

[2022 기업멤버십 SW캠프]

클라우드 활용 빅데이터서비스 개발자 부트캠프

Semi-Project for Part 1

Training & testing traditional ML algorithms (Titanic survival analysis)

Daeyeon Jo
repositivator@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

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

* Part 1 에서 배운 지식들을 최대한 빠짐없이 활용하는데 초점을 맞춰주세요.

* “Titanic prediction” 등과 같이 관련 코드에 대한 직접적인 검색은 피해주세요.

* 발표 시 포함할 사항 : 데이터 전처리 방법 & 이유 / 모델 적용 프로세스 / 모델 적용 결과
발표 시 제출할 사항 : 상세 주석이 포함된 전체 코드 (.ipynb 제출, PPT 발표자료 필수 X)

* **목~월** : 팀별 분석 작업 -> **3/21 화요일 16:20** : 팀별 발표 및 질의응답 (15분 내외/팀)
: **3/21 (화) 16:10 까지** Jupyter notebook 제출 (여러 파일 제출 시 파일이름 내 넘버링) @ 슬랙 DM

* 도움이 필요할 경우 슬랙 채널에서 호출

수업 관련 공지사항

1팀 : 김현수, 신주용, 이재용, 이화정

2팀 : 박은영, 김찬욱, 진광환, 최소윤, 허우영

3팀 : 이도원, 김주환, 김효경, 이동민, 이영재

4팀 : 최민정, 송재원, 이지원, 조현민

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

0. Blank notebook for this semi-project

ML for Titanic survival prediction (Blank)

Logout

FileEditViewInsertCellKernelWidgetsHelp

TrustedPython 3

Save

+

Undo

Redo

Run

Code

```
In [1]: 1 import warnings
2 warnings.filterwarnings("ignore")
3
4 import pandas as pd
5 import numpy as np
6 import matplotlib.pyplot as plt
7
8 # from sklearn import ?
9 # from sklearn.metrics import ?
```

1. Preparing dataset (2번부터 실습)

```
In [12]: 1 data_df = pd.read_csv('titanic.csv')
2 data_df.head(3)
```

Data info

- **PassengerId** : Unique ID of passenger
- **Survived** : 0 = No, 1 = Yes
- **pclass** : Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- **sibsp** : # of siblings & spouses aboard the Titanic
- **parch** : # of parents / children aboard the Titanic
- **ticket** : Ticket number
- **cabin** : Cabin number
- **embarked** : Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S

1. 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 (except 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)	정도(발생)	정도(검거)	폭력
0	계	126401	82688	163	156	276	257	5449	5069	55307	21842	652
1	중부서	2860	1716	2	2	3	2	105	65	1395	477	135
2	중로서	2472	1589	3	3	6	5	115	98	1070	413	127
3	남대문서	2094	1226	1	0	6	4	65	46	1153	382	869
4	서대문서	4029	2579	2	2	5	4	154	124	1812	738	205

