머신러닝 톺아보기_ 3주차. 분류모델

2023170832 고지원

1. MNIST

- 7만개의 숫자 손글씨 데이터셋. 0~9의 숫자가 적힌 이미지.
- 각 이미지는 28*28=784 픽셀로 구성, 1차원 array로 제공
- 훈련세트 : 앞쪽 6만개 이미지 / 테스트세트 : 나머지 1만개 이미지

2. 이진 분류기 훈련

- 해당 이미지가 설정해둔 숫자가 맞는지 아닌지를 판단하는 분류기.
- 레이블 : 해당 숫자가 맞다면 1, 아니라면 0
- SGDClassifier (확률적 경사 하강법)
  - 매우 큰 데이터셋을 효율적으로 처리.
  - 한 번에 훈련 샘플 하나씩 독립적으로 처리.

3. 성능 측정

1) 교차 검증
   - 기준 : 정확도
2) 오차 행렬
   - 클래스 별 예측 결과를 정리한 행렬
   - 행 : 실제 클래스 / 열 : 예측된 클래스
   - 정확도 : (TP + TN) / (TP + TN + FP + FN)
3) 정밀도(precision)와 재현율(recall)
   - **정밀도** : 양성 예측의 정확도. (X라고 예측된 값 중에 진짜 X인 비율)
   - Precision = TP / (TP + FP)
   - **재현율** : 양성 생플에 대한 정확도 (=민감도, 참 양성 비율)
   - Recall = TP / (TP + FN)
   - **F1 점수** : 정밀도와 재현율의 조화 평균
   - F1 = 2*precision*recall / precision+recall
4) 정밀도 / 재현율 트레이드오프
   - 결정 함수 : 분류기가 각 샘플의 점수를 계산할 때 사용
   - 결정 임계값 : 결정 함수의 값이 이 값보다 크거나 같으면 양성 클래스, 작으면 음성 클래스
   - 임계값이 커질수록 정밀도 증가, 재현율 감소
5) ROC 곡선
   - **ROC**(Receiver operating characteristic) : 수신기 조작 특성
   - **FPR**(false positive rate) : 결정 임계값에 따른 거짓 양성 비율
   - **ROC 곡선** : 거짓 양성 비율에 대한 참 양성 비율의 관계를 나타낸 곡선
   - FPR = FP / FP + TN
   - 재현율을 높이고자 하면 거짓 양성 비율도 증가 -> ROC 곡선이 y축에 최대한 근접하는 결과 나오게 해야함
   - **AUC** : ROC 곡선 아래의 면적. 1에 가까울수록 성능 좋은 분류기

4. 다중 분류

- **다중 클래스 분류기** : 세 개 이상의 클래스로 샘플을 분류하는 예측기

- **다중 클래스 분류 지원 분류기** : SGD 분류기, 랜덤 포레스트 분류기, 나이브 베이즈 분류기
- **이진 분류만 지원하는 분류기** : 로지스틱 회귀, 서포트 벡터 머신
- **이진 분류기를 활용해 다중 클래스 분류** : 일대다(OvR, OvA), 일대일(OvO)
- **일대다 방식** : 모든 숫자에 대해 이진 분류 실행(총 10번) 후 가장 높은 점수 받은 클래스 선택
- **일대일 방식** : 조합가능한 모든 일대일 분류 방식 진행해 가장 높은 점수 얻은 숫자 선택

5. 에러 분석

- 오차행렬, 오차율 이미지 활용

6. 다중 레이블 분류

- 샘플마다 여러 개의 클래스 출력.
- 다중 레이블 분류 지원 모델 : k-최근접 이웃 분류기, 사이킷런의 KNeighborsClassfier

7. 다중 출력 분류

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