

# 3장 분류 (2부)

# 주요 내용

- MNIST
- 이진 분류기 훈련
- 분류기 성능 측정
- 다중 클래스 분류
- 오류 분석
- 다중 레이블 분류와 다중 출력 분류

### 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(SVM)

## 이진 분류기 활용 클래스 분류

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

## 손글씨 분류: 일대다 방식 활용법

- 숫자-5 예측기와 동일한 방식으로 모든 숫자에 대해 이진 분류기 실행
- 각 샘플에 대해 총 10개의 이진 분류기 훈련
- 각 샘플에 대해 가장 높은 결정 점수를 주는 이진 분류기에 해당하는 클래스 선택

## 손글씨 분류: 일대일 방식 활용법

- 각 샘플에 대해 가능한 모든 조합의 일대일 대결 분류기 훈련 진행 후 가장 승률이 높은 숫자 선택
- MNIST의 경우, 0과 1 구별, 0과 2 구별, ..., 1과 2 구별, 1과 3 구별, ..., 8과 9 구별 등 총 45개의 분류기 활용.
- 각각의 분류기는 해당되는 샘플만 훈련에 사용. 예를 들어, 0과 1을 구별하는 분류기는 0과 1에 해당하는 샘플만으로 훈련.
- 각각의 훈련 샘플에 대해 가장 많은 결투를 이긴 숫자의 클래스를 예측값으로 사용함. 예를 들어, 어떤 샘플에 대해 숫자 1이 9번의 결투를 모두 이기면 숫자 1을 해당 샘플의 예측값으로 지정.

## 예제: 서포트 벡터 머신

- 훈련 세트의 크기에 민감하여 작은 훈련 세트에서 많은 분류기를 훈련시키는 쪽이 훨씬 빠름. 따라서 다중 클래스 분류에 일대일 전략을 사용함.
- 대부분의 이진 분류기는 일대다 전략 선호

## 일대일 또는 일대다 전략 선택

- 이진 분류기를 일대일 전략 또는 일대다 전략으로 지정해서 학습하도록 만들 수 있음.
- 사이킷런의 경우: `OneVsOneClassifier` 또는 `OneVsRestClassifier` 사용
- 예를 들어, SVC 모델을 일대다 전략으로 훈련시키려면 `OneVsRestClassifier` 활용

```
from sklearn.multiclass import OneVsRestClassifier
... ovr_clf = OneVsRestClassifier(SVC())
... ovr_clf.fit(X_train, y_train)
```

## 다중 클래스 지원 분류기

- `SGDCClassifier` 또는 `RandomForestClassifier` 는 다중 클래스 분류를 직접 지원함.
- 따라서 사이킷런의 OvR, OvO 등을 적용할 필요 없음

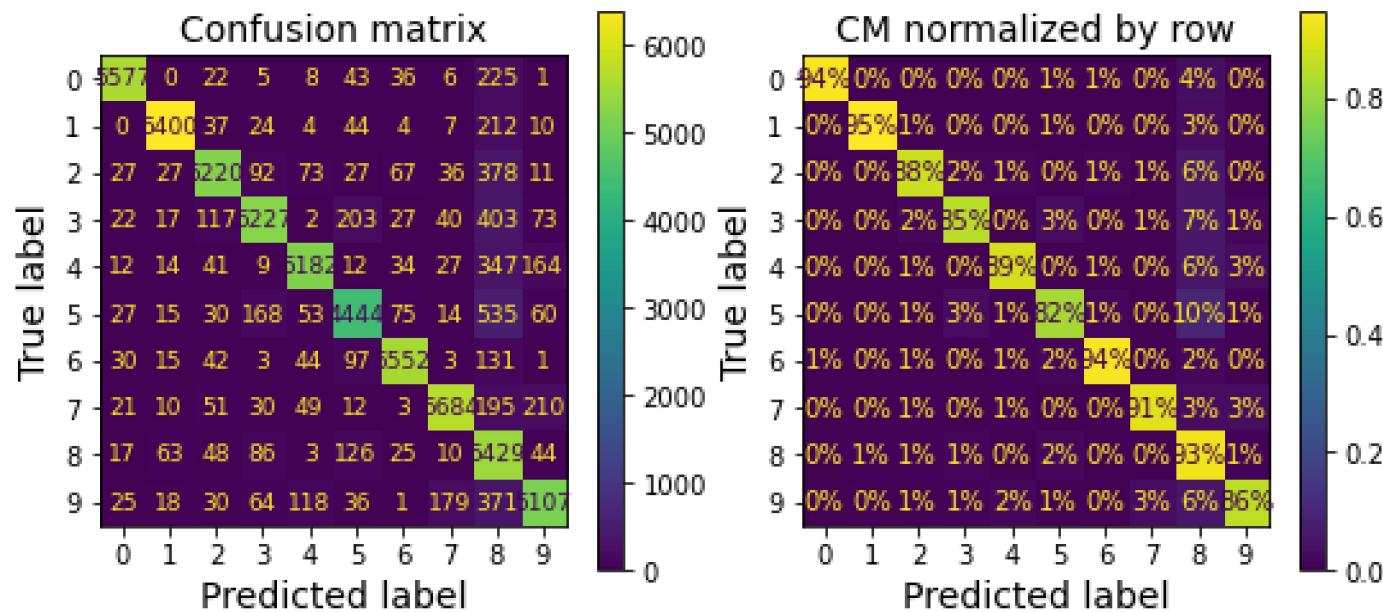
## 다중 클래스 분류기 성능 측정

- 다중 클래스 분류기의 성능 평가는 교차검증을 이용하여 정확도를 측정
- MNIST의 경우 0부터 9까지 숫자가 균형 있게 분포되어 있어서 데이터 불균형의 문제가 발생하지 않음.

