

3장 분류

감사의 글

자료를 공개한 저자 오렐리앙 제롱과 강의자료를 지원한 한빛아카데미에게 진심어린 감사를 전합니다.

주요내용

- MNIST
- 이진 분류기 훈련
- 분류기 성능 측정
- 다중 클래스 분류기 훈련

3.1 MNIST

MNIST 데이터셋

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

5	0	4	1	9	2	1	3	1	4
3	5	3	6	1	7	2	8	6	9
4	0	9	1	1	2	4	3	2	7
3	8	6	9	0	5	6	0	7	6
1	8	7	9	3	9	8	5	9	3
3	0	7	4	9	8	0	9	4	1
4	4	6	0	4	5	6	1	0	0
1	7	1	6	3	0	2	1	1	7
8	0	2	6	7	8	3	9	0	4
6	7	4	6	8	0	7	8	3	1

문제 정의

- 지도학습: 각 이미지가 담고 있는 숫자가 레이블로 지정됨.
- 분류: 이미지 데이터를 분석하여 0부터 9까지의 숫자로 분류
 - 이미지 그림을 총 10개의 클래스로 분류하는 **다중 클래스 분류**(multiclass classification) **다항 분류**(multinomial classification)라고도 불림
- 배치 또는 온라인 학습: 둘 다 가능
 - 모델에 따라 처리 방법이 다름
 - 확률적 경사하강법(stochastic gradient descent, SGD): 배치와 온라인 학습 모두 지원
 - 랜덤 포레스트 분류기: 배치 학습

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 학습

- 확률적 경사 하강법(stochastic gradient descent) 분류기라고 불림.
- 한 번에 하나씩 훈련 샘플 처리 후 파라미터 조정
- 매우 큰 데이터셋 처리에 효율적이며 온라인 학습에도 적합함.

3.3 성능 측정

성능 측정 세가지 방법

- 정확도 측정
- 정밀도/재현율 조율
- AUC 측정

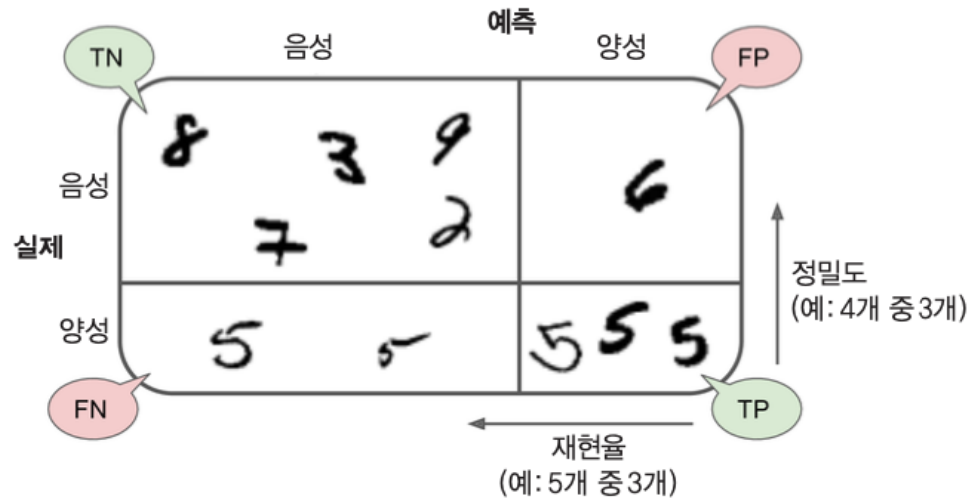
3.3.1 교차 검증을 사용한 정확도 측정

- 성능 측정 기준: 정확도, 즉 정확히 예측한 비율

3.3.2 오차 행렬

- 교차 검증 결과가 95% 이상으로 매우 우수한 것으로 나옴.
 - 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 행렬을 조사하여 분류기의 성능을 다르게 평가할 수 있음

- **오차 행렬**: 클래스별 예측 결과를 정리한 행렬
- 예제: '숫자 5-감지기'에 대한 오차 행렬은 (2, 2) 모양의 2차원 (넘파이) 어레이로 생성됨.
 - 레이블의 값이 0과 1 두 개의 값으로 구성되기 때문



