Classification with k-NN

2016314726 정영준

와인 데이터셋을 사용하여 k-NN Classification(k-최근접 이웃 알고리즘)을 수행하였습니다. 먼저 와인 데이터셋을 load\_wine을 통해 load하고 data/target으로 나누고 shape를 사용하여 열의 개수를 파악하여 데이터셋의 input (X)의 feature는 13개라는 것을 알 수 있었습니다. Target의 class의 개수는 np.unique()를 사용하여 알 수 있었습니다. 학습세트와 테스트세트로 데이터를 나누기 위하여 sklearn의 train\_test\_split 함수를 사용하였고 stratify를 y로 지정하여 target값이 균등하게 학습데이터와 테스트데이터로 나뉠 수 있도록 하였습니다. 먼저 Scaling을 하지않고 k-NN 모델을 n\_neighbors=5 (default)로 설정하고 학습 데이터로 학습을 시켰습니다accuracy score를 사용하여 학습 데이터의 target값과 모델이 예측한 학습 데이터의 target의 일치를 확인한 결과 0.74라는 값을 얻을 수 있었습니다. 테스트 데이터도 동일하게 예측하였더니 0.71의 accuracy score를 기록했습니다. 그 후 n\_neighbors= 2,4,6,8 로 변경하고 같은 학습과정을 진행 하였습니다.

|  |  |  |
| --- | --- | --- |
| n\_neighbors | Train set accuracy score | Test set accuracy score |
| 2 | 0.82 | 0.60 |
| 4 | 0.78 | 0.68 |
| 6 | 0.78 | 0.75 |
| 8 | 0.81 | 0.73 |

이번엔 Standard Scaling을 통해 학습 feature를 통한 Scaling을 마친 후 n\_neighbors를 변화시키면서 다시 accuracy score를 측정하였습니다

|  |  |  |
| --- | --- | --- |
| n\_neighbors | Train set accuracy score | Test set accuracy score |
| 2 | 0.97 | 0.96 |
| 4 | 0.95 | 0.98 |
| 6 | 0.95 | 0.96 |
| 8 | 0.97 | 0.98 |

Train set와 Test set에서 Accuracy score가 골고루 높은 StandardScaler를 적용하였고 n\_neighbors = 8이 최고의 hyperparameter라고 생각합니다.













