Part. 02

회귀분석

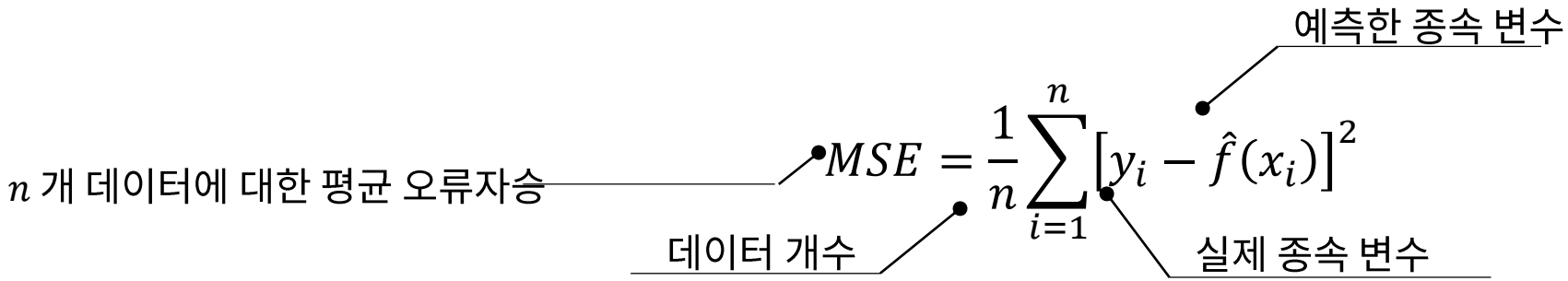
모형의 성능 지표

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택

I 모형의 성능지표

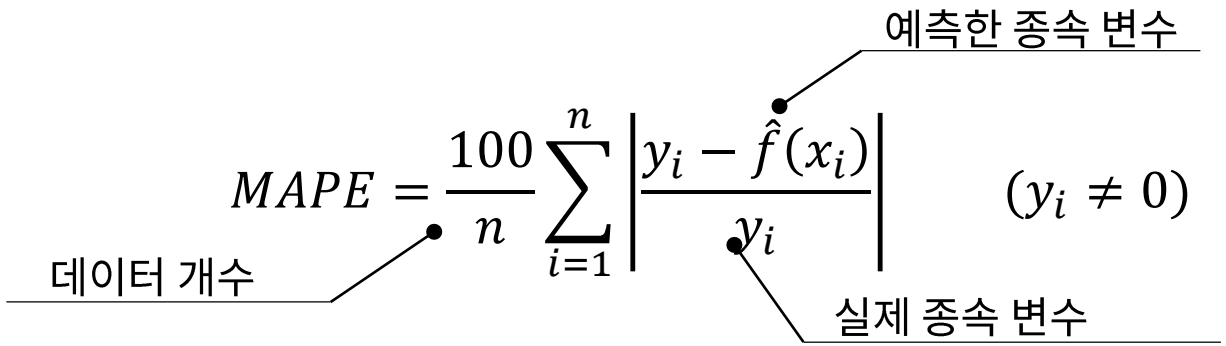
- MSE(Mean Squared Error)
  - f가 제대로 추정되었는지 평가하기 위해, 예측한 값이 실제 값과 유사한지 평가하는 척도가 필요함



- MSE는 실제 종속 변수와 예측한 종속 변수간의 차이
- MSE가 작을 수록 좋지만, MSE를 과도하게 줄이면 과적합의 오류를 범할 가능성이 있음
- 따라서, 검증 집합의 MSE를 줄이는 방향으로 f를 추정

I 모형의 성능지표

- MAPE(mean absolute percentage error)
  - f가 제대로 추정되었는지 평가하기 위해, 예측한 값이 실제 값과 유사한지 평가하는 척도가 필요함



- MAPE는 퍼센트 값을 가지며 0에 가까울수록 회귀 모형의 성능이 좋다고 해석할 수 있음
- 0%~100% 사이의 값을 가져 이해하기 쉬우므로 성능 비교 해석이 가능

I 모형의 성능지표

- 정확도 (Accuracy)

클래스 ={정상, 불량}		예측한 클래스	
		정상	불량
실제 클래스	정상	TN	FP
	불량	FN	TP

- 정확도(accuracy)는 전체 데이터 중에서 모형으로 판단한 값이 실제 값과 부합하는 비율
- 분모는 전체 데이터가 되고 분자는 모형이 실제 정상을 정상으로 그리고 실제 이상을 이상으로 옳게 분류한 데이터임

$$Accuracy = \frac{\text{옳게 분류된 데이터의 수}}{\text{전체 데이터의 수}} = \frac{TP + TN}{TP + FN + FP + TN}$$

