혁신성장 청년인재 집중양성

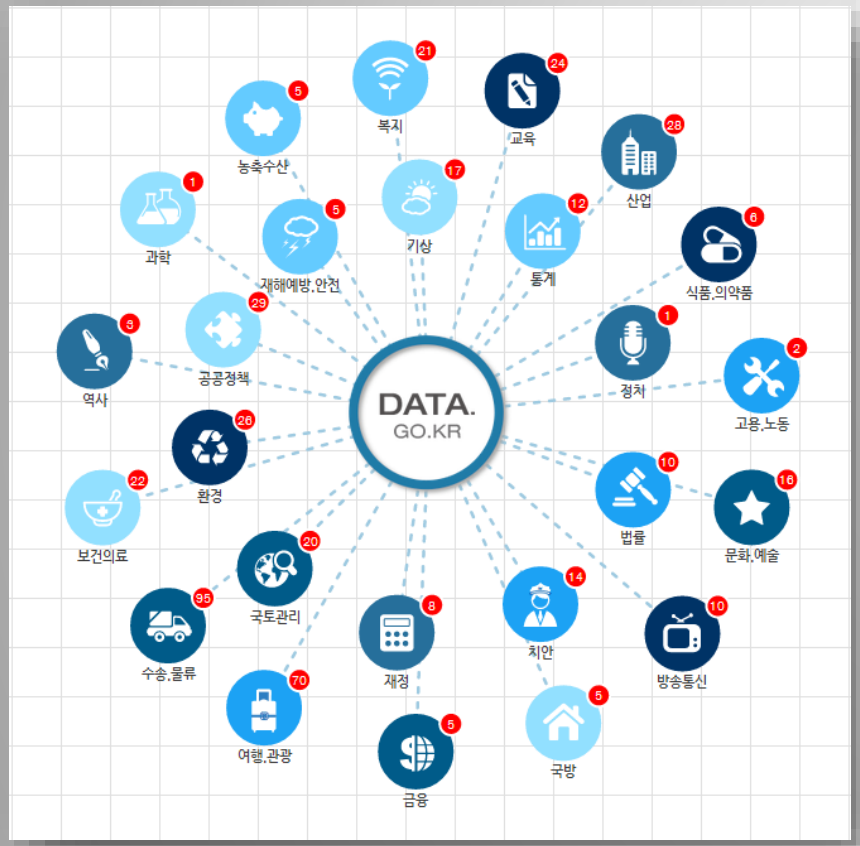
이미지 분석 인공지능 서비스 개발 실무 과정

Semi-Project for Part 1~5

- Data collection / exploration / visualization
- Train & test traditional ML algorithms
- Train & test deep learning models

Daeyeon Jo
repositivator@gmail.com

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 – 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)

https://www.dataquest.io/blog/free-datasets-for-projects – **19 Places to Find Free Data Sets for Data Science Projects**
http://www.aihub.or.kr/ – **AI 오픈이노베이션 허브 (한국어 음성 & 대화, 한국인 안면, 법률/특허/헬스케어/관광/농업/이미지 데이터)**
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://data.seoul.go.kr – **서울 열린 데이터 광장**
http://quandl.com – Financial Data
http://aws.amazon.com/datasets – AWS dataset
https://search.datacite.org – Locate, identify, and cite research data
https://opendatainception.io – 2600+ Open Data Portals around the World
http://figshare.com – Help academic institutions store, share and manage their research

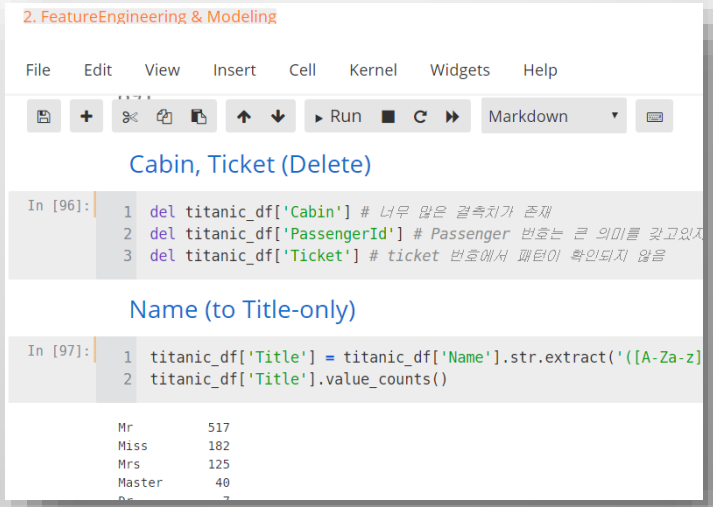
* 각종 데이터분석 관련 공모전/대회/프로젝트사례 모음 @ http://j.mp/2MPDfON
* 해외 기업의 인공지능 데이터 개방과 활용 현황 (구글 사례를 중심으로) @ http://j.mp/2paKt7j
* 딥러닝 학습을 위한 국내외 데이터셋 현황 (이미지 & 동영상) @ http://j.mp/2BSShNy & http://j.mp/2roFj8i

