

(Decision) Tree model

심동녘

건국대학교 산업경영공학부 신산업융합학과

sk4me@konkuk.ac.kr

ANALYSIS & DATA

0 학습 할 주요 주제

의사결정 나무(Decision Tree) 강의 개요

- 의사결정 나무(Decision Tree)

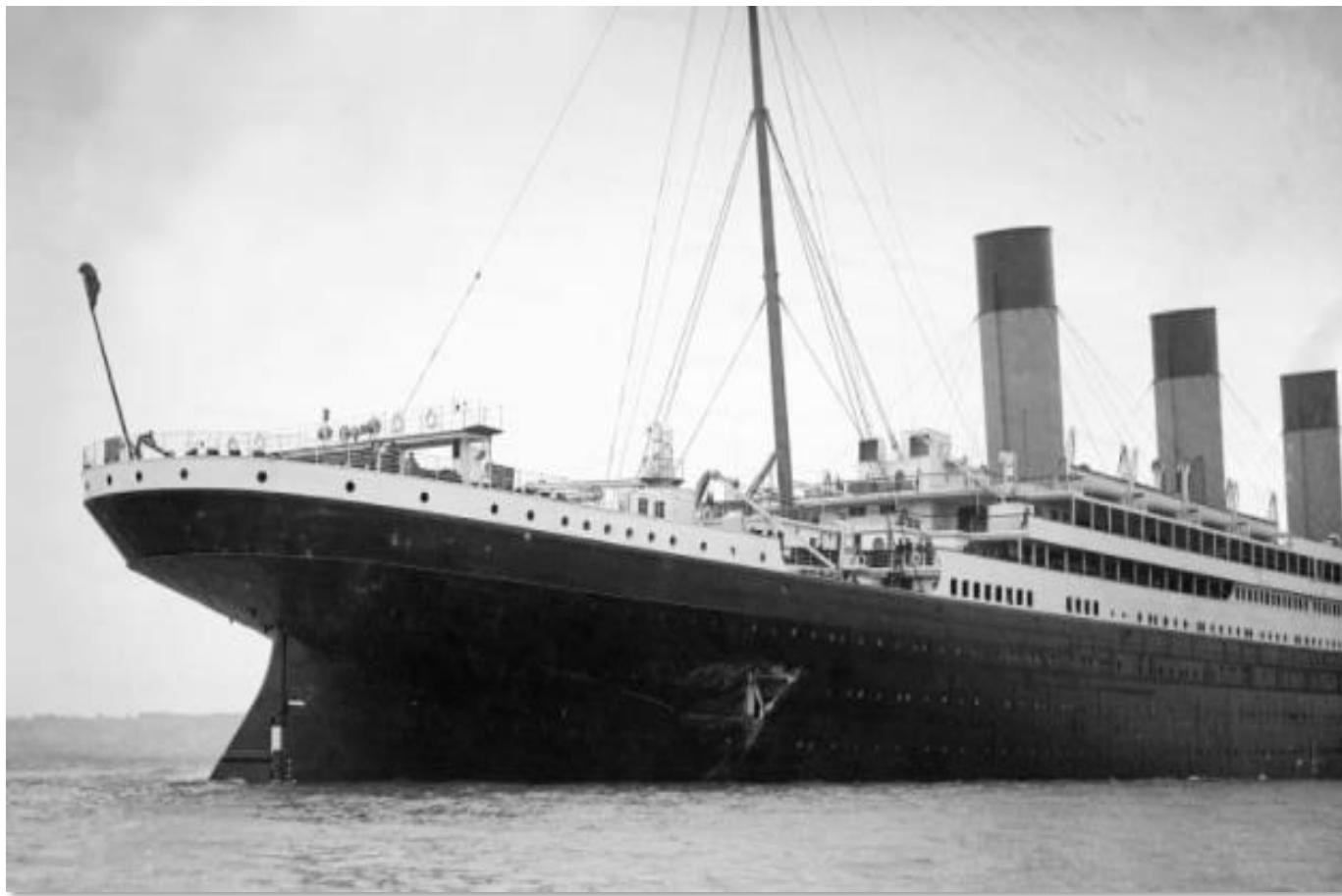
- 목적
- 모형 생성 방법
 - 재귀적 분기 + 사후적 가지치기(post-pruning)
 - 사전적 가지치기
- 의사결정나무 예시

- 회귀 나무(Regression Tree)

- 목적과 배경
- 모형생성방법
- 로지스틱 회귀모델 예제

0 학습 할 주요 주제

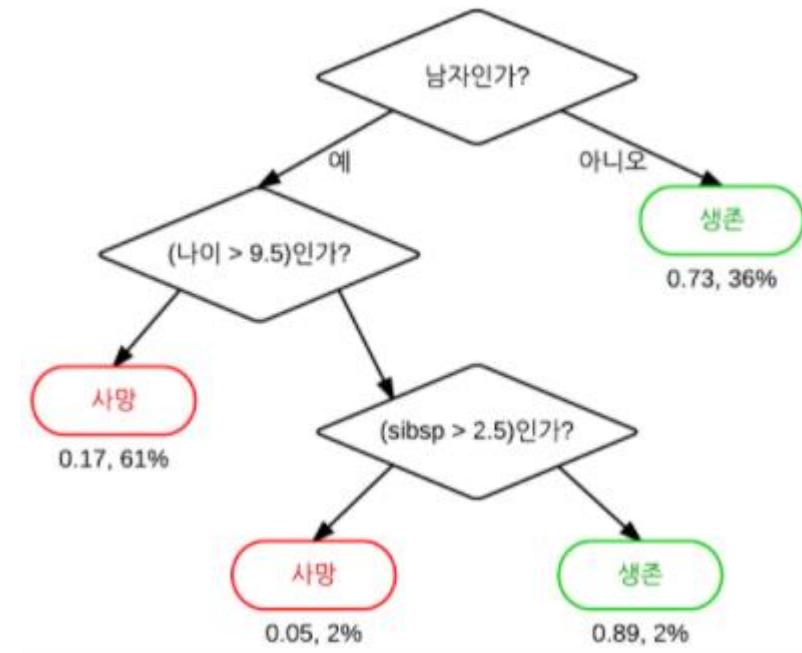
Motivation



0 학습 할 주요 주제

Classification

- survived : 목표변수, 종속변수, 생존여부(0:사망, 1:생존)
- pclass : 좌석 등급
- name : 탑승객 이름
- gender : 성별
- age : 나이
- sibsp : 함께 탑승한 형제수
- parch : 함께 탑승한 부모수
- ticket : 티켓 번호
- fare : 탑승권 가격
- cabin : 선실 번호
- embarked : 탑승 승착장

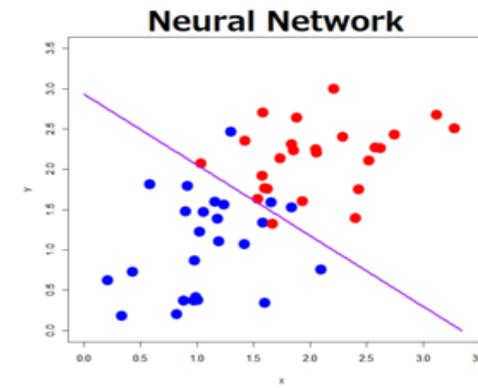
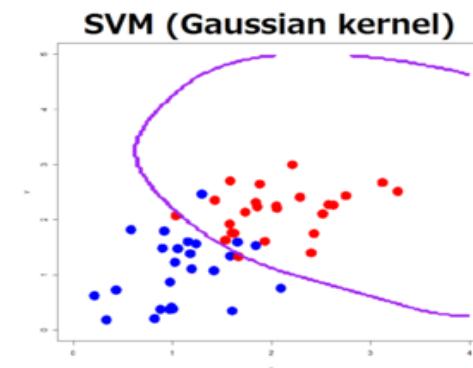
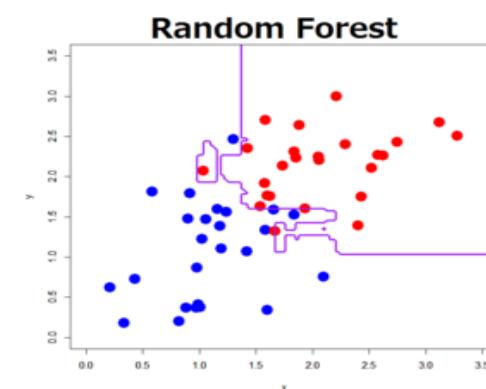
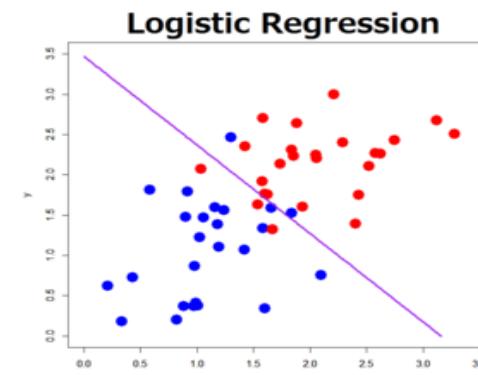
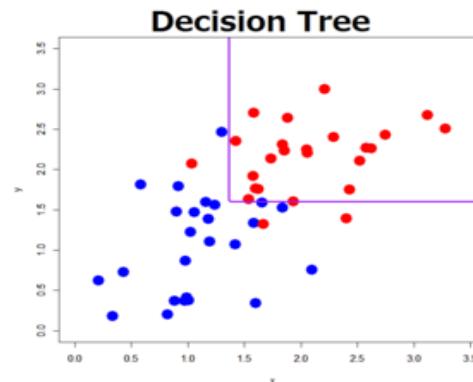


의사결정 나무

1 의사결정 나무 개요

목적

- 여러 가지 머신러닝 알고리즘이 필요한 이유는 무엇인가?



1 의사결정 나무 개요

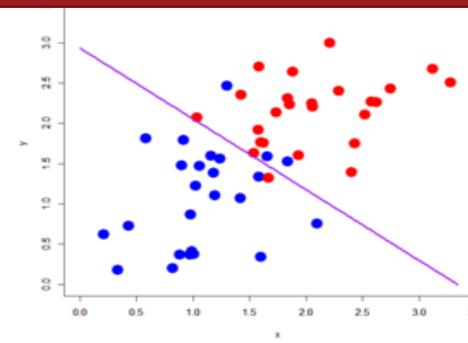
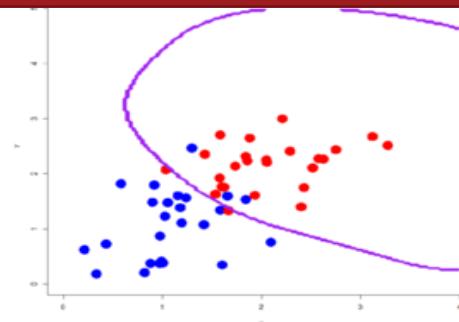
목적

- 여러 가지 머신러닝 알고리즘이 필요한 이유는 무엇인가?



특정 알고리즘이 모든 상황에서 다른 알고리즘 보다 우월하다는 결론을 내릴 수 없음

그때 그때 달라요~

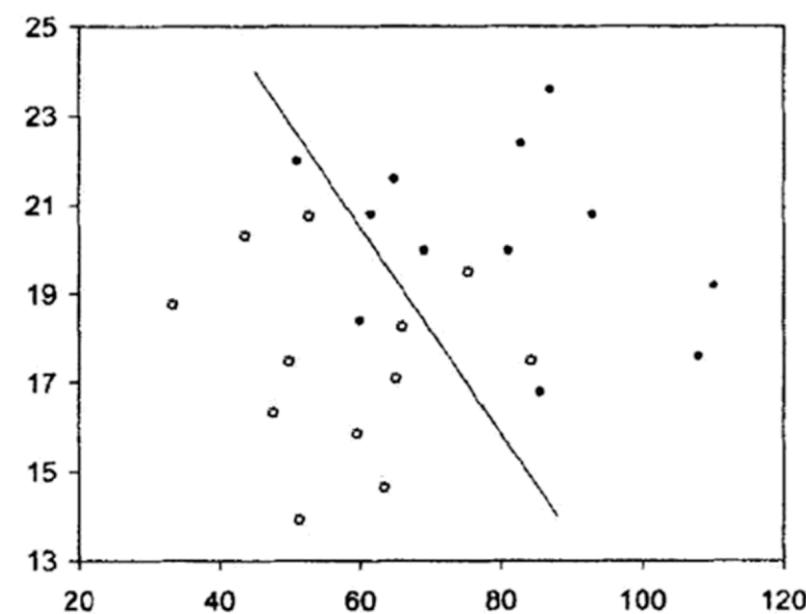
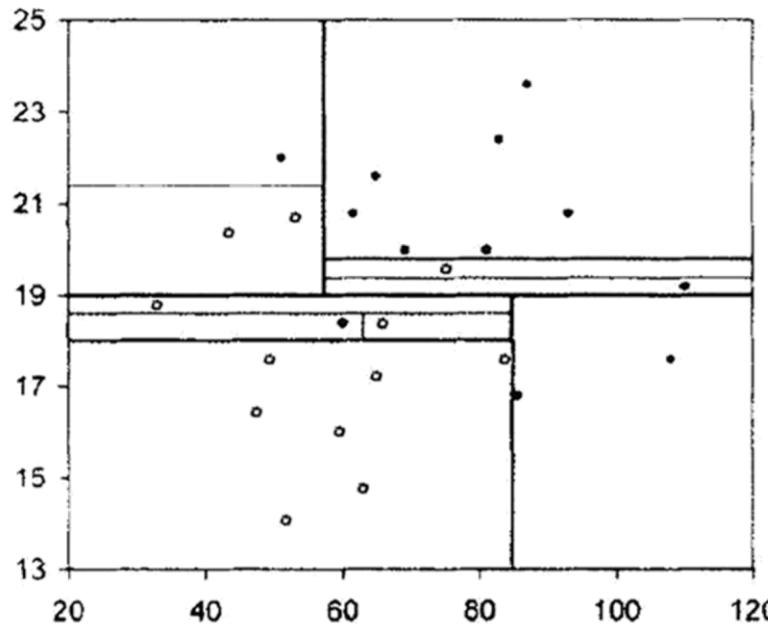


1 의사결정 나무 개요

분류 알고리즘

- 머신러닝 알고리즘은 여러 가지가 존재
- 동일한 결과를 얻기 위한 다양한 ‘길’이 존재하기 때문

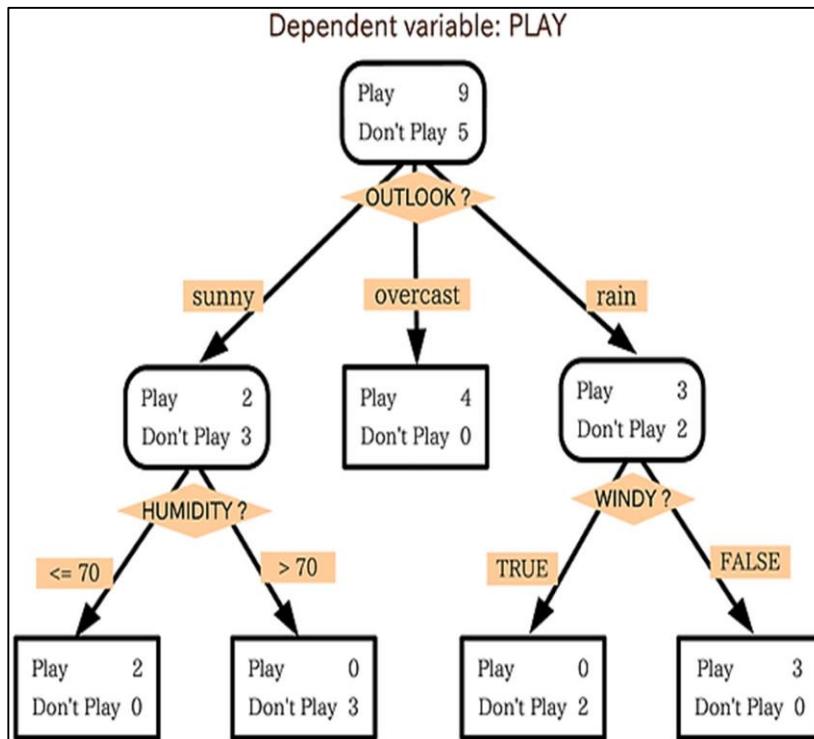
“Separate the riding mower buyers(●) from non-buyers(○)”



1 의사결정 나무 개요

목적

- 한 번에 하나씩의 설명변수를 사용하여 정확한 예측이 가능한 규칙들의 집합을 생성
- 최종 결과물은 나무를 뒤집어 놓은 형태인 규칙들의 집합



규칙 예시

만일 내일 날씨가 맑고 습도가 70% 이하이면 아이는 밖에 나가서 놀 것이다.

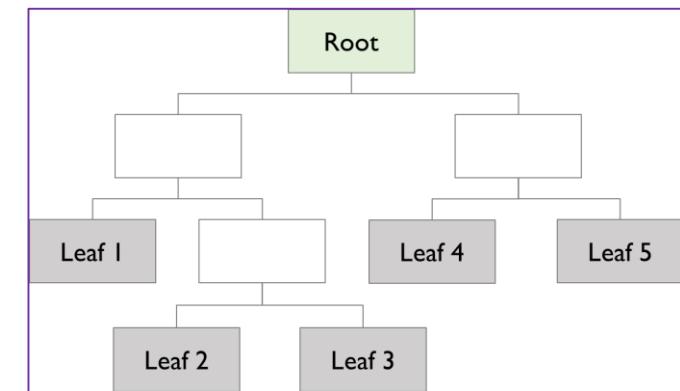
Or

만일 내일 비가 오고 바람이 불면 아이는 밖에 나가 놀지 않을 것이다.

