03
평가

평가(evaluation)

Machine learning 은 특징 추출, 모델 학습, 예측, 그리고 평가의 프로세스로 구성됩니다. Machine learning 모델은 여러 가지 방법으로 예측 성능을 평가할 수 있습니다. 성능 평가 방법은 모델의 과업(task)가 분류인지, 또는 회귀인지에 따라 달라지게 됩니다. 분류의 경우 간단하게 생각하면 전체 데이터 중 모델이 맞춘 데이터의 개수를 의미하는 정확도(accuracy)를 가장 좋은 평가 지표로 생각할 수 있습니다. 정확도는 클래스마다 데이터의 개수가 비슷한 경우에는 좋은 평가 지표가 될 수 있지만, 데이터의 클래스 분포가 불균형한 경우에는 좋은 지표로 보기 어렵습니다. 예를 들어 100 개의 데이터를 평가할 때 99 개의 데이터가 A 클래스, 1 개의 데이터가 B 클래스에 속하는 경우 모든 데이터를 A 클래스로 간주하는 단순한 모델의 정확도는 99%로 매우 높아지게 됩니다. 이런 문제 때문에 조금 더 균형 있는 모델의 평가 지표가 필요합니다. 본 장에서는 정확도를 포함해 machine learning 모델의 다양한 평가 지표에 대해서 살펴볼 예정입니다.

정확도(accuracy)

정확도는 가장 기본적이면서 일반적으로 널리 사용되는 지표로, 실제 데이터 중에서 잘 예측된 데이터의 개수를 의미합니다.

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

```python
from sklearn.metrics import accuracy_score

y_test = [1, 1, 1, 1, 1, 0, 0, 0, 0, 0]
y_pred = [1, 0, 0, 0, 1, 0, 0, 0, 0, 1]
acc = accuracy_score(y_test, y_pred)
print(acc)
```

혼동행렬(confusion matrix)

분류 모델에서 성능 지표로 활용 할 수 있는 혼동행렬은 어떤 클래스의 데이터가 어떻게 잘못 예측되고 있는지를 함께 보여줄 수 있는 지표입니다. 혼동행렬은 이진 분류로 가정할 때 아래와 같은 4 분면 행렬에서 실제(actual) 클래스와 예측(predicted) 클래스 값이 어떤 유형으로 분류되는지 확인할 수 있습니다.

		Predicted class	
		Negative	Positive
Actual class	Negative	True negative	False positive
	Positive	False negative	True positive

```python
from sklearn.metrics import confusion_matrix

y_test = [1, 1, 1, 1, 1, 0, 0, 0, 0, 0]
y_pred = [1, 0, 0, 0, 1, 0, 0, 0, 0, 1]
cm = confusion_matrix(y_test, y_pred)
print(cm)
```

일반적으로 데이터 클래스 분포가 불균형한 이진 분류 모델에서는, 많은 데이터 중에서 중점적으로 찾아야 하는 적은 수의 클래스를 positive 로 설정해 1 의 값을 부여하고, 그렇지

않은 클래스에 대해 negative 로 설정해 0 의 값을 부여하는 경우가 많습니다. 이제 다시 TP, TN, FP, FN 을 이용해 정확도를 표현해 보면 다음과 같습니다.

$$정확도(Accuracy) = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

정밀도와 재현율(Precision and Recall)

정밀도와 재현율은 positive 데이터 세트의 예측 성능에 초점을 맞춘 평가 지표입니다. 수식은 아래와 같습니다.

$$정밀도(Precision) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$재현율(Recall) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

정밀도는 예측을 positive 로 한 데이터 중 실제로 positive 인 데이터의 비율입니다. Positive 예측 성능을 더욱 정밀하게 측정하기 위한 평가 지표입니다. 재현율은 실제 클래스가 positive 인 데이터 중 예측도 positive 로 맞았던 데이터의 비율입니다. 두 지표 모두 TP 를 분자로 하고 있지만, 분모가 positive 로 예측한 데이터(정밀도) 또는 실제 클래스가 positive 인 데이터(재현율)에 따라 달라집니다.

```
from sklearn.metrics import precision_score, recall_score

y_test = [1, 1, 1, 1, 1, 0, 0, 0, 0, 0]
y_pred = [1, 0, 0, 0, 1, 0, 0, 0, 0, 1]
ps = precision_score(y_test, y_pred)
rs = recall_score(y_test, y_pred)
print(ps)
print(rs)
```

정밀도와 재현율은 한 값이 높아지면 다른 한 값이 낮아지는 trade-off 관계에 있습니다. 분류하려는 업무의 특성에 맞게 분류 임계 값(threshold)를 조정해 정밀도 또는 재현율의 수치를 높일 수 있습니다.

