

기계학습 4주차 과제

은하계 종류 예측

컴퓨터공학과
17011584 정재경

개요

주어진 데이터셋은 은하계에 대한 벡터화된 사진정보가 포함되어있어, 각 정보에 따른 은하계 종류 구분을 나타낸다. 제공된 정보에 따라서 은하계의 종류를 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 알고리즘이다. 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.
4. class_weight - 각각의 피쳐값에 가중치값을 의미한다. default는 모든 클래스가 같은 가중치를 갖는다고 판단하고, balanced 는 y값에 등장하는 빈도의 비율에 따라 자동으로 가중치를 지정해준다.

실험

이번 과제에서 변경을 해보며 결과값의 차이 추이를 본 하이퍼파라미터는 solver, penalty, c 이다.

Solver	Penalty	C	Score
newton-cg	l2	0.01	0.79466
newton-cg	l2	1	0.79666
newton-cg	l2	100	0.79666
newton-cg	none	default - penalty가 없어서 의미 없음	0.79666
liblinear	l1	0.01	0.78600
liblinear	l1	1	0.79400
liblinear	l1	100	0.79466

liblinear	l2	0.01	0.79133
liblinear	l2	1	0.79466
liblinear	l2	100	0.79466

결론

이 데이터셋은 newton-cg solver, l2 penalty, c=1 or 100 일 때의 성능이 가장 좋았다. Scikit learn의 공식 docs에 따르면, liblinear 방식이 작은 데이터셋에 적합하다는 언급이 되어있어서 사용했지만, 실제로는 newton-cg방식이 더 성능이 좋았다. 주어진 데이터 자체가 작다기보단 벡터화된 이미지 데이터라서 그럴 수도 있다는 추측이 된다. 예측성능을 더 높이기 위해선 다른 방식을 채택하던지, 이미지 벡터화 과정부터 조금씩 변화가 있어야 할 것으로 판단된다.