1 의사결정 나무 개요

용어(Terminologies)

- 한 번에 하나씩의 설명변수를 사용하여 정확한 예측이 가능한 규칙들의 집합을 생성
- 노드(Node):**
 - 입력 데이터 공간의 특정 영역
- 부모 노드(Parent node):**
 - 분기split 전 노드
- 자식 노드(Child node):** 부모 노드로부터 분기 후 파생된 노드
- 분기 기준(Split criterion):**
 - 한 부모 노드를 두 개 이상의 자식 노드들로 분기하는데 사용되는 변수 및 기준 값
- 시작/뿌리 노드(Root node):**
 - 전체 데이터를 포함하는 노드, 자식 노드만 존재하며 부모 노드가 존재하지 않음
- 말단/잎새 노드(Leaf node):**
 - 더 이상 분기가 수행되지 않는 노드, 부모 노드만 존재하며 자식 노드가 존재하지 않음



1 의사결정 나무 개요

왜 의사결정 나무인가?

- 결과를 사람이 이해할 수 있는 규칙의 형태로 제공함
- 데이터의 사전 전처리를 최소화함 (정규화/결측치 처리 등을 하지 않아도 됨)
- 수치형 변수와 범주형 변수를 함께 다룰 수 있음

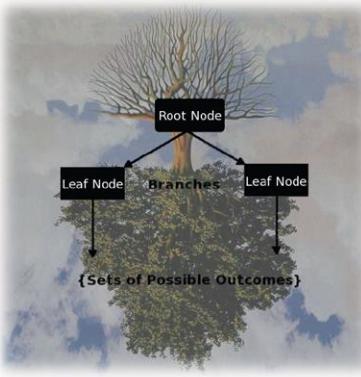
핵심 아이디어(Idea)

• 재귀적 분기: Recursive Partitioning

- 입력 변수의 영역을 두 개로 구분 → 구분하기 전보다 구분된 뒤에 각 영역의 순도(purity, homogeneity)가 증가하도록

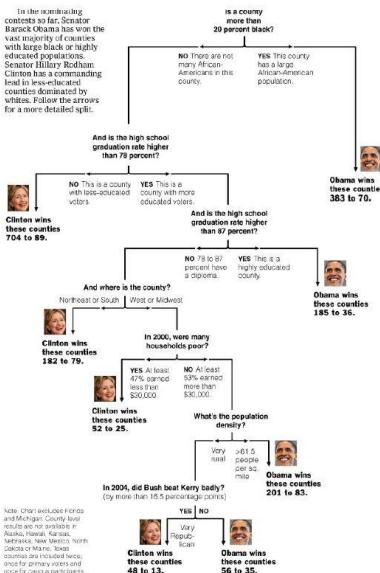
• 가지치기: Pruning the Tree

- 과적합(Overfitting)을 방지하기 위하여 너무 자세하게 구분된 영역을 통합



Classification and Regression Tree (CART)

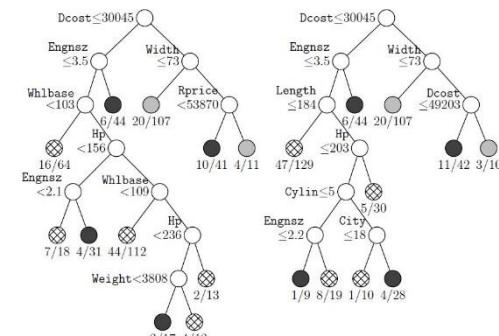
- ✓ 개별 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 규칙을 생성하는 지도학습 기법 (Breiman, 1984)
- ✓ If-then 형식으로 표현되는 규칙(rules)을 생성함으로써, 결과에 대한 예측과 함께 그 이유를 설명할 수 있는 장점이 있음
- ✓ 수치형 변수와 범주형 변수에 대한 동시 처리 가능

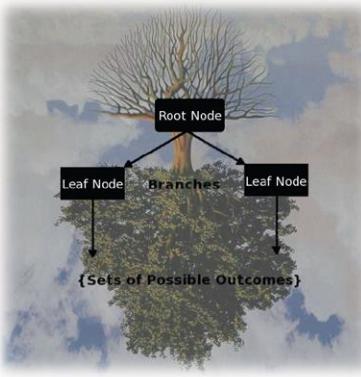


재귀적 분기 (Recursive Partitioning)

- 특정 영역(부모 노드)에 속하는 개체들을 하나의 기준 변수 값의 범위에 따라 분기
- 분기에 의해 새로 생성된 자식 노드의 동질성이 최대화 되도록 분기점 선택
- 불순도를 측정하는 기준으로는 범주형 변수에 대해서는 지니계수, 수치형 변수에 대해서는 분산을 이용

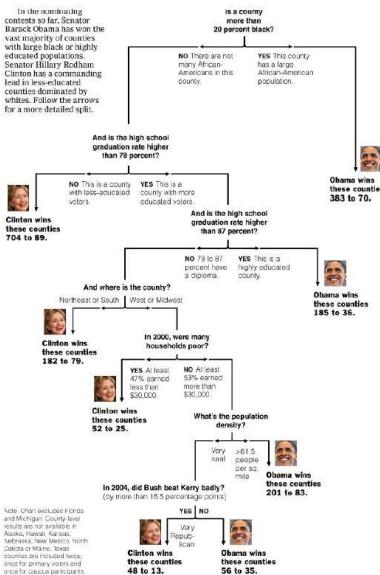
가지치기 (Pruning)



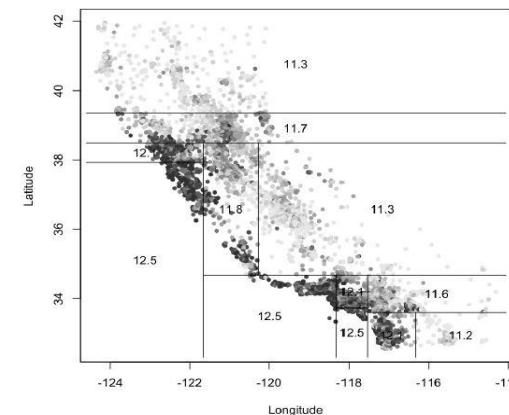


Classification and Regression Tree (CART)

- ✓ 개별 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 규칙을 생성하는 지도학습 기법 (Breiman, 1984)
- ✓ If-then 형식으로 표현되는 규칙 (rules)을 생성함으로써, 결과에 대한 예측과 함께 그 이유를 설명할 수 있는 장점이 있음
- ✓ 수치형 변수와 범주형 변수에 대한 동시 처리 가능



재귀적 분기 (Recursive Partitioning)



가지치기 (Pruning)

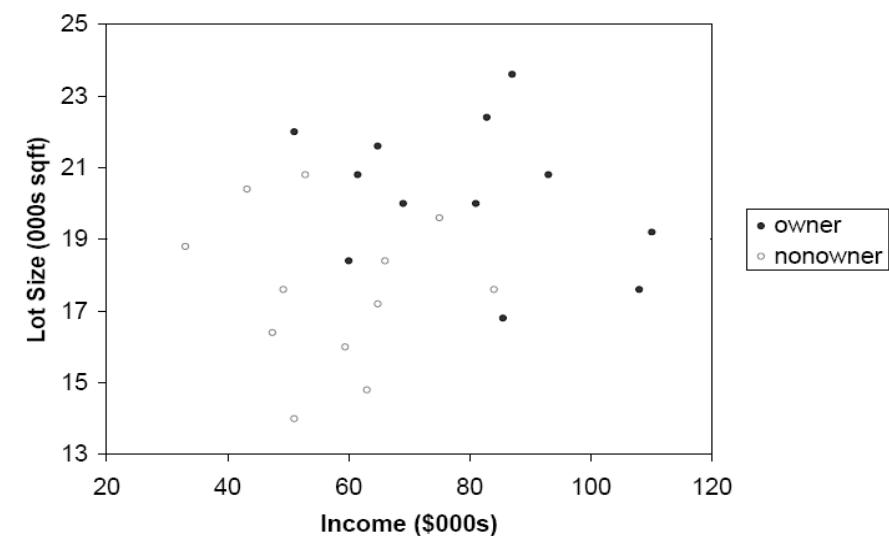
- 과적합(Over-fitting)을 방지하기 위하여 하위 노드들을 상위 노드로 결합
- Pre-pruning: Tree를 생성하는 과정에서 최소 분기 기준을 이용하는 사전적 가지치기
- Post-pruning: Full-tree 생성 후, 검증 데이터의 오분류율과 Tree의 복잡도(말단 노드의 수) 등을 고려하는 사후적 가지치기

2 모델 생성 방법

예시: 잔디깎기 기계 구입 예측

- 목적: 24개의 가정에 대해 잔디깎기 기계 구입 여부를 분류
- 설명변수: 수입(Income), 집의 크기(Lot size)

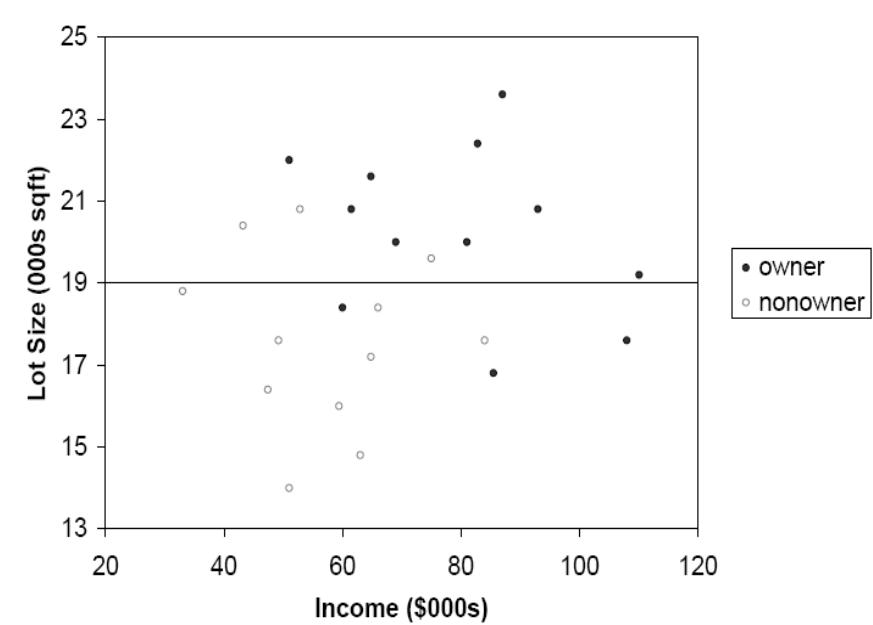
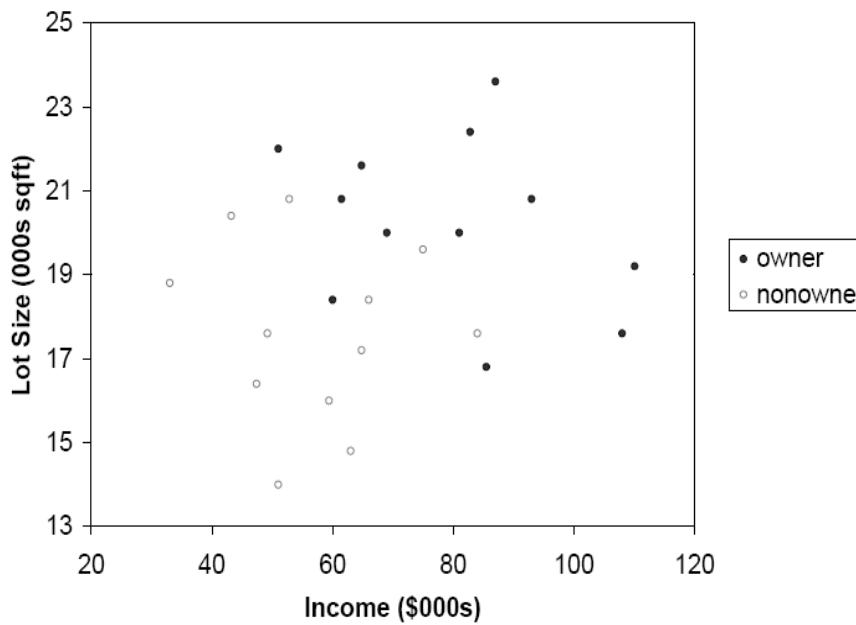
Income	Lot Size	Ownership	Income	Lot Size	Ownership
60.0	18.4	Owner	75.0	19.6	Non-owner
85.5	16.8	Owner	52.8	20.8	Non-owner
64.8	21.6	Owner	64.8	17.2	Non-owner
61.5	20.8	Owner	43.2	20.4	Non-owner
87.0	23.6	Owner	84.0	17.6	Non-owner
110.1	19.2	Owner	49.2	17.6	Non-owner
108.0	17.6	Owner	59.4	16.0	Non-owner
82.8	22.4	Owner	66.0	18.4	Non-owner
69.0	20.0	Owner	47.4	16.4	Non-owner
93.0	20.8	Owner	33.0	18.8	Non-owner
51.0	22.0	Owner	51.0	14.0	Non-owner
81.0	20.0	Owner	63.0	14.8	Non-owner



2 모델 생성 방법

예시: 잔디깎기 기계 구입 예측

- 전체 영역을 어떻게 분할해야 하는가?
 - 분할 후 각 노드에는 한 범주에 속하는 객체들이 다수 존재하도록 분할하는 것이 좋음
- 이러한 개념을 하나의 숫자로 어떻게 정량화할 것인가?
 - 아래 예시에서 분기 기준 1과 분기 기준 2의 효과는 어떻게 측정하며 둘 중 어느 분기 기준이 더 좋은 기준인가?
 - E.g.) 회귀분석에서 RSS(Residual sum of square)와 유사한 개념을 만들어 내야 함



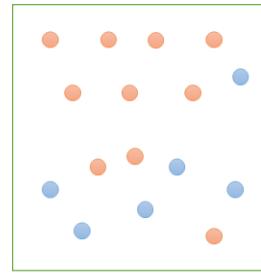
2 모델 생성 방법

불순도 지표1: 지니 계수(Gini Index)

- c개의 범주가 존재하는 A 영역에 대한 지니 계수

$$I(A) = 1 - \sum_{k=1}^c p_k^2$$

- p_k = A영역에 속한 객체들 중 k 범주에 속하는 레코드의 수



$$\begin{aligned}
 I(A) &= 1 - \sum_{k=1}^c p_k^2 \\
 &= 1 - \left(\frac{6}{16}\right)^2 - \left(\frac{10}{16}\right)^2 = 0.4688
 \end{aligned}$$

- 두 범주만 존재하는 상황에서 영역 A의 모든 객체들이 동일한 범주에 속할 경우: $I(A) = 0$
- 두 범주만 존재하는 상황에서 영역 A에 속하는 객체들이 1/2의 비중으로 존재할 경우: $I(A)=0.5$

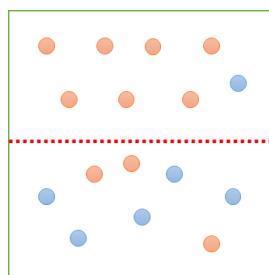
2 모델 생성 방법

불순도 지표1: 지니 계수(Gini Index)

- 두 개 이상의 영역에 대한 지니 계수

$$I(A) = \sum_{i=1}^d \left(R_i \left(1 - \sum_{k=1}^c p_{ik}^2 \right) \right)$$

- R_i = 분할 전 레코드 중 분할 후 i영역에 속하는 레코드의 비율



$$\begin{aligned} I(A) &= \frac{8}{16} \times \left(1 - \left(\frac{7}{8} \right)^2 - \left(\frac{1}{8} \right)^2 \right) + \frac{8}{16} \times \left(1 - \left(\frac{3}{8} \right)^2 - \left(\frac{5}{8} \right)^2 \right) \\ &= 0.3438 \end{aligned}$$

- 분기 후의 정보 획득(Information gain): $0.4688 - 0.3438 = 0.1250$

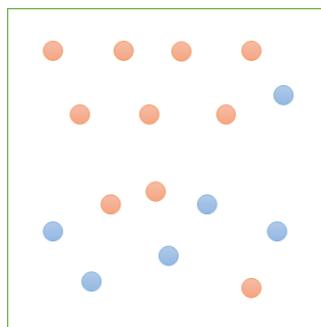
2 모델 생성 방법

불순도 지표2: Deviance

- 특정 영역에 대한 Deviance 개념

$$D_i = -2 \sum_k n_{ik} \log(p_{ik})$$

- i: node index, k: class index, pik: probability of class k in node i



$$\begin{aligned} D_i &= -2 \times \left(10 \times \log \left(\frac{10}{16} \right) + 6 \times \log \left(\frac{6}{16} \right) \right) \\ &= 21.17 \end{aligned}$$

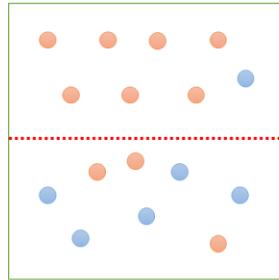
2 모델 생성 방법

불순도 지표2: Deviance

- 특정 영역에 대한 Deviance 개념

$$D_i = -2 \sum_k n_{ik} \log(p_{ik})$$

- i: node index, k: class index, pik: probability of class k in node i



$$D_1 = -2 \times \left(7 \times \log\left(\frac{7}{8}\right) + 1 \times \log\left(\frac{1}{8}\right) \right) = 6.03$$

$$D_2 = -2 \times \left(3 \times \log\left(\frac{3}{8}\right) + 5 \times \log\left(\frac{5}{8}\right) \right) = 10.59$$

