BOAZ Spring Study : Weekly Assignment
(Due date : 2023.04.02)

[필기]

1. 다음 중 머신러닝 모델에 대한 설명이 옳으면 O, 틀리면 X를 하고, 그렇게 생각한 이유를 적으세요. **(20 points)**

(1) K-최근접 이웃(K-Nearest Neighbor) 알고리즘을 이용하여 분류 문제를 풀 때, 파라미터로 K값을 사용한다. 이때, K값이 달라지면 같은 데이터를 예측하더라도 예측값이 달라질 수 있다.

(2) K-최근접 이웃(K-Nearest Neighbor) 알고리즘은 미리 분류 모형을 만들지 않기 때문에, 계산량이 많다는 단점이 존재한다.

(3) 서포트 벡터 머신(Support Vector Machine) 알고리즘에서 하드 마진(hard margin)은 소프트 마진(soft margin)보다 어느 정도의 오류를 더 허용하면서, 일반화 성능을 강화한다.

(4) 결정 트리(Decision Tree)는 규칙 기반으로 한 머신러닝 기법으로, 머신러닝의 결과를 해석할 때 유용할 수 있다.

(5) 어떤 사람이 아이템을 구매할 확률을 확인하고 싶을 때, 로지스틱 회귀보다는 선형 회귀를 이용하여 더 잘 확인할 수 있다.

2. 아래는 어떤 머신러닝 문제를 풀었을 때 나온 오차 행렬(Confusion Matrix)의 결과입니다. 아래 물음에 답하세요. (단, 현재는 이진 분류(Binary Classification)을 가정합니다.) **(30 points)**

		예측 값	
		Positive(양성)	Negative
실제 값	Positive	50	150
	Negative	150	650

(1) 정확도(Accuracy), 재현율(Recall), 정밀도(Precision), F1-Score를 각각 구하세요. **(5 points)**

(2) 위에서 Type I Error와 Type II Error의 비율을 구하세요. **(5 points)**

(3) 만약, 위 머신러닝 문제가 암 환자를 판별해내는 문제라고 가정합시다. 그렇다면, 정확도가 꼭 유의미한 결과를 내포하나요? 그렇지 않다면, 정확도 보다는 어떤 지표를 중점으로 보는 것이 더 좋을지 쓰고, 이유를 밝혀주세요. **(7 points)**

(4) 만약, 위 머신러닝 문제가 스팸 메일을 판별해내는 문제라고 가정합시다. 그렇다면, 정확도가 꼭 유의미한 결과를 내포하나요? 그렇지 않다면, 정확도 보다는 어떤 지표를 중점으로 보는 것이 더 좋을지 쓰고, 이유를 밝혀주세요. **(7 points)**

(5) (3)과 (4)에서는 정확도만으로 성능 평가를 하는 것은 한계가 있다는 것을 알았습니다. 즉, 정확도 뿐만 아니라 다른 지표들에 대해서도 같이 판단을 해야 한다는 것인데, 데이터의 어떤 특성 때문에 이러한 결과가 나왔을까요? **(6 points)**

[Python Coding - Assignment]
< -- 실습 과제를 시작하기 전, Notion에 있는 실습 파일을 다운받아주세요. -- >

3. 다음은 여러 머신러닝 알고리즘을 이용하여 분류 문제를 푸는 것에 대한 내용입니다. 아래 물음에 답하세요. **(20 points)**
(1) 아래 요구조건을 참고하여, K-Nearest Neighbors 알고리즘을 skicit-learn을 이용하여 구현하세요. **(10 points)**

데이터를 적절하게 전처리를 한 후(결측치 처리, scaling, 파생변수 생성 등), 학습 데이터와 테스트 데이터를 7 : 3의 비율로 분할하세요. KNeighborsClassifier에서 neighbors 파라미터의 수를 3으로 지정하고 그 때의 정확도를 출력하는 코드를 작성하세요. (학습 데이터와 테스트 데이터 분할 시, 0과 1의 비율은 기존과 동일하게 나누세요.)

(2) (1)에서 K값을 1부터 15까지 순환하면서, k = 3일 때보다 더 좋은 성능을 가지는 K값을 찾는 코드를 작성하고, 그 때의 정확도를 출력하세요. **(10 points)**

4. K-Nearest Neighbors를 제외한 나머지 분류 알고리즘을 하나 선택하고, 아래 물음에 답하세요. **(30 points)**
(1) 아래 요구조건을 참고하여, 각자 선택한 모델의 정확도와 f1_score를 출력하세요. **(10 points)**

■ 학습 데이터와 테스트 데이터의 비율을 8 : 2로 나누어 모델을 학습시키세요. ■ 학습 데이터와 테스트 데이터에서 0과 1의 비율을 기존과 동일하게 나누세요. ■ 하이퍼 파라미터를 넣지 않은 상태에서 모델의 정확도와 f1_score를 출력하세요.

(2) 선택한 모델에 하이퍼파라미터를 추가하여, 모델의 정확도와 f1_score를 (1)에서보다 더 높게 나오도록 하세요. **(5 points)**

(3) 아래 설명을 참고하여 물음에 답하세요.

sklearn에서 test data를 예측할 때는 model.predict(X_test)를 이용합니다. 이 함수는 말 그대로, X_test 데이터에 대해서 예측한 값을 반환합니다. 이진 분류에서는 0 또는 1을 반환하는데, 기본적으로 X_test에서 1로 예측할 확률이 0.5보다 크면 1로, 작으면 0으로 예측합니다. 이때, sklearn에서 threshold(예측할 확률의 기준)을 바꿔서 예측할 수 있는데, model.predict_proba 함수를 이용할 수 있습니다. predict_proba는 이진분류에서 데이터를 0으로 예측할 확률과 1로 예측할 확률을 array의 형태로 반환합니다.

선택한 모델을 이용하여, threshold를 0.15부터 0.85까지 0.05의 간격으로 설정하여 threshold가 몇일 때 f1 score가 높게 나오는 지 코드를 이용하여 검사하세요.
(15 points)