<머신러닝 톺아보기> 6주차 과제

박민서

1. 의사결정트리: 각 데이터들이 가진 속성들로부터 패턴을 예측 가능한 규칙들의 조합으로 나타내며, 이를 바탕으로 분류를 수행할 수 있도록 하는 지도학습 모델, 의사결정나무는 분류와 회귀 모두 가능한 모델로, 범주형과 연속형 수치 모두 예측 가능
2. 의사결정트리의 분류와 회귀 결과: 의사결정트리의 분류와 회귀는 끝마디의 어떤 값을 반환하느냐의 차이점을 가진다. 분류의 경우, 새로운 데이터가 속하는 해당 끝마디에서 가장 빈도가 높은 범주로 새로운 데이터를 분류한다. 회귀의 경우, 해당 끝마디의 종속변수의 평균을 예측값을 반환한다. (예측값의 종류는 끝마디의 개수와 일치함)
3. 의사결정트리 수행 과정
   (1) 의사결정나무 형성: 분석의 목적과 데이터 구조에 따라 적절한 분리기준과 정지규칙을 지정하여 의사결정나무 모형을 생성
   -분리기준: 하나의 부모마디로부터 자식마디가 형성될 때, 어떤 입력변수로 어떻게 분리하는 것이 목표변수를 가장 잘 분류하는지를 파악하는 기준(ex. 순수도, 불순도), 부모마디의 순수도에 비해, 자식마디들의 순수도가 증가하도록(불순도가 감소하도록) 자식마디를 형성
   -정지규칙: 더 이상 분리가 일어나지 않고 현재의 마디가 끝마디가 되도록 하는 규칙, 깊이나 끝마디의 개수가 몇 개가 될 때까지 나눌 것인지 정하여 규칙 설정
   (2) 가지치기: 분류오류를 크게 할 위험이 높거나 부적절한 추론 규칙을 가지고 있는 가지를 제거하는 단계, 의사결정나무의 분기 수가 증가하면, 새로운 데이터에 대한 오분류율이 감소하지만, 일정 수준 이상이 되면 오분류율이 증가하는 현상 발생=과적합 발생, 즉 적절한 가지치기를 통해 과적합을 막을 수 있음.
   (3) 타당성 평가, 해석 및 예측
4. 의사결정트리 장단점
   (1) 장점: 사용자가 모형을 쉽게 이해 가능(해석의 용이성), 두 개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 미치는지 알 수 있음(교호작용의 해석), 선형성, 정규성, 혹은 등분산성 등의 가정이 필요하지 않음(비모수적 모형)
   (2) 단점: 연속형 변수를 비연속적 값으로 취급하기 때문에 분리의 경계 부근에서 예측 오류가 클 가능성이 있음(비연속성), 각 변수의 목표변수에 대한 영향력을 알 수 없음(선형성/주효과의 결여), 훈련용 데이터에 의존하므로 새로운 데이터의 예측에서 불안정할 수 있음(비안정성)

5. 결정트리 훈련과 활용: 결정트리 방식은 일반적으로 데이터 전처리가 불필요
   (1) 결정트리 시각화
   (2) 지니 불순도
   (3) 범주 예측
   (4) 범주에 속할 확률: 주어진 샘플에 대해 예측된 노드에 속한 샘플들의 범주별 비율
6. CART 알고리즘(classification and regression tree): 각 노드에서 비용함수를 최소화하는 특성 k와 특성의 임계값 tk를 결정해서 사용함, 비용함수가 작을수록 불순도가 낮은 두 개의 부분집합으로 분할됨. 분할 과정 반복: 규제의 한계에 다다르거나 더 이상 불순도를 줄이는 분할이 불가능할 때까지 진행
7. 엔트로피: 지니 불순도 대신에 샘플들의 무질서 정도를 측정하는 엔트로피 사용, 지니 불순도를 사용할 때와 비교해서 큰 차이가 나지 않음, 엔트로피 방식이 노드를 보다 균형 잡힌 두 개의 자식 노드로 분할함, 두 방식이 큰 차이가 없고 지니 불순도 방식보다 빠르게 훈련되어 기본값으로 사용함.
8. 비파라미터 모델: 결정트리 모델은 데이터에 대한 어떤 가정도 하지 않음, 노드를 분할할 때 어떤 제한도 가해지지 않으며, 노드를 분할할 때마다 새로운 파라미터가 학습되기 때문에 학습되어야 하는 파라미터의 개수를 미리 알 수 없다. 이러한 모델을 비파라미터 모델이라 부른다. 비파라미터 모델의 자유도는 제한되지 않기에 과대적합될 가능성이 높다.
9. 파라미터 모델: 선형 모델과 같은 모델은 파라미터 수가 훈련 시작 전에 규정되기에 과대적합 가능성이 상대적으로 적어진다.
10. 회귀 결정트리: 결정트리 알고리즘 아이디어를 그대로 이용하여 회귀 문제에 적용 가능, 잡음이 포함된 2차 함수 형태의 데이터셋을 이용하여 결정트리 회귀 모델을 훈련시킬 수 있음.
11. 규제: 분류의 경우처럼 규제가 없으면 과대적합이 발생할 수 있음.
12. 결정트리의 단점
    (1) 훈련셋 회전 민감도: 결정트리 알고리즘은 성능이 매우 우수하지만 기본적으로 주어진 훈련셋에 민감하게 반응함, 결정트리는 항상 축에 수직인 분할을 사용, 따라서 조금만 회전을 가해도 결정 경계가 많이 달라짐.
    (2) 높은 분산: 훈련 데이터의 작은 변화에도 매우 민감함, random_state를 지정하지 않음ㄴ서 동일한 모델을 훈련시키면 다른 결과가 나옴, 높은 분산 문제는 여러 개의 결정트리를 동시에 훈련시킨 후 평균값을 활용하는 랜덤 포레스트 모델을 이용하면 해결할 수 있다.