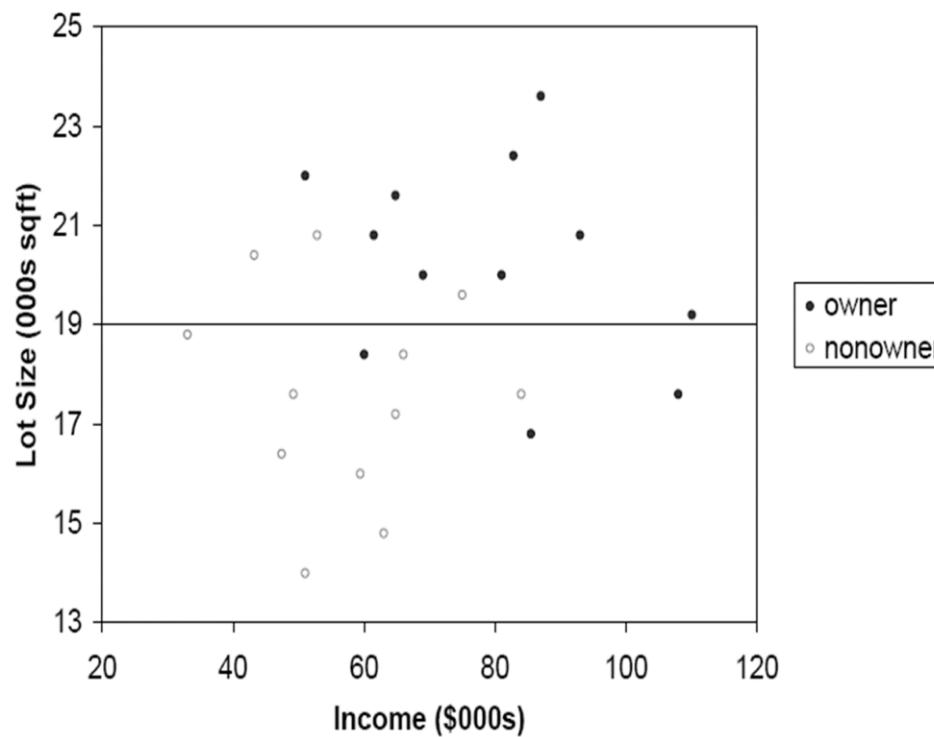
$$D_1 + D_2 = 16.62$$

분기 후의 정보 획득(Information gain): $21.17 - 16.62 = 4.55$

2 모델 생성 방법

재귀적 분기 수행방법

- 한 변수를 기준으로 하여 정렬
 - 본 예시에서는 Lot Size 변수를 먼저 정렬 기준 변수로 사용
 - 변수의 정렬 순서, 정렬 기준(오름차순/내림차순)은 최종 모델에 영향을 미치지 않음



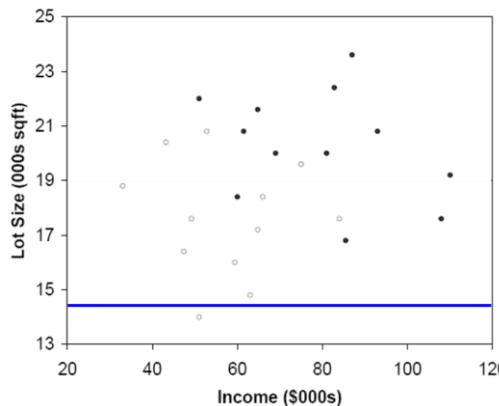
Income	Lot Size	Ownership
51.0	14.0	Non-owner
63.0	14.8	Non-owner
59.4	16.0	Non-owner
47.4	16.4	Non-owner
85.5	16.8	Owner
64.8	17.2	Non-owner
108.0	17.6	Owner
84.0	17.6	Non-owner
49.2	17.6	Non-owner
60.0	18.4	Owner
66.0	18.4	Non-owner
33.0	18.8	Non-owner
110.1	19.2	Owner
75.0	19.6	Non-owner
69.0	20.0	Owner
81.0	20.0	Owner
43.2	20.4	Non-owner
61.5	20.8	Owner
93.0	20.8	Owner
52.8	20.8	Non-owner
64.8	21.6	Owner
51.0	22.0	Owner
82.8	22.4	Owner
87.0	23.6	Owner

2 모델 생성 방법

재귀적 분기 수행방법

- 순차적으로 가능한 분기점에 대한 정보 획득 계산

- 첫 번째 후보 분기점 = $14.4 (0.5 * (14.0 + 14.8))$
- Lot size < 14.4 와 Lot size > 14.4 인 두 영역으로 구분
- 불순도(impurity) 계산: 지니 계수(Gini index)
- 분기 전: $1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 0.5000$
- 분기 후: $\frac{1}{24} \times \left(1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2\right) + \frac{23}{24} \times \left(1 - \left(\frac{12}{23}\right)^2 - \left(\frac{11}{23}\right)^2\right) = 0.4783$
- 정보 획득(Information gain): $0.5000 - 0.4783 = 0.0217$



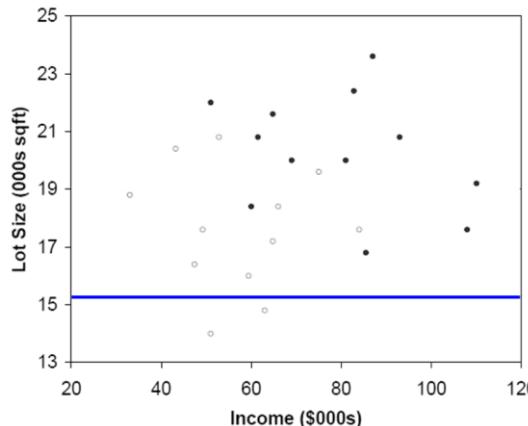
Income	Lot Size	Ownership
51.0	14.0	Non-owner
63.0	14.8	Non-owner
59.4	16.0	Non-owner
47.4	16.4	Non-owner
85.5	16.8	Owner
64.8	17.2	Non-owner
108.0	17.6	Owner
84.0	17.6	Non-owner
49.2	17.6	Non-owner
60.0	18.4	Owner
66.0	18.4	Non-owner
33.0	18.8	Non-owner
110.1	19.2	Owner
75.0	19.6	Non-owner
69.0	20.0	Owner
81.0	20.0	Owner
43.2	20.4	Non-owner
61.5	20.8	Owner
93.0	20.8	Owner
52.8	20.8	Non-owner
64.8	21.6	Owner
51.0	22.0	Owner
82.8	22.4	Owner
87.0	23.6	Owner

2 모델 생성 방법

재귀적 분기 수행방법

- 순차적으로 가능한 분기점에 대한 정보 획득 계산

- 두 번째 후보 분기점 = 15.6 ($0.5 * (14.8 + 16.0)$)
- Lot size < 15.6 와 Lot size > 14.5 인 두 영역으로 구분
- 불순도(impurity) 계산: 지니 계수(Gini index)
- 분기 전: $1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 0.5000$
- 분기 후: $\frac{1}{24} \times \left(1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2\right) + \frac{22}{24} \times \left(1 - \left(\frac{12}{22}\right)^2 - \left(\frac{10}{22}\right)^2\right) = 0.4545$
- 정보 획득(Information gain): $0.5000 - 0.4545 = 0.0455$



Income	Lot Size	Ownership
51.0	14.0	Non-owner
63.0	14.8	Non-owner
59.4	16.0	Non-owner
47.4	16.4	Non-owner
85.5	16.8	Owner
64.8	17.2	Non-owner
108.0	17.6	Owner
84.0	17.6	Non-owner
49.2	17.6	Non-owner
60.0	18.4	Owner
66.0	18.4	Non-owner
33.0	18.8	Non-owner
110.1	19.2	Owner
75.0	19.6	Non-owner
69.0	20.0	Owner
81.0	20.0	Owner
43.2	20.4	Non-owner
61.5	20.8	Owner
93.0	20.8	Owner
52.8	20.8	Non-owner
64.8	21.6	Owner
51.0	22.0	Owner
82.8	22.4	Owner
87.0	23.6	Owner

2 모델 생성 방법

재귀적 분기 수행방법

- 순차적으로 가능한 분기점에 대한 정보 획득 계산

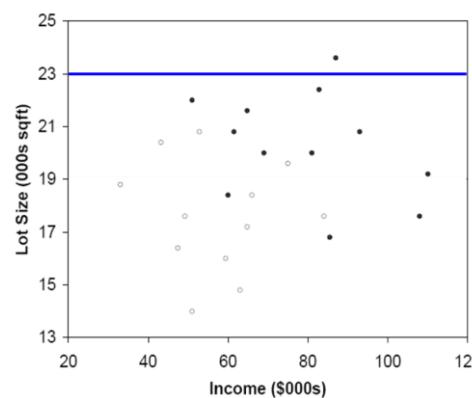
- 마지막 후보 분기점 = 23.0 ($0.5 * (22.4 + 23.6)$)
- Lot size < 23.0 와 Lot size > 23.0 인 두 영역으로 구분

- 불순도(impurity) 계산: 지니 계수(Gini index)

- 분기 전: $1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 0.5000$

- 분기 후: $\frac{23}{24} \times \left(1 - \left(\frac{11}{23}\right)^2 - \left(\frac{12}{23}\right)^2\right) + \frac{1}{24} \times \left(1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right) = 0.4783$

- 정보 획득(Information gain): $0.5000 - 0.4783 = 0.0217$

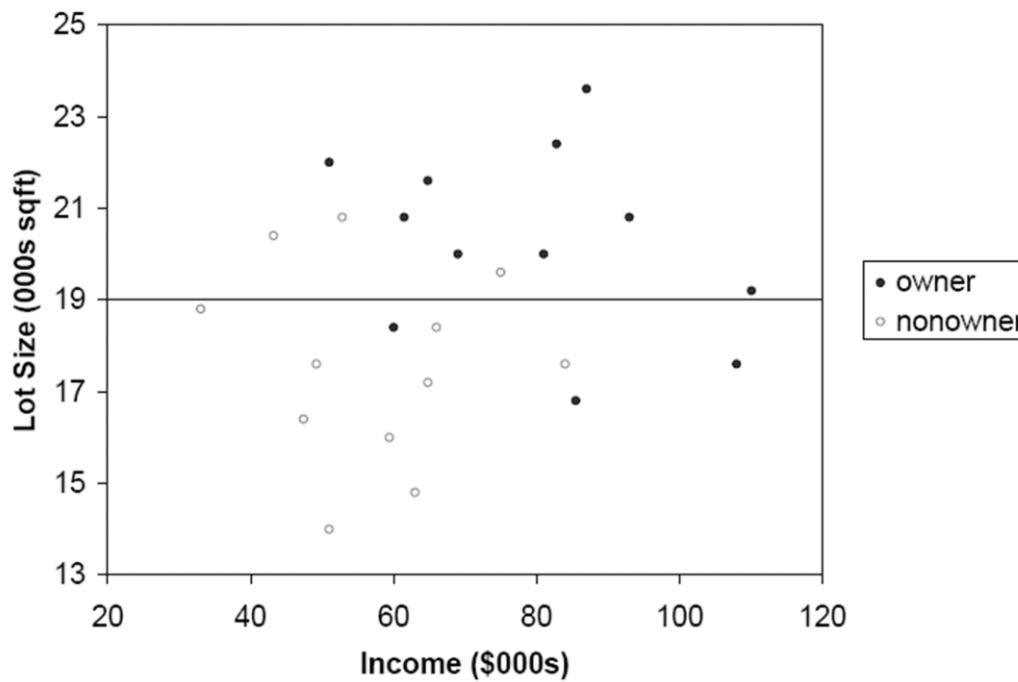


Income	Lot Size	Ownership
51.0	14.0	Non-owner
63.0	14.8	Non-owner
59.4	16.0	Non-owner
47.4	16.4	Non-owner
85.5	16.8	Owner
64.8	17.2	Non-owner
108.0	17.6	Owner
84.0	17.6	Non-owner
49.2	17.6	Non-owner
60.0	18.4	Owner
66.0	18.4	Non-owner
33.0	18.8	Non-owner
110.1	19.2	Owner
75.0	19.6	Non-owner
69.0	20.0	Owner
81.0	20.0	Owner
43.2	20.4	Non-owner
61.5	20.8	Owner
93.0	20.8	Owner
52.8	20.8	Non-owner
64.8	21.6	Owner
51.0	22.0	Owner
82.8	22.4	Owner
87.0	23.6	Owner

2 모델 생성 방법

재귀적 분기 수행방법

- 한 변수를 기준으로 하여 정렬
 - Income 변수를 기준으로 다시 정렬



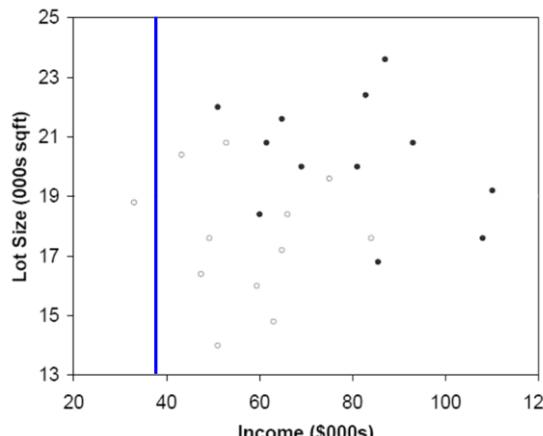
Income	Lot Size	Ownership
33.0	18.8	Non-owner
43.2	20.4	Non-owner
47.4	16.4	Non-owner
49.2	17.6	Non-owner
51.0	14.0	Non-owner
51.0	22.0	Owner
52.8	20.8	Non-owner
59.4	16.0	Non-owner
60.0	18.4	Owner
61.5	20.8	Owner
63.0	14.8	Non-owner
64.8	17.2	Non-owner
64.8	21.6	Owner
66.0	18.4	Non-owner
69.0	20.0	Owner
75.0	19.6	Non-owner
81.0	20.0	Owner
82.8	22.4	Owner
84.0	17.6	Non-owner
85.5	16.8	Owner
87.0	23.6	Owner
93.0	20.8	Owner
108.0	17.6	Owner
110.1	19.2	Owner

2 모델 생성 방법

재귀적 분기 수행방법

- 순차적으로 가능한 분기점에 대한 정보 획득 계산

- 첫 번째 후보 분기점 = $38.1 (0.5 * (33.0 + 43.2))$
- Income < 38.1 과 Income > 38.1 인 두 영역으로 구분
- 불순도(impurity) 계산: 지니 계수(Gini index)
- 분기 전: $1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 0.5000$
- 분기 후: $\frac{1}{24} \times \left(1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2\right) + \frac{23}{24} \times \left(1 - \left(\frac{12}{23}\right)^2 - \left(\frac{11}{23}\right)^2\right) = 0.4783$
- 정보 획득(Information gain): $0.5000 - 0.4783 = 0.0217$



Income	Lot Size	Ownership
33.0	18.8	Non-owner
43.2	20.4	Non-owner
47.4	16.4	Non-owner
49.2	17.6	Non-owner
51.0	14.0	Non-owner
51.0	22.0	Owner
52.8	20.8	Non-owner
59.4	16.0	Non-owner
60.0	18.4	Owner
61.5	20.8	Owner
63.0	14.8	Non-owner
64.8	17.2	Non-owner
64.8	21.6	Owner
66.0	18.4	Non-owner
69.0	20.0	Owner
75.0	19.6	Non-owner
81.0	20.0	Owner
82.8	22.4	Owner
84.0	17.6	Non-owner
85.5	16.8	Owner
87.0	23.6	Owner
93.0	20.8	Owner
108.0	17.6	Owner
110.1	19.2	Owner

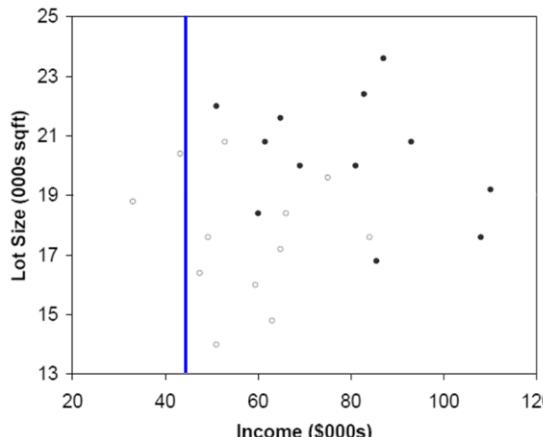
2 모델 생성 방법

재귀적 분기 수행방법

- 순차적으로 가능한 분기점에 대한 정보 획득 계산

- 두 번째 후보 분기점 = $45.3 (0.5 * (43.2 + 47.4))$
- Income < 45.3 과 Income > 45.3 인 두 영역으로 구분
- 불순도(impurity) 계산: 지니 계수(Gini index)
- 분기 전: $1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 0.5000$
- 분기 후: $\frac{1}{24} \times \left(1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2\right) + \frac{22}{24} \times \left(1 - \left(\frac{12}{22}\right)^2 - \left(\frac{10}{22}\right)^2\right) = 0.4545$

