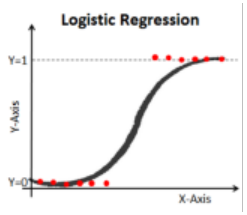
이진분류 문제에서 로지스틱 회귀 모델의 **최적 하이퍼 파라미터 도출**

1) 로지스틱 회귀 모델

이진 분류 문제에 널리 사용되는 강력하고도 활용성이 높은 모델로서 고객 이탈 예측, 고객 신용도 평가, 대출 상환 능력 판단 등 다양한 사회과학 분야의 문제 해결에 널리 사용되고 있다.

로지스틱 회귀 모델은 범주형 종속변수를 예측하는 모델로서, y=0(실패), y=1(성공)로 표현된 클래스를 예측하여 분류하는 모델이다.

하이퍼 파라미터로 규제 강도를 조절하는 C, 최대 반복 횟수를 제어하는 max_iter, 최적화 알고리즘의 종류인 solver 등이 있다. solver 종류로 lbfgs, newton, saga, sag, liblinear이 있다.



<로지스틱 회귀 모델 그래프>

기존 하이퍼 파라미터 튜닝 방식

사용자가 하이퍼 파라미터의 값을 직접 리스트로 입력하여 입력값의 모든 조합에 대한 예측 성능 평가와 비교를 진행하는 Scikit-Learn에서 제공하는 GridSearchCV가 사용되었다.

GridSearchCV 방식의 한계점

사용자가 제공하지 않은 파라미터 이외의 값에서 최적의 값이 존재할 가능성이 있으며, 광범위의 하이퍼 파라미터 값을 입력할 경우 모델 학습에 소요되는 시간이 기하급수적으로 늘어난다.

연구 방향

최대한 전역적인 하이퍼 파라미터의 최적값을 찾기 위해 범위를 어떻게 지정해주어야 하는지, 회귀 모델의 정확도와 직접적 연관이 있는 C를 우선적으로 고려하여, 설정 가능한 옵션에 대한 가이드라인을 검토해 보았다.

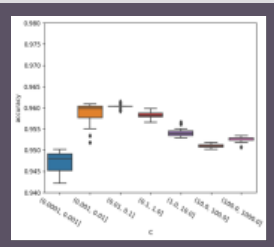
2) MNIST 손글씨 데이터

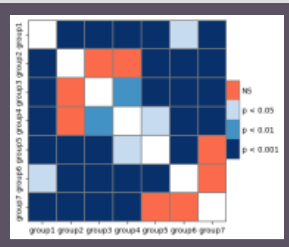
MNIST데이터는 손으로 쓴 숫자 이미지 데이터셋으로, 60000개의 Train Set과 10000개의 Test Set으로 구성된 머신러닝의 학습에 가장 대표적인 데이터이다. 이진 분류 모델의 적용을 위해 두 숫자만을 구분하도록, 분류 정확도가 가장 낮게 도출된 5와 8을 Target으로 분석을 진행하였다.


<MNIST 데이터 출력>

3) 비모수적 방법론을 활용한 최적의 C값 범위 한정

C값을 0.0001부터 1000까지의 범위에서 7개의 그룹으로 분할하여 각 그룹에서 랜덤 추출된 C값들로 각각 로지스틱 회귀 모델을 학습하여 정확도를 확인하였다.


<C그룹 별 Accuracy Box plot>


<Dunnett 사후검정 히트맵>

순위	그룹	
1	Group 2, 3, 4	
2	Group 5	Group 7
3	Group 6	
4	Group 1	

<C그룹 별 순위>

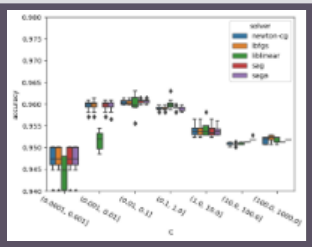
Kruskal-Wallis 검정

그룹별 정확도의 위치모수를 Kruskal-Wallis 방법으로 검정하였고, Test statistic: 190.36, P-Value: 0.000 으로 7개의 그룹 간에 유의한 차이가 있음을 확인하였다.

Dunnett 사후검정

Dunnett 사후검정을 실시하였고, 사후검정 히트맵를 통해 Group2: (0.001, 0.01], Group3: (0.01, 0.1], Group4: (0.1, 1] 간에는 차이가 없으며, 가장 높은 Accuracy를 얻는 그룹임을 확인하였다.

4) C값과 solver 조합을 고려한 최적의 하이퍼 파라미터 찾기


<solver에 따른 C그룹 별 Accuracy Boxplot>

liblinear를 제외한 나머지 solver에서 최적의 C그룹은 2,3,4로 관찰되었다. GridSearchCV를 사용할 때 제시할 C값들은 0.001, 0.01, 0.1, 1 사이의 값들이 적절하다.

solver를 liblinear로 지정할 경우 최적의 C그룹은 3,4로 관찰되었다. GridSearchCV를 사용할 때 제시할 C값들은 0.01, 0.1, 1 사이의 값들이 적절하다.

5) 해석 및 결론

로지스틱 회귀모델을 학습시킬 때, C값으로는 0.001~1 사이의 범위를 우선하여 GridSearchCV에 적용할 수 있다. 단, 데이터의 크기 및 특성을 고려한 적절한 solver를 GridsearchCV에 포함함으로써 최적의 C값과 solver, max_iter을 적용할 수 있다. 이와 같은 최적의 하이퍼 파라미터 튜닝을 통해 연산비용과 수행시간을 효과적으로 단축하여 높은 Accuracy를 얻을 수 있다.