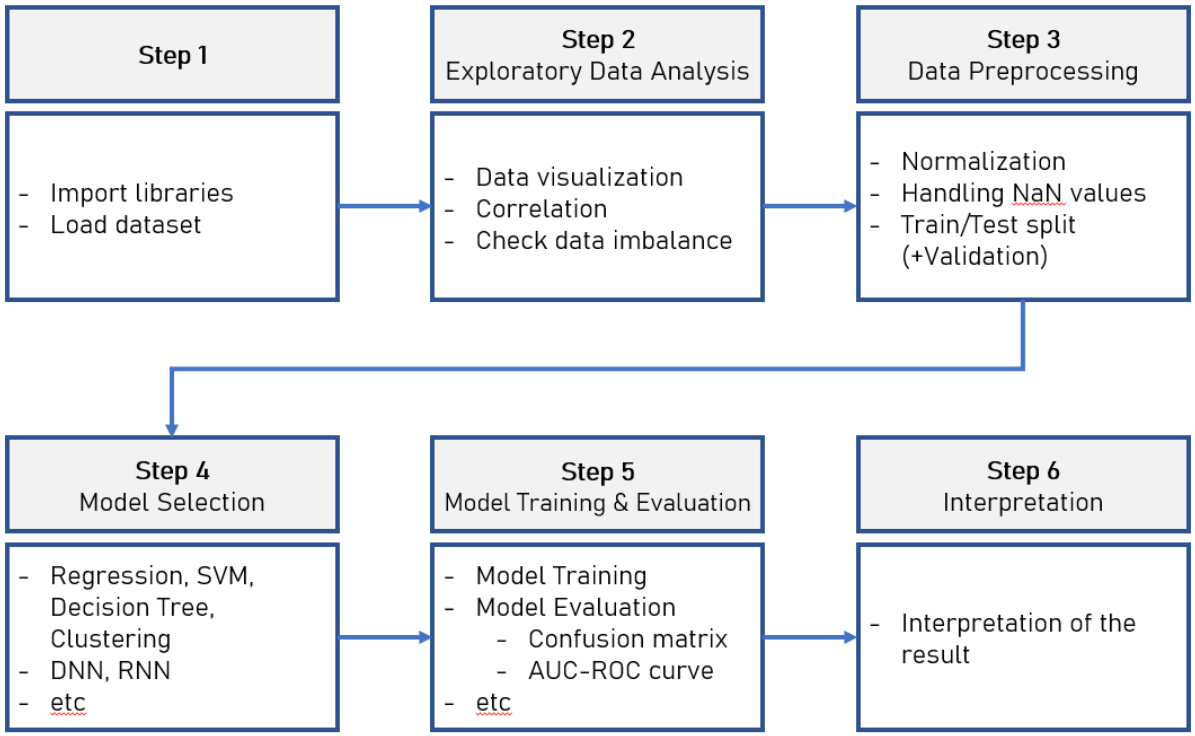
Assignment 2 Machine Learning Part

Total: 100 points (5 problems X 20 points)

0. Introduction

Data Analysis Workflow



일반적인 데이터 분석의 흐름을 정리해보았고, 주어지는 코드 역시 대략 위의 과정을 따르므로 익혀 두길 바란다.

1. Linear Regression

Problem Definition

You are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

(a) (5pt) car_dataset_revised.csv를 car_data 변수에 넣어 읽어오고, 데이터의 구조를 파악하여라. 또한, 해당 데이터에서의 numerical variable 과 가격 간의 각각의 상관관계를 시각화하고 경향성을 파악하여라.

(b) (5pt) one-hot encoding이 필요한 변수를 처리하여 final_data 변수에 저장하고, 이 과정이 왜 필요한지 (주석 달고) 설명하여라 (자신이 찾아본 url 을 같이 첨부하여도 됨).