2. 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

서울시 범죄현황 통계자료 분석 및 시각화

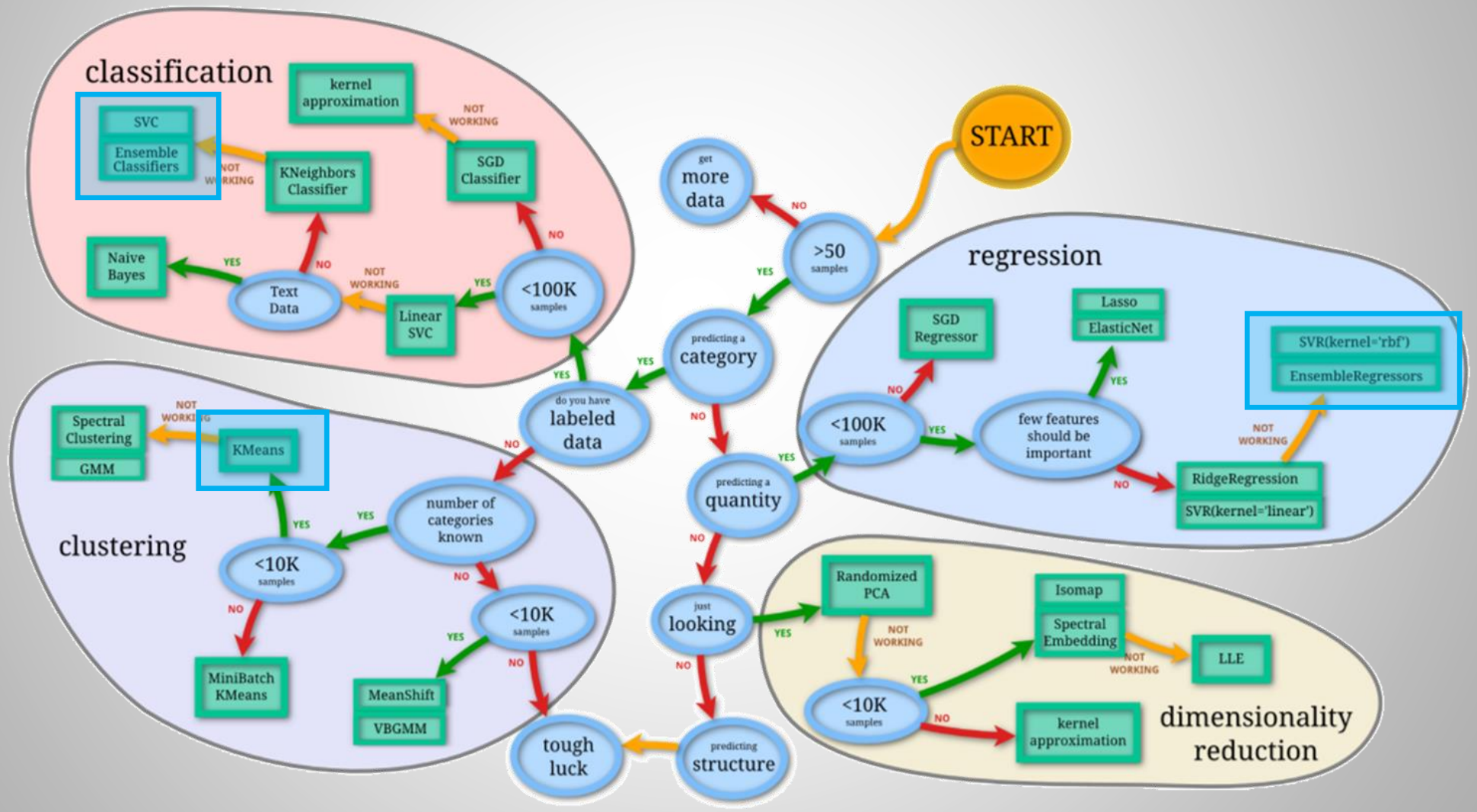
Titanic dataset preprocessing



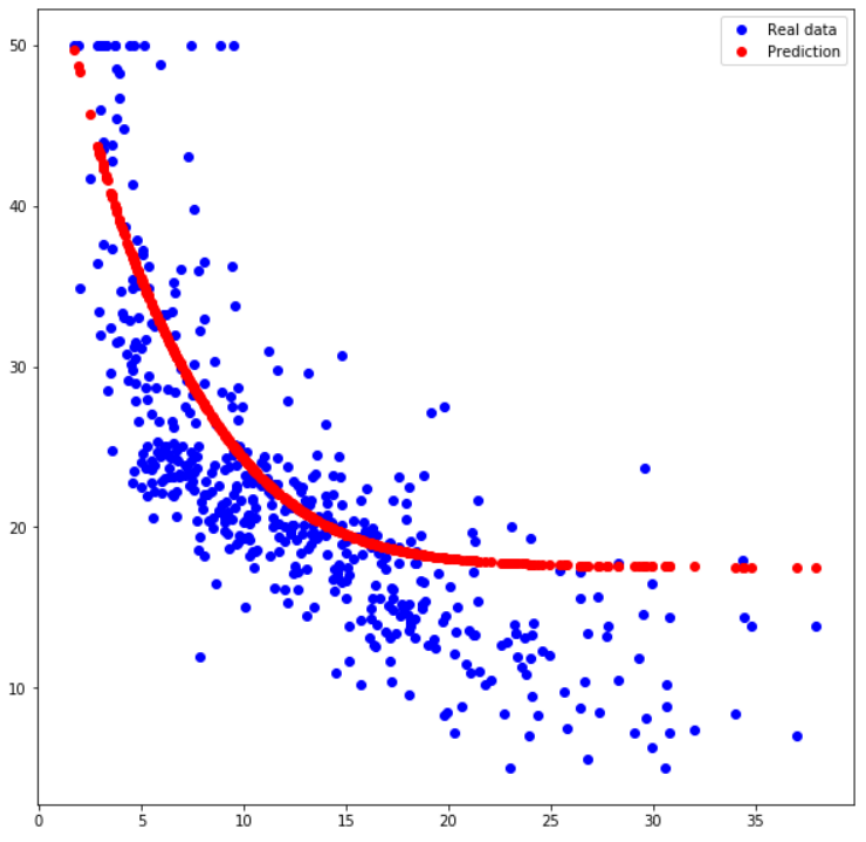
파이썬을 활용한 기초 통계분석

