♥ TITLE : 3_평가 ♥ DATE : 2022.08.29

1 정확도(Accuracy)

↳ $정확도 = \frac{예측\ 결과가\ 동일한\ 데이터\ 건수}{전체\ 예측\ 데이터\ 건수}$

- ML 모델의 성능을 왜곡할 가능성이 O … label 값이 불균형 할 때
  ↳ 무조건 특정한 결과로 찍어도 데이터 분포가 균일하지 않은 경우 높은 수치가 나타날 수 O

2 오차 행렬

- 이진 분류의 예측 오류가 얼마인지와 더불어 어떠한 유형의 예측 오류가 발생하고 있는지를 함께 나타내는 지표

실제 클래스 \ 예측 클래스	(0)	(1)
(0)	TN	FP
(1)	FN	TP

- confusion_matrix( ) 사용 ⇒ 결과 array([TN, FP], [FN, TP])

3 정밀도와 재현율

① 정밀도 [예측을 positive로 한 대상 중 실제 값이 positive로 일치한 데이터의 비율
TP/(TP+FP)

② 재현율 (=민감도) [실제 값이 positive인 대상 중 예측과 실제 값이 positive로 일치한 데이터의 비율
TP/(FN+TP)
실제 positive 데이터를 negative로 잘못 판단 시 업무상 큰 영향이 발생하는 경우 중요함.

- 사이킷런에서 정밀도 계산은 precision_score( ), 재현율 계산은 recall_score( )로 수행

③ 정밀도/재현율 트레이드오프

- 정밀도와 재현율은 상호 보완적인 평가 지표 ⇒ 어느 한쪽을 강제로 높이면 다른 하나의 수치는 떨어지기 쉬움. ⇒ Trade-off
- predict_proba( ) [학습이 완료된 사이킷런 Classifier 객체에서 호출 가능
  입력 parameter: 테스트 feature 데이터셋
  개별 클래스의 예측 확률을 ndarray(mxn)로 반환
  ⇒ [클래스 0, 클래스 1]
- 분류 결정 임곗값을 조정해 정밀도/재현율의 성능 수치를 상호 보완적으로 조정할 수 O
- Binarizer 클래스 [threshold: 분류 결정 임계값 설정
  binarizer.fit_transform( )을 이용해 넘파이 ndarray 입력 시 값 ≤ threshold 이면 0, 값 > threshold 이면 1로 변환
- 임계값(threshold) ↓ [재현율 ↑ ↳ True 비율이 증가
  정밀도 ↓
- precision_recall_curve( ) [parameters [y_true: 실제 클래스값 배열
  probas_pred: positive 예측 확률 배열
  returns [정밀도(precision)
  재현율(recall)
  ↳ 시각화

④ 정밀도와 재현율의 맹점

- positive 예측의 임계값에 따라 정밀도와 재현율의 수치가 변동

i) 정밀도 100% 만들기
↳ 확실한 경우에만 positive로 예측하고 나머지는 모두 negative로 예측하기

ii) 재현율 100% 만들기
↳ 모두 positive로 예측

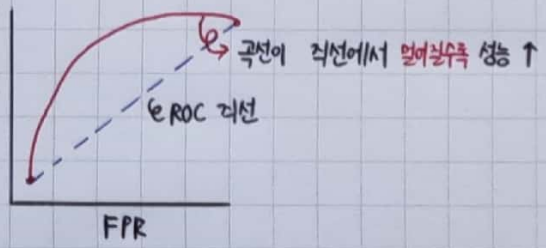
♥ TITLE : 3_평가 ♥ DATE : 2022.08.29

4 F1 score

- 정밀도와 재현율을 결합한 지표 ⇒ 정밀도나 재현율이 어느 한쪽으로 치우치지 않을 때 ↑
- $F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 \times \frac{precision \times recall}{precision + recall}$
- f1_score (실제값, 예측값)

5 ROC 곡선과 AUC

① ROC 곡선

- FPR(False Positive Rate)가 변할 때 TPR의 변화 양상 파악
- 민감도: TP/(TP+FN) TPR
  특이도: TN/(TN+FP) TNR
- 

- 임계값을 변화시켜 FPR 값을 0에서 1까지 변경 ⇒ TPR 구하기
- roc_curve( ) (1→0)
  - parameters>
    - y_true: 실제 클래스 값
    - y_score: 보통 predict_proba( )의 반환 값 array에서 positive 칼럼의 예측 확률이 주로 사용됨
  - returns>
    - fpr: FPR을 array로 반환
    - tpr: TPR 〃
    - thresholds: threshold를 〃

② AUC

- ROC 곡선 밑의 면적 → 1에 가까울수록 good
- 가운데 직선에서 멀어지고 왼쪽 상단 모서리에 가까울수록 AUC ↑
- 예측 확률값을 기반으로 계산됨.