

3장. 평가(Evaluation)

- 분류(Classification) 성능 평가지표 개요와 정확도(Accuracy) 소개

: ML모델의 예측 성능을 평가하는데 다양한 지표들이 존재한다

(정확도(Accuracy), 오차행렬(Confusion Matrix), 정밀도(Precision), 재현율(Recall), F1 스코어, ROC-AUC)

① 정확도(Accuracy)

$$\text{정확도(Accuracy)} = \frac{\text{예측 결과가 동일한 데이터 건수}}{\text{전체 예측 데이터 건수}}$$

정확도는 직관적으로 모델 예측 성능을 나타내는 평가 지표이다. 하지만 이진 분류의 경우 데이터의 구성에 따라 ML 모델의 성능을 왜곡할 수 있기 때문에 정확도 수치 하나만 가지고 성능을 평가하지 않는다.

특히 정확도는 불균형한(imbalanced) 레이블 값 분포에서 ML 모델의 성능을 판단할 경우, 적합한 평가 지표가 아니다.

*정확도의 문제점

(예) 타이타닉 생존자 예측에서 배에서 사고가 발생하면 첫번째로 구명보트에 태워야 할 사람은 여성과 노약자이다. 그렇기 때문에 복잡한 ML 알고리즘을 사용하지 않고도 단지 '성별이 여성이다'라고만 해도 상당히 높은 정확도를 얻을 수 있다.

```
import numpy as np
from sklearn.base import BaseEstimator

class MyDummyClassifier(BaseEstimator):
    # fit( ) 메소드는 아무것도 학습하지 않음.
    def fit(self, X, y=None):
        pass

    # predict( ) 메소드는 단순히 Sex feature가 1 이면 0 , 그렇지 않으면 1 로 예측함.
    def predict(self, X):
        pred = np.zeros((X.shape[0], 1))
        for i in range(X.shape[0]):
            if X['Sex'].iloc[i] == 1:
                pred[i] = 0
            else:
                pred[i] = 1
        return pred
```

Dummy Classifier의 정확도는: 0.7877

→ 성별이 남성인지 여성이지만 구별해서 판단해도 약 78%의 정확도를 얻을 수 있다.

- 오차행렬(Confusion Matrix), 정밀도(Precision), 재현율(Recall) 소개

② 오차 행렬(Confusion Matrix)

오차행렬은 이진 분류의 예측 오류가 얼마인지와 더불어 어떠한 유형의 예측 오류가 발생하고 있는지를 함께 나타내는 지표이다.

		예측 클래스(Predicted Class)	
		Negative(0)	Positive (1)
실제 클래스 (Actual Class)	Negative(0)	TN (True Negative)	FP (False Positive)
	Positive(1)	FN (False Negative)	TP (True Positive)

③ 정밀도(Precision)과 재현율(Recall)

- **정밀도 = $TP / (FP + TP)$**

정밀도는 예측을 Positive로 한 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율을 뜻한다.

정밀도의 값이 높아지기 위해서는 Positive에 대한 예측이 높아져야 하기 때문에 FP(예측을 Positive로 했는데 실제로는 Negative인 것)의 값이 낮아져야 한다.

- **재현율 = $TP / (FN + TP)$**

재현율은 실제 값이 Positive인 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율을 뜻한다.

재현율의 값이 높아지기 위해서는 실제 값이 Positive인 대상이 틀리면 안되기 때문에 FN(예측을 Negative로 했는데 실제로는 Positive인 것)의 값이 낮아져야 한다.

*사이킷런의 정밀도, 재현율

정밀도는 `precision_score()`, 재현율은 `recall_score()` 제공

- 정밀도와 재현율의 트레이드오프(Trade-off)

*업무에 따른 재현율과 정밀도의 상대적 중요도

재현율이 상대적으로 더 중요한 지표인 경우에는 실제 Positive 양성인 데이터 예측을 Negative로 잘못 판단하게 되면 업무상 큰 영향이 발생하는 경우

→ 암 진단, 금융사기 판별

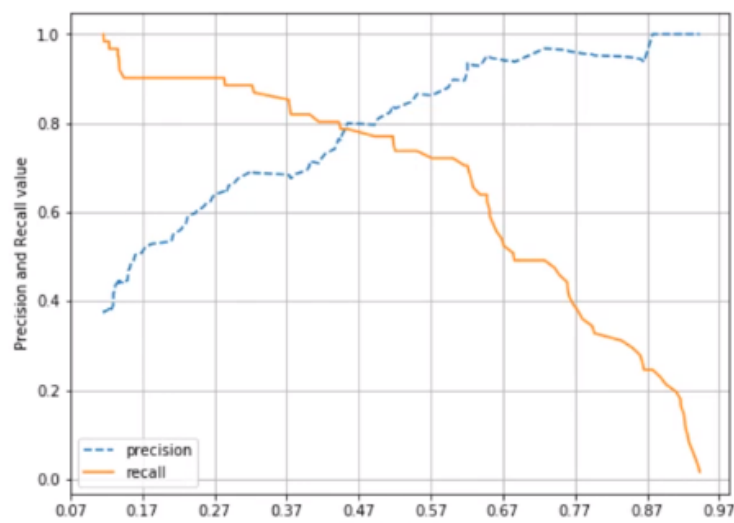
정밀도가 상대적으로 더 중요한 지표인 경우는 실제 Negative 음성인 데이터 예측을 Positive 양성으로 잘못 판단하게 되면 업무상 큰 영향이 발생하는 경우

→ 스팸 메일

*정밀도/재현율 Trade-off

분류하는 업무의 특성상 정밀도 또는 재현율이 특별히 강조돼야 할 경우 분류의 결정 임계값 (Threshold)을 조정해 정밀도 또는 재현율의 수치를 높일 수 있다.