- 정보 획득(Information gain): $0.5000 - 0.4545 = 0.0455$



Income	Lot Size	Ownership
33.0	18.8	Non-owner
43.2	20.4	Non-owner
47.4	16.4	Non-owner
49.2	17.6	Non-owner
51.0	14.0	Non-owner
51.0	22.0	Owner
52.8	20.8	Non-owner
59.4	16.0	Non-owner
60.0	18.4	Owner
61.5	20.8	Owner
63.0	14.8	Non-owner
64.8	17.2	Non-owner
64.8	21.6	Owner
66.0	18.4	Non-owner
69.0	20.0	Owner
75.0	19.6	Non-owner
81.0	20.0	Owner
82.8	22.4	Owner
84.0	17.6	Non-owner
85.5	16.8	Owner
87.0	23.6	Owner
93.0	20.8	Owner
108.0	17.6	Owner
110.1	19.2	Owner

2 모델 생성 방법

재귀적 분기 수행방법

- 순차적으로 가능한 분기점에 대한 정보 획득 계산

 - 마지막 후보 분기점 = $109.05 (0.5 * (108.0 + 110.1))$

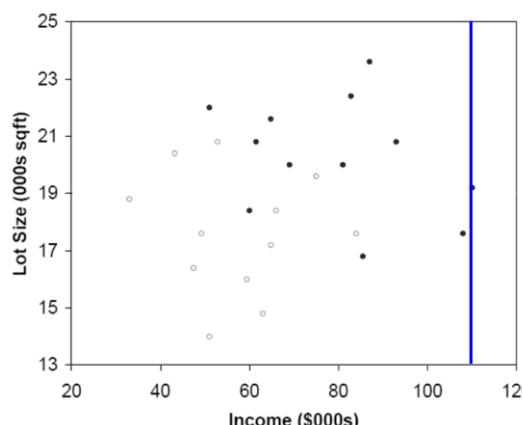
 - Income < 109.05 과 Income > 109.05 인 두 영역으로 구분

 - 불순도(impurity) 계산: 지니 계수(Gini index)

- 분기 전: $1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 0.5000$

- 분기 후: $\frac{23}{24} \times \left(1 - \left(\frac{11}{23}\right)^2 - \left(\frac{12}{23}\right)^2\right) + \frac{1}{24} \times \left(1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right) = 0.4783$

- 정보 획득(Information gain): $0.5000 - 0.4783 = 0.0217$



Income	Lot Size	Ownership
33.0	18.8	Non-owner
43.2	20.4	Non-owner
47.4	16.4	Non-owner
49.2	17.6	Non-owner
51.0	14.0	Non-owner
51.0	22.0	Owner
52.8	20.8	Non-owner
59.4	16.0	Non-owner
60.0	18.4	Owner
61.5	20.8	Owner
63.0	14.8	Non-owner
64.8	17.2	Non-owner
64.8	21.6	Owner
66.0	18.4	Non-owner
69.0	20.0	Owner
75.0	19.6	Non-owner
81.0	20.0	Owner
82.8	22.4	Owner
84.0	17.6	Non-owner
85.5	16.8	Owner
87.0	23.6	Owner
93.0	20.8	Owner
108.0	17.6	Owner
110.1	19.2	Owner

2 모델 생성 방법

범주형 변수의 경우 재귀적 분류

- 범주형 변수에 대해서는 분기가 가능한 모든 경우의 수를 조사
- 예시: 3개의 범주 A, B, and C가 존재하는 경우
 - {A} vs. {B,C}
 - {B} vs. {A,C}
 - {C} vs. {A,B}
- 위 경우의 수에 대한 정보 획득을 각각 계산

2 모델 생성 방법

재귀적 분기 수행방법

- 순차적으로 가능한 분기점에 대한 정보 획득 계산

 - 최적 분기점 = Lot size 19.0

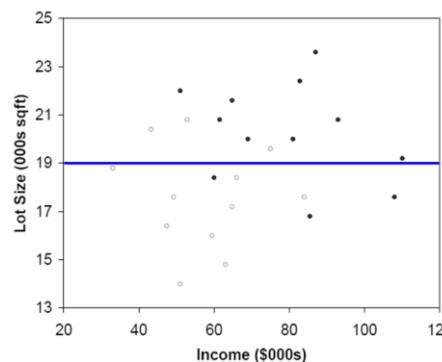
 - Lot size < 19.0 과 Lot size > 19.0 인 두 영역으로 구분

 - 불순도(impurity) 계산: 지니 계수(Gini index)

- 분기 전: $1 - \left(\frac{12}{24}\right)^2 - \left(\frac{12}{24}\right)^2 = 0.5000$

- 분기 후: $\frac{12}{24} \times \left(1 - \left(\frac{3}{12}\right)^2 - \left(\frac{9}{12}\right)^2\right) + \frac{12}{24} \times \left(1 - \left(\frac{9}{12}\right)^2 - \left(\frac{3}{12}\right)^2\right) = 0.3750$

- 정보 획득(Information gain): $0.5000 - 0.3750 = 0.1250$

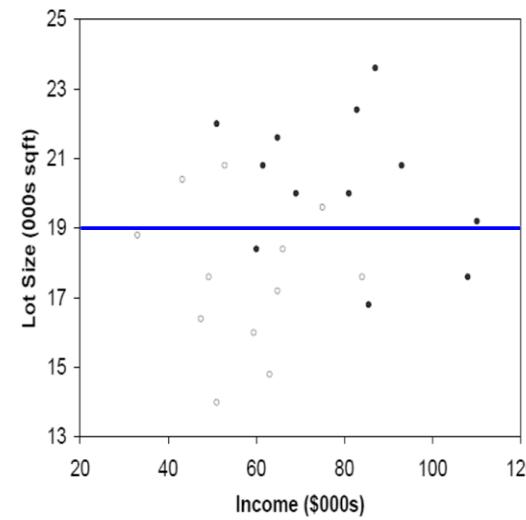
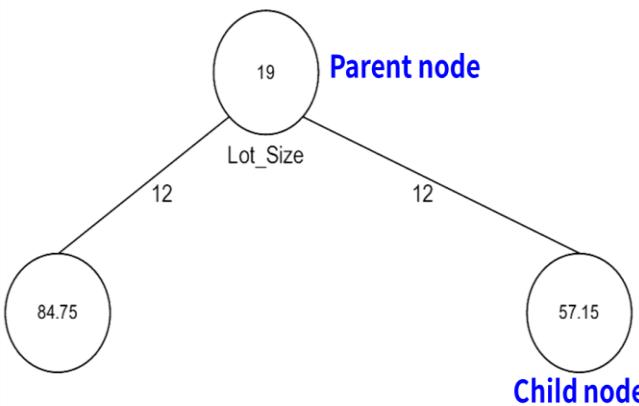


Income	Lot Size	Ownership
51.0	14.0	Non-owner
63.0	14.8	Non-owner
59.4	16.0	Non-owner
47.4	16.4	Non-owner
85.5	16.8	Owner
64.8	17.2	Non-owner
108.0	17.6	Owner
84.0	17.6	Non-owner
49.2	17.6	Non-owner
60.0	18.4	Owner
66.0	18.4	Non-owner
33.0	18.8	Non-owner
110.1	19.2	Owner
75.0	19.6	Non-owner
69.0	20.0	Owner
81.0	20.0	Owner
43.2	20.4	Non-owner
61.5	20.8	Owner
93.0	20.8	Owner
52.8	20.8	Non-owner
64.8	21.6	Owner
51.0	22.0	Owner
82.8	22.4	Owner
87.0	23.6	Owner

2 모델 생성 방법

재귀적 분기 수행방법

- 의사결정나무 구조
- 분기점은 나무의 노드가 되며 원 안의 숫자는 분기 기준이 되는 변수의 값임
- 사각형은 말단 노드(leaf node, 더 이상 분기가 수행되지 않는 노드)가 됨
- 선 아래 숫자는 분기로 인해 나눈 객체의 수를 의미



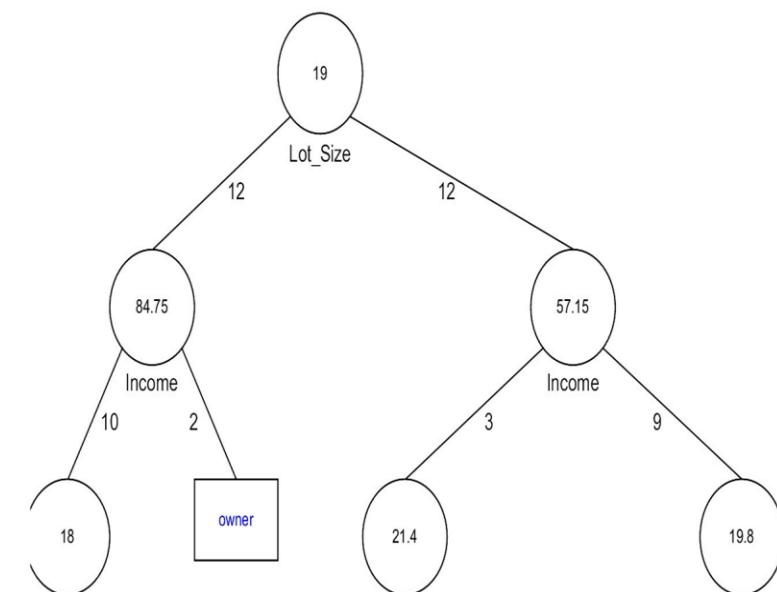
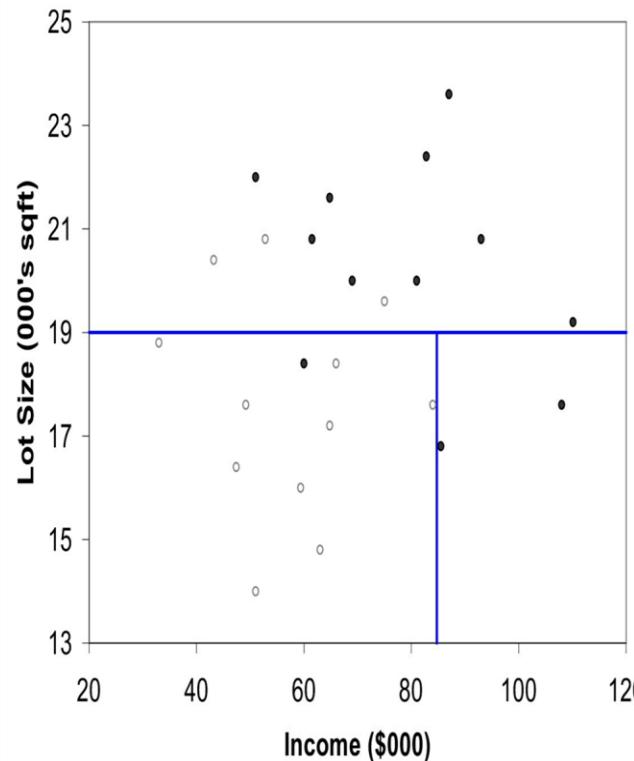
Income	Lot Size	Ownership
51.0	14.0	Non-owner
63.0	14.8	Non-owner
59.4	16.0	Non-owner
47.4	16.4	Non-owner
85.5	16.8	Owner
64.8	17.2	Non-owner
108.0	17.6	Owner
84.0	17.6	Non-owner
49.2	17.6	Non-owner
60.0	18.4	Owner
66.0	18.4	Non-owner
33.0	18.8	Non-owner
110.1	19.2	Owner
75.0	19.6	Non-owner
69.0	20.0	Owner
81.0	20.0	Owner
43.2	20.4	Non-owner
61.5	20.8	Owner
93.0	20.8	Owner
52.8	20.8	Non-owner
64.8	21.6	Owner
51.0	22.0	Owner
82.8	22.4	Owner
87.0	23.6	Owner

2 모델 생성 방법

재귀적 분기 수행방법

- 모든 노드의 순도가 100%가 될 때까지 분기를 수행
- 정보획득이 0이 되는 시점까지 수행
- 두 번째 분기점: $\text{income} = 84.75$

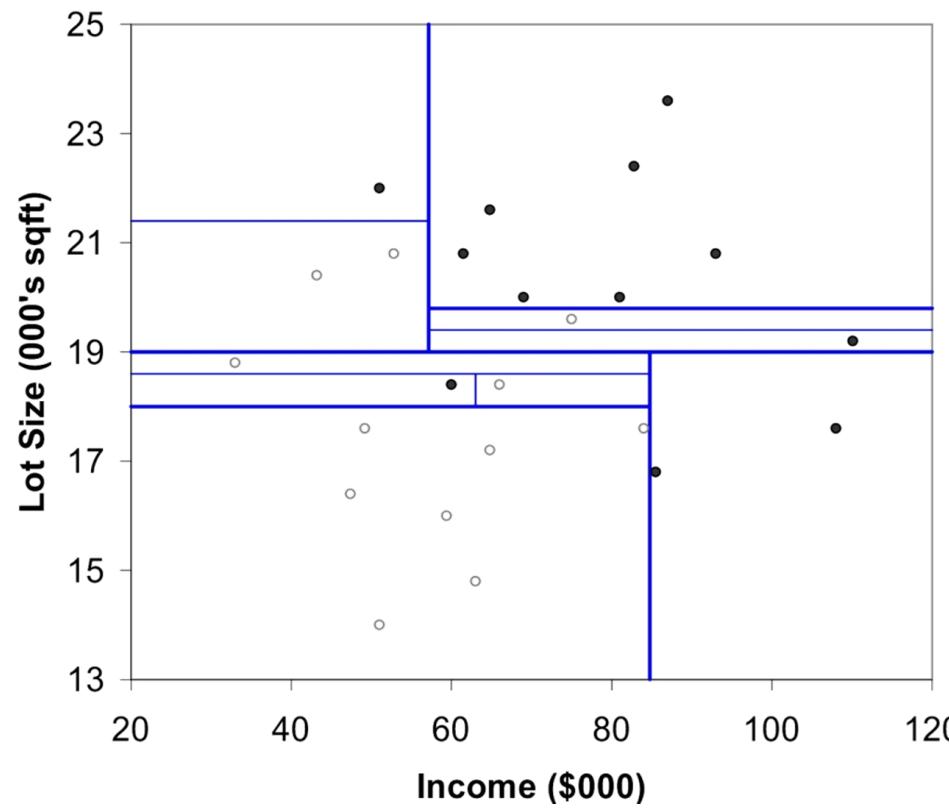
Income	Lot Size	Ownership
33.0	18.8	Non-owner
47.4	16.4	Non-owner
49.2	17.6	Non-owner
51.0	14.0	Non-owner
59.4	16.0	Non-owner
60.0	18.4	Owner
63.0	14.8	Non-owner
64.8	17.2	Non-owner
66.0	18.4	Non-owner
84.0	17.6	Non-owner
85.5	16.8	Owner
108.0	17.6	Owner



2 모델 생성 방법

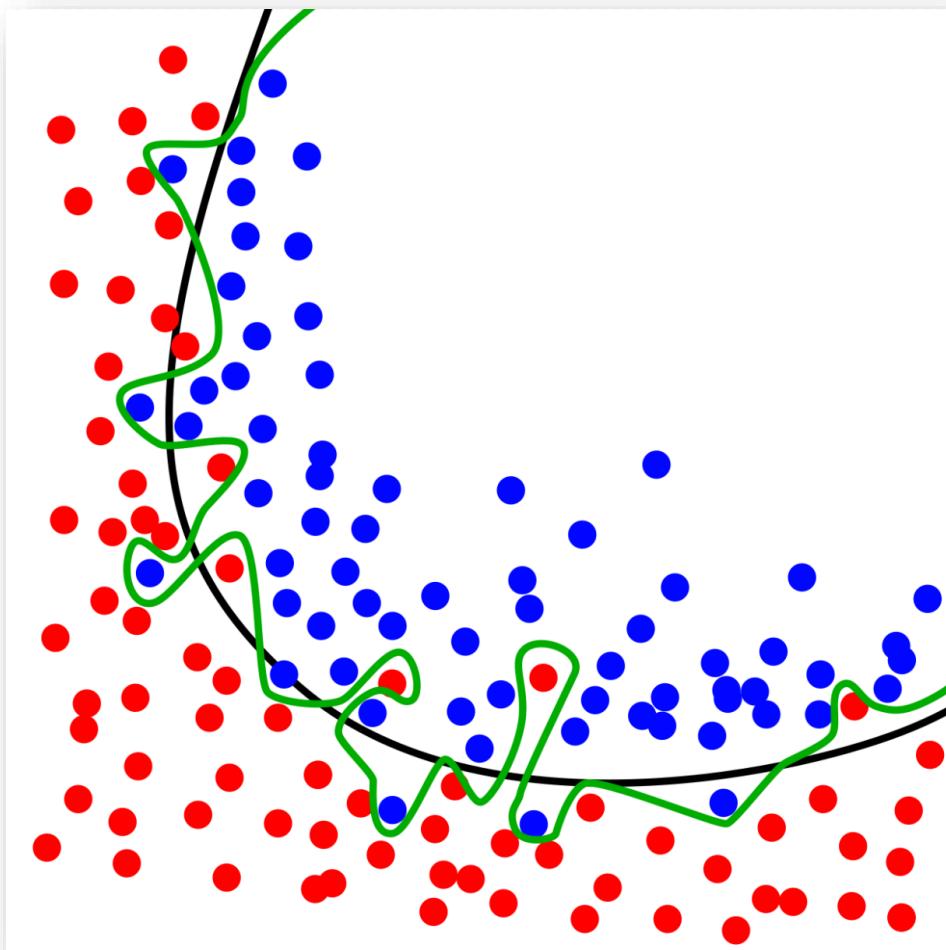
재귀적 분기 완료

- 모든 말단 노드에서 한 범주의 객체들만 존재
- Full node 완성
- 순도 100%



2 모델 생성 방법

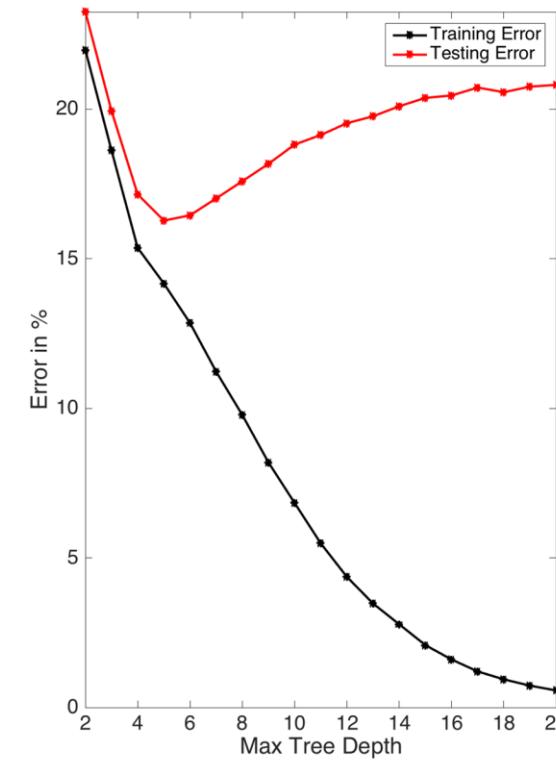
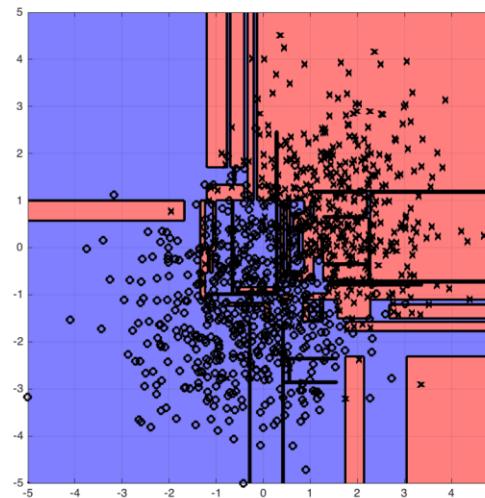
과적합(Overfitting)



