**Advance Data Mining**

**경희대학교 산업경영공학과**

Reading Assignment

**Chapter 6**



고급데이터마이닝

진창호

산업경영공학과

2016100937

김성수

과목명

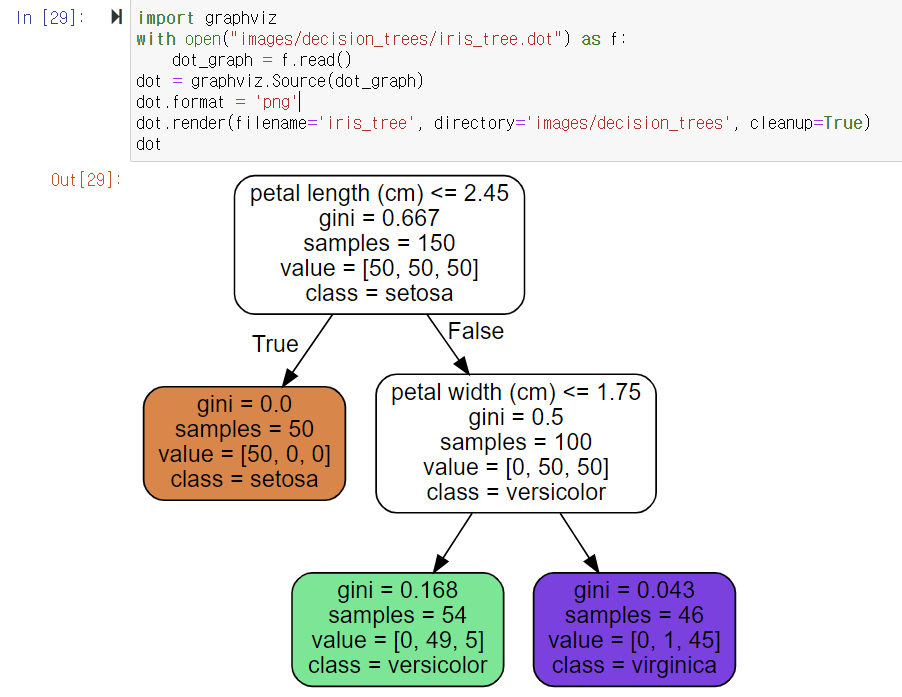
담당교수

학과

학번

이름

**Chapter6**

1. 결정 트리(Decision Tree)  
   ■ 데이터의 전처리가 거의 필요하지 않으며, 특성의 스케일을 맞추거나 평균을 원점에  
    맞추는 작업이 필요 없음  
   ■ 사이킷런은 CART 알고리즘을 사용하여 이진 트리만 만들지만 ID3는 둘 이상의 자식  
    노드를 가진 결정 트리를 만드는 것이 가능  
   ■ 직관적이고 결정 방식을 이해하기 쉬운 화이트박스 모델  
   ■ 각 클래스에 속할 확률을 각각 추정 가능  
   ■ 계단 모양의 결정 경계를 만들어 훈련 세트의 회전 및 작은 변화에도 매우 민감  
    🡪 PCA나 랜덤 포레스트로 이러한 불안정성 극복 가능
2. 결정 트리의 시각화  
     
   ■ 해석: 꽃잎의 길이가 2.45cm보다 짧은가? 🡪 짧으면 왼쪽으로 이동 후 ‘setosa’로 분류, 그렇지 않으면 오른쪽으로 이동 후 ‘꽃잎의 너비가 1.75cm보다 작은지’ 검사 🡪 작다면 ‘versicolor’로 분류, 그렇지 않다면 ‘virginica’로 분류  
   ■ 용어  
    ①gini: 불순도를 측정하여, 모든 샘플이 같은 클래스에 속해 있다면 0의 값을 가짐  
    ②samples: 얼마나 많은 훈련 샘플이 적용되었는가 기입한 것  
    ③value: 노드에서 각 클래스에 얼마나 많은 훈련샘플이 있는지 알려줌
3. CART 알고리즘  
   ■ 훈련 세트를 특성 k의 임계값 를 사용해 두 개의 서브셋으로 나누어 이진 트리 생성  
    🡪 여기서 (k, )의 선정 기준은 가장 순수한 서브셋으로 나눌 수 있는 값  
   ■ 미리 정의한 최대 깊이가 되면 중지하거나 불순도를 줄이는 분할이 불가능할 때 중지  
   ■ greedy 알고리즘이며, 최적의 트리를 찾는 것은 NP-Complete 문제
4. 지니 불순도 VS 엔트로피  
   ■ 지니 불순도: 계산이 빠르나 가장 빈도 높은 클래스를 한쪽으로 고립시키는 경향이 있음  
   ■ 엔트로피: 지니 불순도에 비해 균형 잡힌 트리를 만듦
5. 규제 매개변수  
   ■ 결정 트리는 훈련되기 전에 파라미터 수가 결정되지 않는 비파라미터 모델로, 모델 구조가  
    자유로워 훈련 데이터에 과대적합 되기 쉬움  
   ■ 훈련 데이터에 과대적합을 피하기 위한 규제 방안  
    (1) 파라미터를 활용한 규제  
    ① max\_depth: 결정 트리의 최대 깊이 제한  
    ② min\_samples\_split: 분할되기 위해 노드가 가져야 하는 최소 샘플 수  
    ③ min\_samples\_leaf: 리프 노드가 가지고 있어야 할 최소 샘플 수  
    ④ min\_weight\_fraction\_leaf: 가중치가 부여된 전체 샘플 수에서의 비율  
    ⑤ max\_leaf\_nodes: 리프 노드의 최대 수  
    ⑥ max\_features: 각 노드에서 분할에 사용할 특성의 최대 수  
    (2) 가지치기  
    - 제한 없이 결정 트리를 훈련시키고 불필요한 노드를 제거하는 방법  
    - 가지치기 후 통계적 검정으로 불필요한 노드가 모두 없어질 때까지 계속 실시
6. 결정 트리의 회귀에 활용  
   ■ 각 노드에서 클래스를 예측하는 대신 어떤 값을 예측하는 것으로 활용 가능  
   ■ 불순도를 최소화하는 방향으로 분할하는 대신 MSE를 최소화하도록 분할