2019150445 신백록

1. CART에서 criterion을 gini와 entropy를 각각 적용하고 max\_depth를 변경하여 결과를 비교하고 해석하라.

Iris data를 random\_state=1을 주고 train set과 test set으로 분리한 후, criterion=’gini’, max\_depth=2인 decision tree를 학습해 결과를 확인해보았다.

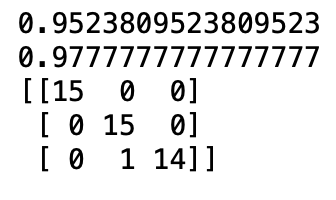
텍스트이(가) 표시된 사진

자동 생성된 설명

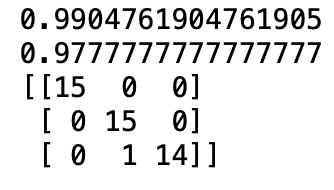
Train acc, test acc 둘다 95% 정도로 성능도 나쁘지 않고 overfitting도 없게 결과가 나왔다.

Criterion=’entropy’로 주고 tree를 그려 확인해 본 결과, max\_depth=2인 상태에서는 criterion=’gini’와 같은 기준에 의해 분리되었고 같은 결과가 나왔다.

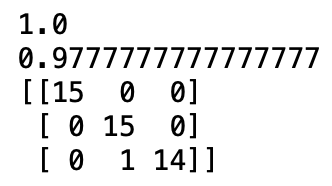
이제 max\_depth를 점점 늘려보고 결과를 확인해 보았다. 일단 max\_depth=3인 상태에서는 train acc는 같은 반면, test acc는 커졌다. 하지만 이는 겨우 하나의 데이터 차이이기에 모델의 성능에 큰 영향을 미치지는 못한다.



Criterion=’gini’, Max\_depth를 5를 준 결과, overfitting이 약간 발생하기 시작하였고, tree도 점점 복잡해졌다.



마지막으로 max\_depth를 주지 않았을 때, tree는 완벽하게 학습이 되었고, test set에 대한 성능은 똑같았다.



이를 토대로 봤을 때, criterion=’gini’, ‘entropy’는 모델에 큰 영향을 미치지는 못한다. 중요한 것은 max\_depth와 min\_impurity\_decrease같은 pruning을 위한 하이퍼 파라미터 같다. 이 iris dataset에서는 tree의 depth가 깊어질수록 train acc는 높아지지만 test acc는 그대로인 것을 확인할 수 있다. 그러므로 test acc는 최대이면서 model은 덜 복잡한 Max\_depth=2나 3이 적정할 것 같다.

하지만 이 dataset에서는 data의 수가 너무 적고, 특히 test set의 데이터가 적기에 test set의 편중이 발생할 수 있다. 따라서 cross-validation과 Grid search를 활용해 tree를 다시 한 번 학습시켜 보았다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그랬더니 max\_depth=2, criterion=’gini’로 가장 simple한 model이 모든 dataset에 대하여 성능이 좋은 것으로 판명 났고, 실제로 특성변수들에 대한 그림을 그려서 봤을 때도 class별로 well-separated 되어있는 상태이기에 그리 깊은 Depth가 필요해 보이지도 않았다.