EURON 1주차 과제_3장

평가

정확도(Accuracy)

- 실제 데이터에서 예측 데이터가 얼마나 같은지 판단하는 지표
- 정확도 = 예측 결과가 동일한 데이터 건수 / 전체 예측 데이터 건수
- 이진 분류의 경우 데이터의 구성에 따라 정확도 지표가 ML 모델 성능을 왜곡할 수 있음
- 정확도 평가는 불균형한 레이블 세트에서 성능 수치로 사용X
- 여러 가지 분류 지표와 함께 적용하여 평가해야 함

오차 행렬

- 예측 오류가 얼마인지와 더불어 어떠한 유형의 예측 오루가 발생하는지 함께 나타냄
- TN, FP, FN, TP 형태로 나타냄 : 예측 클래스와 실재 클래스의 Positive 결정 값(값 1)과 Negative 결정 값(값 0)의 결합에 따라 결정
- 사이킷런은 오차 행렬을 구하기 위해 confusion_matrix() API 제공
- 정확도 = (TN+TP)/(TN+FP+FN+TP)
- 불균형한 이진 분류 데이터 세트에서는 Positive 데이터 건수가 매우 작기 때문에 데이터에 기반한 ML 알고리즘은 Positive보다는 Negative로 예측 정확도가 높아지는 경향 발생 → 수치 판단 오류

정밀도와 재현율

- Positive 데이터 세트의 예측 성능에 초점을 맞춘 평가 지표

▼ 정밀도 = TP / (FP+TP)

  - 예측을 Positive로 한 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율
  - 양성 예측도라고도 불림

- 실제 Negative 음성 데이터를 Positive 로 잘못 판단하게 되면 업무상 큰 영향이 발생하는 경우 중요 지표가 됨
- FP를 낮추는데 초점

▼ 재현율 = TP / (FN+TP)

- 실제 값이 Positive인 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율
- 민감도 또는 TPR라고도 불림
- 실제 Positive 양성 데이터를 Negative로 잘못 판단하게 되면 업무상 큰 영향이 발생하는 경우 중요 지표가 됨
- FN을 낮추는데 초점

정밀도/재현도 트레이드오프

- 분류의 결정 임곗값을 조정해 정밀도 또는 재현율의 수치를 높일 수 있음
- 하지만 둘은 보완적인 평가 지표이기 때문에 어느 한쪽을 높이면 다른 한 쪽의 수치는 떨어지기 쉬움
- 사이킷런의 분류 알고리즘은 예측 데이터가 특정 레이블에 속하는지 계산하기 위해 먼저 개별 레이블별로 결정 확률을 구함
- 개별 데이터별로 예측 확률을 반환하는 메서드인 predict_proba() 제공

정밀도와 재현율의 맹점
상호 보완할 수 있는 수준에서 적용돼야 함

F1 스코어

- 정밀도와 재현율을 결합한 지표
- 둘 중 어느 한 쪽으로 치우치지 않는 수치를 나타낼 때 상대적으로 높은 값을 가짐

ROC 곡선과 AUC

- 이진 분류의 예측 성능 측정에서 중요하게 사용됨

- ROC 곡선은 수신자 판단 곡선, 머신러닝 이진 분류 모델 예측 성능 판단하는 중요한 평가 지표
- AUC 값은 ROC 곡선 밑의 면적을 구한 것으로서 일반적으로 1에 가까울수록 좋은 수치
- 보통의 분류는 0.5 이상의 AUC를 가짐

피마 인디언 당뇨병 예측

ipynb 파일 참고