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



- 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)

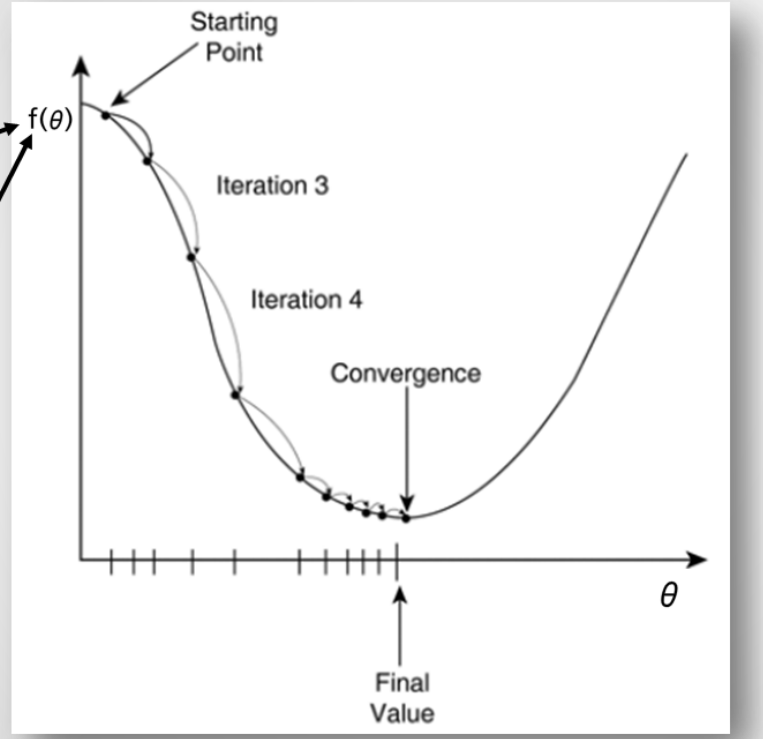
4. Test the model with appropriate metrics

