Chapter 6 결정 트리 문제

6.1 dot 파일을 만들지 않고 바로 트리를 그릴 수 있는 함수는?

답: plot_tree( )

6.2 결정 트리의 장점 중 한 가지를 설명하세요.

답: 데이터 전처리가 거의 필요하지 않다. 특성의 스케일을 맞추거나 평균을 원점에 맞추는 작업이 필요없다. (예시)

6.3 결정 트리는 한 샘플이 특정 클래스 k에 속할 확률을 추정할 수 있다(O/X)

답: O

6.4 CART 알고리즘은 최적의 솔루션을 보장한다. (O/X)

답: X, CART 알고리즘은 그리디 알고리즘으로 노드에서 최적의 분할을 찾지만, 현재 단계의 분할이 몇 단계를 거쳐 가장 낮은 불순도로 이어질 수 있는지에 대한 문제는 고려하지 않는다. 그리디 알고리즘은 최적의 솔루션을 보장하지 않는다.

6.5 훈련 세트의 개수와 상관없이 미리 데이터를 정렬하면 훈련 속도를 높일 수 있다. (O/X)

답: X, 수천 개 이하의 샘플 정도로 작은 경우 미리 데이터를 정렬하면 훈련 속도를 높일 수 있지만, 훈련 세트가 크다면 속도가 많이 느려진다.

6.6 지니 불순도와 엔트로피 불순도를 사용할 때의 차이점을 장단점을 중심을 설명하세요.

답: 지니 불순도가 조금 더 계산이 빠르기 때문에 기본값으로 좋다. 그러나 다른 트리가 만들어지는 경우에는 지니 불순도가 가장 빈도 높은 클래스를 한쪽 가지로 고립시키는 경향이 있는 반면에 엔트로피는 조금 더 균형 잡힌 트리를 만든다.

6.7 비파라미터 모델과 파라미터 모델에 대해 설명하세요.

비파라미터 모델: 훈련되기 전에 파라미터 수가 결정되지 않는 모델

파라미터 모델: 미리 정의된 모델 파라미터 수를 가지므로 자유도가 제한되고 과대 적합될 위험이 줄어든다. 반면 과소 적합될 위험은 커진다.

6.8 결정 트리에서 회귀 작업을 할 때 과대적합되기 쉬워서 규제가 필요합니다. 이 때 해결할 수 있는 방법 중 한 가지 방법에 대해서 설명하세요.

답: min_samples_leaf = 10 과 같이 리프노드가 가지고 있어야 할 최소 샘플 수를 지정해 주면 과대적합을 완화시켜준다.

6.9 결정 트리의 제한 사항에 대해(두 가지) 설명하세요.

답: 결정 트리는 계단 모양의 결정 경계를 만들어서 훈련 세트의 회전에 민감하다. 훈련 데이터에 있는 작은 변화에도 민감하다.