2025.3.27 2024170928 노연경

머신러닝 세션 3주차

# MNIST 데이터셋 : 손으로 쓴 숫자 데이터셋

- 28\*28 필셀로 구성된 이미지 데이터 (1차원 array로 제공)

- 지도학습, 다중 클래스 분류, 배치 또는 온라인학습

- X에 숫자 이미지가 필셀값으로 구성된 array, y에 각 이미지에 해당하는 숫자 레이블 저장

- some\_digit\_image = some\_digit.to\_numpy().reshape(28,28) : 1차원의 array를 28\*28 배열로 재구성

- plt.imshow(some\_digit\_image, cmap = “binary”) : 2차원 배열에 따라 이미지로 시각화. cmap = “binary” : 흑백 이미지로 표시하기 위한 설정

- y = y.astype(np.uint8) : y에 담긴 데이터를 8비트 부호없는 정수형으로 변환

- X\_train, X\_test, y\_train, y\_test = X[:60000], X[60000:], y[:60000], y[60000:] : 훈련 셋과 데이터 셋 나누기

# 이진 분류기 훈련

- 레이블을 0 또는 1로 수정하여 원하는 숫자를 가리키는 이미지 레이블을 1, 이외의 수를 가리키는 이미지 레이블을 0으로 설정

- SGDClassfier : 확률적 경사하강법

- fit() : 모델을 학습시키는 매서드. 입력 데이터와 실제 정답을 사용하여 모델의 내부 파라미터를 조정

- predict() : 학습된 모델을 사용해 새로운 데이터에 대해 예측을 수행하는 매서드

- sgd\_clf.fit(X\_train, y\_train\_5) 로 모델을 학습시키고 sgd\_clf.predict([X.iloc[0]])로 예측

- cross\_val\_score() 을 사용하여 교차검증

# 오차 행렬

- 분류모델의 예측 결과와 실제 정답을 비교하는 행렬

- 행은 실제 클래스(정답), 열은 모델이 예측한 클래스

- TP, TN, FP, FN으로 나뉨

- y\_train\_pred = cross\_val\_predict(sgd\_clf, X\_train, y\_train\_5, cv = 3) : X\_train과 y\_train\_5를 3개의 폴드로 나누어 하나를 검증셋, 나머지 두 개를 학습셋으로 하여 sgd\_clf 모델을 학습시키는 과정을 3번 반복하여 성능을 평가. 결과적으로 0 또는 1이 저장됨.

- cm = confusion\_matrix(y\_train\_5, y\_train\_pred) : 오차행렬 계산

# 정밀도와 재현율

- 정밀도 = TP/(TP+FP) : ex) 5라고 예측된 값 중 진짜 5인 숫자들의 비율

- 재현율 = TP/(TP+FN)

- F1 점수 : 정밀도와 재현율의 조화 평균

- 보통 F1 점수가 높을수록 분류기의 성능을 좋게 평가하지만 경우에 따라 재현율과 정밀도 둘 중 하나에 높은 가중치를 두어야 할 때도 있음

- 정밀도/재현율 트레이드 오프 : 임계값을 조정해서 FP를 줄이려 하면 FN이 늘고 FN을 줄이려 하면 FP가 늘어나는 관계로 인해 트레이드 오프 발생. 임계값이 커질수록 양성이라고 판단하는 경우가 줄어들고 따라서 FP가 줄어 정밀도가 올라가고 재현율은 떨어진다.

# ROC 곡선

- x축 : 실제로는 음성인데 모델은 양성으로 예측한 비율 : FPR

- y축 : 실제로도 양성이고 모델도 양성으로 예측한 비율 : TPR = 재현율

- 왼쪽 위 모서리에 가까울수록 좋은 모델. 오탐비율(FPR)이 낮고 정탐비율(TPR)이 높음

- AUC : ROC 곡선 아래의 면적. AUC가 1에 가까울수록 성능이 좋은 분류기

# 다중 분류

- SGD, 랜덤 포레스트, 나이브 베이즈 분류기는 다중 클래스 분류 지원

- 로지스틱 회귀, 서포트 벡서 머신은 이진 분류만 지원함

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

# 에러 분석

- 오차 행렬을 이미지로 표현 🡪 대각선이 밝으면 대체로 잘 분류되었다는 뜻

- 오차율 이미지 : 각 행을 전체 샘플 수로 나누어 어떤 숫자가 다른 숫자로 잘못 예측 되었는지 비율을 표시. 어두우면 잘 분류했다는 뜻

# 다중 레이블 분류

- k-최근접 이웃 분류기 : 가장 가까운 k개의 이웃을 찾고, 다수결로 클래스를 정하는 방법

# 다중 출력 분류

- 다중 레이블 분류에서 한 레이블이 다중 클래스가 될 수 있도록 일반화한 것