3.3.3 정밀도와 재현율

정밀도(precision)

- 6만개의 훈련 세트에 대한 이진 분류 결과의 오차 행렬

53057		1522

1325		4096

- 양성 예측의 정확도
- 여기서는 숫자 5라고 예측된 값들 중에서 진짜로 5인 숫자들의 비율

$$\text{precision} = \frac{TP}{TP + FP} = \frac{4096}{4096 + 1522} = 0.729$$

재현율(recall)

- 정밀도 하나만으로 분류기의 성능을 평가할 수는 없음
 - 숫자 5를 가리키는 이미지 중에 숫자 5라고 판명한 비율인 **재현율**을 고려하지 않기 때문
- 양성 샘플에 대한 정확도, 즉, 분류기가 정확하게 감지한 양성 샘플의 비율
- 재현율을 **민감도(sensitivity)** 또는 **참 양성 비율(true positive rate)**로도 부름

$$\text{recall} = \frac{TP}{TP + FN} = \frac{4096}{4096 + 1325} = 0.756$$

F₁ 점수

- 정밀도와 재현율의 조화 평균인 F₁ 점수를 이용하여 분류기의 성능을 평가하기도 함.

$$F_1 = \frac{2}{\frac{1}{\text{정밀도}} + \frac{1}{\text{재현율}}}$$

- F₁ 점수가 높을 수록 분류기의 성능을 좋게 평가하지만 경우에 따라 재현율과 정밀도 둘 중의 하나에 높은 가중치를 두어야 할 때가 있음.
 - 앞서 정의된 F₁ 점수는 재현율과 정밀도의 중요도가 동일하다고 가정하였음.

정밀도 대 재현율

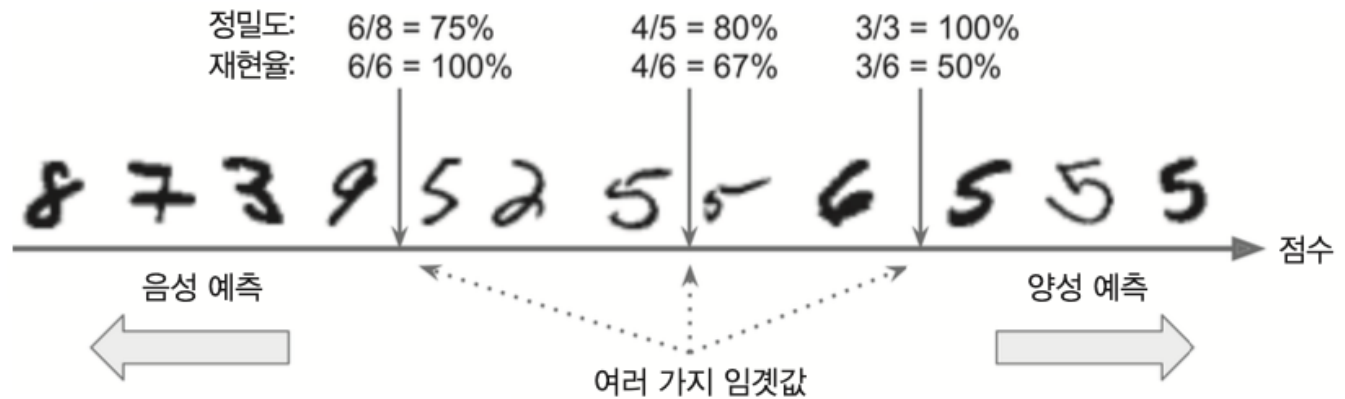
- 모델 사용의 목적에 따라 정밀도와 재현율의 중요도가 다를 수 있음.
- 재현율이 보다 중요한 경우: 암 진단 기준
 - 정밀도: 양성으로 진단된 경우 중에 실제로도 양성인 경우의 비율
 - 재현율: 실제로 양성인 경우 중에서 양성으로 진단하는 경우의 비율
- 정밀도가 보다 중요한 경우: 아이에게 보여줄 안전한 동영상 선택 기준
 - 정밀도: 안전하다고 판단된 동영상 중에서 실제로도 안전한 동영상의 비율
 - 재현율: 실제로 좋은 동영상 중에서 좋은 동영상이라고 판단되는 동영상 비율

