케어/관광/농업/이미지 데이터)**
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**
http://figshare.com – Help academic institutions store, share and manage their research
https://opendatainception.io – 2600+ Open Data Portals around the World
https://search.datacite.org – Locate, identify, and cite research data
http://aws.amazon.com/datasets – AWS dataset
http://quandl.com – Financial Data

* 각종 데이터분석 관련 공모전/대회/프로젝트사례 모음 @ http://j.mp/2MPDfON
* 해외 기업의 인공지능 데이터 개방과 활용 현황 (구글 사례를 중심으로) @ http://j.mp/2paKt7j
* 딥러닝 학습을 위한 국내외 데이터셋 현황 (이미지/동영상/음성) @ http://j.mp/2BSShNy / http://j.mp/2roFj8i / http://j.mp/2VbHXeG

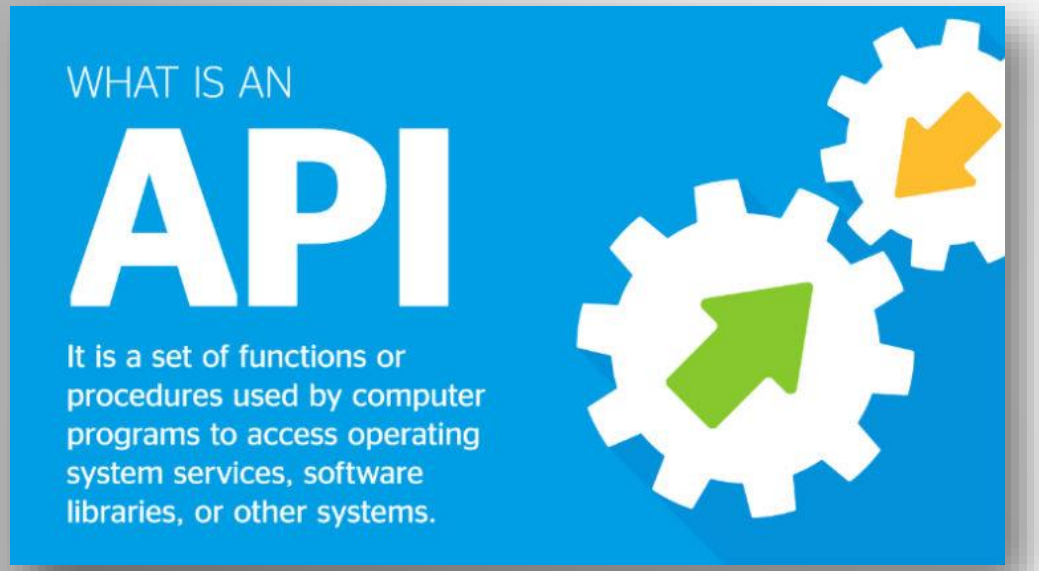
Various data collection – Public data & Open data (APIs & files)



- 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

Various data collection – Unowned data



Use APIs & Web scraper

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

* 네이버 크롤링 라이브러리 Kocrawl (날씨/미세먼지/지도/맛집/맞춤법) @ https://j.mp/2CbdRA8
* 여기어때, 야놀자 DB 무단수집 위법 판결 @ https://j.mp/3fgxi9Q + 크롤링과 저작권 침해 고소 진행 일대기 @ https://j.mp/3k5vbbl
* Web Scraping Tool & Web Data Extractor : **ScrapeStorm** ($49/month) @ http://j.mp/2Y4porj + **Octoparse** @ https://j.mp/3iYQTgX

The process of web scraping (detailed)

URL 분석 (패턴 존재 여부, query 종류)
ex. https://alldic.daum.net/search.do?q=happiness

↓

URL 구성 (str, 추후 자동화 고려)

HTTP Response 얻기 : urlopen(URL) **or** request.get(URL).content

HTML source 얻기 : BeautifulSoup(HTTP Response, 'html.parser')

