<머신러닝 톺아보기> 3주차(분류모델) 과제

박민서

1. MNIST 데이터셋: 미국 고등학생과 인구조사국 직원들이 손으로 쓴 70,000개의 숫자 이미지로 구성된 데이터셋. 사용된 0부터 9까지의 숫자는 28*28=784 크기의 픽셀로 구성된 이미지 데이터. 2차원 array가 아닌 길이가 784인 1차원 array로 제공(레이블: 총 70,000개의 사진 샘플이 표현하는 값)
2. 문제정의
   (1) 지도학습: 각 이미지가 담고 있는 숫자가 레이블로 지정됨.
   (2) 분류: 이미지 데이터를 분석하여 0부터 9까지의 숫자로 분류, 이미지 그림을 총 10개의 클래스로 분류하는 다중 클래스 분류
   (3) 배치 또는 온라인 학습 가능
3. 훈련 셋과 데이터 셋 나누기
   (1) 훈련 세트: 앞쪽 60,000개 이미지
   (2) 테스트 세트: 나머지 10,000개 이미지
4. 이진 분류기 훈련
   (1) SGDClassfier(확률적 경사 하강법): 매우 큰 데이터셋을 효율적으로 처리하는 장점을 가짐. 한 번에 하나씩 훈련 샘플을 독립적으로 처리한다는 특징을 가짐.
5. 성능 측정
   (1) 교차 검증을 사용한 정확도 측정
   정확도: 전체 샘플을 대상으로 정확하게 예측한 비율(숫자 5를 표현하는 이미지를 True로 예측한 비율)
   주의) 훈련 세트의 샘플이 불균형적으로 구성되었다면, 정확도를 분류기의 성능 측정 기준으로 사용하는 것은 피해야 함.
   (2) 오차 행렬: 클래스 별 예측 결과를 정리한 행렬(행: 실제 클래스, 열: 예측된 클래스/클래스 A의 샘플이 클래스 B의 샘플로 분류된 횟수를 알고자 하면 A행 B열의 값을 확인하면 된다.)
   정확도: (TP+TN)/(TP+TN+FP+FN)
   정밀도: 양성 예측의 정확도, TP/(TP+FP)<=정밀도만으로는 분류기의 성능을 평가할 수 없다. 분류기가 정확하게 감지한 양성 샘플의 비율인 재현율을 함께 다루어야 한다.
   재현율: 양성 샘플에 대한 정확도, TP/(TP+FN), 분류기가 정확히 감지한 양성 샘플의 비율

F1 점수: 정밀도와 재현율의 조화 평균, 일반적으로 F1 점수가 높을수록 분류기의 성능을 좋게 평가하지만, 경우에 따라 재현율과 정밀도 둘 중의 하나에 높은 가중치를 두어야 할 때가 있다.

(3) 정밀도/재현율 트레이드오프

결정함수: 분류기가 각 샘플의 점수를 계산할 때 사용

결정 임계값: 결정 함수의 값이 이 값보다 같거나 크면 양성 클래스로 분류, 아니면 음성 클래스로 분류(임계값이 커질수록 정밀도는 올라가고, 재현율은 떨어진다.)

(4) ROC 곡선: 수신기 조작 특성 곡선(ROC 곡선)을 활용하여 이진 분류기의 성능이 측정 가능하다. 결정 임계값에 따른 거짓 양성 비율에 대한 참 양성 비율의 관계를 나타낸 곡선.

AUC와 분류기 성능: 재현율과 거짓 양성 비율 사이에도 서로 상쇄하는 기능이 있기에 재현율을 높이고자 하면 거짓 양성 비율 역시 함께 증가한다. 따라서 좋은 분류기는 재현율은 높으면서 거짓 양성 비율은 최대한 낮게 유지해야만 한다. AUC는 ROC곡선 아래의 면적으로 이 면적이 1에 가까울수록 성능이 좋은 분류기로 평가된다.

6. 다중 분류

(1) 다중 클래스 분류기: 세 개 이상의 클래스로 샘플을 분류하는 예측기

(2) 이진 분류기를 활용한 다중 클래스 분류

일대다 방식(OvA, OvR): 숫자 5 예측하기에서 사용했던 이진 분류 방식을 동일하게 모든 숫자에 대해서 실행. 각 샘플에 대해 10번 각기 다른 이진 분류기를 실행하고, 각 분류기의 결정 점수 중에서 가장 높은 점수를 받은 클래스를 선택

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

7. 다중 레이블 분류: 샘플마다 여러 개의 클래스 출력
8. 다중 출력 분류: 다중 레이블 분류에서 한 레이블이 다중 클래스가 될 수 있도록 일반화한 것