주제분석 1주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 주제분석 1-3주차 패키지 문제의 조건 및 힌트는 Python을 기준으로 하지만, R을 사용해도 무방합니다. 하지만, Python을 사용하는 것을 권장합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 16시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요. 또한 불성실한 제출로 인한 경고를 2회 받으실시도 퇴출이니 유의해주세요.

Chapter 1: 개발 환경 준비(Anaconda, PyCharm, Jupyter Notebook))

주제분석 1주차 패키지부터는 파이썬을 활용한 데이터 분석 및 모델 개발을 해보겠습니다. 데이터 분석에 필요한 파이썬 개발 환경을 만들기 위해서는 몇 가지 준비가 필요합니다.

Anaconda는 프로젝트마다 독립적인 가상 환경을 만들 수 있도록 도와주는 프로그램입니다. 다양한 라이브러리를 설치하여 사용하다 보면 서로 다른 버전으로 인한 충돌이 발생하는 경우가 발생할 수 있으므로, 가상 환경을 적절히 활용하여 개발 환경을 분리해 주어야 합니다. 이외에 PyCharm과 Jupyter Notebook에 대한 설명은 이번 학기 OT 자료를 참고해주세요.

[조건 : requirements.txt에 기재된 목록 이외의 라이브러리는 사용할 수 없습니다.]

문제1. Anaconda를 설치한 후 'WEEK1'이라는 이름의 가상 환경을 생성해주세요. (파이썬 버전 3.8)

(HINT1) Anaconda Prompt에 명령어를 입력하여 가상 환경을 생성할 수 있습니다.

문제2. 'WEEK1' 가상 환경을 열어 requirements.txt의 라이브러리를 모두 설치해주세요.

(HINT2) Anaconda Prompt에 명령어를 입력하여 한 번에 설치할 수 있습니다.

문제3. PyCharm을 열어 'WEEK1' 가상 환경을 사용하는 새로운 프로젝트를 생성해주세요.

(HINT3) File > New Project 탭에서 생성할 수 있습니다. (제목 : WEEK1)

문제4. Anaconda에 Jupyter Notebook을 설치해주세요.

Chapter 1의 문제에 대한 답은 .ipynb 파일의 셀에 각각 주석으로 어떤 명령어를 사용했는지 적어주세요.

Chapter 2: 데이터 전처리 및 시각화

사용할 데이터의 형태를 확인하고 전처리 과정을 진행합니다. 데이터셋을 분리한 후 학습 데이터의 기술통계량과 분포를 확인하겠습니다. 그리고 상관관계 및 다중공선성을 확인하여 제거할 변수가 있다면 제거하는 과정을 거칩니다. 마지막으로 이상치들을 제거하여 Strength 예측 모델링을 위한 최종 데이터셋을 만들어 보겠습니다.

문제1. 주어진 concrete_data.csv 데이터셋을 불러오고, 비어있는 행이 있는지 확인해 주세요.

문제2. 주어진 데이터셋을 train 데이터셋과 test 데이터셋으로 분리하고, 각각의 크기를 출력해 주세요.

(HINT) test 데이터셋의 비율을 0.2로 설정하고, random state를 설정해 주세요.

문제3. 학습 데이터(train)의 기술통계량과 변수별 분포와 관계를 확인하세요.

(HINT) seaborn 라이브러리를 활용하면 편합니다.

문제4. 변수들 간의 상관계수 행렬을 확인하고, 이를 히트맵으로 시각화 하세요.

