핸즈온 머신러닝

Ch 3. 분류모델

분류모델(Classification Model)

- 지도학습의 한 종류(레이블이 사전에 정의되어 있음)
- 입력데이터, 레이블을 이용하여 학습한 후 테스트 데이터의 레이블을 예측하는 모델

Classification Model 의 종류

- 경사하강법
    - Logistic Regression
    - SGD Classifier
    - ANNs(Artificial Neural Networks)(인공신경망)
- 확률기반 알고리즘
    - LDA
    - QDA
    - Naïve Bayse Classifier
- 거리 기반
    - K-NN(K-Nearest Neighbor)
    - SVM(Support Vector Machine)
- 트리기반
    - Decision Tree
    - Ensemble(앙상블 기반)
        - Random Forest
        - AdaBoost
        - GBT(Gradient Boost Tree)
        - XGBoost
        - CatBoost

이진분류기 vs 다중분류기

- 이진분류기: ex) 이미지 샘플이 5 를 표현하는가, 아닌가?
- 다중분류기: ex) 이미지 샘플이 0-9 중 어떤 숫자를 표현하는가?

분류 모델의 성능 측정(1): 오차 행렬(Confusion Matrix)

- TP(True Positive), FP(False Positive), TN(True Negative), FN(False Negative)
- Accuracy(정확도): (TP+TN)/(TP+FP+TN+FN)
- Precision(정밀도): TP/(TP+FP) *양성이라고 예측한 것 중 실제로 양성이었던 비율

- Recall(재현율=민감도=참 양성 비율): TP/(TP+FN) *실제로 양성인 것 중 분류기가 양성이라고 정확하게 예측한 비율
- F1 Score: (2*Precision*Recall)/(Precision+Recall) *F1 Score 가 높을수록 모델의 성능이 좋다고 평가함
- Precision(정밀도)와 Recall(재현율)의 Trade-off: 결정 임계값(Threshold)을 높이면 정밀도가 높아지고 재현율이 낮아짐

분류 모델의 성능 측정(2): ROC(Receiver Operating Characteristic) 곡선

- 재현율(참 양성 비율, TPR), 거짓 양성비율(FPR)(FP/(TN+FP)) 사이의 관계를 나타낸 곡선 *거짓 양성비율: 원래 음성인 것 중 양성이라고 잘못 분류된 비율
- AUC(ROC 곡선의 아래 면적)이 1 에 가까울 수록 좋은 성능을 가진 모델
- TPR 이 높이면 FPR 도 함께 높아짐

다중분류기

- SGD Classifier, Random Forest, Naïve Bayes Classifier 사용 가능(Logistic Regression, SVM 사용 불가)
- 이진 분류기 활용하여 다중 분류
    - 일대다(OvR or OvA)
    - 일대일(OvO)
- 다중 클래스 지원 분류기 사용
    - SGD Classifier
    - K-NN