3장 분류 (1부)

주요내용

- **MNIST**
- **이진 분류기 훈련**
- **분류기 성능 측정**
- 다중 클래스 분류
- 에러 분석

3.1 MNIST

MNIST 데이터셋

- 미국 고등학생과 인구조사국 직원들이 손으로 쓴 70,000개의 숫자 이미지로 구성된 데이터셋
- 사용된 0부터 9까지의 숫자는 각각 28x28= 784크기의 픽셀로 구성된 이미지 데이터.
- 2차원 어레이가 아닌 길이가 784인 1차원 어레이로 제공
- 레이블: 총 70,000개의 사진 샘플이 표현하는 값

문제 정의

- 지도학습: 각 이미지가 담고 있는 숫자가 레이블(타깃)로 지정됨.
- 분류: 이미지 데이터를 분석하여 0부터 9까지의 숫자로 분류
- 이미지 그림을 총 10개의 클래스로 분류하는 **다중 클래스 분류**multiclass classification
- 배치 학습 활용

훈련셋과 데이터셋

- MNIST 데이터셋 이미 6:1 분류되어 있음
- 훈련 세트: 앞쪽 60,000개 이미지
- 테스트 세트: 나머지 10,000개의 이미지

3.2 이진 분류기 훈련

예제: 숫자 5-감지기

- 이미지 샘플이 숫자 5를 표현하는지 여부를 판단하는 이진 분류기
- 모든 레이블을 0 또는 1로 수정해야 함
    - 0: 숫자 5 이외의 수를 가리키는 이미지 레이블
    - 1: 숫자 5를 가리키는 이미지 레이블

SGD 분류기 활용

- 확률적 경사 하강법stochastic gradient descent 분류기
- 한 번에 하나씩 훈련 샘플 처리 후 파라미터 조정
- 매우 큰 데이터셋 처리에 효율적이며 온라인 학습에도 적합

- 훈련: `fit()` 메서드 호출

```
from sklearn.linear_model import SGDClassifier

sgd_clf = SGDClassifier(max_iter=1000, tol=1e-3,
random_state=42)
sgd_clf.fit(X_train, y_train_5)
```

3.3 분류기 성능 측정

성능 측정 기준

- 정확도
- 정밀도/재현율
- ROC 곡선의 AUC

교차 검증 활용 정확도 측정

- 숫자 5를 표현하는 이미지를 정확하게 예측한 비율.
- `cross_val_score` 모델의 `scoring="accuracy"` 키워드 인자 지정

```
from sklearn.model_selection import cross_val_score

cross_val_score(sgd_clf, X_train, y_train_5, cv=3, scoring="accuracy")
```

95%의 정확도를 갖는 분류기 이해

- 교차 검증 결과: 95% 이상의 정확도
- 하지만 무조건 '5 아님'이라고 찍는 분류기도 90%의 정확도를 보임.
- 훈련 세트의 샘플이 불균형적으로 구성되었다면, 정확도를 분류기의 성능 측정 기준으로 사용하는 것은 피해야 함

오차 행렬, 정밀도, 재현율

- 오차 행렬을 이용하여 분류기의 또다른 성능 측정 기준인 정밀도와 재현율 설명

오차행렬

- **오차 행렬**confusion matrix: 클래스별 예측 결과의 참/거짓을 정리한 행렬

- 숫자-5 감지기에 대한 오차 행렬

```
array([[53892,   687],
       [ 1891,  3530]])
```

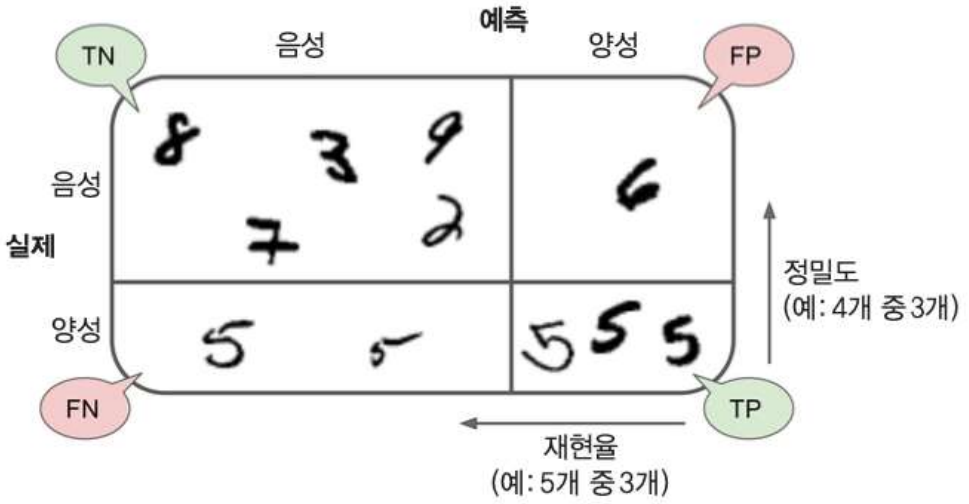
오차 행렬 해석

- 이진 분류기의 오차 행렬 내용
    - TN(참 음성): 음성을 음성으로 잘 예측한 경우
    - FP(거짓 양성): 음성을 양성으로 잘못 예측한 경우
    - FN(거짓 음성): 양성을 음성으로 잘못 예측한 경우
    - TP(참 양성): 양성을 양성으로 잘 예측한 경우