↓

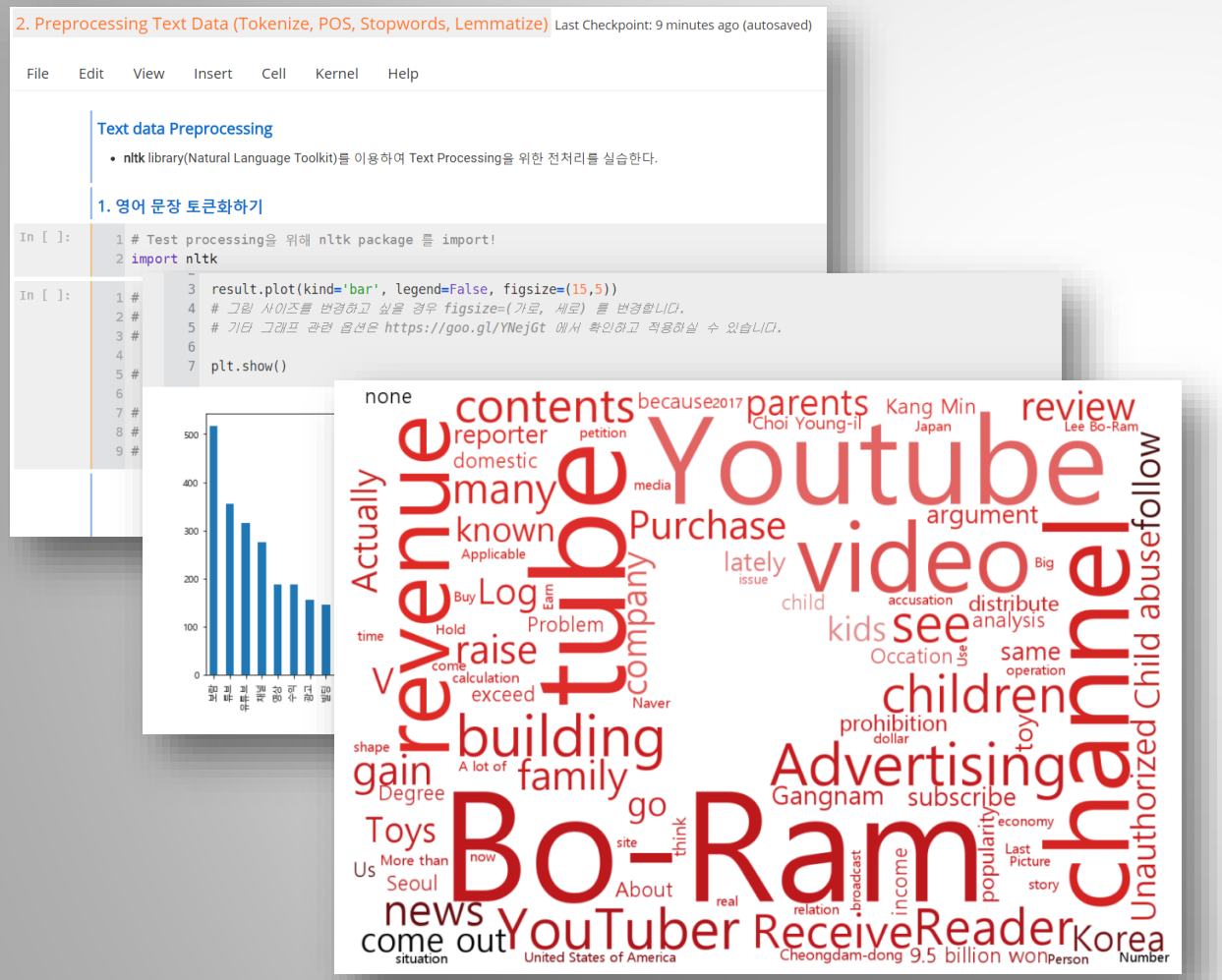
HTML Tag 꺼내기 (Tag 이름, Attr 이름, Attr 값 지정)