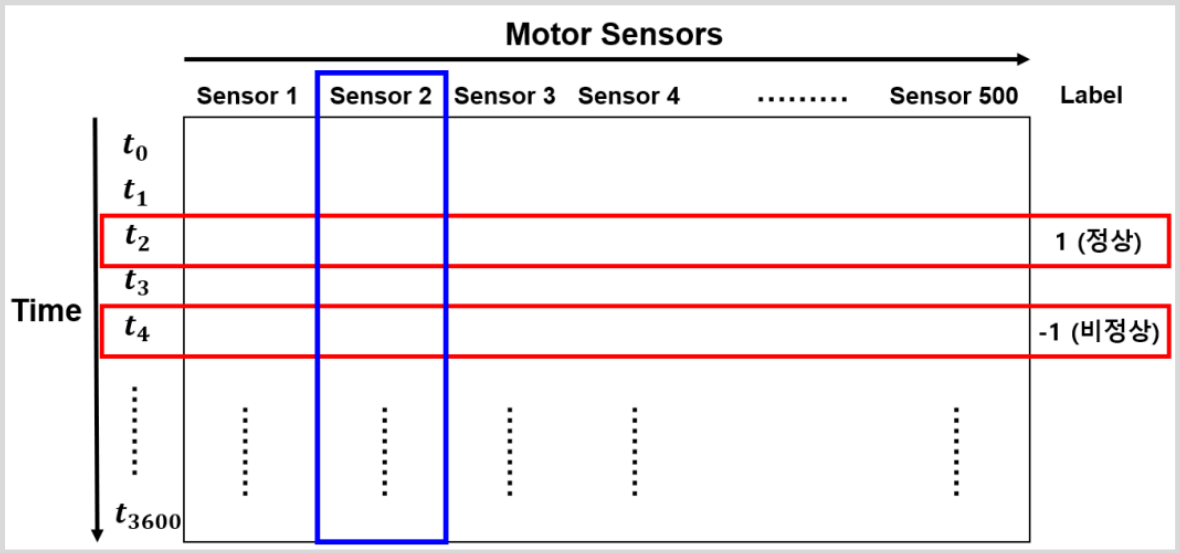
(c) (10pt) scikit learn의 LinearRegression module을 사용하여 훈련 데이터를 통해 학습하고, 테스트 데이터를 통해 예측하여라. 그리고 MSE와 R-square 값을 계산하여라.

2. Logistic Regression

Problem Definition

The provided data belongs to the type of time series data known to have the highest analysis demand in the manufacturing industry and contains values measured over time from 500 sensors in the automobile system.

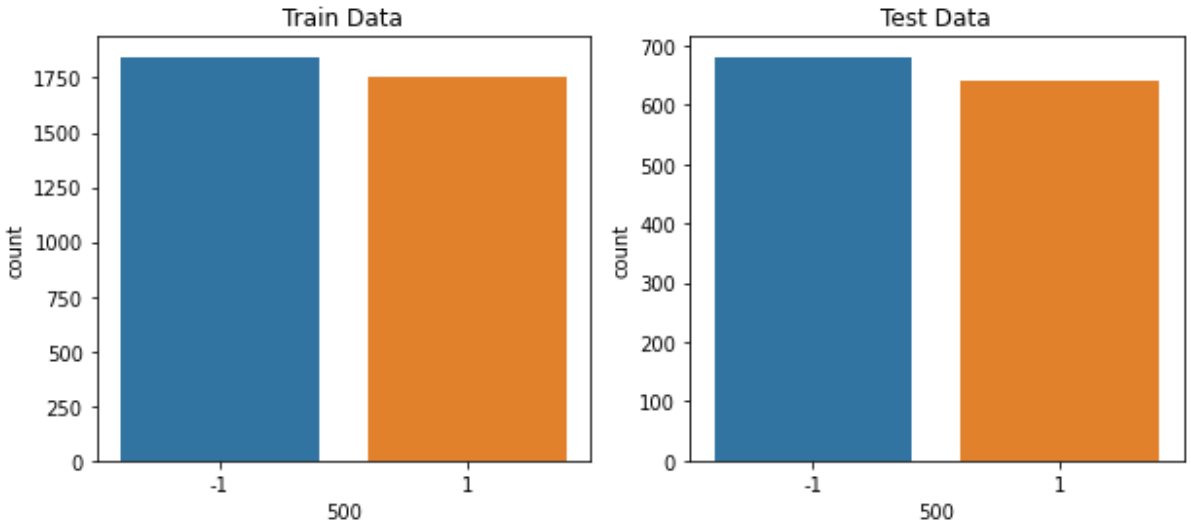
The structure of the data is as follows.