2 모델 생성 방법

Full node: 과적합(Overfitting) 문제 발생

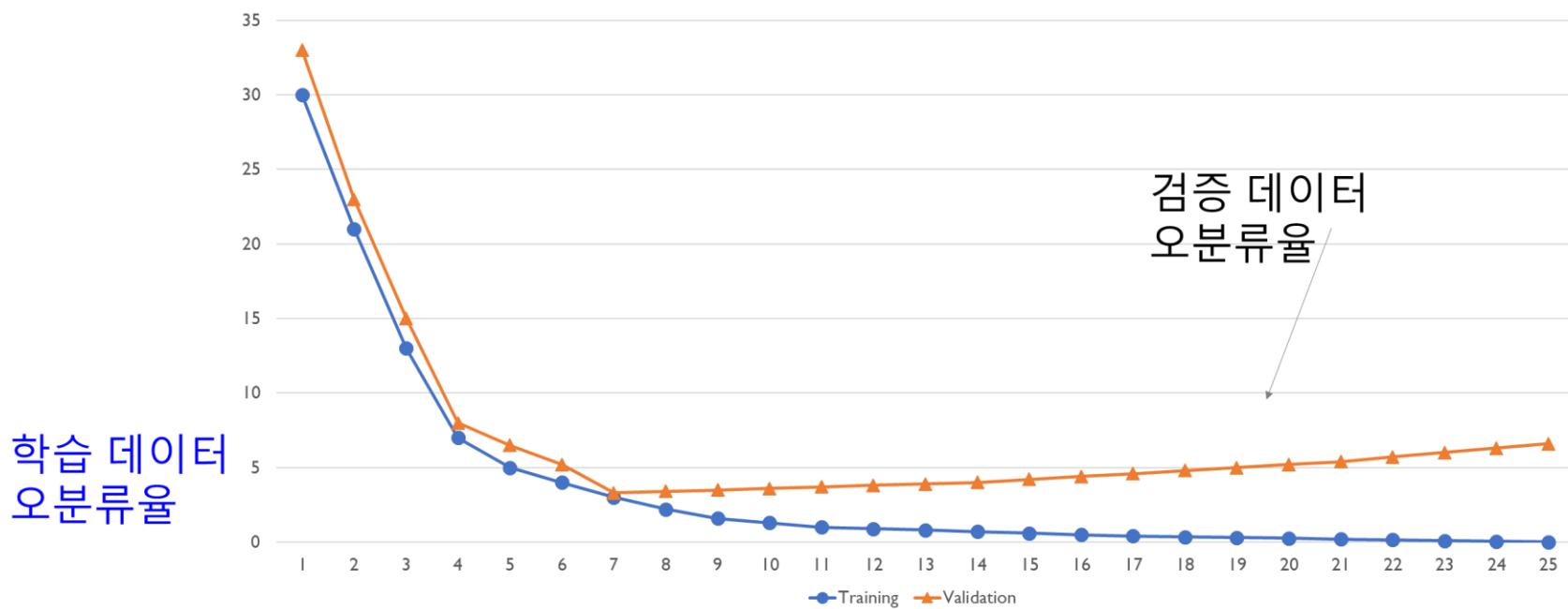
- 재귀적 분기는 모든 말단노드의 순도가 100%일 때 종료됨
- 그런데, 의사 결정나무의 목적은 새로운 데이터에 대한 예측 (분류)



2 모델 생성 방법

과적합(Overfitting) 문제 발생

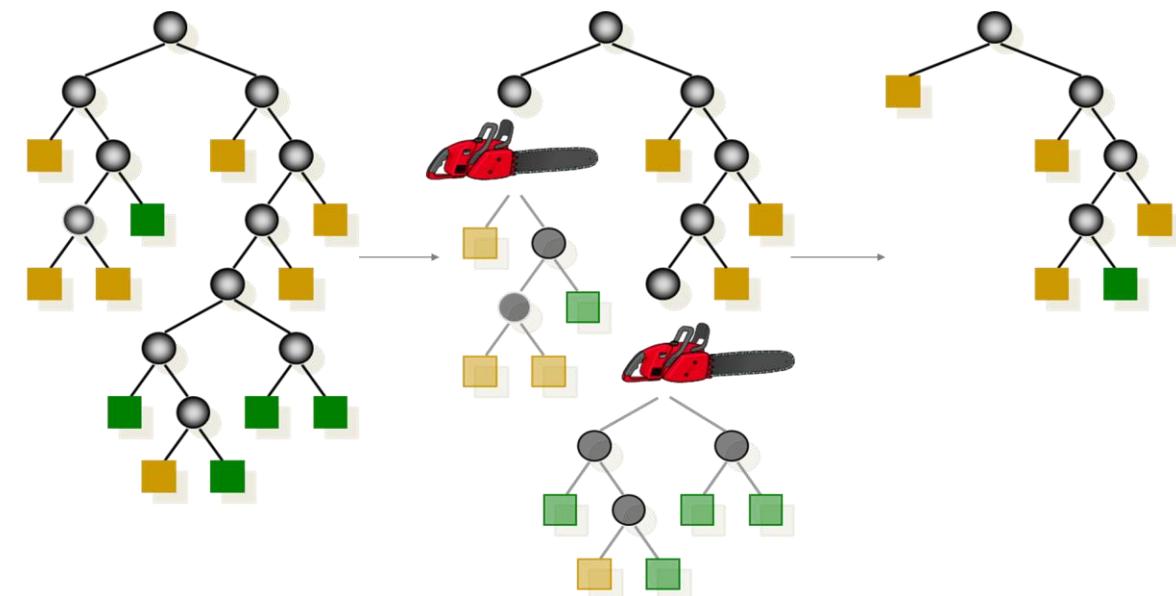
- 일반적으로 이러한 Full tree는 과적합Overfitting의 문제를 내포하고 있으며, 이는 새로운 데이터에 대한 예측 성능 저하의 위험을 안고 있음
- 의사결정나무의 노드 수가 증가할 때, 처음에는 새로운 데이터에 대한 오분류율이 감소하나, 일정 수준 이상이 되면 오분류율이 증가하는 현상 발생



2 모델 생성 방법

가지치기(Pruning)

- 의사결정나무는 Full Tree를 생성한 뒤 적절한 수준에서 말단 노드를 결합하는 가지치기 수행 (Post-Pruning)
- 검증데이터에 대한 오분류율이 증가하는 시점에서 가지치기
- Full tree에 비해 구조가 단순한 의사결정나무가 생성됨
- 비용 복잡도(cost complexity)를 사용하여 최적의 의사결정나무 구조 선택



2 모델 생성 방법

가지치기(Pruning)

- 비용 복잡도

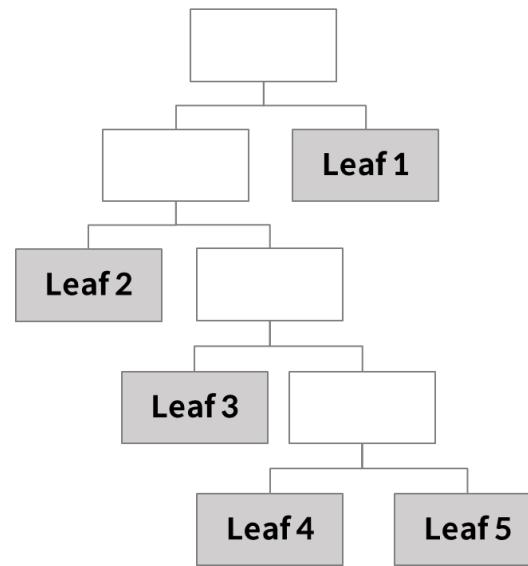
$$CC(T) = Err(T) + \alpha \times L(T)$$

- $CC(T)$ = 의사결정나무의 비용 복잡도
 - 낮을수록 우수한 의사결정나무
- $ERR(T)$ = 검증데이터에 대한 오분류율
- $L(T)$ = 말단 노드의 수 (구조의 복잡도)
- α = $ERR(T)$ 와 $L(T)$ 를 결합하는 가중치
 - 사용자(연구자)에 의해 부여됨, 일반적인 소프트웨어 패키지에서는 default 값이 존재

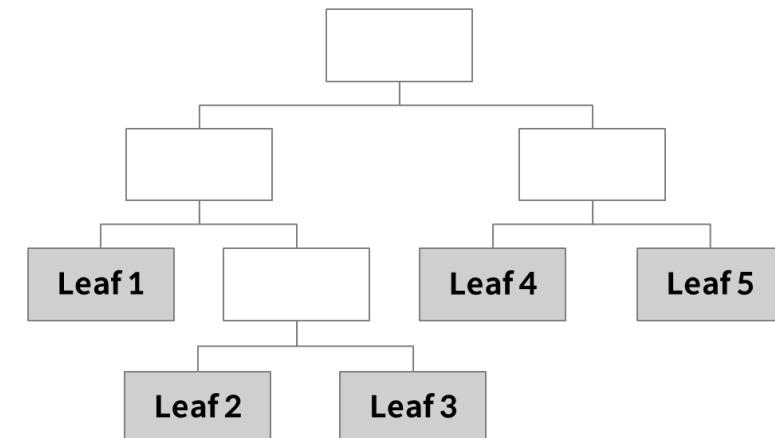
2 모델 생성 방법

비용 복잡도 예시 1)

Tree 1



Tree 2



검증 데이터셋 오류율 = 10%

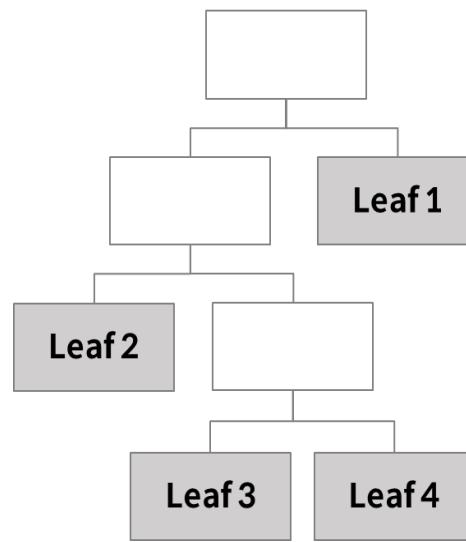
검증 데이터셋 오류율 = 15%

같은 수의 말단 노드를 갖는 Tree라면 검증 데이터셋 오류율이 낮은 Tree 1을 선택

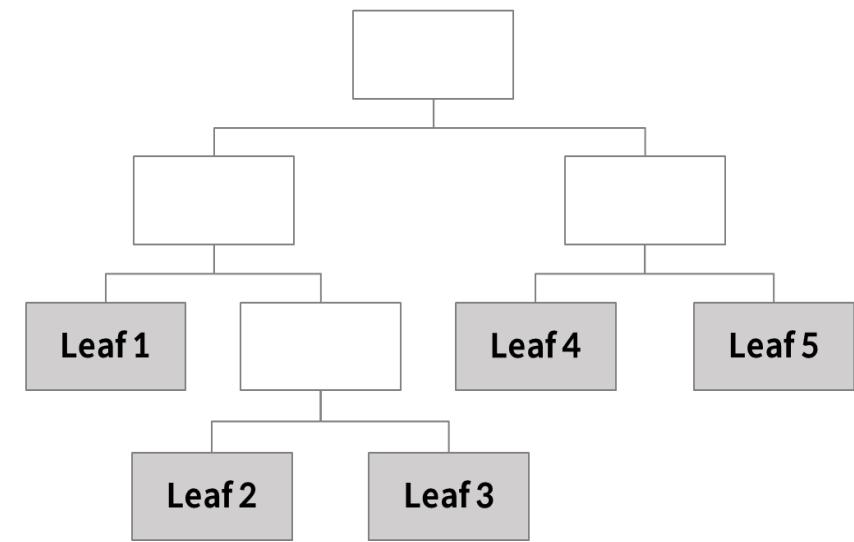
2 모델 생성 방법

비용 복잡도 예시 2)

Tree A



Tree B



검증 데이터셋 오류율 = 15%

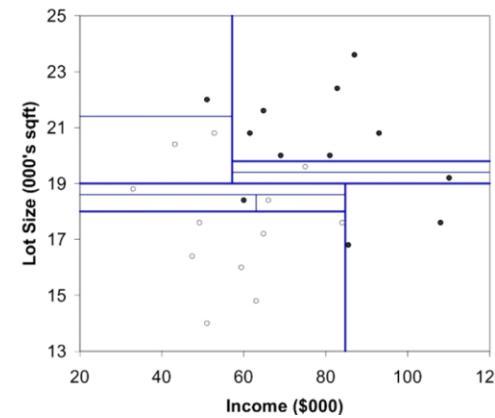
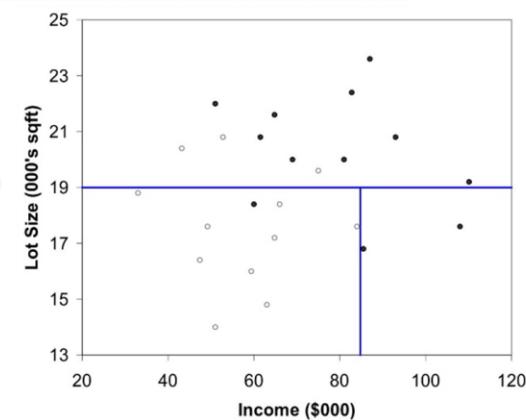
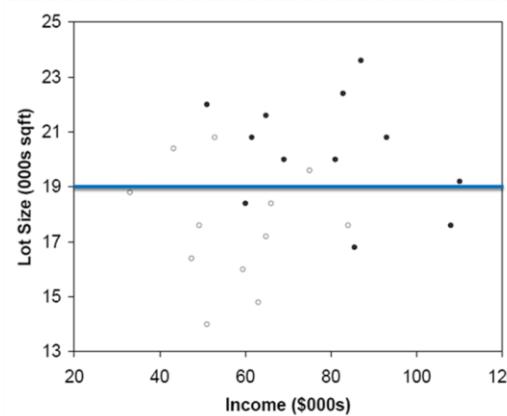
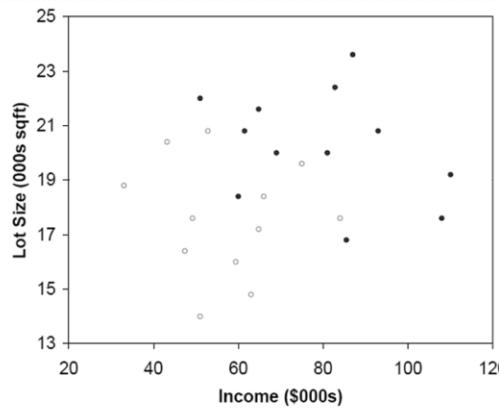
검증 데이터셋 오류율 = 15%