.find(~) : 1개의 Tag (조건이 동일한 Tag가 여러 개일 경우에는 첫번째 Tag)
.find_all(~) : 여러 개의 Tag를 찾은 다음 for 문으로 Tag 단위로 꺼내어 활용

↓

Tag 로부터 텍스트 혹은 Attribute values 꺼내기 : Tag.get_text() **or** Tag.attrs

The process of data analysis for text data



텍스트 데이터를 str 자료형으로 준비

↓

Tokenize (형태소 분석)

POS Tagging (Part-of-speech, 품사 표시)

Stopwords 제거 (불용어 제거)

↓

단어 갯수 카운팅 & 단어 사전 생성

단어 사전 기반 데이터 시각화

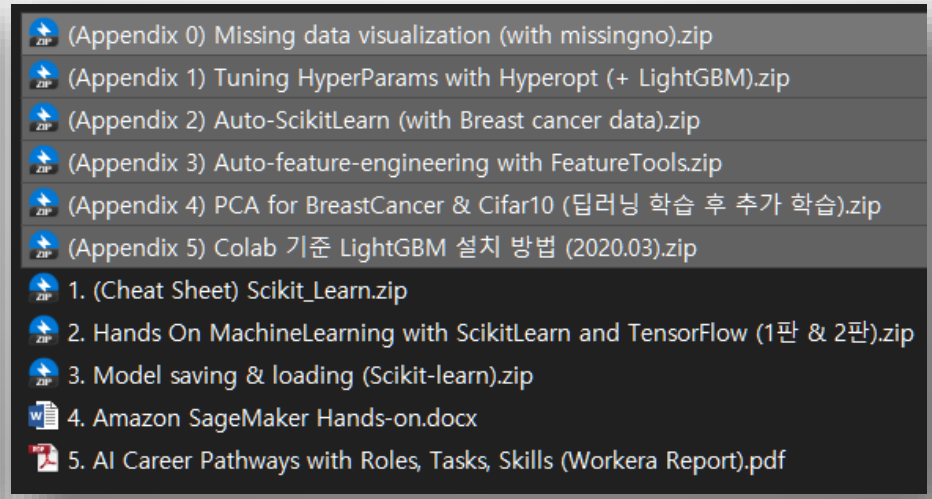
(+ 머신러닝/딥러닝 모델 적용)

2. Possible pathways for data preprocessing

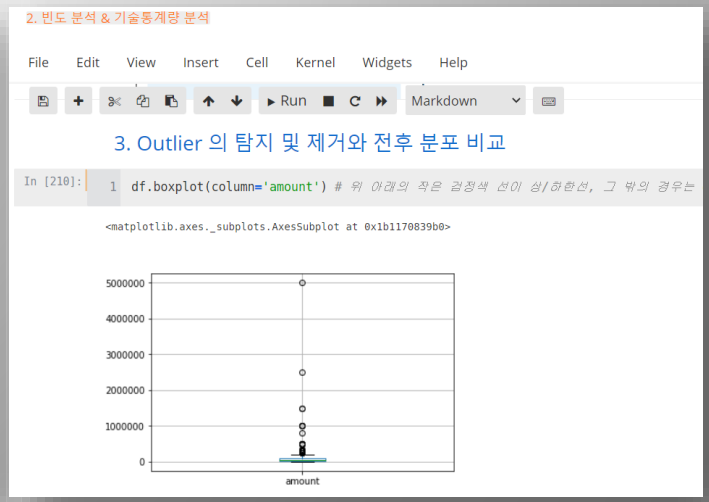
+ Check & adjust data for handling Missing data & Outlier
+ Select important columns (or just use all columns & improve your model later)
+ Change characters to numbers (Binary num, Class num, One-hot vector, etc.)
+ (If applicable & useful) Select features with Tree-based models
+ (If applicable & useful) Modify the scale of features with StandardScaler / MinMaxScaler
+ (If applicable & useful) Reduce dimension with PCA
+ (If applicable & useful) Try other traditional ML Models for enhancing the result (except DL/NN)

