Multivariate Data Analysis Assignment #3

Dimensionality Reduction for Multivariate Linear Regression

Dataset: Weather Ankara

(https://sci2s.ugr.es/keel/dataset.php?cod=41)

해당 데이터셋은 Annkara 지역의 날씨에 대한 정보를 포함하고 있는 데이터셋이다. Mean_temperature 항목이 종속변수이며 나머지 9개의 변수들이 설명변수이다. 총 321건의 관측치가 존재한다(General information에서 1,609건의 관측치는 잘못 기재된 정보임).

General information

Weather Ankara data set			
Type	Regression	Origin	Real world
Features	9	(Real / Integer / Nominal)	(9 / 0 / 0)
Instances	1609	Missing values?	No

Attribute description

Attribute	Domain
Max_temperature	[23.0, 100.0]
Min_temperature	[-7.1, 65.5]
Dewpoint	[-3.1, 57.6]
Precipitation	[0.0, 4.0]
Sea_level_pressure	[29.46, 30.6]
Standard_pressure	[26.3, 27.18]
Visibility	[0.2, 11.5]
Wind_speed	[0.0, 18.0]
Max_wind_speed	[2.19, 57.4]
Mean_temperature	[7.9, 81.8]

전체 데이터셋을 임의로 250 개의 Training dataset 과 71 개의 Validation dataset 으로 구분한 뒤 다음 각 물음에 답하시오.

[Q1] 모든 변수를 사용하여 Multiple Linear Regression (MLR) 모형을 학습하시오. 학습된 모형의 Adjusted R^2는 얼마인가? 또한, 유의수준 1% (significance level = 0.01)에서 통계적으로 유의미한 변수는 어떤 것이 있는가? 학습한 모형을 이용하여 Validation dataset 에 대한 RMSE, MAE, MAPE 를 산출해 보시오.

[Q2] Exhaustive Search를 수행하는 함수를 직접 구현하고 Training Dataset에 대한 Adjusted R^2 기준으로 가장 높은 값이 산출된 변수 집합을 제시하시오. 또한 Exhaustive Search에 소요된 시간을 산출하시오. 학습한 모형을 이용하여 Validation dataset에 대한 RMSE, MAE, MAPE를 산출하고 모든 변수를 사용한 MLR 모형의 결과와 비교하시오.

[Q3] Forward Selection, Backward Elimination, Stepwise Selection 방식을 사용하여 MLR변수 선택 과정을 수행해 보시오. 각 방법론마다 [Q2]와 같이 Training dataset에 대한 Adjusted R^2 및 소요 시간, Validation dataset에 대한 RMSE, MAE, MAPE를 산출하고 [Q1] 및 [Q2]의 결과와 비교하시오.

[Q4] Adjusted R^2를 Fitness function으로 하는 Genetic Algorithm 기반의 변수 선택 함수를 작성해 보시오. 작성한 함수를 이용하여 GA를 이용한 변수 선택을 수행한 결과를 소요 시간, Adjusted R^2, Validation dataset에 대한 RMSE, MAE, MAPE 관점에서 앞선 결과들과 비교해보시오.

[Q5] Genetic Algorithm에서 변경 가능한 하이퍼파라미터들(population size, Cross-over rate, Mutation rate 등)에 대해 최소한 세 가지 이상의 후보 값들을 선정하고 각 조합에 대한 변수 선택 결과를 비교해 보시오. 최종 결과에 가장 큰 영향을 미치는 하이퍼파라미터는 무엇으로 나타났는가? 왜 그런 결과가 나타났다고 생각하는가?