

머신러닝과 빅데이터분석®

9주차 의사결정트리와 랜덤포레스트

박길식 교수



고려사이버대학교
THE CYBER UNIVERSITY OF KOREA



학습 목표

 의사결정트리를 이해하고 구현할 수 있다.

 랜덤포레스트를 이해하고 구현할 수 있다.



학습 목차

- 1 의사결정트리와 랜덤포레스트
- 2 의사결정트리와 랜덤포레스트 실습

CHAPTER

의사결정트리와 랜덤포레스트

의사결정트리

일련의 분류 규칙을 통해 데이터를 분류, 회귀하는
지도 학습 모델

- 의사결정 규칙을 나무 구조로 나타내어 여러 가지 규칙을
순차적으로 적용하면서 독립 변수 공간을 분할하는 분류 모델

⌘ 분류(Classification)과 회귀(Regression) 문제에 모두 사용할 수
있으므로, CART(Classification And Regression Tree)라고도 부름

[01] 의사결정트리



분류 문제 해석하기

» 운동 경기가 열렸다면,

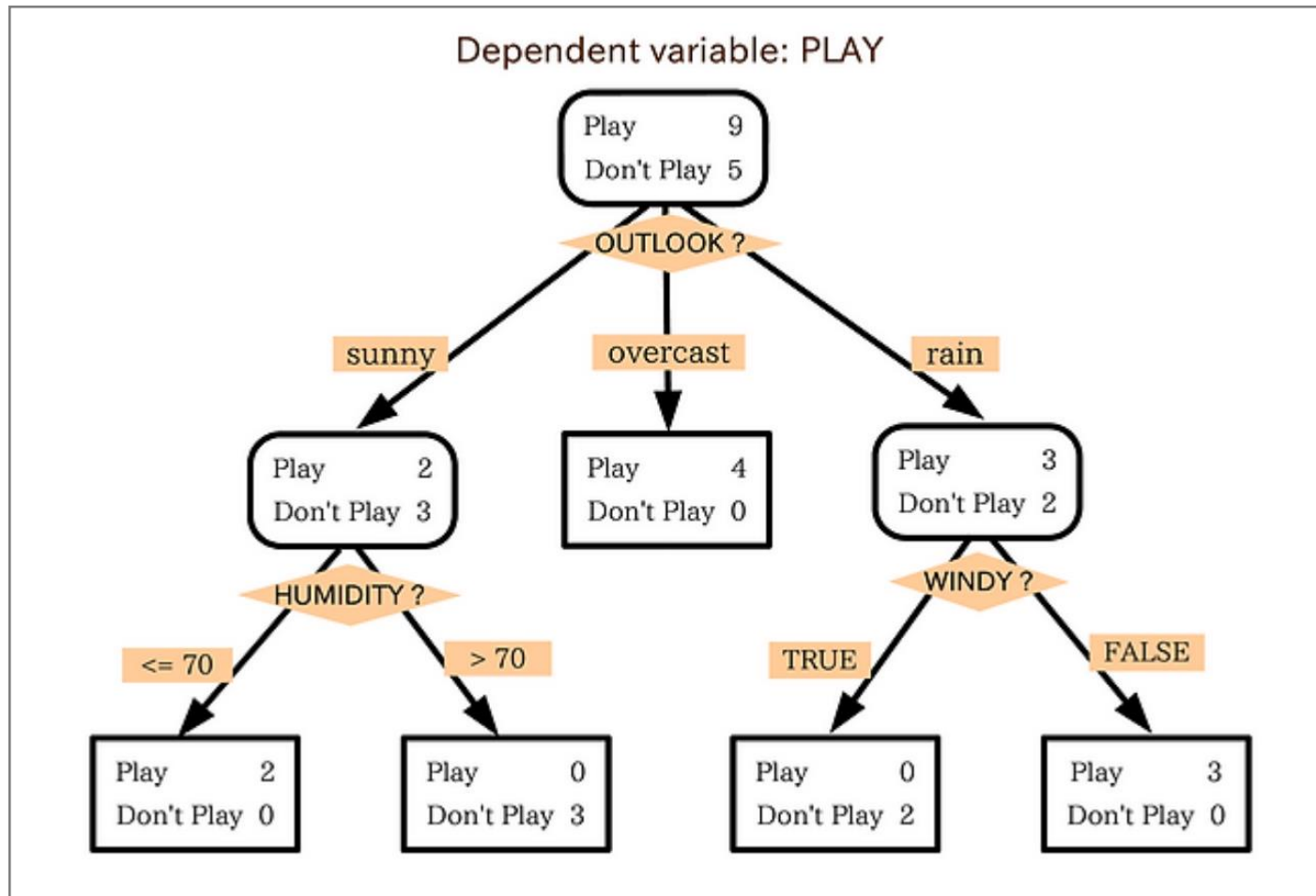
Play = 1

» 운동 경기가 열리지

Play = 0 (Don't Play)

