[5주차 과제2] Logistic Regression의 하이퍼파라미터 변경에 따른 성능 결과 분석

수면시간에 따른 우울증 예측을 통한 분석

18011480 이진형 I 컴퓨터공학과

문제 분석

문제에서 주어진 데이터 셋(trainX.csv, testX.csv)는 수면의 양과 질에 대한 데이터 셋이다. sleep_time, sleep_quality 등 14개의 컬럼으로 이루어져 있다. 해당 데이터 셋으로 Logistic Regression 모델을 학습시킨 뒤, 우울증 유/무를 분류하는 문제이다. 캐글 문제 Description에서 데이터 전처리 없이 위 feature들을 사용해 학습하라고 주어졌기 때문에, 추가적인 전처리를 통하지 않고, standardscaler만 이용하였다.

실험 조건

scikit-learn 문서[1]에 따르면 Logistic Regression의 파라미터는 penalty, dual, tol, C, fit_intercept, intercept_scaling, class_weight, random_state, solver, max_iter, multi_class, verbose, warm_start, n_jobs, l1_ratio 가 존재한다. 이번 실험에선 이들 중 penalty, C, solver 세 가지 파라미터의 변화를 통해 실험을 진행했다.
실험에 사용할 파라미터의 종류를 나타내면 다음과 같다.

Solver

liblinear	Library for large linear classification. 공식 문서에 따르면 작은 크기의 데이터셋에 유리하다.	l1, l2
saga	Stochastic average gradient descent인 sag에 L1 정규화가 가능하도록 한 solver.	elasticnet, l1, l2, none
lbfgs	Limited-memory Broyden-Fletcher-Goldfarb-Shanno. 0.22.0 버전부터 기본값이다.	l2, none

Penalty

l1	Lasso. 영향을 크게 미치는 핵심적인 피처(x_i)들만 반영하도록 한다.
l2	Ridge. 전체적인 W값을 감소시켜 그래프의 곡률을 줄인다.

C

C는 규제 강도를 나타내는 alpha 값의 역수이다. C값이 작을 수록 규제 강도가 크다.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

실험

진행한 실험의 조건과 결과는 다음과 같다. 이전 1번 과제에서 penalty를 사용하지 않았을 때 C의 값 변화가 무의미함을 확인했기에, Penalty가 none일 때는 C 값을 부여하지 않았다.

Solver	Penalty	C	Score
liblinear	l1	0.1	0.52173
liblinear	l1	1.0	0.65217
liblinear	l1	10.0	0.56521
liblinear	l2	0.1	0.65217
liblinear	l2	1.0	0.65217
liblinear	l2	10.0	0.65217
saga	none	null	0.65217
saga	**l1**	**0.1**	**0.69565**
saga	l1	1.0	0.65217
saga	l1	10.0	0.69565
saga	l2	0.1	0.60869
saga	l2	1.0	0.65217
saga	l2	10.0	0.65217
lbfgs	none	null	0.56521
lbfgs	l2	0.1	0.60869
lbfgs	l2	1.0	0.65217
lbfgs	l2	10.0	0.65217

실험 결과

위와 같은 조건으로 실험을 진행한 결과, 최고 점수인 0.69565를 기록한 파라미터는 다음과 같았다.
saga/l1/0.1

개선사항

규제정도를 나타내는 C값에 따른 개선사항을 더 효과적으로 확인하기 위해서는 학습 데이터와 테스트 데이터의 점수를 확인하여 생성된 모델이 fit된 정도를 확인해야 하는데, 해당 추이를 확인하기엔 무리가 있는 결과가 나와 아쉬웠다.