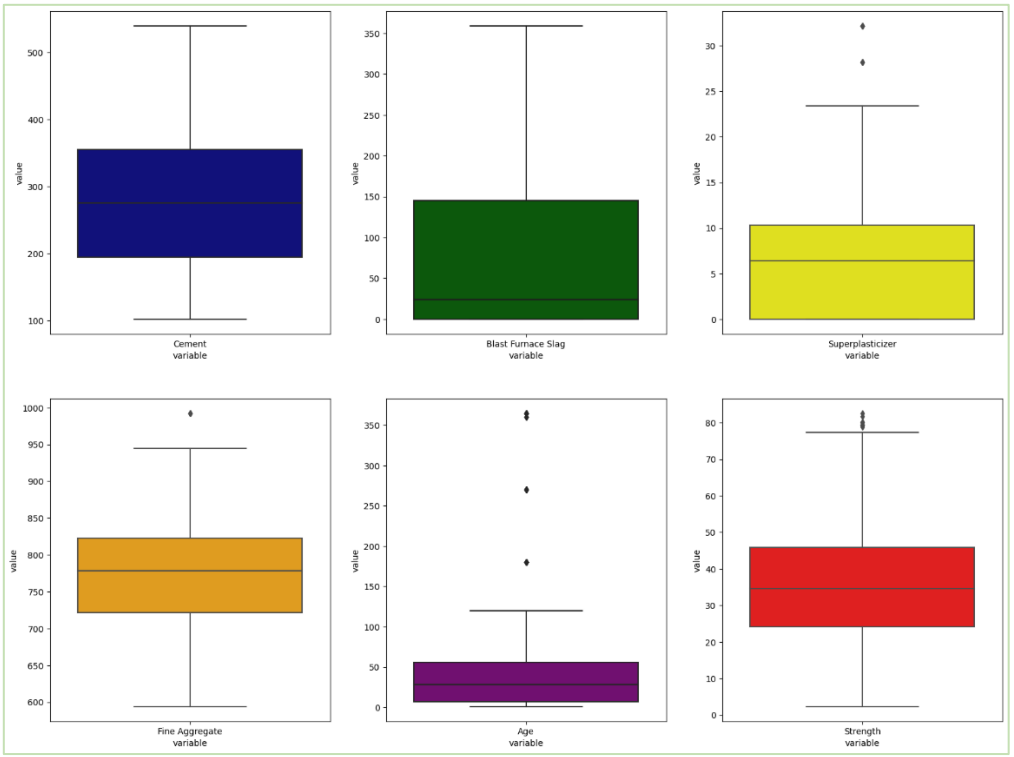
(HINT) seaborn 라이브러리를 활용하면 편합니다.

문제5. VIF(Variance Inflation Factors)가 무엇인지 간략히 설명한 후, VIF를 확인하여 제거해야 할 변수가 있다면 제거해 주세요.

(HINT) statsmodels 라이브러리를 사용하면 편합니다.

문제6. 변수별로 이상치가 있는지 파악하기 위해 Boxplot을 아래와 같이 시각화 하세요.

- 그래프의 크기는 (20, 15)입니다.



문제7. 주어진 outlier.py를 설명에 따라 완성하고, del_outlier 함수를 활용하여 이상치를 제거해 주세요.

문제8. 주어진 transform.py를 설명에 따라 완성하고, Scaling 클래스를 활용하여 train 데이터셋과 test 데이터셋 모두에 대해 minmax scaling을 진행해 주세요.

- Minmax Scaling 된 학습 데이터의 기술통계량을 확인하여 scaling이 잘 되었는지 확인하세요.
- Minmax Scaling 된 학습 데이터의 VIF를 확인하고, 어떻게 달라졌는지 설명해 주세요.

Chapter 3: 모델링 – Normal Equation

크게 2가지로 회귀모형을 적합해보고, 그 결과를 비교해볼 것입니다. 첫 번째로, 행렬 연산을 통한 최소제곱법을 구현해보도록 하겠습니다. 주어진 zip 파일에 들어있는 model.py와 matrix.py를 설명에 따라 완성하고, 이 모듈들을 활용하여 다중회귀모형을 적합해 봅시다.

문제1. 주어진 model.py를 완성하세요.

문제2. 주어진 matrix.py를 완성하세요.

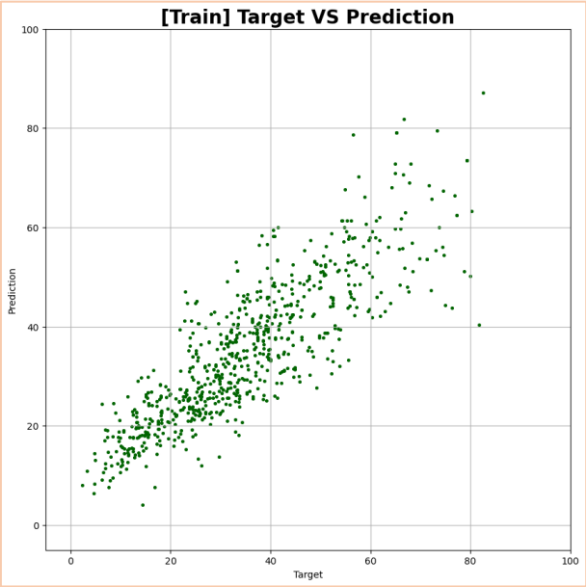
문제3. 완성된 matrix.py를 import한 후, Chapter 2에서 만든 최종 데이터셋을 활용하여 LSE 클래스의 객체를 생성해 주세요.