Let's make a model to classify defective (abnormal) products!

(a) (5 pt) 1) train과 test 데이터 셋으로부터 설명 변수와 종속 변수를 구분하여 저장하고 (X_train, y_train, X_test, y_test), 2) 훈련과 테스트 데이터 내에 정상/비정상 집단의 비율을 시각화 하여 확인하여라.

예시)



(b) (5 pt) 모델을 학습 시키고 테스트 데이터로 예측하여라.

(c) (10 pt) 모델의 성능을 평가하기 위해 'conf_matrix' 그리고 'model_evaluation' 함수를 정의하였다.

1) 'model_evaluation' 함수 내의 TP, TN, FP, FN 값을 각각 정의하고 이 값들을 적절히 사용하여 accuracy, precision, recall, f1_score를 출력하는 함수를 정의하여라.

2) 위에서 정의한 두가지 함수를 이용해 (모델의) 성능을 평가하여라.

3) scikitplot 모듈을 사용해 ROC curve를 그려라.

한편, scikit-learn이 제공하는 함수들을 통해, 위에서 정의한 'model_evaluation' 함수를 통해 구현한 네 가지 값을 쉽게 얻을 수 있는데, 이 값이 일치하는지 살펴보고, 다르다면 이유를 생각해보아라.

참고) https://stackoverflow.com/questions/35178590/scikit-learn-confusion-matrix

3. Support Vector Machine

Problem Definition

Breast cancer is the most common cancer amongst women in the world. It accounts for 25% of all cancer cases, and affected over 2.1 Million people in 2015 alone. Early diagnosis significantly increases the chances of survival. The key challenges against it's detection is how to classify tumors into malignant (cancerous) or benign (non cancerous).

Your task is to classify tumors into malignant (cancerous) or benign (non-cancerous) using features obtained from several cell images.

(a) (2 pt) 학습과 테스트 데이터 셋을 7:3의 비율로 나누어라.

(b) (5 pt) Linear kernel을 이용해 훈련 데이터로 모델을 학습시키고 테스트 데이터로 (print_model_score 함수를 이용해서) 평가하여라.

(c) (8 pt) 앞서 'describe' 함수를 통해 데이터의 분포를 확인해보았다. 1) feature간 데이터 분포를 맞추어 주기 위해 standard-scaling을 진행하고, 2) Linear kernel 이용한 SVM 모델을 만들고 평가하여라.

(d) (5 pt) 모델의 성능을 측정하기 위한 ROC curve를 (plot_roc_curve 함수를 이용해서) 시각화하고, AUC를 출력하여라.

4. Decision Tree & Random Forest

Problem Definition

Last time, we implemented a model using logistic regression. This time, let's implement a model using decision trees and ensemble techniques!

Notice: The blank (TO DO) code snippets in the .ipynb remain the same as in logistic regression. So do not forget to fill in the code snippets.

(a) (2.5 pt) Decision Tree를 이용해 max_depth 등 hyperparameter를 조정해가며 모델을 생성하고 훈련용 데이터로 학습시키고, 테스트 데이터로 예측한 결과를 위에서 정의한 'conf_matrix'와 'model_evaluation' 함수를 통해 평가하여라.