3.3.4 정밀도/재현율 트레이드오프

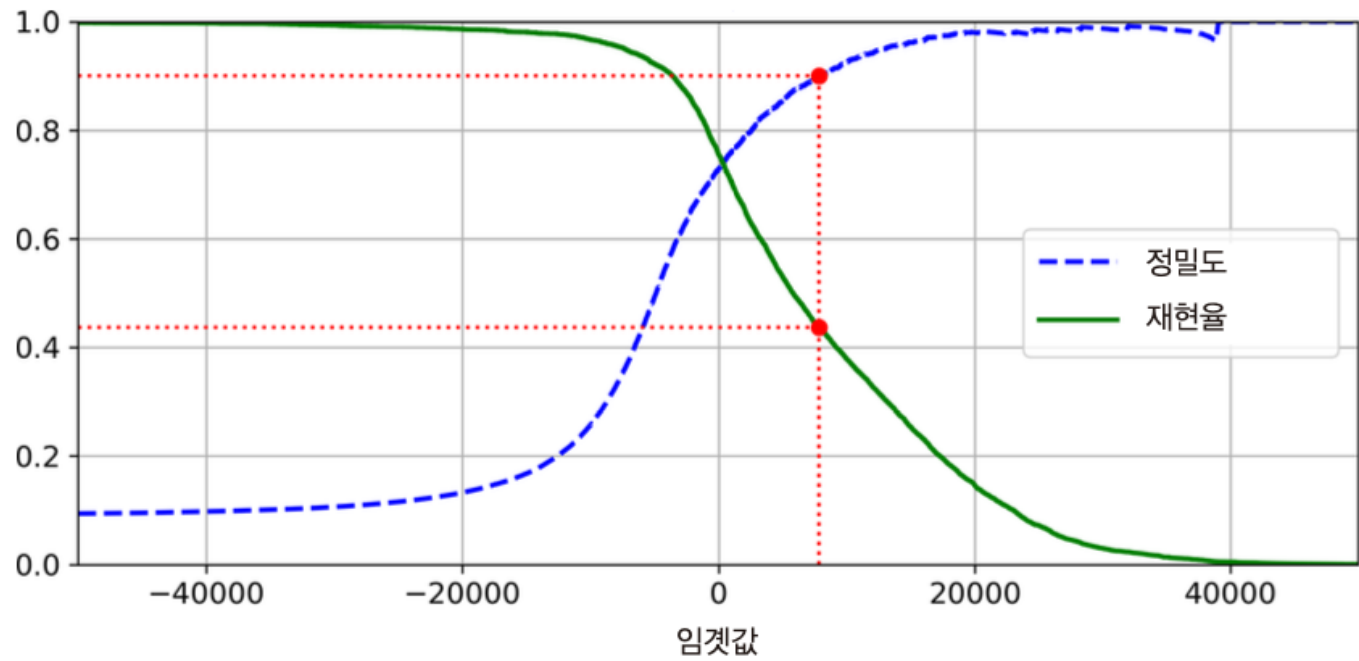
- 정밀도와 재현율은 상호 반비례 관계임.
- 정밀도와 재현율 사이의 적절한 비율을 유지하는 분류기를 찾아야 함.
- 적절한 **결정 임계값**을 지정해야 함.