문제4. Model 클래스 내에 정의된 describe() 함수의 Exception이 정상적으로 출력되는지 확인해 주세요.

(HINT) describe() 함수의 인자로 'X', 'y'가 아닌 string을 아무거나 넣으면 됩니다.

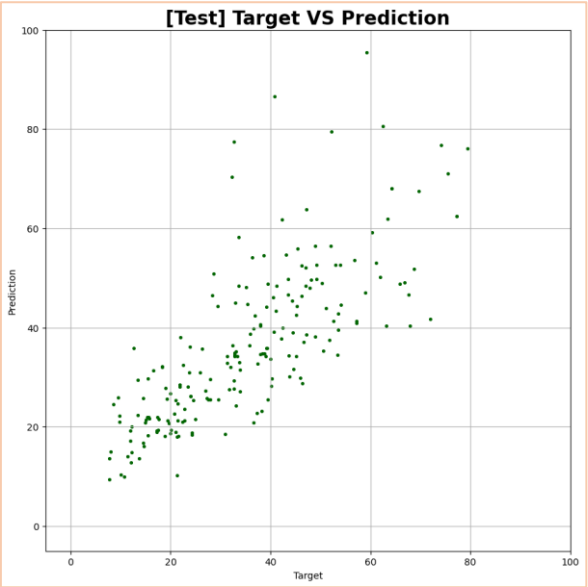
문제5. LSE 클래스의 normal_eq() 함수를 사용하여 회귀모형을 적합하고, 학습 데이터에 대한 예측값과 실제값을 아래와 같이 비교해 주세요.

- 그래프의 크기는 (10, 10) 점의 크기는 7, 제목의 크기는 20, 볼드체입니다.



문제6. Model 클래스 내에 정의된 predict() 함수를 활용하여 test 데이터셋에 대한 예측을 진행하고, 아래와 같이 시각화 해주세요.

- 그래프의 설정값은 문제 5와 동일합니다.



Chapter 4: 모델링 – Gradient Descent

두 번째로, 경사하강법을 통해 다중회귀모형을 적합해볼 것입니다. 경사하강법은 딥러닝의 핵심적인 학습 방법이기도 하지만, 딥러닝 이외의 고전적인 머신러닝 모델들을 학습할 때도 사용할 수 있습니다. Numpy를 통해 구현하는 것도 가능하지만, 코드의 가독성과 코드 작성의 편의를 위해 딥러닝 프레임워크인 pytorch를 사용하여 모델을 구현해 보겠습니다.

문제1. 경사하강법이란 무엇인지, 수식이 어떻게 되는지 설명해 주세요.

문제2. 주어진 gradient.py를 완성하세요.

문제3. 완성된 gradient.py를 import한 후, Chapter 2에서 만든 최종 데이터셋을 활용하여 GradientDescent 클래스의 객체를 생성해 주세요.

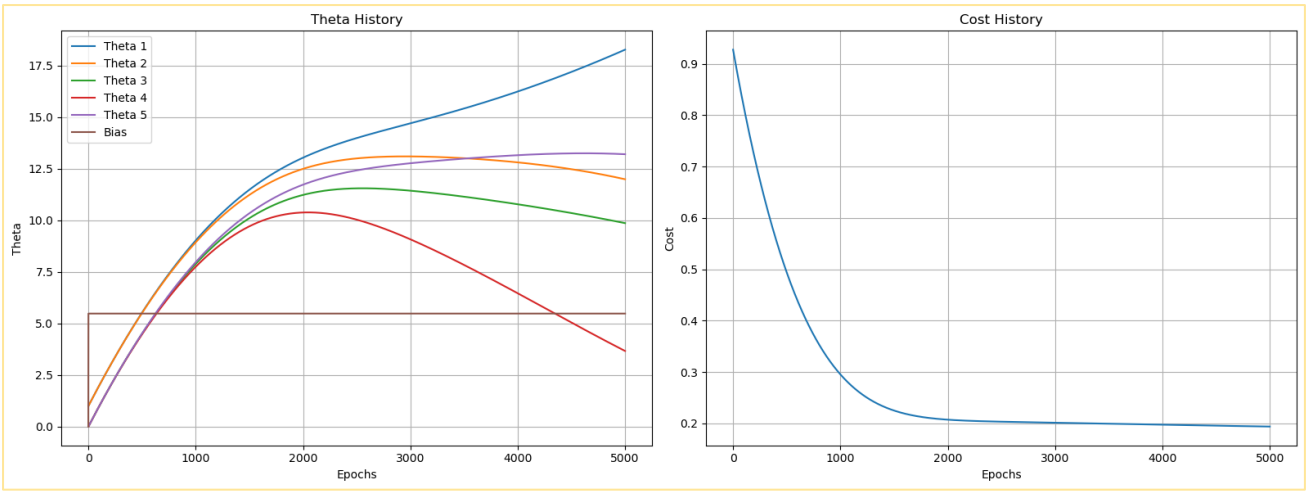
문제4. 주어진 gradient.py를 완성하세요.