X	1	2	3	4	5	6	7	8	9	10
Y	P	P	N	P	P	N	P	N	N	N

- Threshold = 1: precision = 1/1 (100%), recall = 1/5 (20%)
- Threshold = 4: precision = 3/4 (75%), recall = 3/5 (60%)
- Threshold = 7: precision = 5/7 (71%), recall = 5/5 (100%)
- Threshold = 10: precision = 5/10 (50%), recall = 5/5 (100%)

F1 score

F1 score는 정밀도와 재현율을 결합한 지표입니다. 정밀도와 재현율이 어느 한 방향으로 치우치지 않고 균형 있게 높은 값을 가질수록 높은 값을 가지게 됩니다(정밀도와 재현율의 조화평균).

$$F1\ score = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

```
from sklearn.metrics import f1_score
y_test = [1, 1, 1, 1, 1, 0, 0, 0, 0, 0]
y_pred = [1, 0, 0, 0, 1, 0, 0, 0, 0, 1]
ps = f1_score(y_test, y_pred)
print(ps)
```

민감도와 특이도(Sensitivity and specificity)

민감도와 특이도는 원래 의학 분야에서 많이 사용되던 용어로, 민감도의 경우 질병이 있는 사람을 얼마나 잘 찾아내는가를 나타내는 지표이며 특이도는 질병이 없는 정상인 사람을 얼마나 잘 찾아내는지를 나타내는 지표입니다.

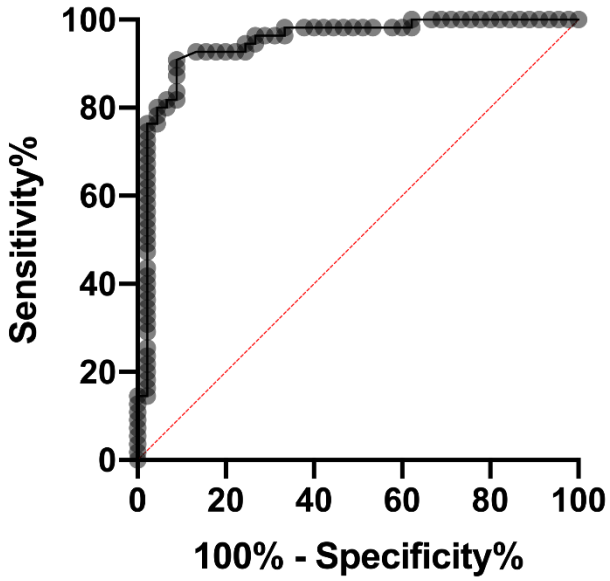
$$민감도(Sensitivity) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$특이도(Specificity) = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}$$

이때 민감도와 재현율의 수식이 같은데, 실제로 같은 값이며 서로 다른 분야에서 동일한 의미로 사용되는 지표입니다. 정밀도와 재현율은 모두 TP 를 분자로 가지고 있기 때문에 negative 를 얼마나 잘 찾아내는지는 확인하기 어렵습니다. 이때 활용 가능한 것이 특이도입니다. 민감도와 특이도 역시 trade-off 관계를 가집니다.

수신자 조작 특성(Receiver Operating Characteristic, ROC)

임계 값에 따라 민감도와 특이도는 달라지는데, 이 달라지는 모든 경우에 대한 대략적인 모델의 성능을 알고 싶을 때 사용하는 방법입니다.



ROC 곡선은 왼쪽 하단과 오른쪽 상단을 대각선으로 이은 빨간 선과 가까울수록 성능이 나쁜 것을 의미하며, 멀수록 성능이 좋을 것을 의미합니다. 이때 ROC 곡선의 아래 영역의 넓이(적분값)을 Area under curve (AUC)라고 부르며 이 값이 1 또는 0 에 가까울수록 좋은 분류 모델, 0.5 에 가까울수록 나쁜 분류 모델로 평가할 수 있습니다.