동일한 검증 데이터셋 오류율이라면 말단 노드의 수가 적은 Tree A를 선택

2 모델 생성 방법

가지치기

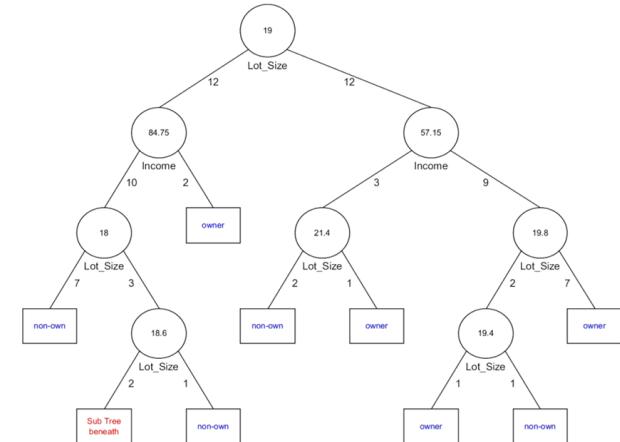
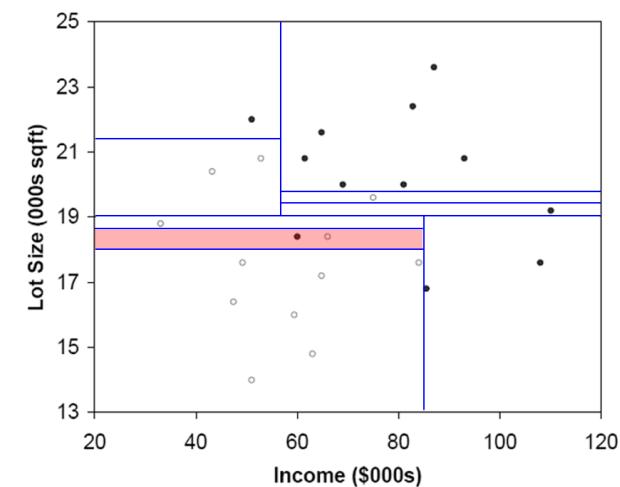
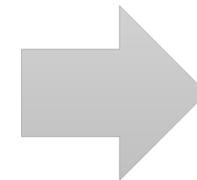
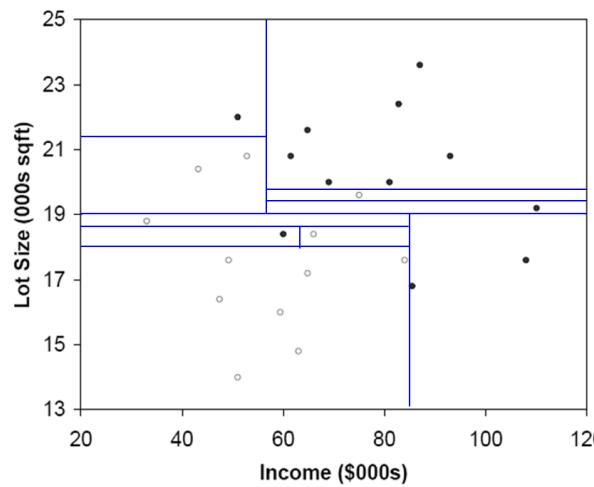
- 재귀적 분기를 통한 Full Tree 생성



2 모델 생성 방법

가지치기

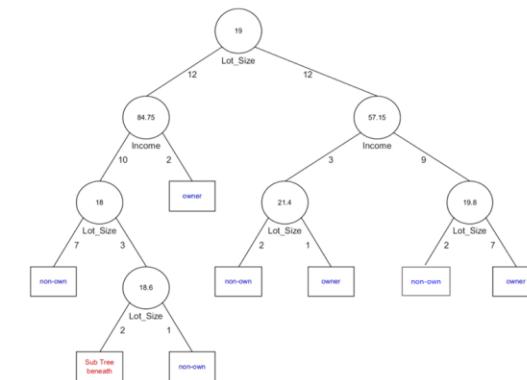
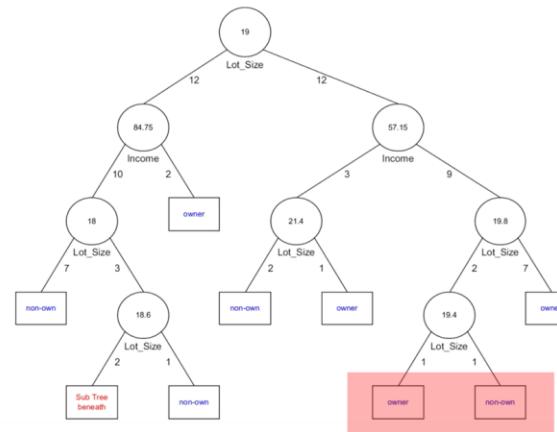
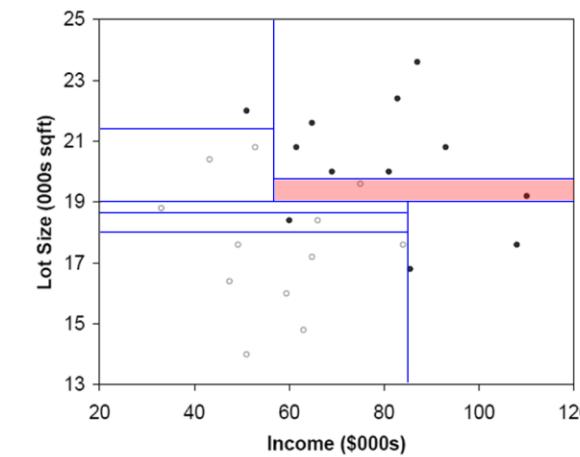
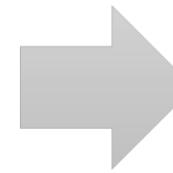
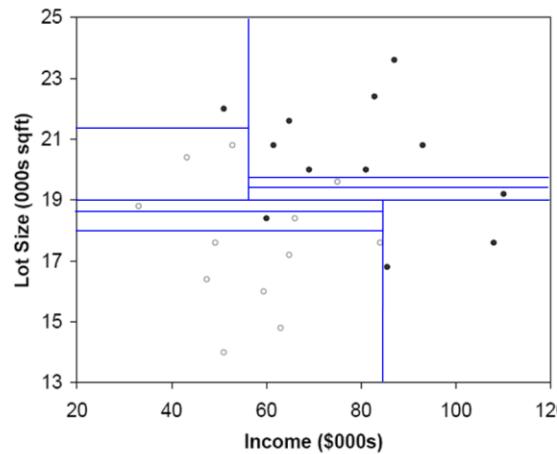
- 가지치기 1단계



2 모델 생성 방법

가지치기

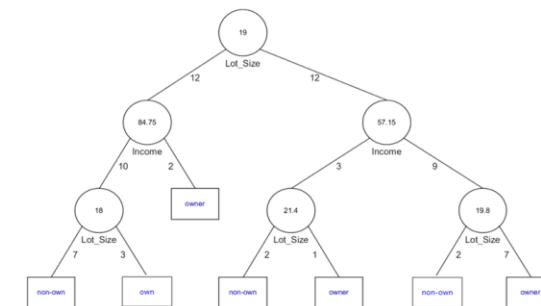
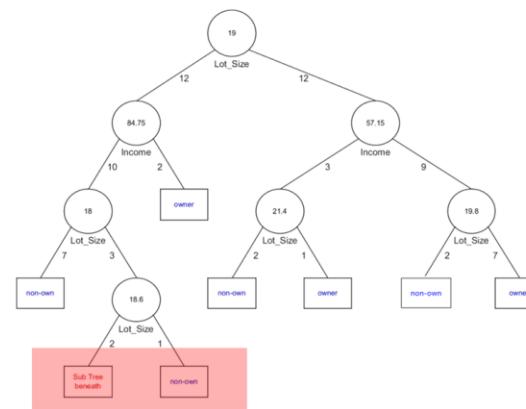
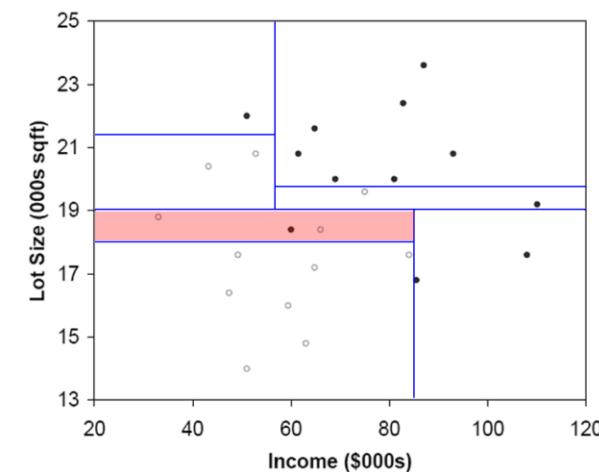
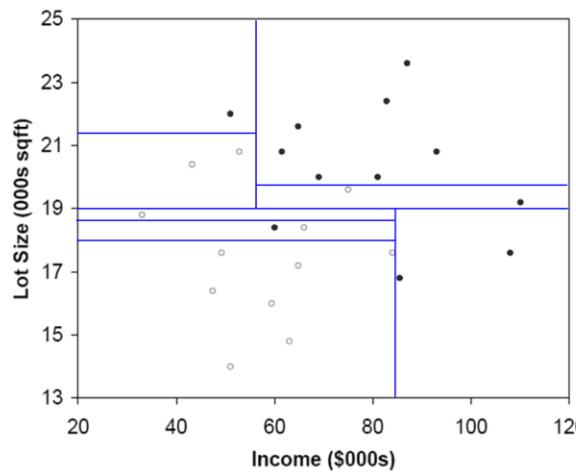
- 가지치기 2단계



2 모델 생성 방법

가지치기

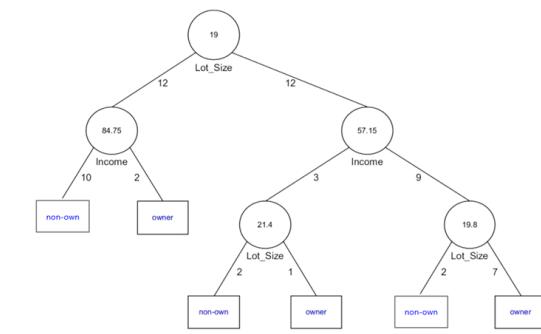
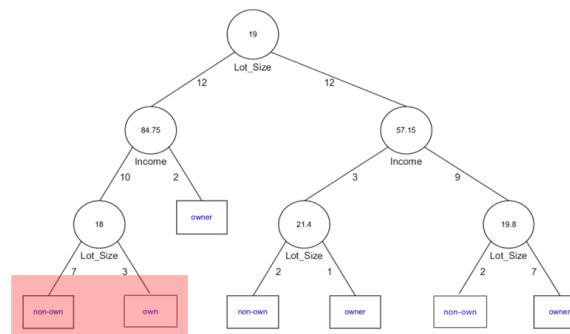
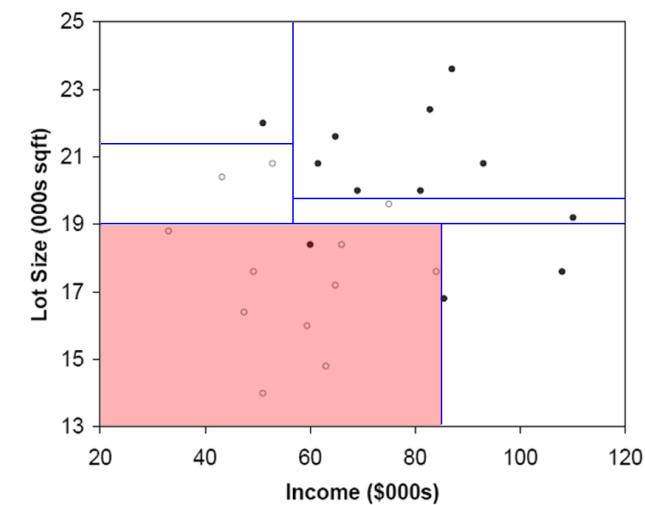
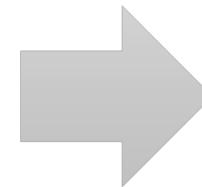
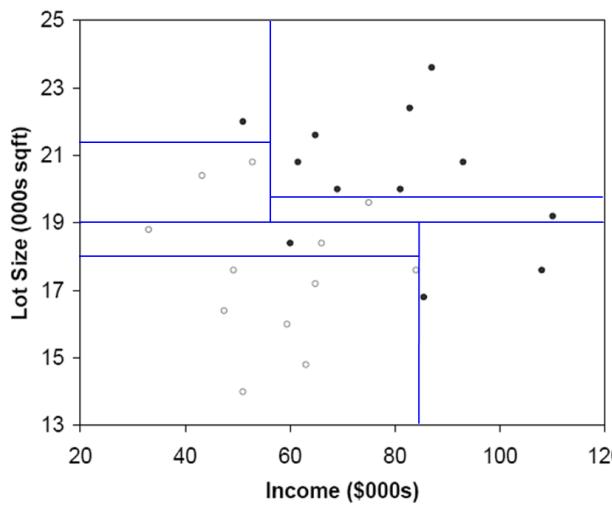
- 가지치기 3단계



2 모델 생성 방법

가지치기

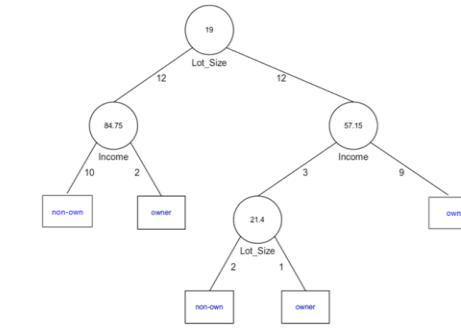
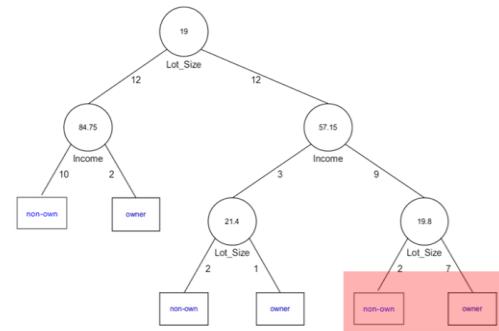
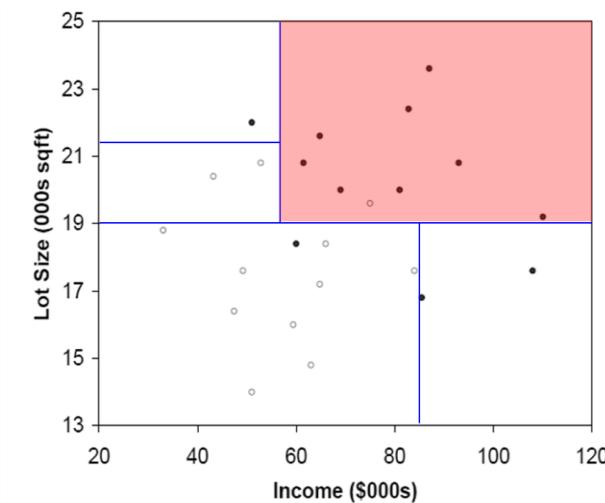
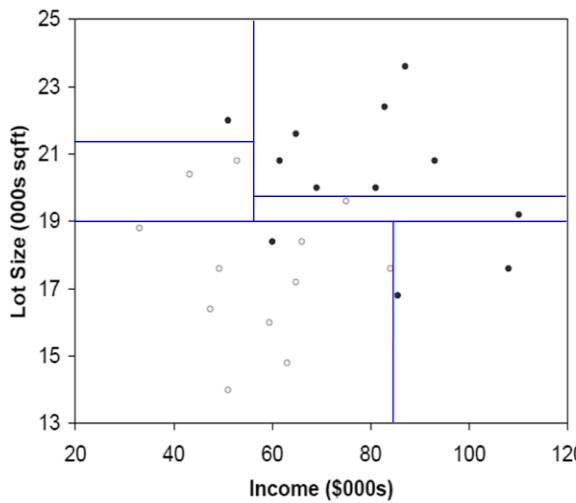
- 가지치기 4단계



2 모델 생성 방법

가지치기

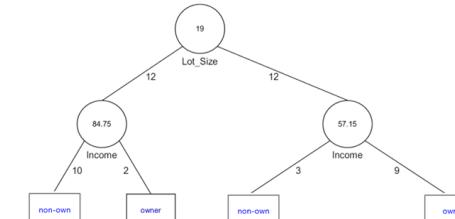
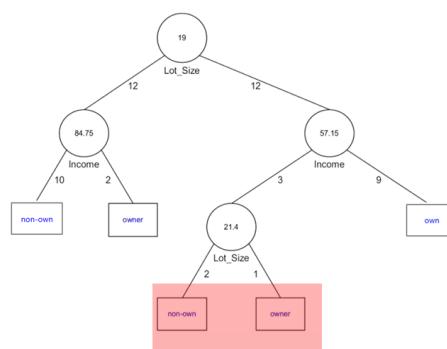
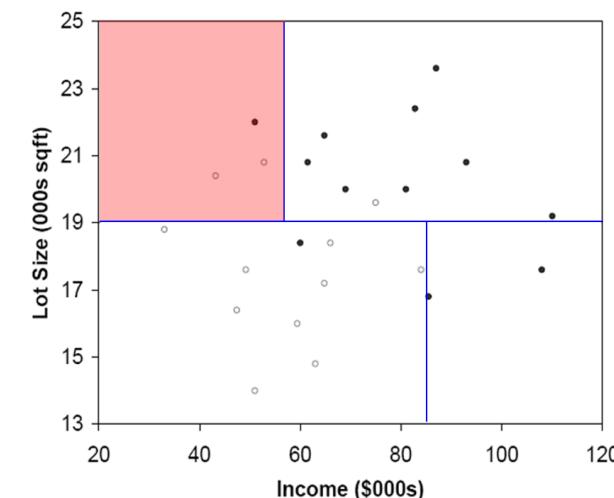
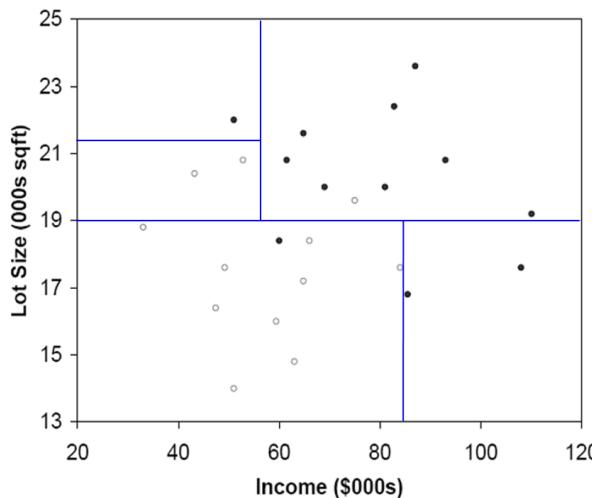
- 가지치기 5단계



