**1. MNIST**

데이터셋 구성: ‘0~9’ 손글씨 숫자 이미지 70,000개 (28x28 pixel, 784 길이의 1D array)

학습 유형: 지도학습. 다중 클래스(10개의 클래스) 분류

학습 방법: 배치 또는 온라인 학습 가능.

ex. SGD는 둘 다 지원, 랜덤 포레스트는 배치 학습만 가능

데이터 분할: 60,000개 훈련용, 10,000개 테스트용

**2. 이진 분류기 훈련**

“숫자 5 감지기. 5인지 아닌지를 분류” (Binary Classification)

타깃 값 변환: 5 → 1, 나머지 → 0

사용 모델: SGDClassifier. 큰 데이터셋에서 효율적으로 작동하며 샘플을 독립적으로 처리

**3. 성능 측정**

1) 교차검증

정확도 기준사용. 높은 정확도(예: 95%)라도 데이터 불균형 문제로 신뢰 어려움

2) 오차 행렬 (Confusion Matrix)

실제 클래스와 예측 클래스의 조합을 정리한 행렬

주요 용어:

TN, FP, FN, TP

3) 정밀도와 재현율

정밀도: 양성 예측의 정확도

재현율: 양성 샘플을 정확히 감지한 비율

F1 점수: 정밀도와 재현율의 조화 평균. 중요도 동일 가정

4) 정밀도-재현율 트레이드오프

결정 함수 / 임계값에 따라 정밀도와 재현율이 반비례

임계값 증가 → 정밀도 ↑, 재현율 ↓

5) ROC 곡선과 AUC

ROC 곡선: TPR(재현율) vs FPR(거짓 양성 비율)

AUC: 면적이 클수록 성능 우수

모델 비교: SGD vs 랜덤 포레스트의 AUC 비교 가능

**4. 다중 분류**

정의: 3개 이상의 클래스로 분류하는 다항 분류

지원 모델:

다중 분류 직접 지원: SGD, 랜덤 포레스트, 나이브 베이즈

이진 분류만 지원: 로지스틱 회귀, SVM

이진 분류기로 다중 분류

OvR (One-versus-Rest): 각 클래스마다 이진 분류기 수행, 최고 점수 클래스 선택

OvO (One-versus-One): 클래스 간 조합된 모든 결투를 통해 결정. 예: MNIST는 45개의 분류기 필요

사이킷런 구현: OneVsOneClassifier, OneVsRestClassifier

SGD: OvR/OvO 없이도 다중 분류 직접 지원

**5. 에러 분석**

오차 행렬 시각화: 대각선이 밝으면 잘 분류된 것

숫자 5는 상대적으로 정확도 낮음, 3과 5가 자주 혼동됨

원인: 선형 모델(SGD)의 한계. 픽셀 강도에만 의존하며, 위치나 회전에 민감

개선 방법: 이미지 중앙 정렬, 회전 제거 등의 전처리 필요

**6. 다중 레이블 분류**

정의: 하나의 샘플이 여러 클래스에 속할 수 있음

예제: 얼굴 인식(엘리스, 밥, 찰리), 숫자 분류(홀수 여부 + 7 이상 여부)

모델: KNeighborsClassifier지원

평가: F1 점수를 각 레이블별로 구해 평균. 레이블 가중치 조정 가능

**7. 다중 출력 분류**

정의: 다중 레이블 분류의 일반화. 각 레이블이 다중 클래스 값을 가짐

예제: 잡음 제거기

입력: 잡음 있는 이미지

출력: MNIST 스타일의 깨끗한 이미지

각 픽셀이 레이블 역할, 픽셀 강도(0~255)는 다중 클래스