범주형자료분석팀

2팀
이지연
심예진
조장희
조혜현
진효주

INDEX

혼동 행렬

분류 모델의 **성능을 평가**할 때 사용되는 지표

예측값($\hat{Y}$)이 **실제 관측값(Y)**을 얼마나 정확히 예측했는지 보여주는 행렬

		관측값(Y)	
		Y=1	Y=0
예측값($\hat{Y}$)	$\hat{Y}$=1	TP	FP
	$\hat{Y}$=0	FN	TN

✔ T(True)와 F(false): 실제와 예측이 같은지 혹은 다른지

P(Positive)와 N(Negative): 예측을 긍정 혹은 부정이라 했는지에 대한 여부

혼동행렬

분류 평가지표

범주형 자료분석은 데이터마이닝 또는 머신러닝의 관점에서 **분류모델**

이번 파트에서는 분류 모델의 다양한 성능 평가지표에 대해 알아볼 예정!

경우에 따라 사용해야 하는 평가지표가 달라지므로 적절한 사용이 중요!

- 정확도 (Accuracy)
- 정밀도 (precision)
- 민감도 (Sensitivity)
- 특이도 (Specificity)
- F1-score
- MCC (매튜 상관계수)

혼동행렬

정확도 (Accuracy/ACC/정분류율)

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = 1 - \text{Error rate}$$

전체 경우에서 실제값과 예측값이 같은 경우의 비율

즉, 예측이 실제 정답과 얼마나 정확히 일치하는지 나타내는 지표

		관측값(Y)	
		Y=1	Y=0
예측값($\hat{Y}$)	$\hat{Y}$=1	TP	FP
	$\hat{Y}$=0	FN	TN

- ✔ 직관적이라 자주 쓰이는 지표
- ✔ 1에 가까울수록 좋은 모형
- ✔ Imbalanced data에서 모형 평가하는 경우 문제가 발생

혼동행렬

F1-score

$$\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FN + FP}$$

Precision과 Recall의 조화평균

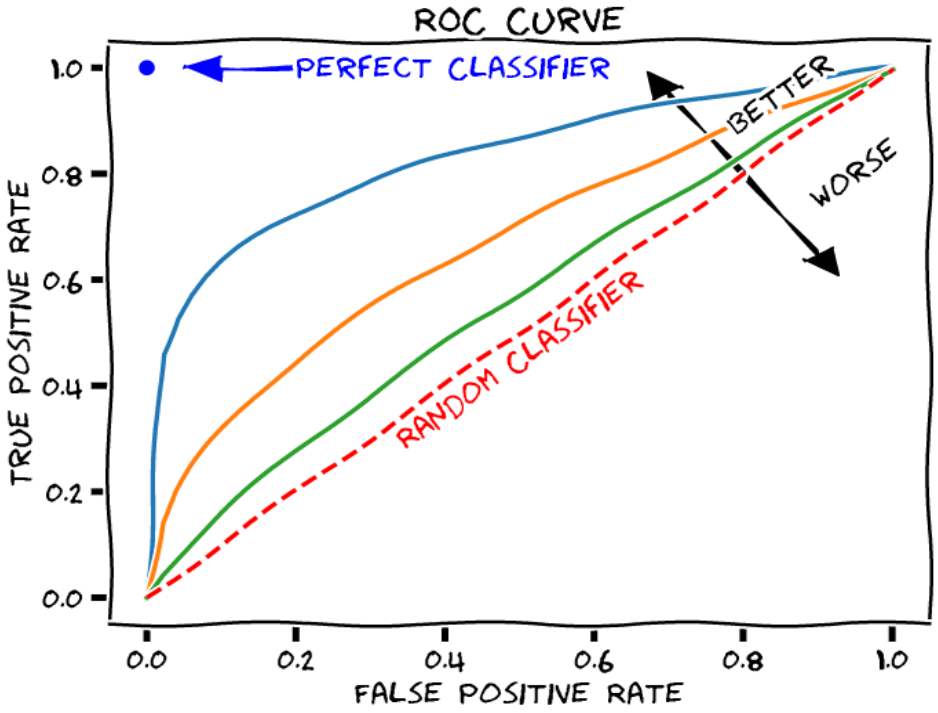
		관측값(Y)	
		Y=1	Y=0
예측값($\hat{Y}$)	$\hat{Y}$=1	TP	FP
	$\hat{Y}$=0	FN	TN

Precision

Recall(Sensitivity)