결정 함수와 결정 임계값

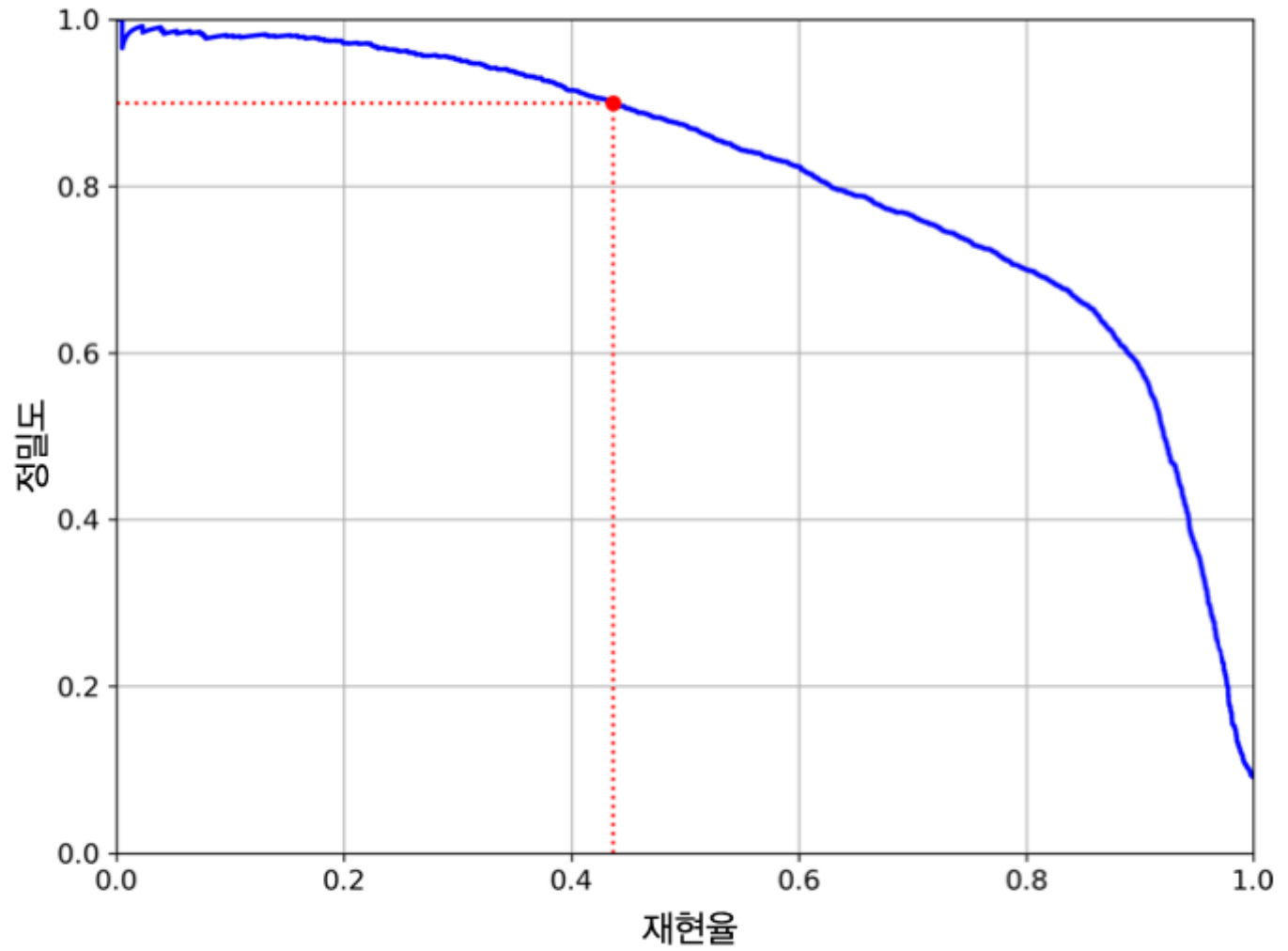
- 결정 함수(decision function): 분류기가 각 샘플의 점수를 계산할 때 사용
- 결정 임계값(decision threshold): 결정 함수가 양성 클래스 또는 음성 클래스로 분류하는 데에 사용하는 기준값
- 임계값이 커질 수록 정밀도는 올라가지만 재현율은 떨어짐.