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

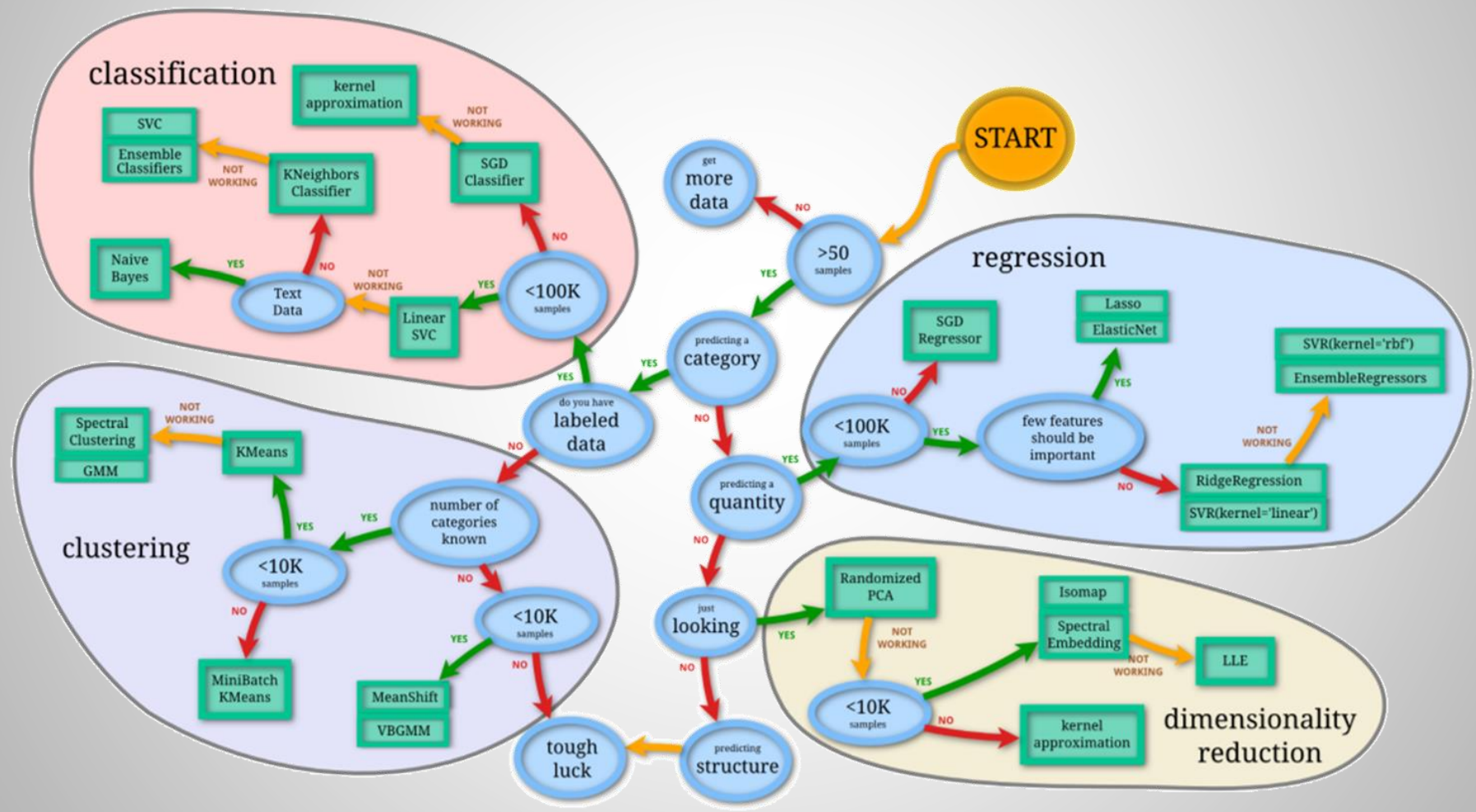
3. Outlier의 탐지 및 제거와 전후 분포 비교

In [210]:
1 df.boxplot(column='amount') # 위 아래의 작은 점(검정색 선이 상/하한선, 그 밖의 경우는

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

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

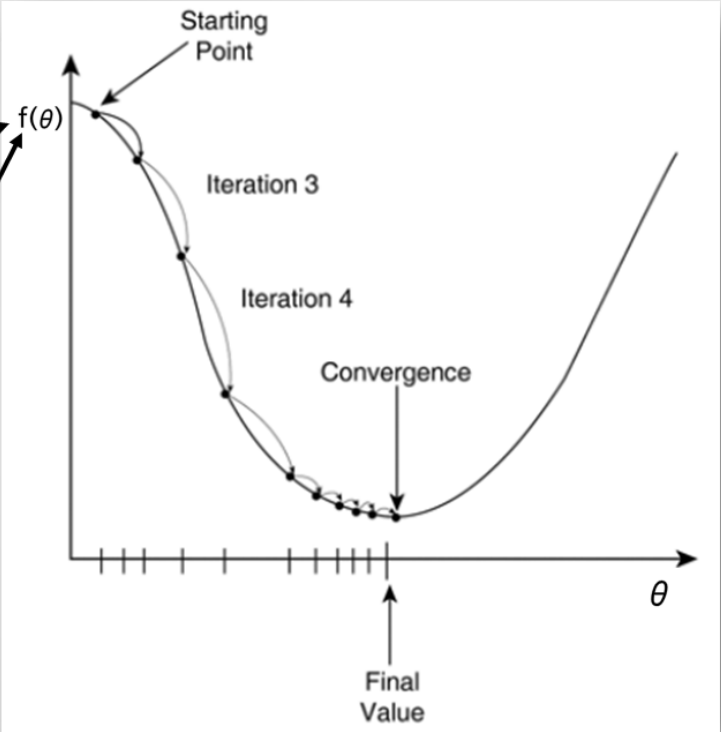
3. Test & compare models with appropriate metrics

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

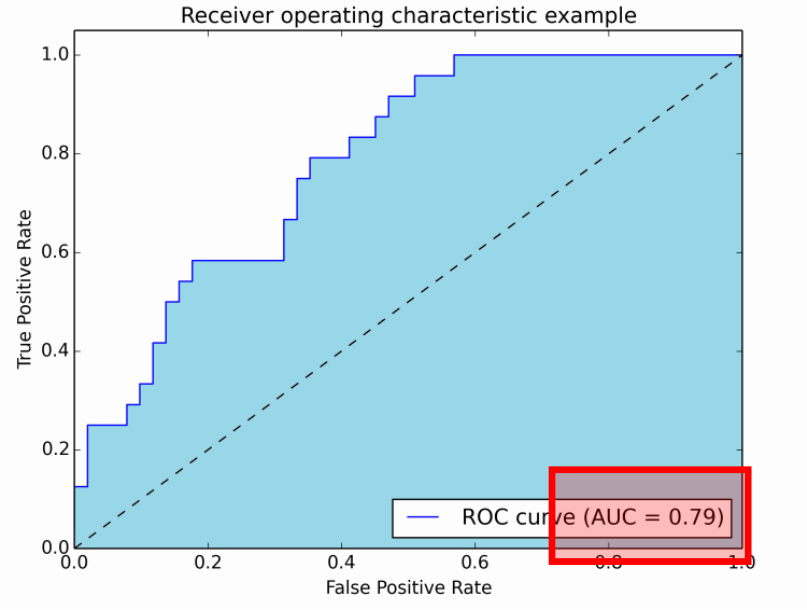
Mean squared error
for regression

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

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>

- + **Model stacking** 적용해보기
- + **AutoML** 적용해보기 (Google AutoML Tables, FeatureTools, Auto-sklearn 등)
- + **Bayesian Hyperparams Optimization** 적용해보기

[2022 기업멤버십 SW캠프]
클라우드 활용 빅데이터서비스 개발자 부트캠프

End of Document