ROC곡선과 AUC

ROC 곡선의 형태



우상향하는 위로 볼록한 곡선

X축 : FPR(1-특이도) *짭인데 맞다고 하는 비율*

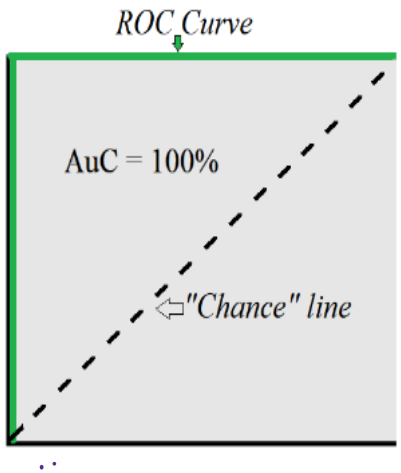
Y축 : TPR(민감도) *찐인데 맞다고 하는 비율*

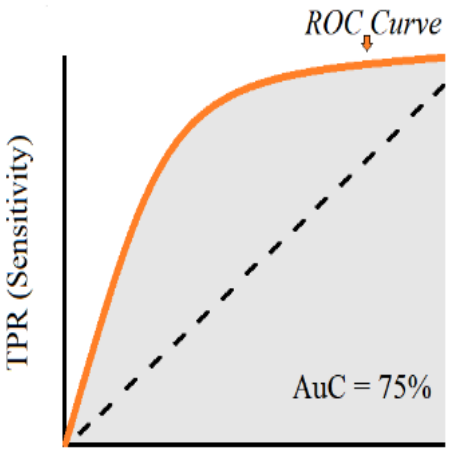
X는 작을수록, Y는 클수록 좋음

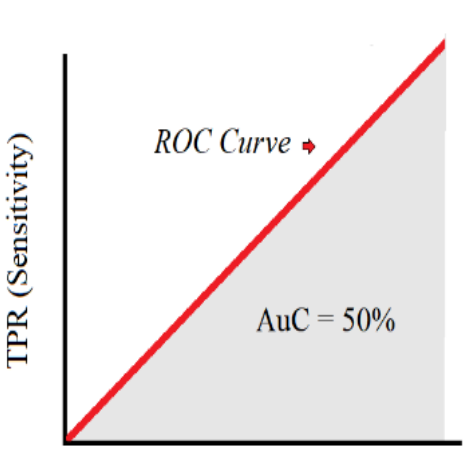
X,Y 둘 다 0~1 사이

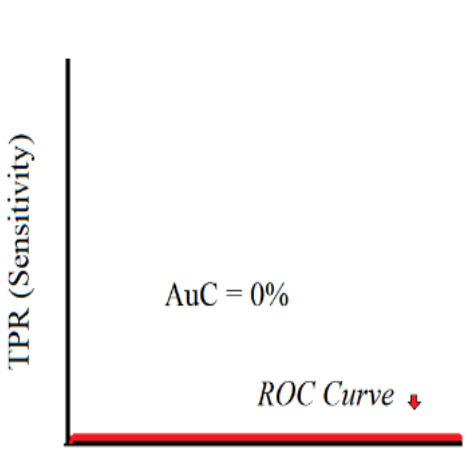
ROC곡선과 AUC

AUC의 특징









✔ AUC = 1

100% 완벽 예측
(Overfitting 의심)

✔ AUC = 0.75

흠 나쁘지 않네

✔ AUC = 0.5

50% 예측
걍 찍은거나 다름없,,

✔ AUC = 0

100% 반대 예측
시험문제 0점은
사실 다 맞은거나 다름 없다,,

클래스 불균형(Class Imbalance)

클래스 균형

소수의 클래스에 특별히 더 관심이 있는 경우에 필요함

Sampling을 통해 클래스 불균형을 해소할 수 있음

Garbage In! Garbage Out!

좋은 모델을 만들려면 좋은 train set이 필요하다!

