**1. MNIST**

**MNIST 데이터셋**

: 0부터 9까지의 손글씨 숫자 이미지. 28\*28=784크기의 픽셀로 구성된 이미지 데이터이며 2차원 array가 아닌 길이가 784인 1차원 array로 제공

**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

배치 또는 온라인 학습 : 확률적 경사하강법(SGD)은 배치와 온라인 학습 모두 지원/랜덤포레스트 분류기는 배치학습만

**훈련 셋과 데이터 셋 나누기**

훈련 세트: 앞쪽 60,000개 이미지

테스트 세트: 나머지 10,000개 이미지

-> MNIST 데이터셋은 이미 6:1로 분류되어 있음

**2. 이진 분류기 훈련(5-감지기)**

레이블을 0 또는 1로 수정

-> 기존에 0부터 9까지의 숫자 대신 0,1 두 값으로만 구성된 새로운 타깃을 생성해서 학습에 사용(5 - 1, 5이외의 수 - 0)

SGDClassifier: 확률적 경사하강법

- 매우 큰 데이터셋을 효율적으로 처리

- 한 번에 하나씩 훈련 샘플을 독립적으로 처리

**3. 성능 측정**

**1. 교차검증을 사용한 정확도 측정**

성능 측정 기준: 정확도

-> 정확도: 전체 샘플을 대상으로 정확하게 예측한 비율(숫자 5를 표현하는 이미지를 True로 예측한 비율)

-> 훈련세트의 샘플이 불균형적으로 구성되었다면, 정확도를 분류기의 성능 측정 기준으로 사용하는 것을 피해야 함

**2. 오차 행렬**

오차 행렬: 클래스 별 예측 결과를 정리한 행렬

오차 행렬의 행은 실제 클래스, 열은 예측된 클래스를 가리킴

텍스트, 폰트, 스크린샷, 영수증이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

**3. 정밀도와 재현율**

정밀도: 양성 예측의 정확도(숫자 5라고 예측된 값들 중에 진짜로 5인 숫자들의 비율)

재현율: 양성 샘플에 대한 정확도(분류기가 정확하게 감지한 양성 샘플의 비율). 민감도라고도 부름

F1점수: 정밀도와 재현율의 조화 평균

-> F1점수가 높을수록 분류기의 성능을 좋게 평가하지만, 경우에 따라서 조심할 필요가 있음(가중치를 다르게 두어야 할 때가 있음)

**4. 정밀도/재현율 트레이드오프**

결정 함수: 분류기가 각 샘플의 점수를 계산할 때 사용

결정 임계값: 결정 함수의 값이 이 값보다 같거나 크면 양성 클래스로 분류, 아니면 음성 클래스

임계값이 커질수록 정밀도는 올라가고, 재현율은 떨어짐.

**5. ROC 곡선**

결정 임계값에 따른 거짓 양성 비율(원래 음성인 샘플 중에서 양성이라고 잘못 분류된 샘플들의 비율, FPR)에 대한 참 양성 비율(재현율, TPR)의 관계를 나타낸 곡선.

AUC와 분류기 성능 - TPR과 FPR사이에도 서로 상쇄하는 기능이 있음(TPR을 높이고자 하면 FPR도 함께 증가

-> 좋은 분류기는 TPR은 높으면서 FPR은 최대한 낮게 유지해야 함.

ROC곡선이 y축에 최대한 근접하는 결과가 나오도록 해야함.

AUC: ROC 곡선 아래의 면적 -> 1에 가까울수록 성능이 좋은 분류가

**4. 다중 분류**

**다중 클래스 분류기(multiclass classifier)**

: 세 개 이상의 클래스로 샘플을 분류하는 예측기. 다항분류기라고도 부름

-> SGD 분류기, 랜덤포레스트 분류기, 나이브베이즈 분류기

**이진 분류기 활용**

이진 분류기를 활용하여 다중 클래스 분류 가능

- 일대다(OvR, OvA) : 숫자 5 예측하기에서 사용했던 이진 분류 방식을 동일하게 모든 숫자에 대해서 실행, 이후 각 분류기의 결정 점수 중에서 가장 높은 점수를 받은 클래스를 선택

-> 대부분의 이진 분류기는 일대다 전략 사용

- 일대일(OvO) : 조합 가능한 모든 일대일 분류 방식을 진행하여 가장 많은 결투를 이긴 숫자를 선택

**5. 에러 분석**

**오차 행렬 활용**

손글씨 클래스 분류 모델의 오차 행렬을 이미지로 표현 가능

대체로 잘 분류됨: 대각선이 밝음

5행은 좀 어두움: 숫자 5의 분류 정확도가 상대적으로 낮음

**오차율 이미지**

8행이 전반적으로 어두움: 8은 잘 분류되었다는 의미임