하지만 정밀도와 재현율은 상호 보완적인 평가 지표이기 때문에 어느 한쪽을 강제로 높이면 다른 하나의 수치는 떨어지기 쉽다. 이를 정밀도/재현율 Trade-off라고 부른다.

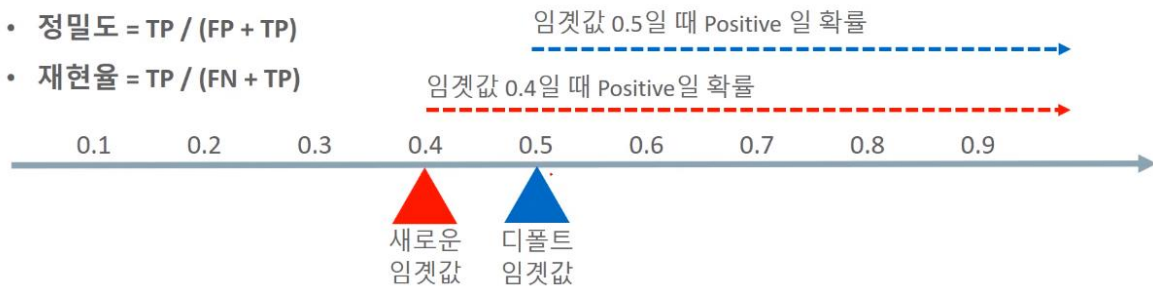


*분류 결정 임계값에 따른 Positive 예측 확률 변화

분류 모델(Classifier)이 양성인지 음성인지 결정할 때 확률로 결정한다.

- $\text{정밀도} = TP / (FP + TP)$

- $\text{재현율} = TP / (FN + TP)$



★분류 결정 임계값이 낮아질수록 Positive로 예측할 확률이 높아진다.

FN(예측을 Negative로 했는데 실제로는 Positive인 것)의 횟수 자체가 작아지기 때문에(Negative로 예측을 하지 않음) 재현율이 증가하게 된다.

사이킷런 Estimator 객체의 `predict_proba()` 메소드는 분류 결정 예측 확률을 반환한다.

이를 이용하면 임의로 분류 결정 임계값을 조정하면서 예측 확률을 변경할 수 있다.

*분류 결정 임계값에 따른 정밀도, 재현율 곡선

사이킷런은 `precision_recall_curve()` 함수를 통해 임계값에 따른 정밀도, 재현율의 변화값을 제공한다.

- F1 Score와 ROC-AUC 소개

- *F1 Score

F1 스코어(Score)는 정밀도와 재현율을 결합한 지표이다. F1 스코어는 정밀도와 재현율이 어느 한 쪽으로 치우치지 않는 수치를 나타낼 때 상대적으로 높은 값을 가진다.

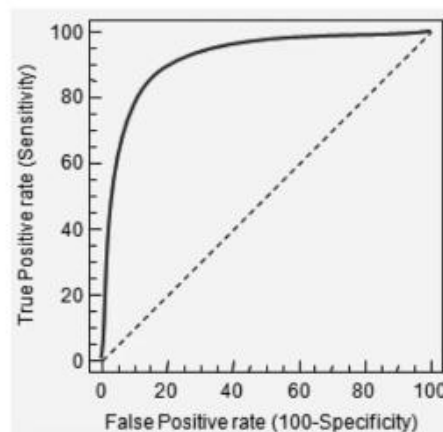
$$F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 * \frac{precision * recall}{precision + recall}$$

사이킷런은 F1 Score를 위해 `f1_score()` 함수를 제공한다.

- *ROC곡선과 AUC

ROC 곡선과 이에 기반한 AUC 스코어는 이진 분류의 예측 성능 측정에서 중요하게 사용되는 지표이다. 일반적으로 의학 분야에서 많이 사용되지만, 머신러닝의 이진 분류 모델의 예측 성능을 판단하는 중요한 평가 지표이기도 하다.

ROC 곡선: FPR(False Positive Rate)이 변할 때 TPR(True Positive Rate)이 어떻게 변하는지를 나타내는 곡선



분류의 성능 지표로 사용되는 것은 ROC 곡선 면적에 기반한 AUC 값으로 결정한다. AUC(Area Under Curve) 값은 ROC 곡선 밑의 면적을 구한 것으로 일반적으로 1에 가까울수록 좋은 수치이다.

- *사이킷런 ROC 곡선 및 AUC 스코어

사이킷런은 임계값에 따른 ROC 곡선 데이터를 `roc_curve()`로, AUC 스코어를 `roc_auc_score()` 함수로 제공

• 평가 Summary

- ① 이진 분류에서 정밀도, 재현율, F1 스코어, AUC 스코어가 주로 성능 평가 지표가 활용된다.
- ② 오차행렬은 실제 클래스 값과 예측 클래스 값의 True, False에 따라 TN, FP, FN, TP로 매핑 되는 4분면 행렬을 제공한다.
- ③ 정밀도와 재현율은 Positive데이터 세트의 예측 성능에 좀 더 초점을 맞춘 지표이며, 분류 결정 임계값을 조정해 정밀도 또는 재현율은 수치를 높이거나 낮출 수 있다.
- ④ F1 스코어는 정밀도와 재현율이 어느 한쪽으로 치우치지 않을 때 좋은 값을 가진다.
- ⑤ AUC 스코어는 ROC 곡선 밑의 면적을 구한 것으로 1에 가까울수록 좋은 수치이다.