Part. 06

Class Imbalanced Problem

| Class Imbalanced Problem이란

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

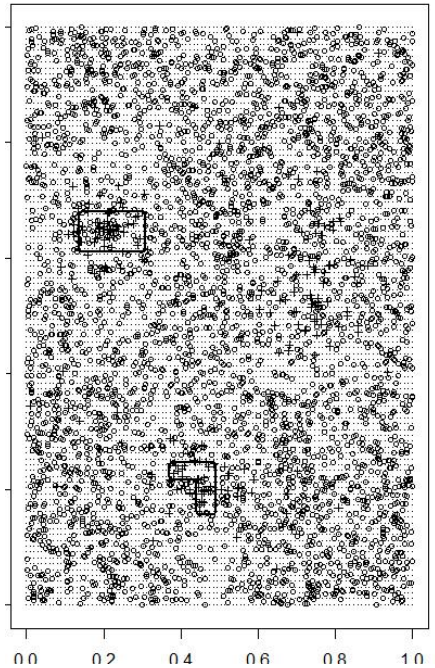
강사. 이경택

I Class Imbalanced Problem이란

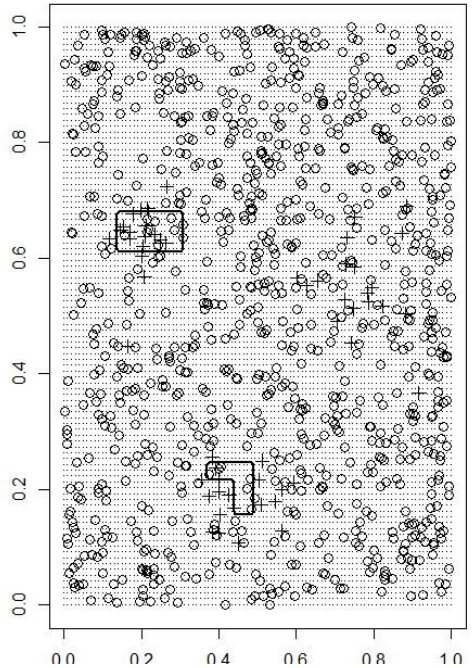
- Class Imbalanced Problem이란
  - 클래스 불균형은 다수 클래스(majority class)의 수가 소수 클래스(minority class)의 수보다 월등히 많은 학습 상황을 의미하며 클래스 불균형 데이터를 이용해 분류 모델을 학습하면 분류 성능이 저하되는 문제가 발생함.
  - 클래스 불균형 데이터는 의료, 반도체, 보험, 텍스트 등 여러 분야에 걸쳐서 발생하고 있는 문제임
  - IR (class imbalanced ratio) $= \frac{\#\ of\ majority\ class}{\#\ of\ minority\ class}$

I Class Imbalanced Problem이란

- Class Imbalanced Problem이란
  - 모델이 소수의 데이터를 무시하는 경향이 생김
  - 아래 예시는 IR = 20인 2차원 데이터를 decision tree로 학습했을 때의 decision boundary



