분류

3.1 MNIST

- 손글씨 식별 데이터 셋

- 트레인 셋과 테스트 셋 만들기

3.2 이진분류기 훈련

- 5 식별하기

- 5는 전체 중 약 10%에 해당(모두 false라고 하더라도 90%정확)

- 싸이킷런의 SGD분류기(stochastic gradient descent classifier)

- 트레인 셋을 이용해서 5모델 훈련

3.3 성능측정

3.3.1 교차검증을 사용한 정확도 측정

- cross\_val\_score() 함수로 폴드가 3개인 교차검증으로 SGDClassifier 모델 정확도 검증: 트레인 셋을 세개(K개)로 나누고 각 폴드를 예측하기 위해 나머지 두개로 훈련시킨 모델 사용

- 정확도는 아주 좋은 성능 측정 지표가 아님, 특히 불균형한 데이터셋일 때

3.3.2 오차 행렬

- cross\_val\_pred() 함수는 K겹 교차 검증 수행 중 각 테스트 폴드에서 얻은 예측 반환, 깨끗한 예측

- confusion\_matrix는 (TP, FP, FN, TN)의 2x2 matrix를 return: 오차행렬

- 오차행렬보다 요약된 지표

정밀도(precision) = TP/(TP+FP)

재현율(recall) = TP/(TP+FN)

3.3.3 정밀도와 재현율

- precision\_score, recall\_score

- 합해서 조화평균을 보여주는 F\_1 score = 2/((1/정밀도)+(1/재현율))

- 정밀도와 재현율이 비슷하면 F\_1 score 높다. 정밀도와 재현율의 중요성에 따라 트레이드 오프를 고려

3.3.4 정밀도/재현율 트레이드 오프

- 임계값

- 분류기의 predict() 대신 decision\_function()을 호출하면 각 샘플의 점수 얻을 수 있음

- 적절한 임계값: 모든 임계값에 대해 정밀도와 재현율 구하여 보기

3.3.5 ROC곡선

- receiver operating characteristic(수신기 조작 특성) 곡선: 거짓양성비율(false positive)에 대한 진짜 양성 비율(true positive)의 곡선

- TPR 높을 수록 FPR도 올라감: 완전 랜덤 분류기는 45도 라인, 좋은 분류기는 최대한 여기서 멀어져야 함