문제5. GradientDescent 클래스의 compute_gradient() 함수를 사용하여 회귀모형을 적합하고, 학습 데이터에 대한 예측값과 실제 값을 Chapter 3의 문제 5, 6과 동일하게 시각화 해주세요.

- num_iter, lr, optimizer는 다양하게 시도해보면서 학습을 진행해 주세요. 정답은 없습니다.

문제6. 학습 과정에서 회귀계수와 손실함수의 값들이 어떻게 변해왔는지 아래와 같이 시각화 해주세요.

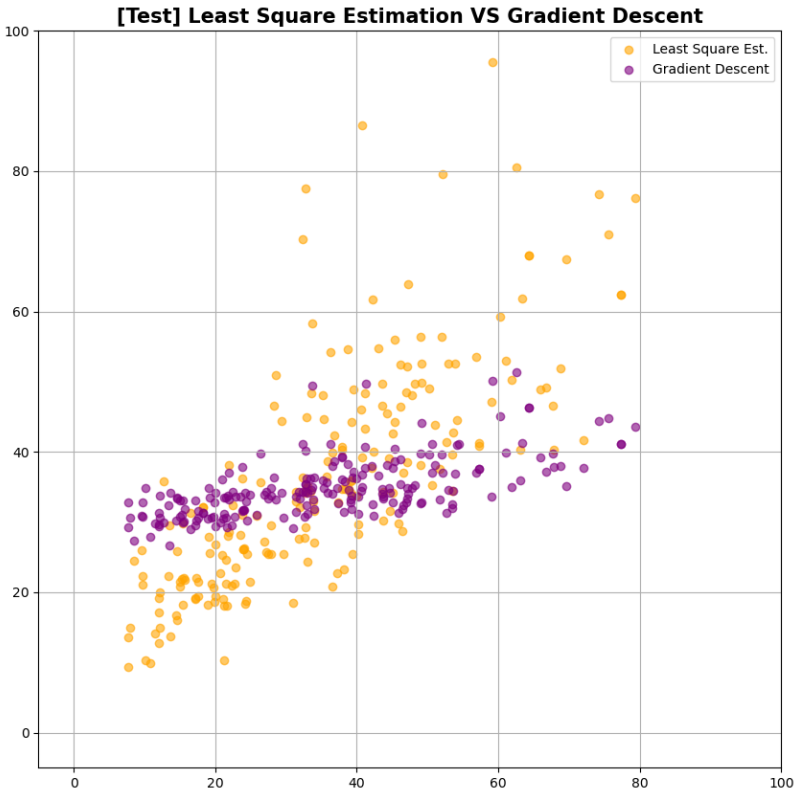
- 그래프의 크기는 (16, 6)입니다.
- 문제 5에서 정답이 없는 것처럼, 문제 6의 결과 또한 정답이 없습니다. 편하게 시각화 해주세요.



Chapter 5: 모델링 결과 비교

최종적으로 2가지 방법을 통해 구현한 회귀모형들을 비교해 보겠습니다. Normal Equation을 사용했을 때와 경사하강법을 사용했을 때 어떤 차이가 있는지, 어떤 장단점이 있는지, 그리고 왜 그런 차이들이 발생하는지 고민해 봅시다.

문제1. 모델별로 test 데이터에 대해 예측한 값들과 실제 값을 아래와 같이 비교해 주세요.



문제2. 결과를 보면, 경사하강법을 활용한 회귀모형의 성능이 최소제곱법을 활용한 회귀모형의 성능보다 떨어지는 것을 알 수 있습니다. 이유가 무엇일까요?