서울시 범죄현황 통계자료 분석 및 시각화

Scikit-learn practices & Appendix

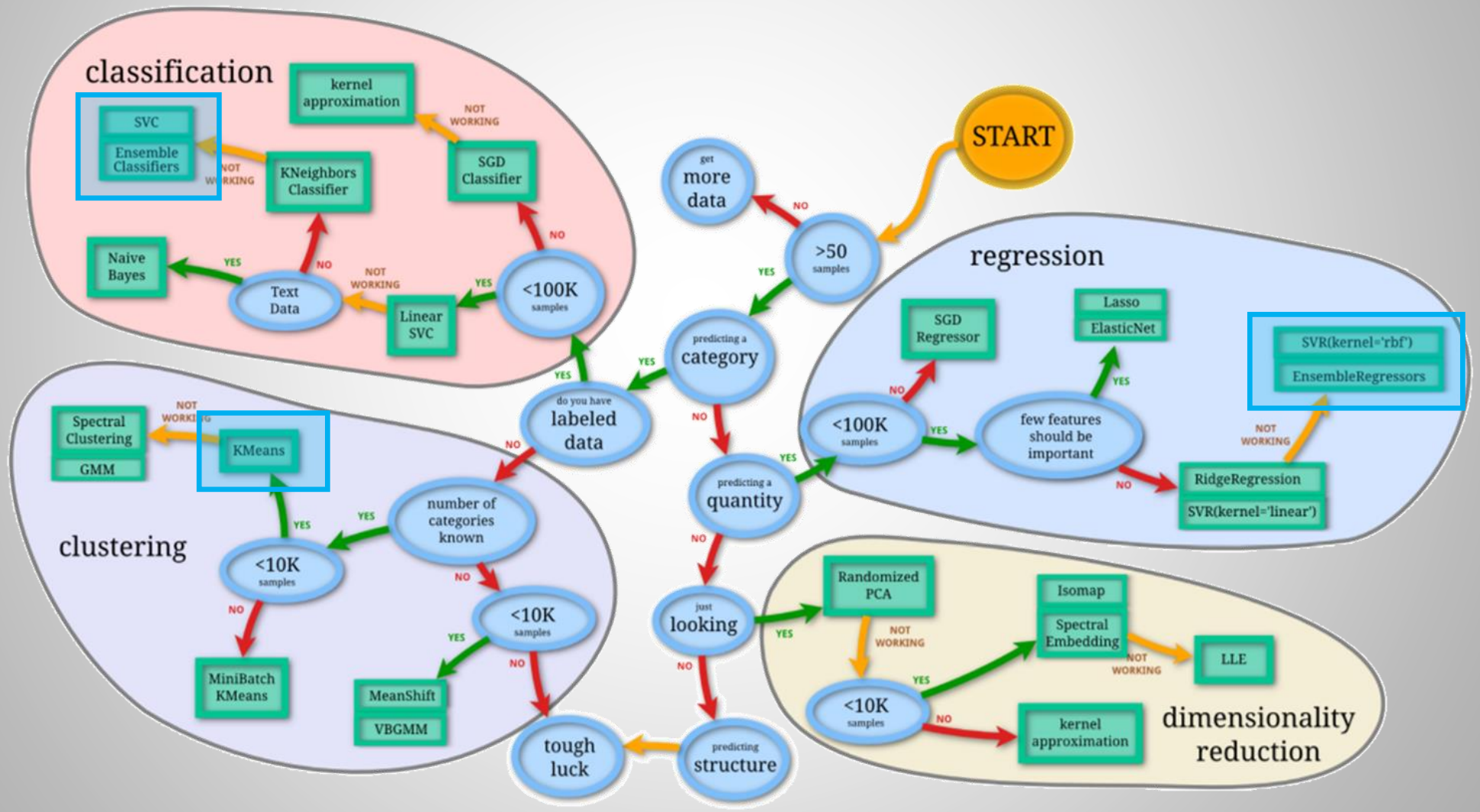


파이썬을 활용한 기초 통계분석

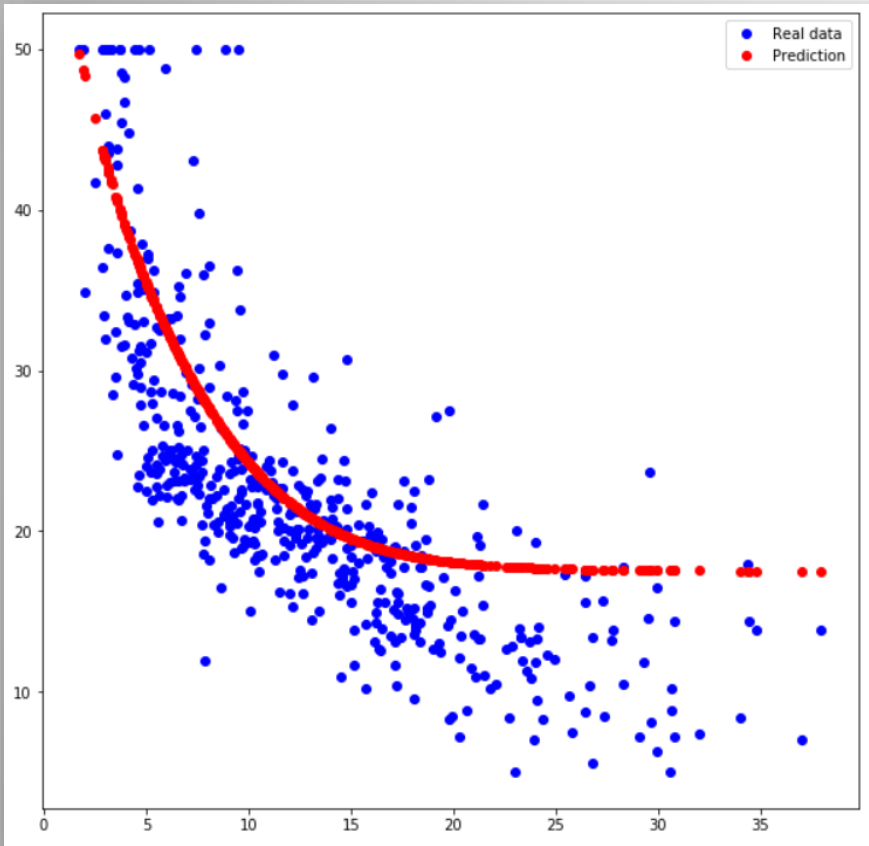


풀어내려는 문제의 종류와 데이터의 타입(형태, 수)에 따른 ML 알고리즘 선택 가이드

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



Neural-network modeling with TensorFlow & Keras



- 0-1. (UseThis) Classification (Titanic dataset).ipynb
- 0-2. (UseThis) Classification with Keras (Titanic dataset).ipynb
- 0-3. (UseThis) Regression (Boston house price dataset).ipynb
- 0-4. (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