I 모형의 성능지표

- 정밀도, 재현율, 특이도
  - 분류 모형의 목적에 따라 다양한 지표를 볼 수 있음

클래스 ={정상, 불량}		예측한 클래스	
		정상	불량
실제 클래스	정상	TN	FP
	불량	FN	TP

$$\text{정밀도}(Precision) = \frac{\text{옳게 분류된 불량 데이터의 수}}{\text{불량으로 예측한 데이터}} = \frac{TP}{FP+TP}$$

$$\text{재현율}(Recall) = \frac{\text{옳게 분류된 불량 데이터의 수}}{\text{실제 불량 데이터의 수}} = \frac{TP}{FN+TP}$$

$$\text{특이도}(Specificity) = \frac{\text{옳게 분류된 정상 데이터의 수}}{\text{실제 정상 데이터의 수}} = \frac{TN}{TN+FP}$$

정밀도(precision)는 분류 모형이 불량을 진단하기 위해 얼마나 잘 작동했는지 보여주는 지표

재현율(recall)은 불량 데이터중 실제로 불량이라고 진단한 제품의 비율 (진단 확률)

특이도(specificity)는 분류 모형이 정상을 진단하기 위해 잘 작동하는지를 보여주는 지표

I 모형의 성능지표

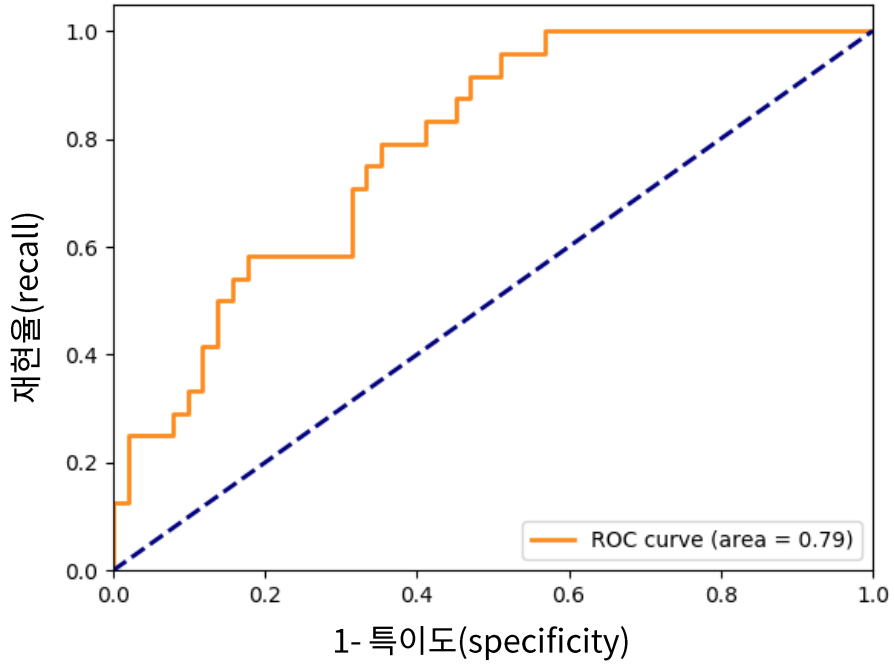
- G-mean, F1 measure
  - 실제 데이터의 대표적인 특성에는 불량(이상) 데이터를 탐지하는 것이 중요하다는 점과 이러한 불량 데이터는 매우 소수의 데이터라는 점임 (class imbalanced 문제)
  - 데이터 1000개 중 불량 데이터가 10개 나머지 990개는 정상 데이터라고 가정했을 때 분류 모형이 모든 데이터를 정상 데이터라고만 예측해도 정확도는 99%이며(accuracy paradox), 만약 우연히 1개만 불량이라고 예측했는데, 실제 불량일 경우 정밀도 지표는 1임
  - 실제데이터의 특성상 정확도보다는 제1종 오류와 제2종 오류 중 성능이 나쁜 쪽에 더 가중치를 주는 G-mean 지표나 불량에 관여하는 지표인 정밀도와 재현율만 고려하는 F_1 measure가 더 고려해볼 수 있는 지표임

$$G-mean = \sqrt{specificity \cdot recall} = \sqrt{(1-\alpha)\cdot(1-\beta)}$$

$$F_1\ measure = \frac{2}{\frac{1}{precision}+\frac{1}{recall}} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

I 모형의 성능지표

- ROC curve, AUC
  - 가로축을 1- 특이도(specificity) 세로축을 재현율(recall)로 하여 시각화한 그래프를 ROC (Receiver Operating Characteristics) curve라고 함.
  - 이때 ROC curve의 면적을 AUC라고 함



Part. 02

회귀분석

변수선택법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택