[4주차 과제1] KNN의 하이퍼파라미터 변경에 따른 성능결과 분석

재배환경 별 작물 종류 예측을 통한 분석

18011480 이진형 I 컴퓨터공학과

문제 분석

문제에서 주어진 데이터 셋(train.csv, test.csv)은 7개의 컬럼으로 구성되어 있다. 재배환경에 해당하는 데이터 셋으로 KNN 모델을 학습시킨 후, 작물 종류에 해당하는 22개의 레이블 중 하나로 분류하는 문제이다.

실험

학습시킬 수 있는 데이터 셋의 크기가 작고, kaggle에 제출 후 정확도를 바로 확인할 수 있어 train_test_split을 진행하지 않아도 된다는 판단 하에 k와 p값 두 가지만을 조정해 하이퍼파라미터 튜닝을 진행했고, 다음과 같은 결과를 얻었다.

p	k	Score
1	3	0.97272
1	7	0.97272
1	5	0.97090
1	9	0.97090
1	6	0.96909
1	8	0.96727
1	1	0.96545
1	4	0.96545
1	2	0.96363

p	k	Score
2	5	0.97090
2	3	0.96727
2	4	0.96727
2	7	0.96727
2	6	0.96545
2	9	0.96545
2	1	0.96181
2	2	0.96181
2	8	0.96181

kaggle상 의 score에 따라 결과를 분석한 결과 k = 3, p = 1일 때와 k = 7, p =1 일때 동률로 가장 높은 정확도로 모델이 만들어 진 것을 알 수 있다.

개선사항

추후 과제를 진행하면서 kaggle을 통해 바로 점수를 확인할 수 없을 때는 학습 데이터와 실험 데이터를 분리해서 정확도를 측정하는 과정이 필요한데, 이번 과제에서 진행했던 것 처럼 직접 튜닝하는 것 보다 scikit-learn 패키지의 GridSearchCV 모듈을 통해 최적화된 값을 도출해 낼 수 있다는 것을 알게 되어 추후 과제 진행시 사용하면 튜닝이 용이할 것 같다.