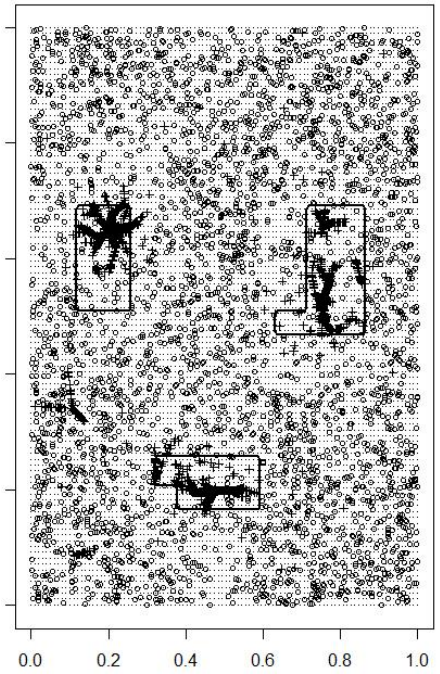
학습데이터



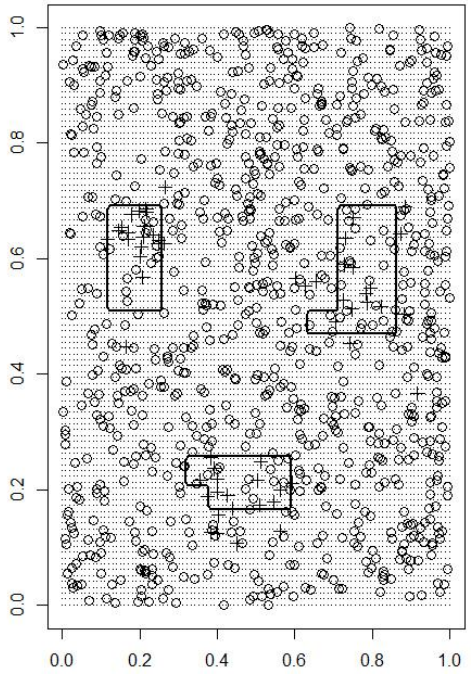
검증데이터

I Class Imbalanced Problem이란

- Class Imbalanced Problem이란
  - 모델이 소수의 데이터를 무시하는 경향이 생김
  - Oversampling 기법을 통해 아래와 같이 minority 데이터를 부풀려서 사용 가능



학습데이터



검증데이터

I Class Imbalanced Problem이란

- Class Imbalanced Problem이란
  - Binary문제에서 일반적으로 모델은 각 데이터에 대한 확률 값을 output으로 함
  - IR이 높은 경우 대부분의 데이터 셋에 대하여 0에 가까운 확률 예측 값을 냄
  - 예측 threshold 값(기본 0.5)이 달라져야하는 문제가 생김

I Class Imbalanced Problem이란

- Class Imbalanced Problem에서 사용하는 모델 성능 지표
  - 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 Class Imbalanced Problem이란

- Class Imbalanced Problem에서 사용하는 모델 성능 지표
  - 정밀도, 재현율, 특이도

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

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

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

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

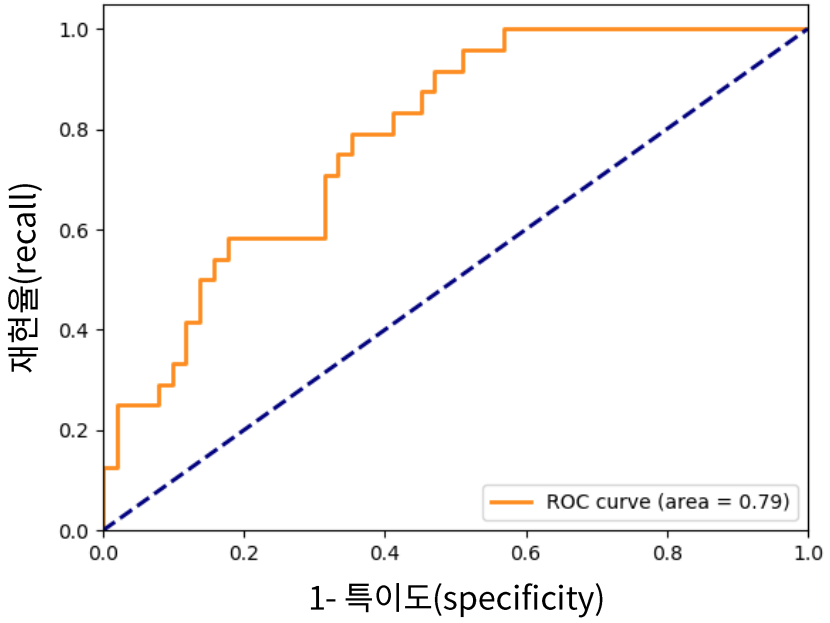
정밀도(precision)는 분류 모형이 불량을 진단하기 위해 얼마나 잘 작동했는지 보여주는 지표

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 Class Imbalanced Problem이란

- Class Imbalanced Problem에서 사용하는 모델 성능 지표
  - 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. 06

Class Imbalanced Problem

| Class Imbalanced Problem을 해결하기 위한 방법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

강사. 이경택