2 모델 생성 방법

가지치기

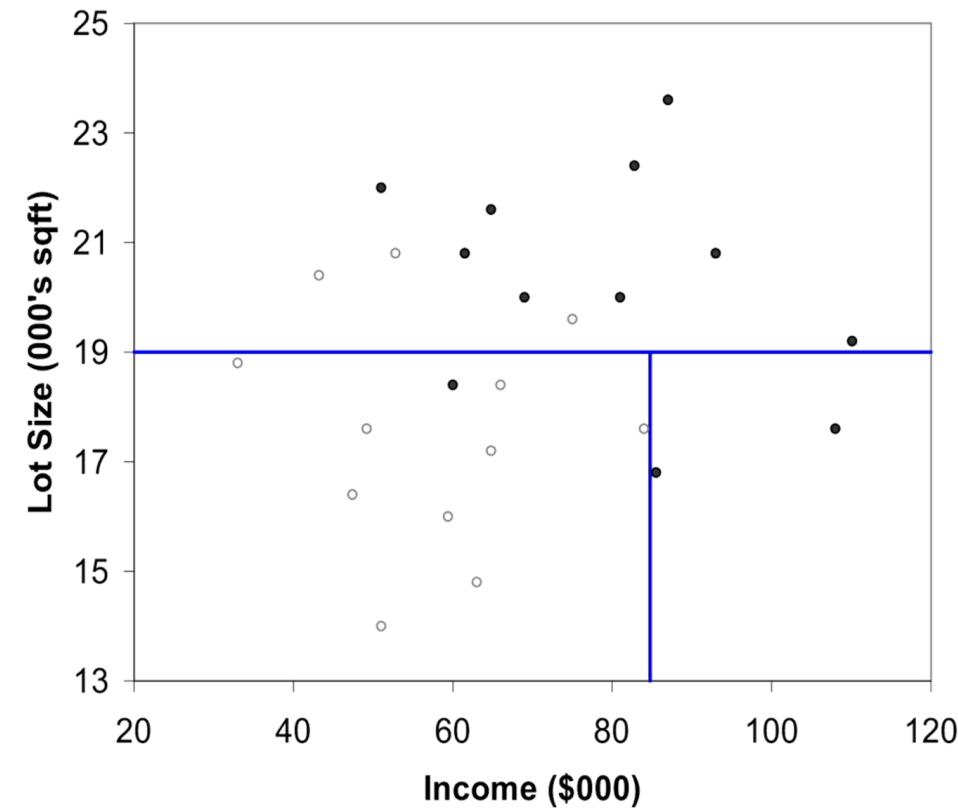
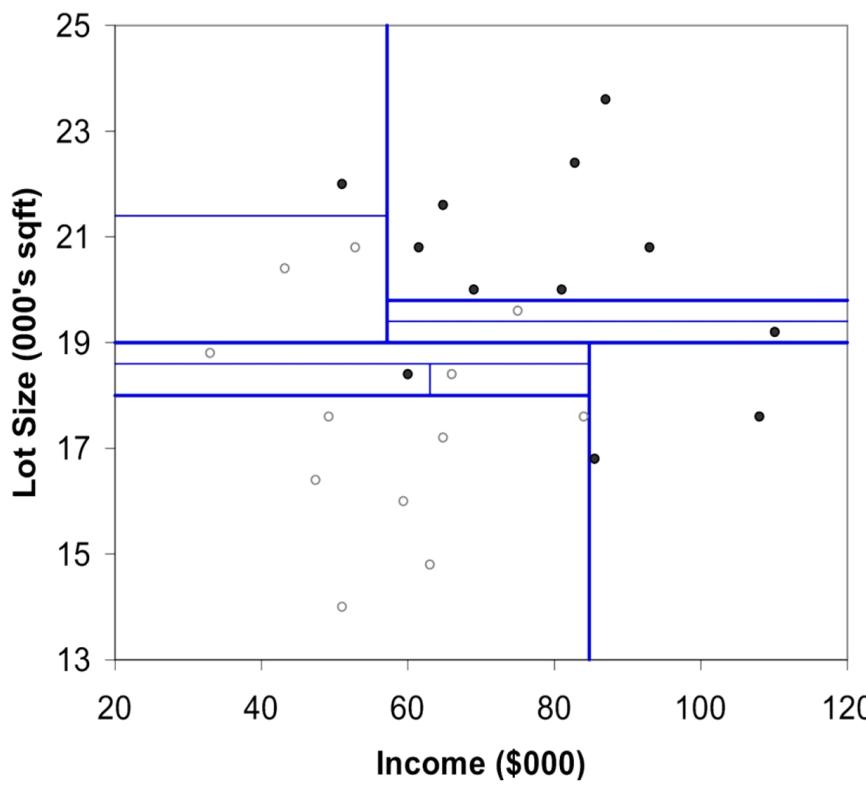
- 가지치기 종료



2 모델 생성 방법

가지치기

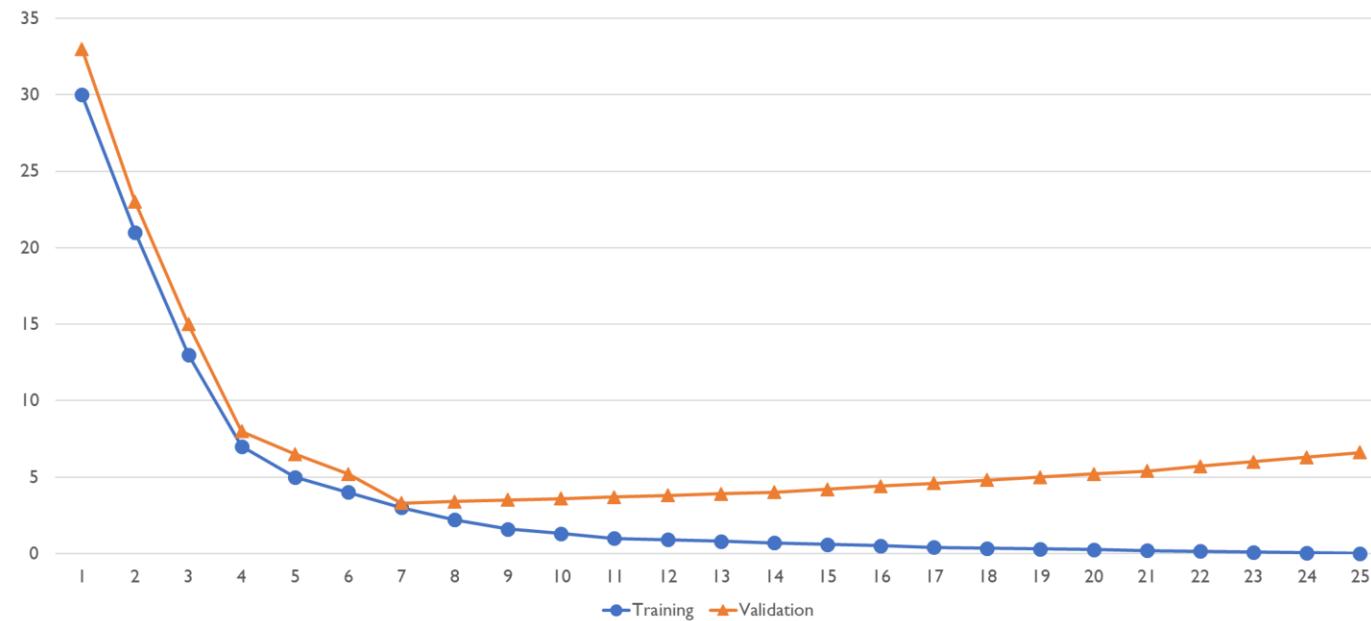
- 가지치기 수행 전(좌측)-후(우측) 비교



2 모델 생성 방법

사전적 가지치기(Pre pruning)

- 사후적 가지치기(Post-Pruning)은 Full Tree를 생성한 뒤 다시 최적의 구조를 찾아 의사결정나무를 단순화시키는 방식



최적의 의사결정나무 구조를 찾기 위해 돌아오는 방식

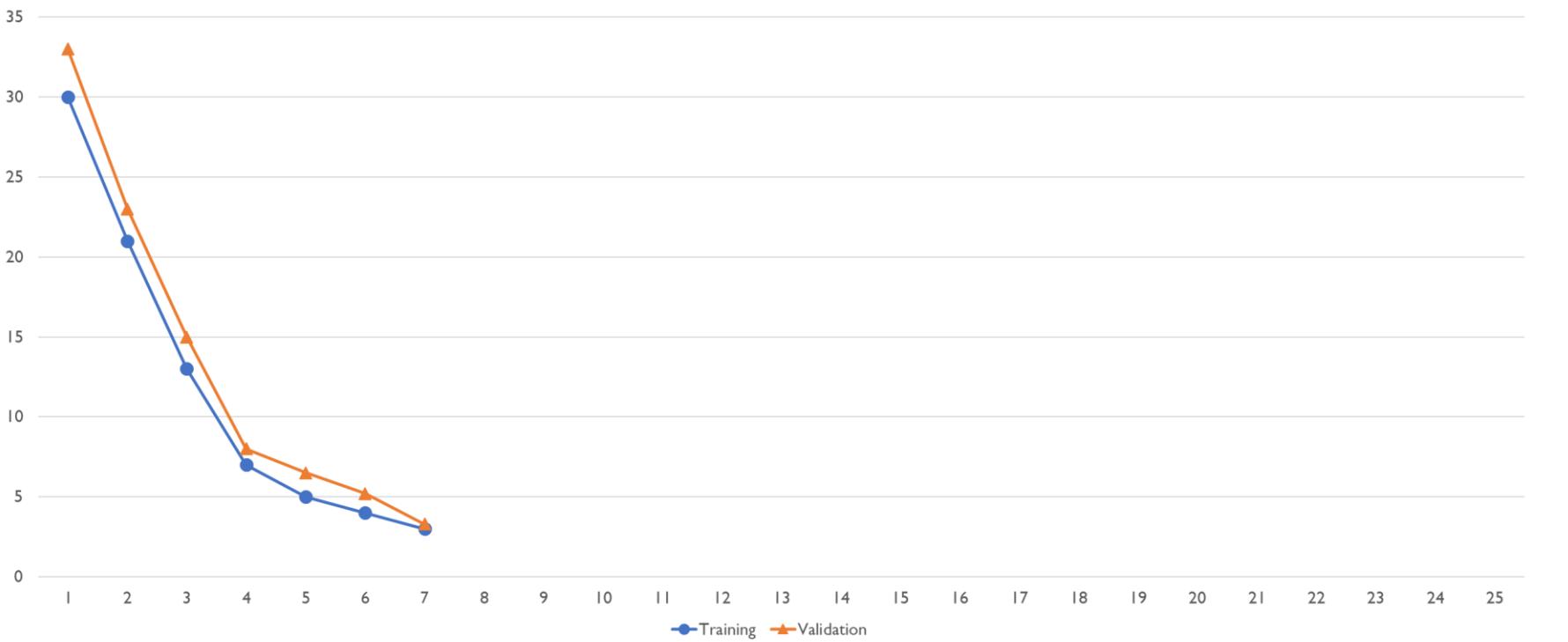
재귀적 분기를 통해 Full Tree를 생성한 후

Pre pruning

2 모델 생성 방법

사전적 가지치기(Pre pruning)

- 처음부터 적정 수준에서 멈추기

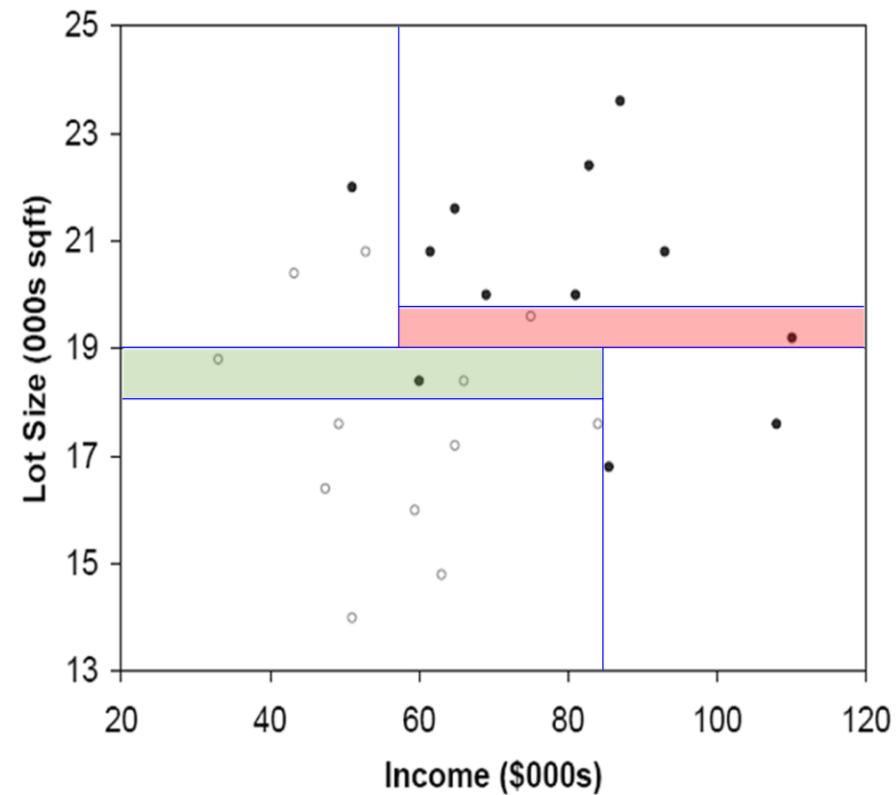


적절한 시점에서 더 이상 의사결정나무를 분기하지 않도록 제약조건을 부여

2 모델 생성 방법

사전적 가지치기(Pre pruning)

- 제약 요건 예시
 - ① 분기 전후 Information Gain의 최저 기준
 - ② 분기 대상이 되는 노드에 속하는 최소 객체 수
 - ③ 의사결정나무가 가질 수 있는 최대 깊이 등
- 오른쪽 그림에서 분기 대상이 되는 노드에 속하는 최소 객체 수를 3으로 설정하면
 - : Owner와 Non-owner가 혼재되어 있는 영역이지만 기준을 충족하지 못하여 더 이상 분기가 되지 않음
 - : Owner와 Non-owner가 혼재되어 있는 영역이며, 기준을 충족하여 분기 가능



3 의사결정 나무 예시

예시 자료

- Universal bank의 개인신용대출 예측
- 고객의 인구통계학적 정보와 은행 이용 행태를 바탕으로 개인신용대출을 이용할 고객 판별

일련 번호	나이	경력	소득	가족 수	월별 신용카드 평균사용액	교육 수준	담보부 채권	개인 대출	증권 계좌	CD 계좌	온라인 뱅킹	신용 카드
1	25	1	49	4	1.60	UG	0	No	Yes	No	No	No
2	45	19	34	3	1.50	UG	0	No	Yes	No	No	No
3	39	15	11	1	1.00	UG	0	No	No	No	No	No
4	35	9	100	1	2.70	Grad	0	No	No	No	No	No
5	35	8	45	4	1.00	Grad	0	No	No	No	No	Yes
6	37	13	29	4	0.40	Grad	155	No	No	No	Yes	No
7	53	27	72	2	1.50	Grad	0	No	No	No	Yes	No
8	50	24	22	1	0.30	Prof	0	No	No	No	No	Yes
9	35	10	81	3	0.60	Grad	104	No	No	No	Yes	No
10	34	9	180	1	8.90	Prof	0	Yes	No	No	No	No
11	65	39	105	4	2.40	Prof	0	No	No	No	No	No
12	29	5	45	3	0.10	Grad	0	No	No	No	Yes	No
13	48	23	114	2	3.80	Prof	0	No	Yes	No	No	No
14	59	32	40	4	2.50	Grad	0	No	No	No	Yes	No
15	67	41	112	1	2.00	UG	0	No	Yes	No	No	No
16	60	30	22	1	1.50	Prof	0	No	No	No	Yes	Yes
17	38	14	130	4	4.70	Prof	134	Yes	No	No	No	No
18	42	18	81	4	2.40	UG	0	No	No	No	No	No
19	46	21	193	2	8.10	Prof	0	Yes	No	No	No	No
20	55	28	21	1	0.50	Grad	0	No	Yes	No	No	Yes