» 날씨가 맑고(Sunny)
습도(Humidity)가
70 이하인 날엔 경기가 열림

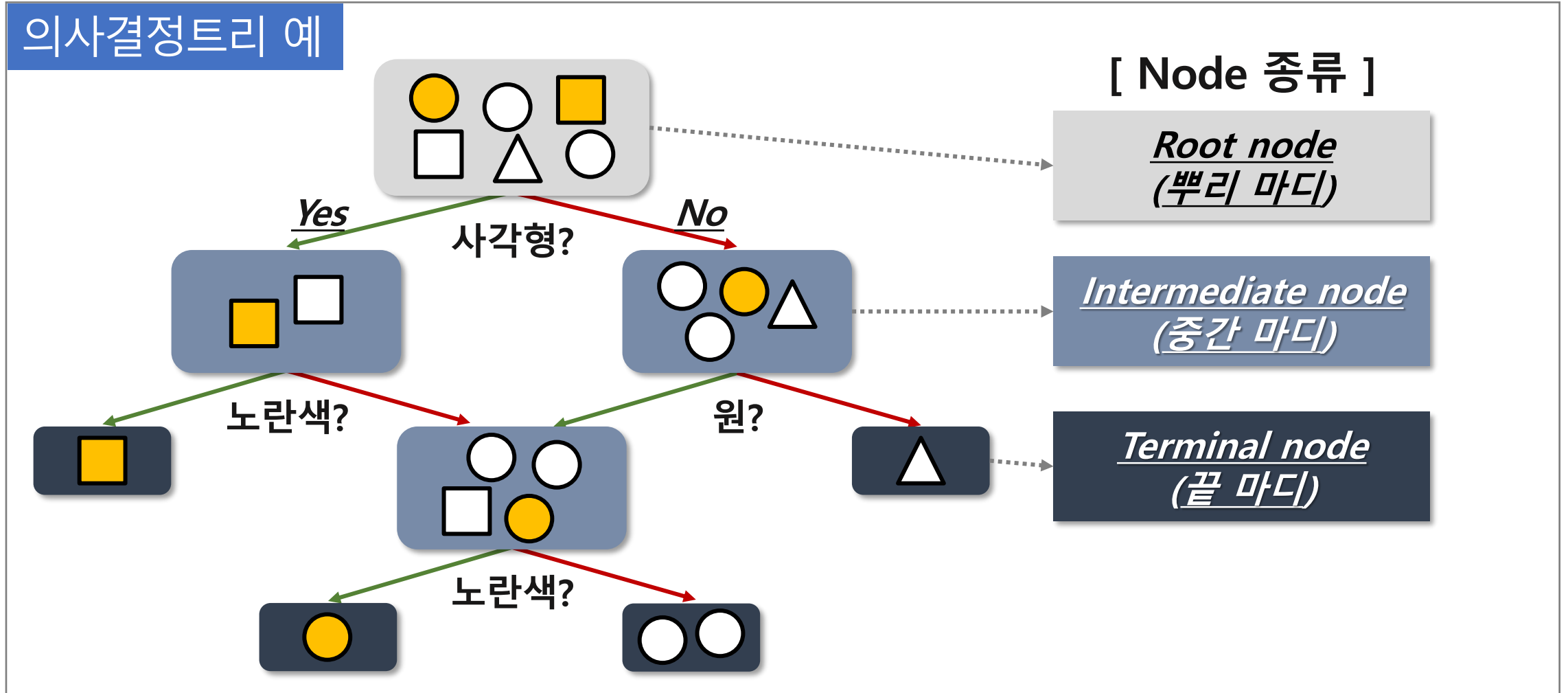
» 비가 오고(Rain)
바람이 부는(Windy) 날엔
경기가 열리지 않음



[01] 의사결정트리



의사결정트리 구조



[01] 의사결정트리



의사결정트리의 장 · 단점

장점

- 모델 구조가 단순하여 해석이 용이하고 유용한 독립변수를 파악하기가 쉬움
- 독립변수 간의 상호작용 및 비선형성을 고려하여 분석이 수행되므로 수학적 가정이 불필요한 비모수적 모델