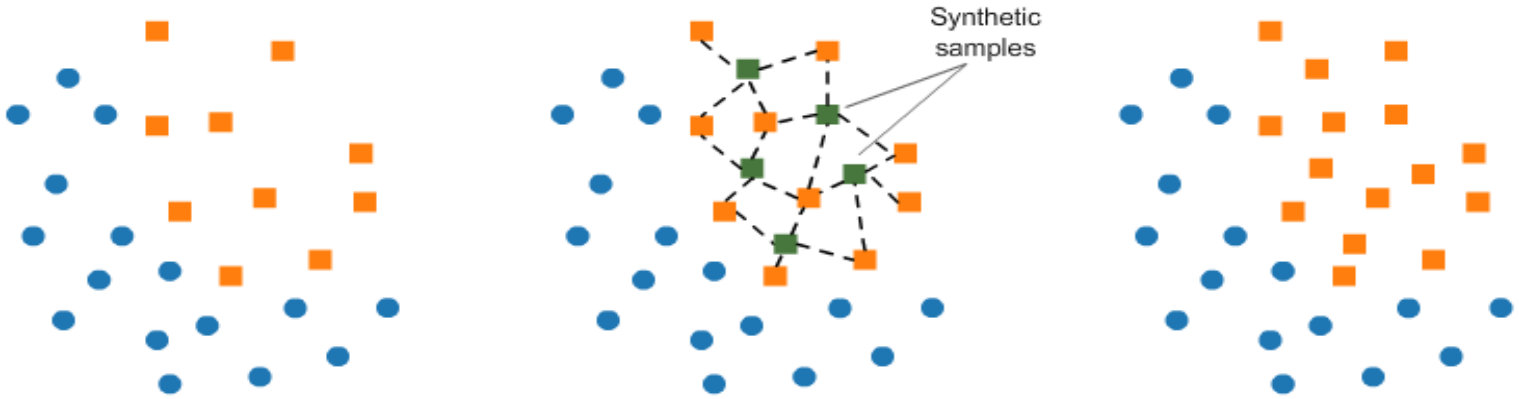
Sampling을 통해 비대칭 문제를 해결하자!

Sampling

SMOTE

알고리즘

1. 소수 클래스의 데이터 중 하나를 선택
2. 선택한 데이터와 가장 가까운 소수 클래스의 데이터 중에서 무작위로 k개의 데이터 선택
3. 선택한 데이터와 k개의 데이터 사이의 직선 상에 가상의 소수 클래스 데이터 생성



One-Hot Encoding(Dummy Encoding)

가변수(dummy variable)를 만들어주는 인코딩 방법

펜트하우스	오윤희	천서진	주단태	나애교
오윤희	1	0	0	0
천서진	0	1	0	0
주단태	0	0	1	0
나애교	0	0	0	1

☞ 해당 범주에는 1, 그 외에는 0 입력

Label Encoding

각 범주를 나누어 주기 위해 단순히 점수를 할당하는 인코딩 방법

(명목형 자료에 많이 사용)

펜트하우스	점수
오윤희	1
천서진	2
주단태	3
나애교	4

할당된 점수의 숫자에는 어떠한 순서나 연관성이 없음

요리보고 저리봐도 알 수 없는
인코딩 인코딩~

Ordinal Encoding

순서형 정보에 대응하는 점수를 할당하는 인코딩 방법

매움 정도	점수
🌶	1
🌶🌶	2
🌶🌶🌶	3
🌶🌶🌶🌶	4

<고추 "매운정도" 표시방법>

구 분				
매운 정도	맵지 않음	약간 매움	보통 매움	매우 매움
캡사이신 함량 (ppm)	100 미만	100~800	800~2,000	2,000이상

농산물 표준규격 기준 고추 맵기 표시

☞ 점수를 할당할 때는 보통 1부터 부여

할당된 점수들은 **순서나 연관성이 있음**

Mean Encoding(Target Encoding)

범주형 변수의 각 수준에 대하여

반응변수(타겟변수)Y의 평균으로 점수를 할당하는 인코딩 방법

[Y] 토익 점수	[X] 팀	Target Encoding
780	선대	855
930	선대	855
850	범주	820
870	범주	820
810	범주	820
750	범주	820
660	데마	863.33
980	데마	863.33
950	데마	863.33

$$\frac{850+870+810+750}{4}$$

=820 (**범주팀 토익 평균**)

실제로 맞는지 아닌지 모름 …
물어보지 마셈 …

Ordered Target Encoding(CatBoost Encoding)

현재 행 이전의 값들을 사용하여 평균을 구하고

이를 점수로 할당하는 인코딩 방법

✔ Target Encoding(Mean Encoding)과 매우 유사

Ordered Target Encoding은 **같은 범주라도 다른 점수 할당** 가능

✔ **CatBoost(부스팅 모델 중 하나)**에서 사용되는 인코딩 방식

장단점은 LOO Encoding과 동일