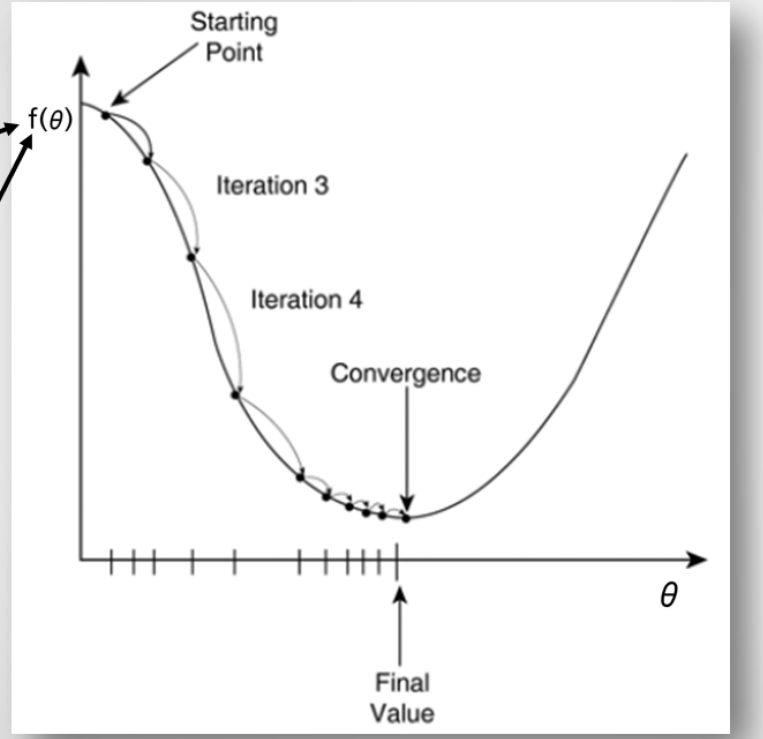
4. Test the model with appropriate metrics

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

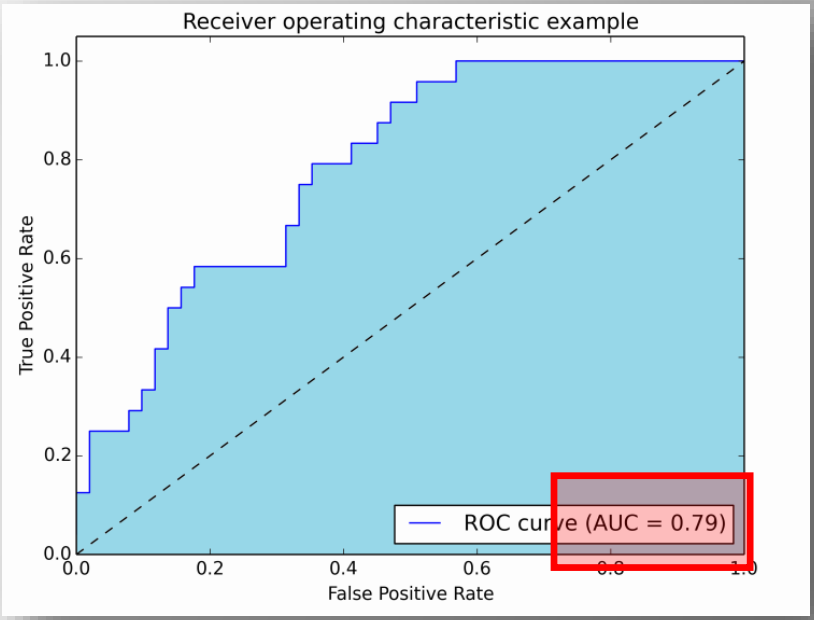
Mean squared error
for regression

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

Cross-entropy
for classification



AUC = Area Under the ROC Curve



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

5. Prepare presentation & Clean your codes

* 발표 시 포함할 사항 :

1. 프로젝트 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 :

전체 코드 with 주석
(.ipynb, 단일 혹은 복수)

발표 자료 필수
(PPT or PDF)

* 매 정시마다 메인 회의실에서 출석 체크 진행합니다.

수업 관련 공지사항

1팀 : 한대영, 고성헌, 문수지, 서경호

2팀 : 강예은, 안현진, 조우람, 최갑주

3팀 : 박광렬, 손부언, 이승재, 이하은

4팀 : 김관주, 강기훈, 조혜빈, 정경훈

광주 ICT 이노베이션스퀘어
인공지능 융복합교육

End of Document