예제

- 아래 그림에 대한 오차 행렬

```
array([[5, 1],
       [2, 3]])
```



정밀도 precision

- 양성 예측의 정확도

- 예제: 숫자 5라고 예측된 값들 중에서 진짜로 5인 숫자들의 비율

$$\text{정밀도} = \frac{TP}{TP + FP} = \frac{3530}{3530 + 687} = 0.837$$

재현율recall

- 양성 샘플에 대한 정확도, 즉, 분류기가 정확하게 감지한 양성 샘플의 비율

- 재현율을 **민감도**(sensitivity) 또는 **참 양성 비율**(true positive rate)로도 부름

$$\text{재현율} = \frac{TP}{TP + FN} = \frac{3530}{3530 + 1891} = 0.651$$

정밀도 vs. 재현율

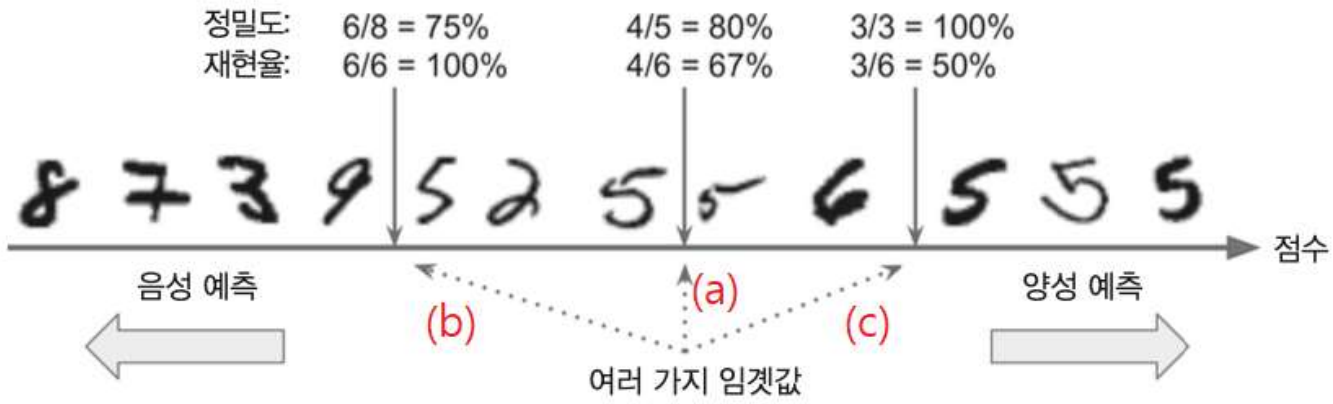
- 목적에 따라 정밀도와 재현율의 중요도가 다름
- 재현율이 보다 중요한 경우: 암 진단 기준
    - 정밀도: 암이라고 진단했는데 진짜 암인 경우의 비율
    - 재현율: 암이 실제로 있는데 암이라고 진단한 경우의 비율
- 정밀도가 보다 중요한 경우: 아동용 동영상 선택 기준
    - 정밀도: 아동용으로 판단된 동영상 중에서 실제로 아동용인 동영상의 비율
    - 재현율: 아동용 동영상 중에서 아동용 동영상이라고 판단된 동영상의 비율

