기계학습 4주차 과제

수면시간에 따른 우울증 예측

컴퓨터공학과
17011584 정재경

개요

주어진 데이터셋은 학생들을 대상으로 수면시간과 우울증에 대한 설문조사정보가 포함돼있어, 각 정보에 따른 우울증의 유무를 나타낸다. 제공된 정보에 따라서 우울증의 발병 여부를 예측하는 문제로, 분류문제에 해당한다. 이번 과제는 Logistic Regression 방식을 사용해서 해결해보았다.

변경 가능한 하이퍼파라미터

scikit learn 에서 제공하는 LogisticRegression 공식문서에 따르면 커스터마이징이 가능한 주요한 하이퍼 파라미터에는 solver , penalty, c, class_weight 등이 있다.

각각 하이퍼 파라미터들의 의미를 살펴보면

1. solver - Logistic Regression 을 할 때 사용하는 최적화 알고리즘인다. 사용 가능한 solver들은 newton-cg, lbfgs, liblinear, sag, saga가 있다. 각 solver들 별로 장단점들이 존재하고, 사용할 수 있는 penalty의 종류에도 제한이 있다. 과제 제출 시 사용한 solver는 newton-cg를 사용했다.
2. penalty - 회귀모델을 fitting할 때, 손실함수에 더해지는 penalty값의 종류를 의미한다. penalty에는 크게 3가지가 있는데, L1, L2, 그리고 이 둘을 합친 ElasticNet이다. L1 Norm을 사용할 경우 중요하지 않은 feature 값들의 coefficient 가 0으로 수렴한다. 이 방식을 LASSO Regression (Least Absolute Shrinkage and Selection Operator) 이라고 부르기도 한다.
3. c - Fitting할 때 정규화 강도의 역을 의미한다. c값이 작아질수록 정규화의 강도가 강해져, 주는 모델에 제약이 강해지고, 반대의 경우는 약해진다.

실험

우선, 이번 과제는 데이터의 전처리 과정을 진행했다. 데이터는 각각 수면시간, 수면 퀄리티에 대해서 mean, std, min, max, 25, 50, 75의 정보가 주어진다. 따라서 수면시간과 퀄리티를 곱한 값으로 학습을 시키는 것이 성능이 더 좋을 것이라 판단하여 다음과 같이 데이터 전처리를 해주었다. 그러나 데이터의 스케일이 다른 std, min, max는 제외하고 학습을 진행했다.

```
# Preprocess data
options = ['mean', '25', '50', '75'] # 'std', , 'min', 'max'

_x, _test = pd.DataFrame(), pd.DataFrame()

for option in options:
    _x['sq_' + option] = x['sleep_time_' + option] * x['sleep_quality_' + option]
    _test['sq_' + option] = test['sleep_time_' + option] * test['sleep_quality_' + option]
```

이번 과제에서 변경을 해보며 결과값의 차이 추이를 본 하이퍼파라미터는 solver, penalty, c, class_weight 이다. 번외로 데이터 전처리를 하지 않은 채 채점을 받은 결과도 포함했다.

Solver	Penalty	Class Weight	Score
newton-cg	l2	None	0.69565
newton-cg	l2	balanced	0.56521
lbfgs	l2	None	0.69565
lbfgs	l2	balanced	0.56521
liblinear	l2	None	0.52173
liblinear	l2	balanced	0.52173
liblinear	l1	None	0.52173
liblinear	l1	balanced	0.43478
lbfgs (데이터 전처리 X)	l2	balanced	0.69565

결론

세 가지 다른 경우에서 최고점을 기록했다. 눈여겨 볼 만한 점은, lbfgs, l2, balanced 옵션은 데이터 전처리를 하지 않았을 때가 전처리를 했을 때보다 더 점수가 높았다. 같은 데이터라도 전처리 방식에 따라 가장 성능이 좋은 하이퍼파라미터들이 다르다는 점을 알 수가 있다.