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

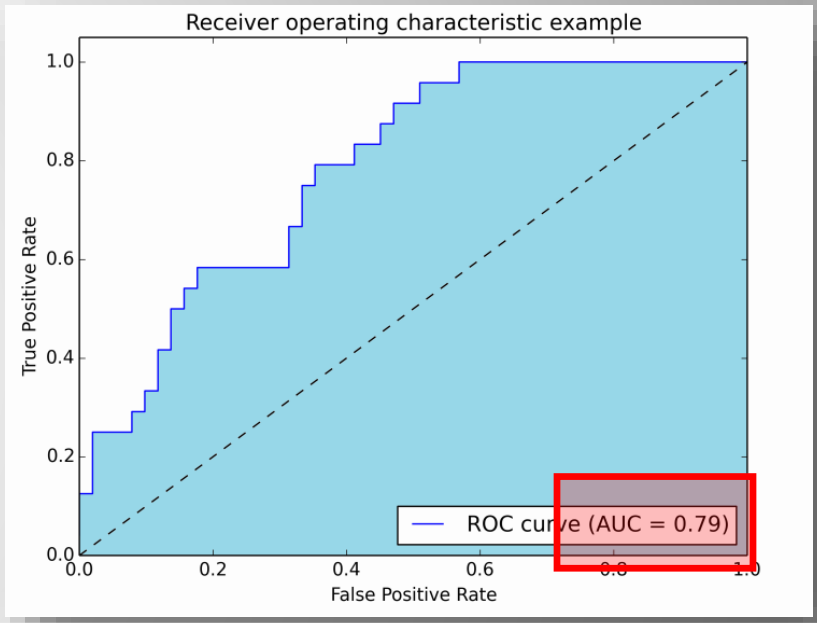
Mean squared error
for regression

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

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 및 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)

(+ 6. 가능 시 추가로 분석하면 좋을 과제 제시)

* 발표 시 제출할 사항 :

전체 코드 with 주석 (.ipynb)

발표 자료 필수 (PPT or PDF)

수업 관련 공지사항

* Part 1~5 에서 **배운 지식들을 최대한 모두 활용**하는데 초점을 맞춰주세요.

* **팀별로 자유로이 3층 내에서 자리를 옮겨서 논의하셔도 무방합니다.**

- 12/31 (화) : **문제 정의 / 데이터 수집** (웹크롤링 활용 권장, 필수 X) **/ 데이터 탐색 & 시각화**
- 1/2 (목) : **데이터 전처리 / 전통적인 머신러닝 & 딥러닝 모델 적용**
- 1/3 (금) : **모델 튜닝 & 최종 모델 선택** / 발표 준비 / **15:55 까지 최종 결과물 제출**

- **1/3 (금) 16:00 부터 팀별 발표 및 질의응답 (질의응답 포함 최대 20분 / 팀)**
  : **금요일 15:55 까지** 발표 자료 & Jupyter notebook 제출 @ repositivator@gmail.com

* **발표 시작 시간은 변동될 수 있습니다**

5팀 5/4팀 4팀
1팀 1/2팀 2팀 3팀

수업 관련 공지사항

* 메인 발표자는 아직 한 번도 발표를 하지 않았던 사람 중에서 선정해주세요.

1팀 : 장재원, 김은아, 문석호, 신광현, 한상국, 한겨레

2팀 : 이경희, 김현하, 배소현, 윤영준, 이재승, 하수진

3팀 : 정수현, 김충희, 송윤성, 이가은, 이예랑, 이철희

4팀 : 김채윤, 문희원, 이동규, 이주환, 이중기, 장준규

5팀 : 한성준, 김지승, 이소은, 장희은, 이대광, 백승환

혁신성장 청년인재 집중양성
이미지 분석 인공지능 서비스 개발 실무 과정

End of Document