정밀도/재현율 트레이드오프

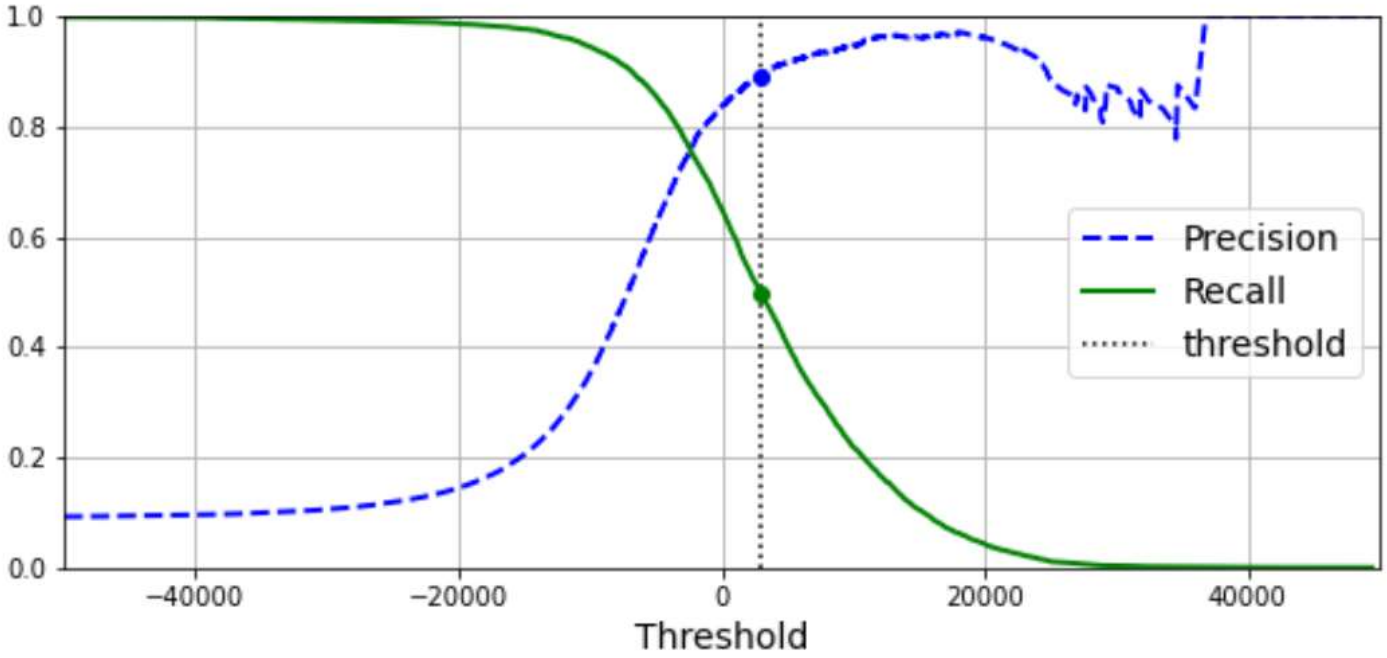
- 정밀도와 재현율은 상호 반비례 관계임.
- 정밀도와 재현율 사이의 적절한 비율을 유지하는 분류기를 찾아야 함.
- 적절한 **결정 임곗값**을 지정해야 함.

결정 함수와 결정 임곗값

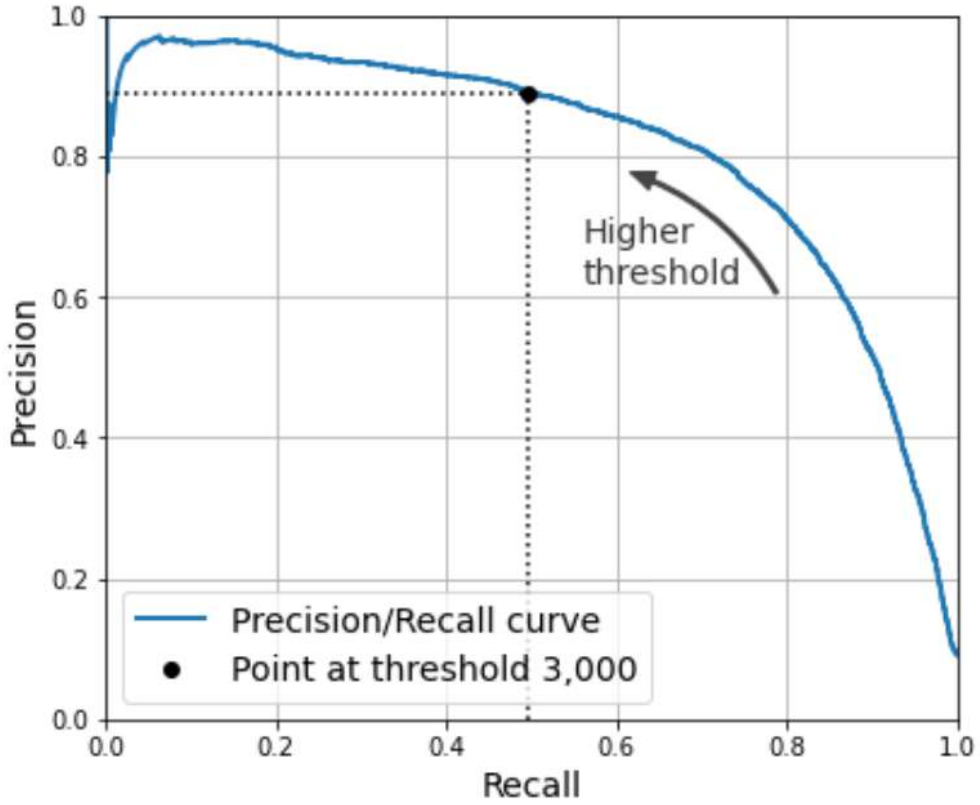
- **결정 함수**decision function: 각 훈련 샘플에 대한 점수를 계산하는 함수
- **결정 임계값**decision threshold: 결정 함수가 양성 클래스 또는 음성 클래스로 분류하는 데에 사용하는 기준값
- 결정 임곗값이 클 수록 정밀도는 올라가지만 재현율은 떨어짐.



