6회차 과제 - ML

Date	@2024년 7월 30일 오후 11:59
Tag	과제

(TODO 2-1) learning curve 및 성능 평가 결과를 참고하여 Decision Tree 모델이 오버피팅 되었는지 판단해주세요. 판단의 근거를 제시하고, ML 모델에서 오버피팅을 완화할 수 있는 방안을 찾아 함께 작성해주세요.

- functions.py 파일에 구현된 plot_learning_curve 의 코드를 바탕으로 learning curve가 의미하는 바가 무엇인지 생각해보세요.
- 오버피팅인지 아닌지의 판단은 성능 평가 결과를 바탕으로 이루어져야 합니다.

plot_learning_curve는 scikit-learn 라이브러리의 learning_curve 함수의 반환값을 시각화한 함수입니다. 즉, 학습 데이터 양의 증가에 따른 train score와 test score를 그래프로 보여줍니다.

Decision Tree의 learning curve를 보면 train accuracy는 일관되게 1에 근접한 값을 보이는 반면, validation accuracy는 학습 초반에 잠깐 상승세를 보이다가 0.81 정도로 수렴하고 있음을 알 수 있습니다. train accuracy가 validation accuracy에 비해 높고, 그 간극이 줄어들지 않는다는 점에서 Decision Tree 모델은 오버피팅되었다고 판단할 수 있습니다.

decision tree는 학습 데이터의 특정 feature에 따라 분기하기에 데이터에 속한 노이즈의 영향을 크게 받을 수 있고, decision tree의 깊이가 깊어짐에 따라 학습 데이터에 대해서 점점 높은 정확도를 가지게 되지만 새로운 데이터에 대한 성능이 떨어질 수 있습니다. 이러한 decision tree의 특성에 따라, 데이터의 feature scaling 및 feature selection을 통해 모델이 특정한 feature에 크게 의존하거나 복잡성이 높아지는 것을 예방하고, DecisionTreeClassifier의 다양한 매개변수(max_depth, min_samples_split, min_samples_leaf, max_leaf_nodes 등)를 활용하여 오버피팅을 완화할 수 있습니다.

💡 (TODO 2-2) 일반적으로 앙상블 모델은 다른 모델에 비해 일반화 성능이 좋습니다. 그 이유가 무엇인지 설명하고, 우리의 성능 평가 결과에서도 XGBoost가 Decision Tree보다 나은 일반화 성능을 보이는지 판단해주세요.

앙상블 모델은 다양한 모델을 결합함으로써 각각의 모델이 데이터를 학습하고 예측하면서 발생한 오류가 서로 상쇄될 수 있게 합니다. 이를 통해 데이터의 특정 노이즈나 이상치에 대한 민감도를 낮추어 오버피팅을 방지함으로써, 새로운 데이터에 대하여도 보다 안정적인 예측 성능을 보입니다.

본 과제의 성능 평가 결과에서도 앙상블 모델인 XGBoost가 Decision Tree에 비해 높은 일반화 성능을 보입니다. training accuracy는 XGBoost가 Decision Tree보다 낮지만, validation accuracy가 높고 두 accuracy 사이의 간극이 점점 좁아진다는 점에서 XGBoost의 일반화 성능이 더 높음을 알 수 있습니다.