Multivariate Data Analysis Assignment #3

Dimensionality Reduction for Multivariate Linear Regression & Logistic Regression

(준비사항)

(1) Dataset for Multivariate Linear Regression – 본인의 과제 1 에서 사용했던 데이터셋

(2) Dataset for Logistic Regression – 본인의 과제 2 에서 사용했던 데이터셋

[Part 1] Multivariate Linear Regression

본인이 보유한 MLR 용 데이터를 Training:Validation 을 70%:30%로 분할하시오.

[Q1] Forward Selection, Backward Elimination, Stepwise Selection 방식을 사용하여 MLR 변수 선택 과정을 수행해 보시오. 각 방법론마다 Training dataset 에 대한 Adjusted R^2 및 소요 시간, Validation dataset 에 대한 RMSE, MAE, MAPE 를 산출하시오.

[Q2] Adjusted R^2를 Fitness function으로 하는 Genetic Algorithm 기반의 변수 선택 함수를 작성해 보시오. 작성한 함수를 이용하여 GA 기반 변수 선택을 수행하고, 선택된 변수를 사용한 MLR의 Validation dataset에 대한 예측 성능(RMSE, MAE, MAPE), 변수 감소율, 수행 시간의 세 가지 관점에서 Forward Selection, Backward Elimination, Stepwise Selection 방식을 사용한 MLR과 비교해보시오.

[Q3] Genetic Algorithm에서 변경 가능한 하이퍼파라미터들(population size, Cross-over rate, Mutation rate 등)에 대해 각각 최소한 세 가지 이상의 후보 값들을 선정(최소 27가지 이상의 조합)하고 각 조합에 대한 변수 선택 결과를 비교해 보시오. 최종 결과에 가장 큰 영향을 미치는 하이퍼파라미터는 무엇으로 나타났는가? 왜 그런 결과가 나타났다고 생각하는지 자신의 생각을 서술해 보시오.

[Part 2] Logistic Regression

본인이 보유한 Logistic Regression 용 데이터를 Training:Validation 을 70%:30%로 분할하시오.

[Q4] Forward Selection, Backward Elimination, Stepwise Selection 방식을 사용하여 Logistic Regression 변수 선택 과정을 수행해 보시오. 각 방법론마다 Training dataset 에 대한 AUROC 및 소요 시간, Validation dataset 에 대한 AUROC, Accuracy, BCR, F1-Measure 를 산출하시오.

[Q5] AUROC를 Fitness function으로 하는 Genetic Algorithm 기반의 변수 선택 함수를 작성해 보시오.

작성한 함수를 이용하여 GA 기반 변수 선택을 수행하고, 선택된 변수를 사용한 Logistic Regression의 Validation dataset에 대한 분류 성능(AUROC, Accuracy, BCR, F1-Measure), 변수 감소율, 수행 시간의 세 가지 관점에서 Forward Selection, Backward Elimination, Stepwise Selection 방식을 사용한 Logistic Regression과 비교해보시오.

[Q6] Genetic Algorithm에서 변경 가능한 하이퍼파라미터들(population size, Cross-over rate, Mutation rate 등)에 대해 각각 최소한 세 가지 이상의 후보 값들을 선정(최소 27가지 이상의 조합)하고 각 조합에 대한 변수 선택 결과를 비교해 보시오. 최종 결과에 가장 큰 영향을 미치는 하이퍼파라미터는 무엇으로 나타났는가? 왜 그런 결과가 나타났다고 생각하는지 자신의 생각을 서술해 보시오.