### 3.5. 오류 분석

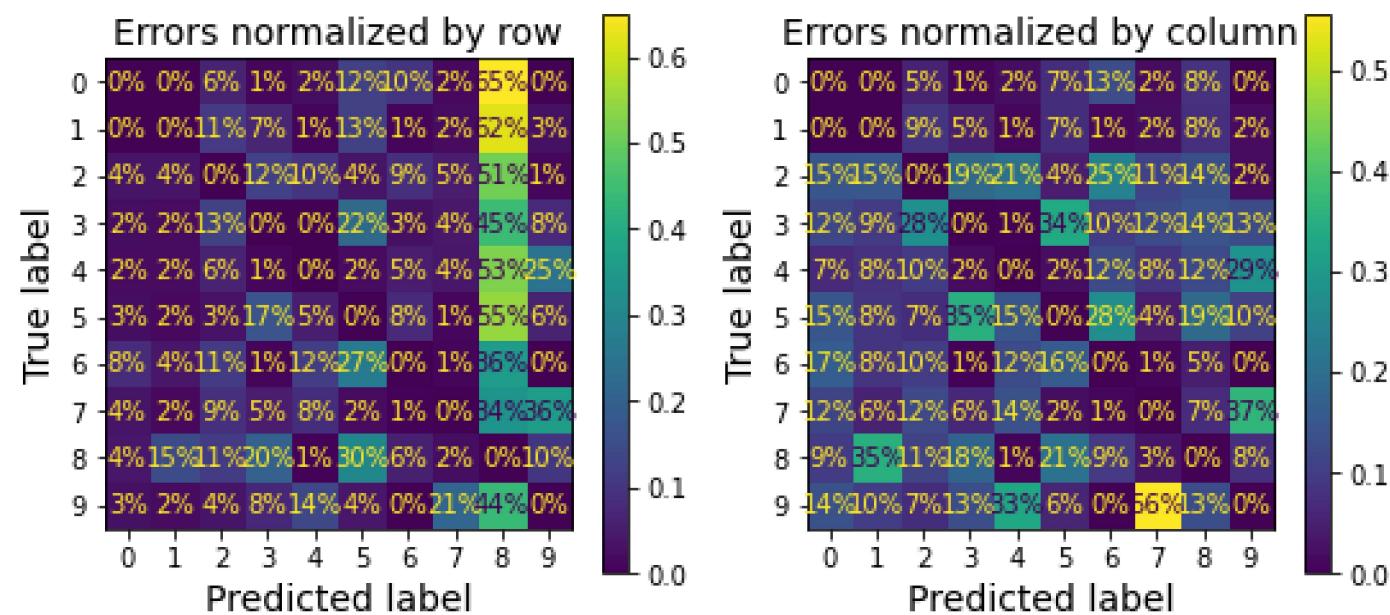
# 오차 행렬 활용

- 왼쪽 이미지: 손글씨 클래스 분류 모델의 오차 행렬을 이미지로 표현 가능
  - 대체로 잘 분류됨: 대각선이 밝음.
  - 5행은 좀 어두움. 숫자 5의 분류 정확도가 상대적으로 낮음
- 오른쪽 이미지: 행별로 100분율 계산



# 오차율 이미지

- 왼쪽 이미지: 8번 열이 밝음. 즉 많은 숫자가 8로 오인됨.
- 오른쪽 이미지:
  - 7로 오인된 숫자중에 9의 비중이 56%로 가장 높음.
  - 5로 오인된 숫자중에 3의 비중이 34%로 가장 높음



## 3과 5 대상 오차행렬

- 음성: 3으로 판정
- 양성: 5로 판정

		Predicted label				
		3	5	3	5	3
True label	3	3 3 3 3 3	3 3 3 3 3	3 3 3 3 3	3 3 3 3 3	3 3 3 3 3
	5	5 5 5 5 5	5 5 5 5 5	5 5 5 5 5	5 5 5 5 5	5 5 5 5 5

### 3.6. 다중 레이블 분류와 다중 출력 분류

# 다중 레이블 분류

multilabel classification

- 샘플마다 여러 종류의 레이블에 대한 값 예측
- 예제
  - 손글씨 사진이 가리키는 숫자가 7 이상인지 여부와 함께 홀수 인지 여부도 함께 예측
  - 숫자 5를 가리키는 이미지의 예측값: [ False, True]

## 다중 출력 분류

multioutput classification

- 다중 출력 다중 클래스 분류라고도 불림
- 다중 레이블 분류를 일반화한 것: 각각의 레이블에 대해 다중 클래스 분류 진행 가능
  - 이전 예제는 각각의 레이블에 대해 이진 분류 진행

## 이미지에서 잡음 제거

- 다중 레이블: 각각의 픽셀에 대해 레이블 예측해야 함.
- 다중 클래스: 각각의 픽셀에서 예측하는 레이블이 0에서 255 사이의 정수 중에 하나.



- 아래 사진: 분류기가 예측한 이미지