임꺽값, 재현율, 정밀도



재현율 대 정밀도

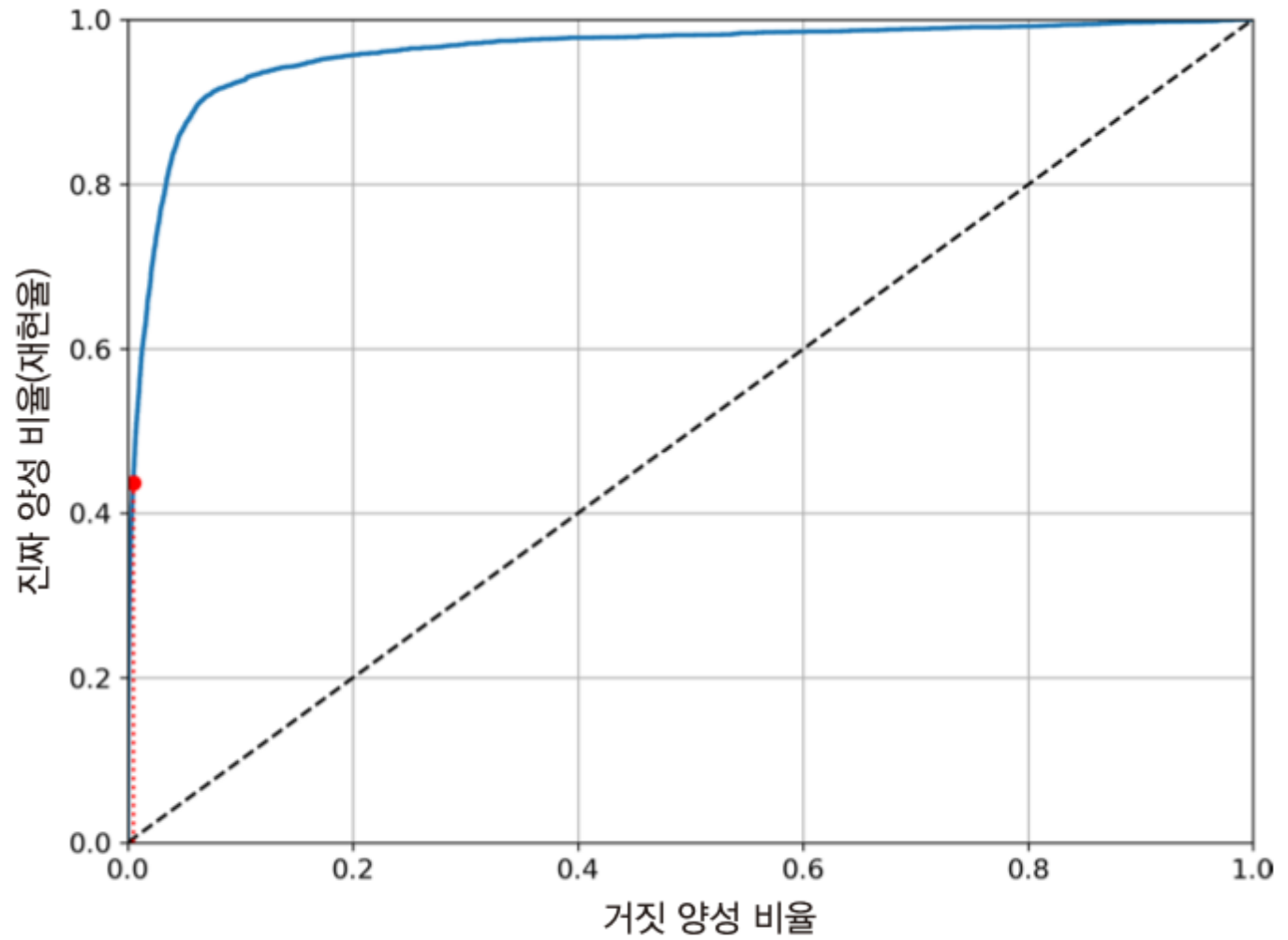


3.3.5 ROC 곡선과 AUC 점수

- 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
 - 결정 임계값에 따른 두 비율의 변화를 곡선으로 보여줌.
 - 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