임곗값, 재현율, 정밀도


1.0
0.8
0.6
0.4
0.2
0.0
−40000
−20000
0
20000
40000
Threshold
Precision
Recall
threshold

재현율 vs. 정밀도

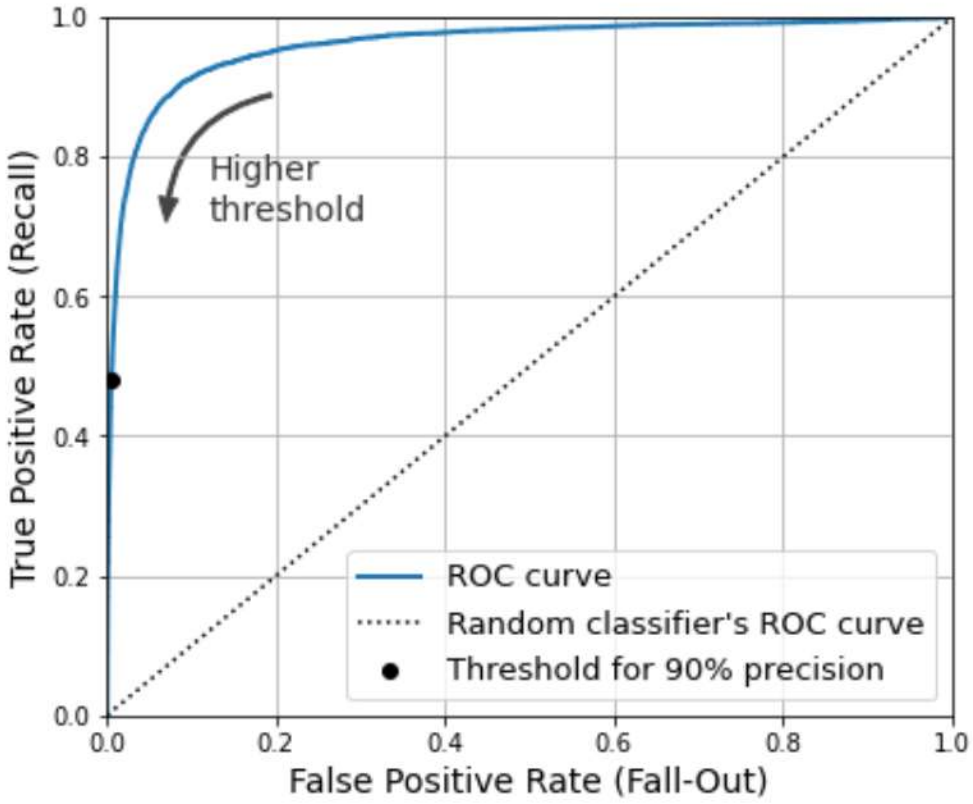

1.0
0.8
0.6
0.4
0.2
0.0
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall
Higher
threshold
Precision/Recall curve
Point at threshold 3,000

ROC 곡선의 AUC

- **수신기 조작 특성**(receiver operating characteristic, ROC) 곡선을 활용하여 이진 분류기의 성능 측정 가능
- **ROC 곡선**: **거짓 양성 비율**false positive rate(FPR)에 대한 **참 양성 비율**true positive rate(TPR)의 관계를 나타내는 곡선
- 참 양성 비율: 재현율
- 거짓 양성 비율: 원래 음성인 샘플 중에서 양성이라고 잘못 분류된 샘플들의 비율. 예를 들어, 5가 아닌 숫자중에서 5로 잘못 예측된 숫자의 비율

$$\mathrm{FPR} = \frac{FP}{FP + TN}$$

참 양성 비율(TPR) vs. 거짓 양성 비율(FPR)



AUC와 분류기 성능

- 재현율(TPR)과 거짓 양성 비율(FPR) 사이에도 서로 상쇄하는 기능이 있다는 것을 확인 가능
- 즉, 재현율(TPR)을 높이고자 하면 거짓 양성 비율(FPR)도 함께 증가
- 좋은 분류기는 재현율은 높으면서 거짓 양성 비율은 최대한 낮게 유지해야함
- ROC 곡선이 y축에 최대한 근접하는 결과가 나오도록 해야함.
- **AUC**(ROC 곡선 아래의 면적)가 1에 가까울 수록 성능이 좋은 분류기로 평가됨.

SGD와 랜덤 포레스트의 AUC 비교