Pair Programing

7주차

1조
201984014
성도윤
201904126
허준혁

Q1. 결정 트리 학습과 시각화에서 random_state 값에 42를 안넣으면 어떻게 될까? (다른 값)

A1. 꽃의 모양이 바뀐다
-> 원하는 값이 나오지 않는다
-> 모델의 학습 및 예측 결과가 다르게 나온다
-> 재현성을 잃게 된다

연습문제)

다음 단계를 따라 moons 데이터셋에 결정트리를 훈련시키고 세밀하게 튜닝하라

1. make_moons(n_sample=1000, noise=0.4)를 사용해 데이터셋을 생성한다
2. 이를 train_test_split()을 사용해 훈련 세트와 테스트 세트로 나눈다
3. DecisionTreeClassifier의 최적의 매개변수를 찾기 위해 교차 검증과 함께 그리드 탐색을 수행한다
(GridSearchCV를 사용하면 됨. 여러가지 max_leaf_nodes 값을 시도)
4. 찾은 매개변수를 사용해 전체 훈련 세트에 대해 모델을 훈련시키고 테스트 세트에서 성능을 측정한다
대략 85~87%의 정확도가 나옴

연습문제)

1. make_moons(n_sample=1000, noise=0.4)를 사용해 데이터 셋을 생성한다

```
1 from sklearn.datasets import make_moons
2
3 X, y = make_moons(n_samples=1000, noise=0.4, random_state=42)
```

2. 이를 train_test_split()을 사용해 훈련 세트와 테스트 세트로 나눈다

```
1 from sklearn.datasets import make_moons
2 from sklearn.model_selection import train_test_split
3
4 X, y = make_moons(n_samples=1000, noise=0.4, random_state=42)
5
6 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

3. DecisionTreeClassifier의 최적의 매개변수를 찾기 위해 교차 검증과 함께 그리드 탐색을 수행한다 (GridSearchCV를 사용하면 됨. 여러가지 max_leaf_nodes 값을 시도)

```python
from sklearn.datasets import make_moons
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV

X, y = make_moons(n_samples=1000, noise=0.4, random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

dt_classifier = DecisionTreeClassifier()

param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [None, 5, 10, 15],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}

grid_search = GridSearchCV(dt_classifier, param_grid, cv=5, scoring='accuracy')

grid_search.fit(X_train, y_train)

print("Best Parameters: ", grid_search.best_params_)
print("Best Score: ", grid_search.best_score_)

best_classifier = grid_search.best_estimator_

test_score = best_classifier.score(X_test, y_test)
print("Test Set Score: ", test_score)
```

4. 찾은 매개변수를 사용해 전체 훈련 세트에 대해 모델을 훈련시키고 테스트 세트에서 성능을 측정한다
대략 85~87%의 정확도가 나옴

```
Test Set Score:  0.84
```