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

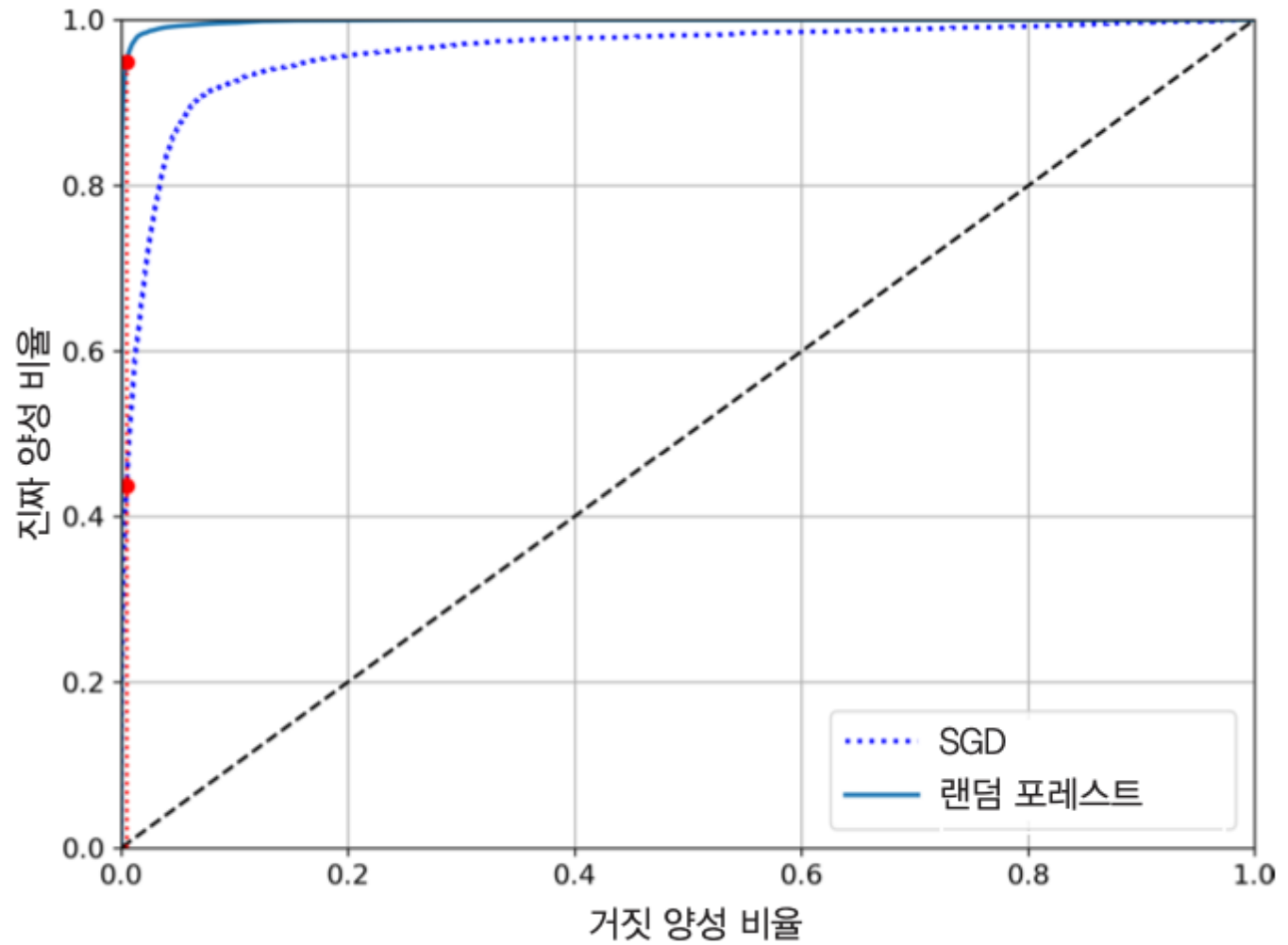
TPR 대 FPR



AUC와 분류기 성능

- 재현율(TPR)과 거짓 양성 비율(FPR) 사이에도 서로 상쇄하는 기능이 있다는 것을 확인 가능
 - 재현율(TPR)을 높이려고 하면 거짓 양성 비율(FPR)도 함께 증가
- 따라서 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



3.4 다중 클래스 분류

다중 클래스 분류기(multiclass classifier)

- 세 개 이상의 클래스로 샘플을 분류하는 예측기
- 다항 분류기(multinomial classifier)라고도 부름
- 예를 들어, 손글씨 숫자 분류의 경우 0부터 9까지 10개의 클래스로 분류해야 함

다중 클래스 분류 지원 분류기

- SGD 분류기
- 랜덤 포레스트 분류기
- 나이브 베이즈(naive Bayes) 분류기

이진 분류만 지원하는 분류기

- 로지스틱 회귀
- 서포트 벡터 머신
 - 사이킷런의 `SVC()` 모델은 다중 클래스 분류도 특별한 기법을 적용하여 지원함.