단점

- 분류 기준값 경계선 근방의 데이터 값에 관해서는 오차가 클 수 있음
- 새로운 데이터에 대한 예측이 불안정함

[02] 랜덤포레스트



의사결정트리의 한계

1 의사결정트리는 예측 성능이 낮음

- 결정경계 특징 때문에 특정 데이터에서만 만족스러운 성능을 도출함 → 과적합이 되기 쉬움

2 과적합(Overfitting)

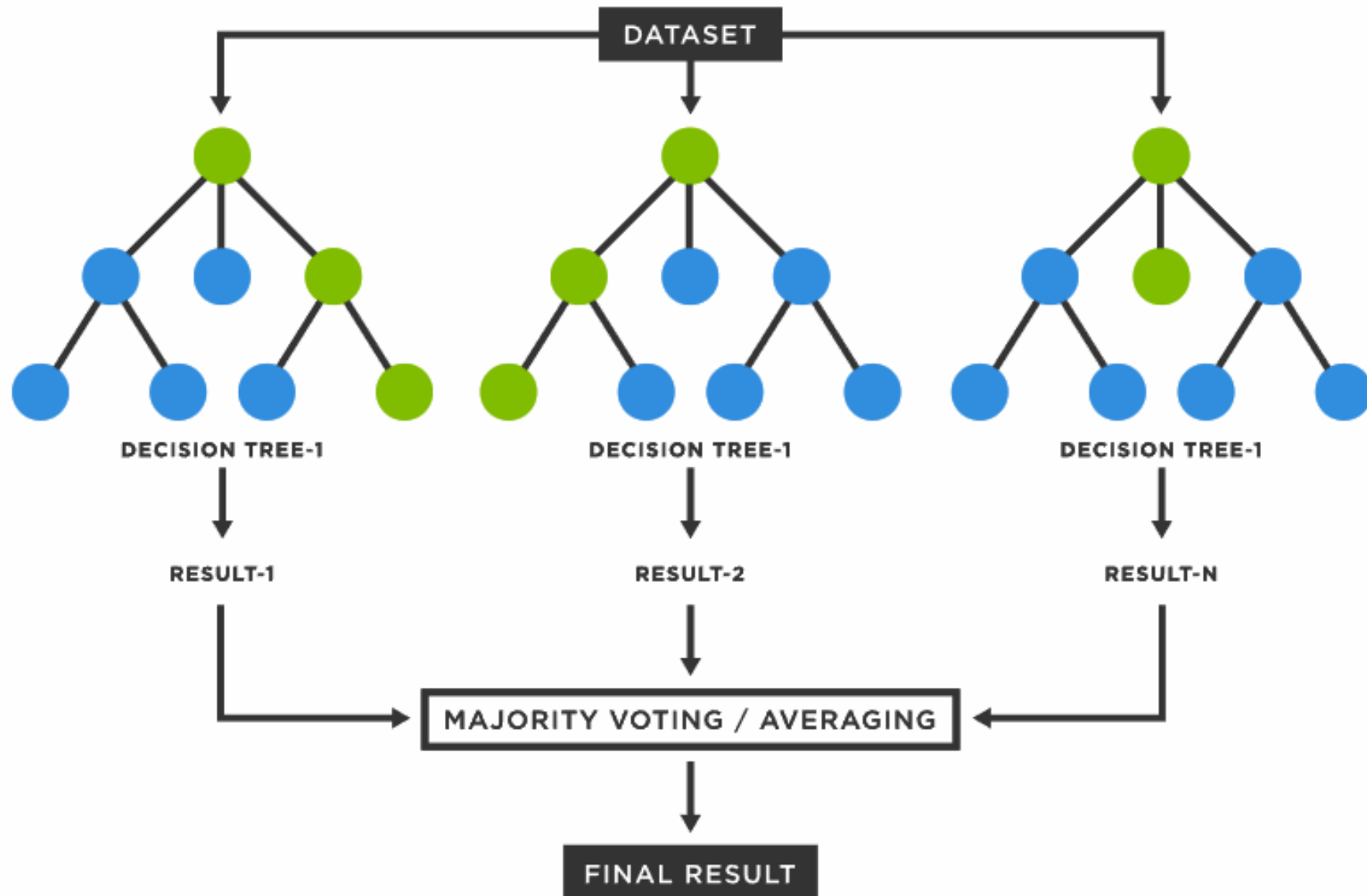
3 지도학습에서 훈련 집합에 대해서는 만족스러운 성능을 도출하나

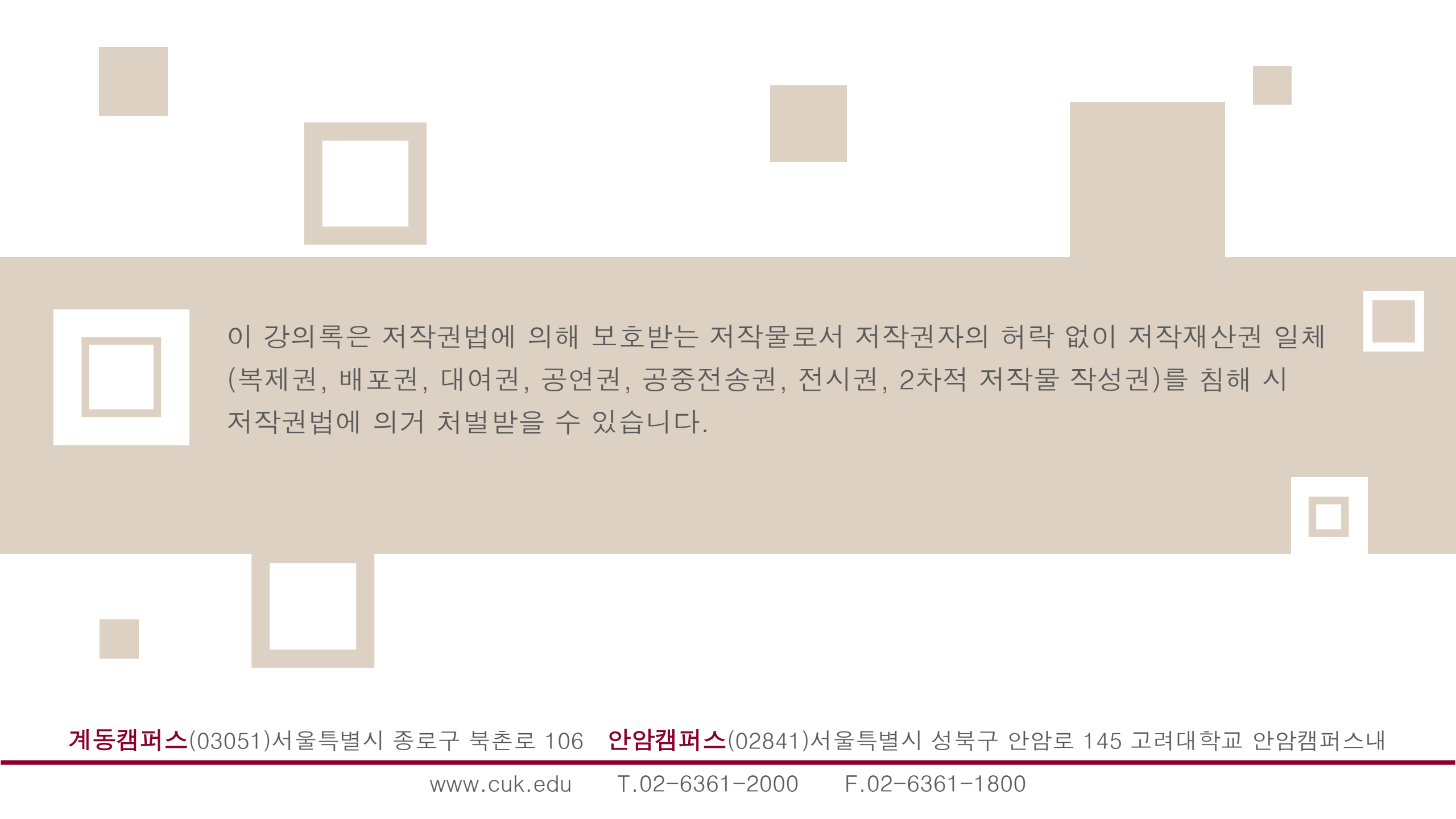
평가 집합에 대해서는 불만족스러운 성능을 도출하는 것
→ 모델의 일반화가 어려움

02 랜덤포레스트

- ② 의사결정트리를 만들 때 난수를 사용하므로 '임의의' 또는 '무작위(Random)'라는 용어 사용
- ② 나무를 여러 개 사용하므로 '숲(Forest)'이라는 용어 사용
- ② 수많은 의사결정트리(Decision Tree)가 모여서 생성
- ② 의사결정트리에서는 가장 좋은 질문을 노드에 부여하는 반면, 랜덤 포레스트는 가장 좋은 k개의 후보 질문 중에서 랜덤하게 선택

[02] 랜덤포레스트





이 강의록은 저작권법에 의해 보호받는 저작물로서 저작권자의 허락 없이 저작재산권 일체 (복제권, 배포권, 대여권, 공연권, 공중전송권, 전시권, 2차적 저작물 작성권)를 침해 시 저작권법에 의거 처벌받을 수 있습니다.