(b) (5 pt) (**Bagging**) scikit learn 에서 제공하는 'BaggingClassifier' module을 이용하여 bagging 통해 모델을 학습시키고 성능을 평가하여라.

(c) (7.5 pt) (**RandomForest**) RandomForest 를 통해 구현한 모델의 hyperparameter 조정 (n_estimators, max_depth 등) 을 해보면서 모델의 성능이 달라지는지 확인해보아라. 다른 hyperparameter를 추가해도 좋다.

(d) (5 pt) (**Boosting**) Boosting 기법 중 하나로서, gradient Boosting 개념을 decision tree에 도입한 알고리즘이지만, gradient boosting 과는 달리 학습을 위한 objective function에 regularization term이 추가되어 모델이 과적합 되는 것을 방지해준다.

xgboost를 통해 구현한 xgb_model을 이용해 test data를 예측하고 성능을 평가하여라.

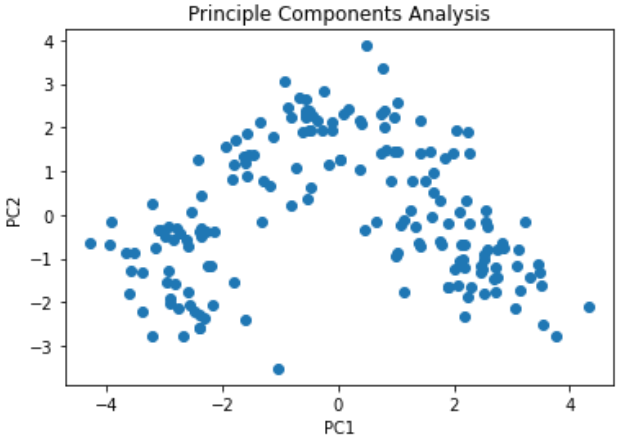
5. Clustering

Problem Definition

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. Let's classify wine cultivars using clustering techniques!

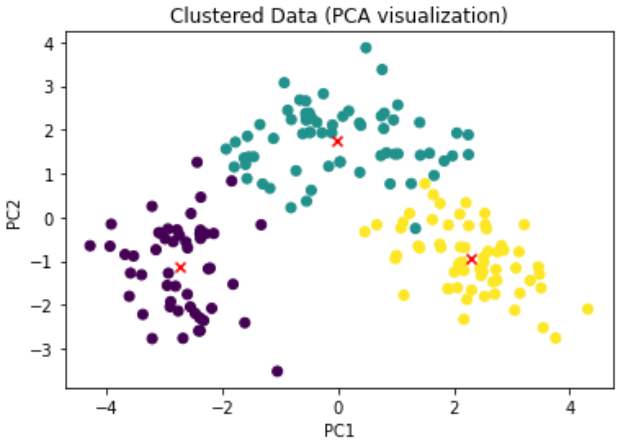
(a) (4 pt) Clustering에 앞서, 데이터의 분포를 확인하기 위하여 PCA를 이용해 multi-dimensional data의 차원을 2차원으로 축소시킨다. 'pca_data'를 활용해 이를 scatter plot으로 시각화 하여라.

예시)



(b) (8 pt) 1) 적절한 k 값의 선정을 위해 Elbow Plot을 그린 것을 확인할 수 있다 (코드 실행하면 됨). 이 그래프가 의미하는 바를 생각해보고 적절하다고 생각되는 k와 그 이유를 설명하여라.

2) 그리고 자신이 고른 k값으로 clustering을 진행하고 다음과 같이 시각화 하여라.



(c) (8 pt) 임의로 난수를 생성하여 대략 5개의 군집을 형성한 그래프를 시각화 하였다. k값을 5로 설정하고 k-means clustering을 진행하였는데 이유가 무엇인지 군집이 잘 형성되지 않았다. 왜 이러한 결과가 나오는지 생각해보고, 이를 해결할 수 있는 방법이 있다면 근거를 간단히 설명하고 다시 군집화 해보자.