이진 분류기 활용

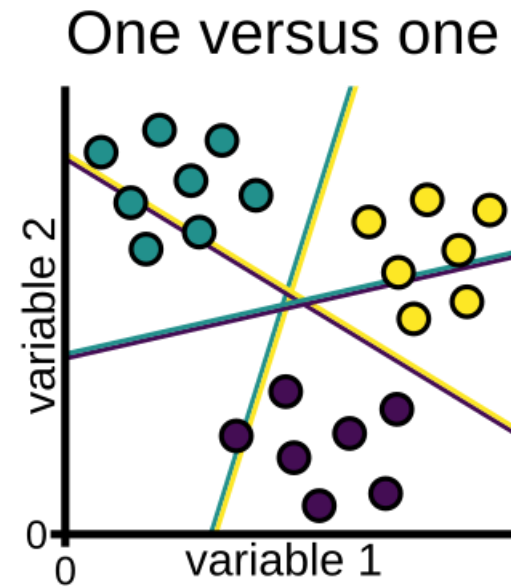
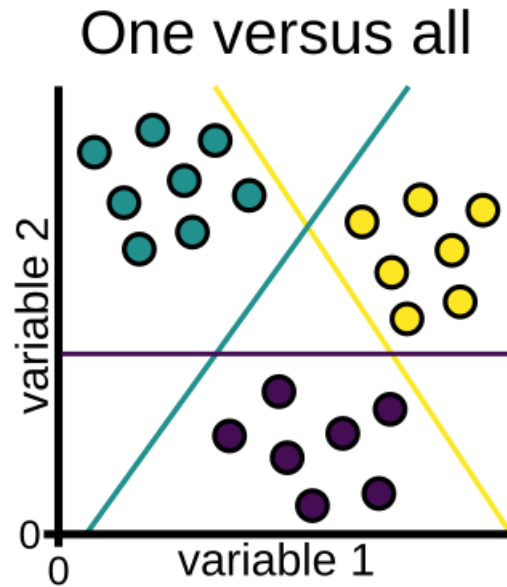
- 이진 분류기를 활용하여 다중 클래스 분류 가능
 - 일대다(OvR 또는 OvA)
 - 일대일(OvO)

- 일대다 방식 활용 예제
 - 숫자 5 예측하기에서 사용했던 이진 분류 방식을 동일하게 모든 숫자에 대해서 실행
 - 각 샘플에 대해 총 10번 각기 다른 이진 분류기를 실행
 - 이후 각 분류기의 결정 점수 중에서 가장 높은 점수를 받은 클래스를 해당 샘플의 클래스로 선택

- 일대일 방식 활용 예제

- 조합 가능한 모든 클래스 일대일 분류 방식을 진행하여 가장 많은 결투(duell)를 이긴 숫자를 선택
- MNIST의 경우, 0과 1 구별, 0과 2 구별, ..., 1과 2 구별, 1과 3 구별, ..., 8과 9 구별 등 $(9+8+\dots+1 = 45)$ 개의 결투를 판별하는 45개의 분류기 활용. 단, 각 결투에 해당되는 데이터 샘플만 훈련에 사용됨.
- 각각의 훈련 샘플에 대해 가장 많은 결투를 이긴 숫자의 클래스를 예측값으로 사용함. 예를 들어, 숫자 1이 9번의 결투를 모두 이기면 숫자 1을 예측값으로 지정함.

일대다 방식 vs. 일대일 방식



<그림 출처: [SVM with the mlr package](#)>

서포트 벡터 머신 활용

- 일대다 방식을 사용하여 다중 클래스 분류 진행
- 예제: 붓꽃(Iris) 데이터 분류. Orange3 활용 참조