3 의사결정 나무 예시

예시 자료

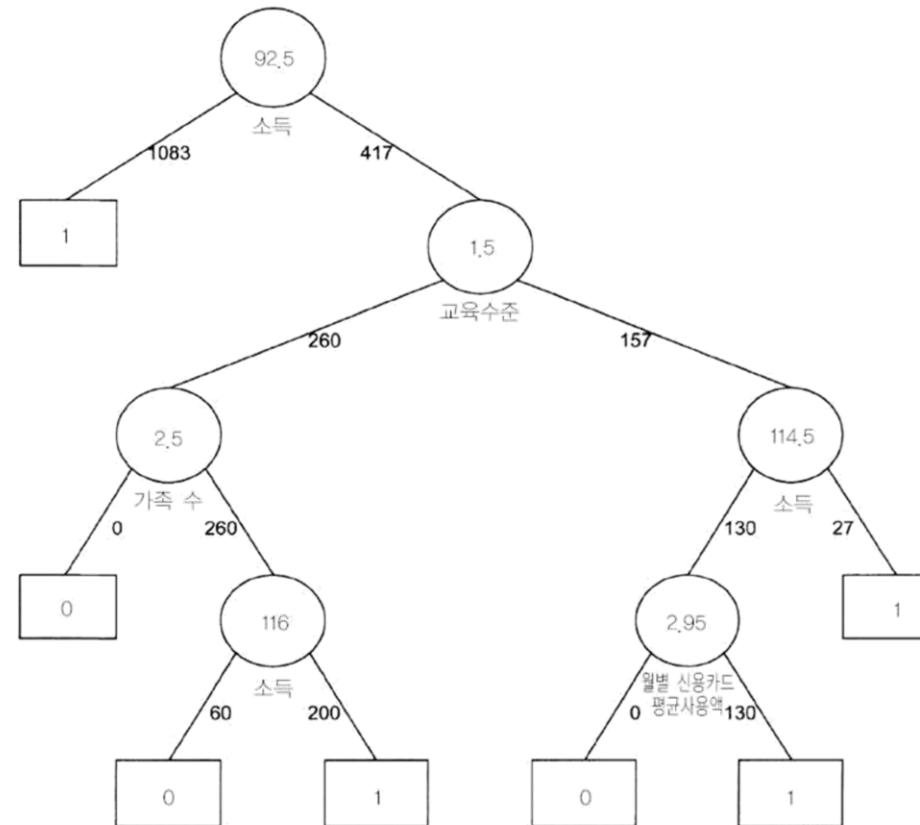
- Universal bank의 개인신용대출 예측
- 고객의 인구통계학적 정보와 은행 이용 행태를 바탕으로 개인신용대출을 이용할 고객 판별

의사결정 마디	학습용 집합의 오차율	평가용 집합의 오차율
41	0	2.133333
40	0.04	2.2
39	0.08	2.2
38	0.12	2.2
37	0.16	2.066667
36	0.2	2.066667
35	0.2	2.066667
34	0.24	2.066667
• • •		
13	1.16	1.6
12	1.2	1.6
11	1.2	1.466667
10	1.6	1.666667
9	2.2	1.666667
8	2.2	1.866667
7	2.24	1.866667
6	2.24	1.6
5	4.44	1.8
4	5.08	2.333333
3	5.24	3.466667
2	9.4	9.533333
1	9.4	9.533333
0	9.4	9.533333

3 의사결정 나무 예시

예시 자료

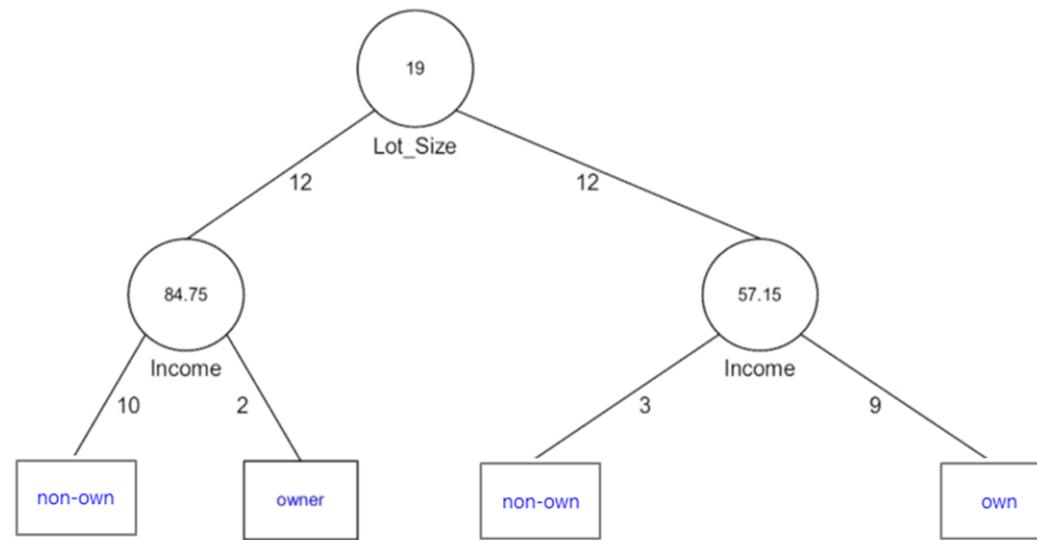
- Universal bank의 개인신용대출 예측
- 고객의 인구통계학적 정보와 은행 이용 행태를 바탕으로 개인신용대출을 이용할 고객 판별



3 의사결정 나무 예시

예시 자료

- 최적 의사결정으로부터 대출/미대출 분류 규칙을 생성



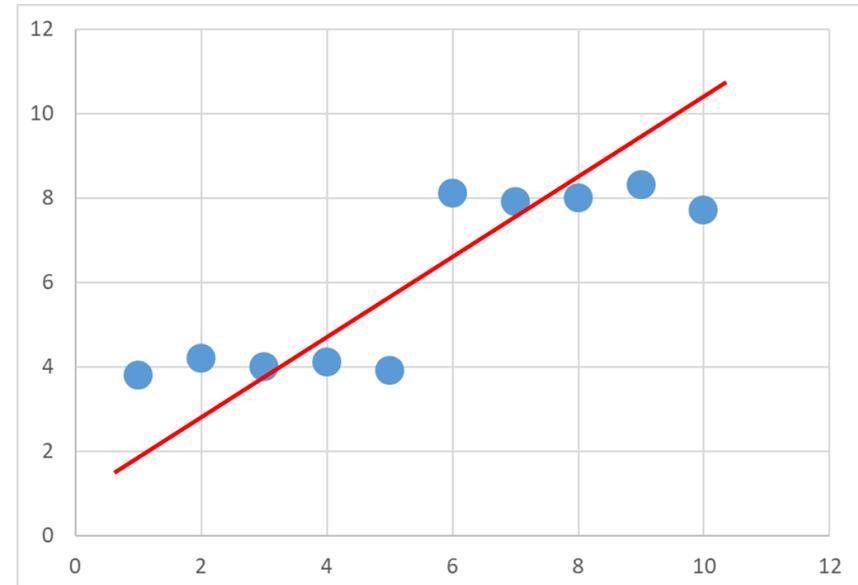
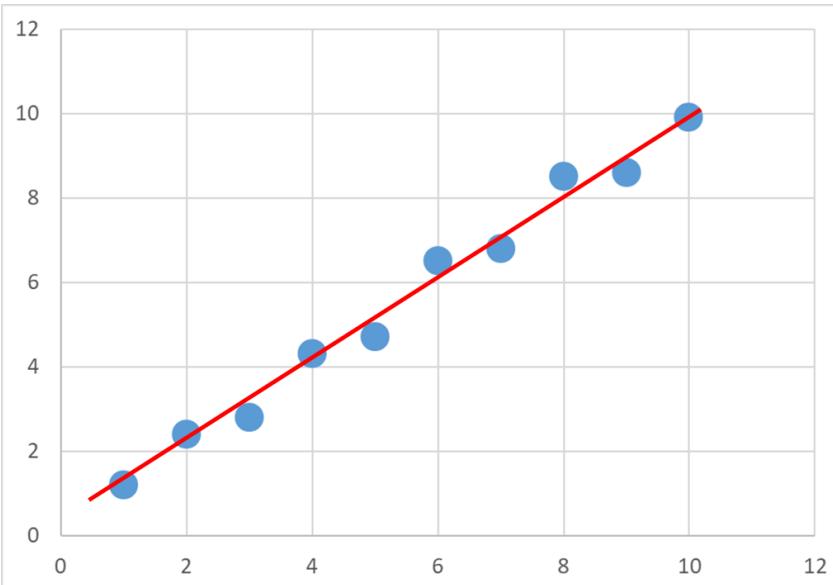
- IF(Lot size < 19) AND IF(Income < 84.75) THEN Owner = No
- IF(Lot size < 19) AND IF(Income > 84.75) THEN Owner = YES
- IF(Lot size > 19) AND IF(Income < 57.15) THEN Owner = NO
- IF(Lot size > 19) AND IF(Income > 57.15) THEN Owner = YES

의사결정나무: 회귀

1 회귀나무 개요

회귀나무의 형태

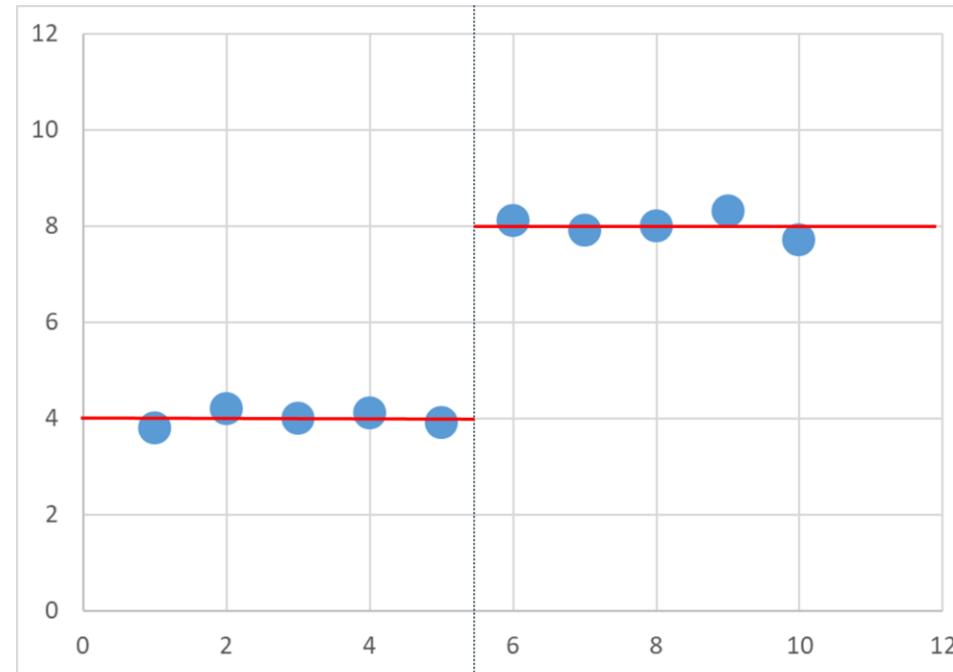
- 아래 그림에서 왼쪽과 같은 형태는 선형 회귀분석으로 추정하는 것이 적합함
- 오른쪽 그림과 같은 형태는 선형 회귀분석으로 추정하는 것이 부적합함



1 회귀나무 개요

말단노드의 예측 값

- 해당 노드에 속하는 모든 개체의 종속변수 값의 평균으로 예측
- 아래 예시에서 split point = 5.5, if $x < 5.5$, $y = 4$, if $x > 5.5$, $y = 8$

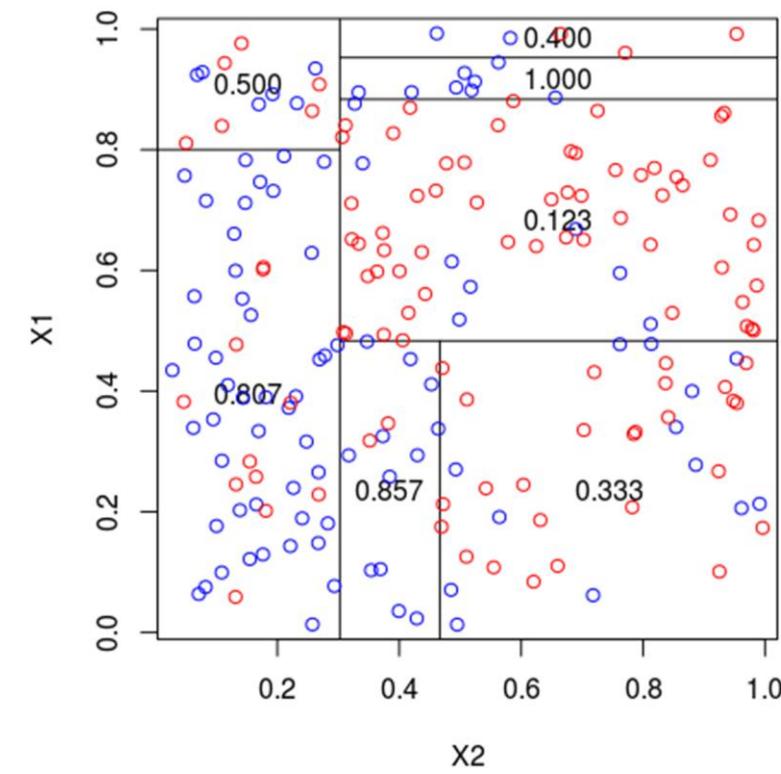
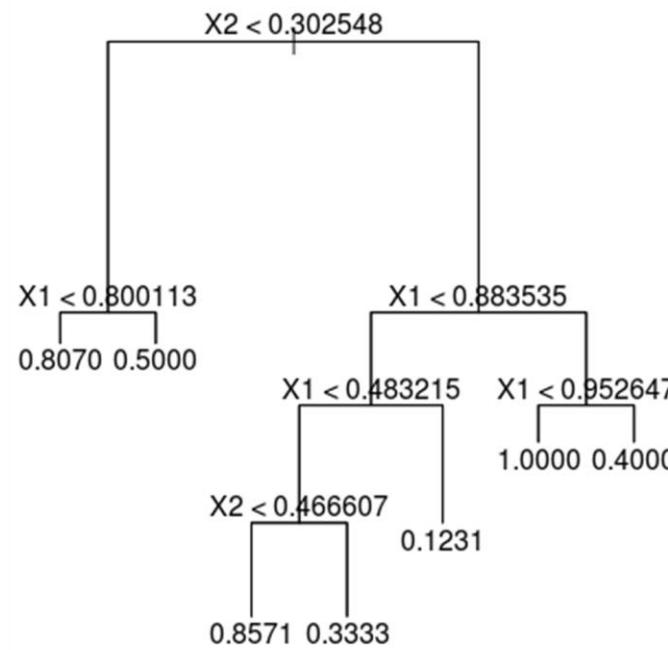


Split point

1 회귀나무 개요

말단 노드의 예측 값

- 해당 노드에 속하는 모든 개체의 종속변수 값의 평균으로 예측



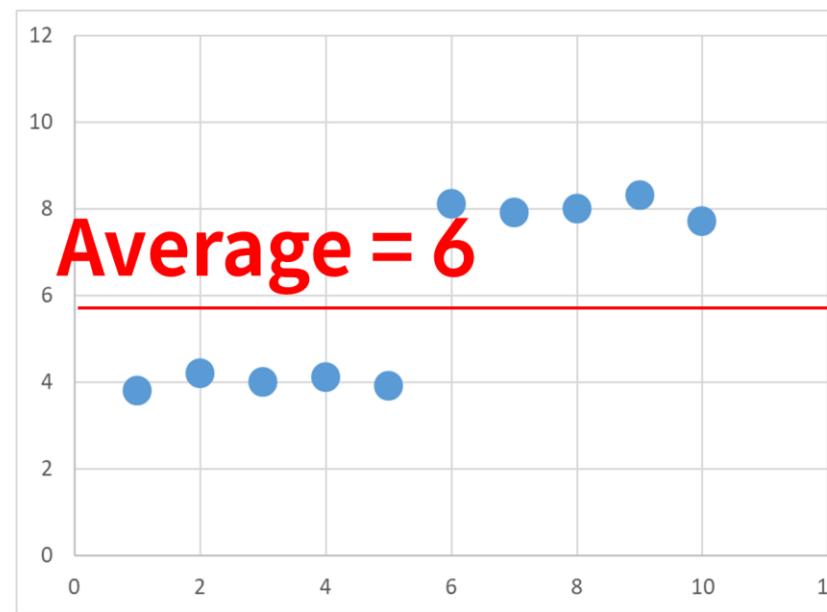
2 회귀나무 모형 추정

회귀모형에서 불순도를 측정하는 과정

- 불순도(impurity) 측정

- Sum of the squared error (SSE: $\sum_{i=1}^n (y_i - \hat{y})^2$)
- 예시

X	Y
1	3.8
2	4.2
3	4.0
4	4.1
5	3.9
6	8.2
7	7.8
8	8.0
9	8.4
10	7.6



$$SSE(\text{Parent}) = (3.8 - 6.0)^2 + (4.2 - 6.0)^2 + \dots + (7.6 - 6.0)^2 = 40.5$$

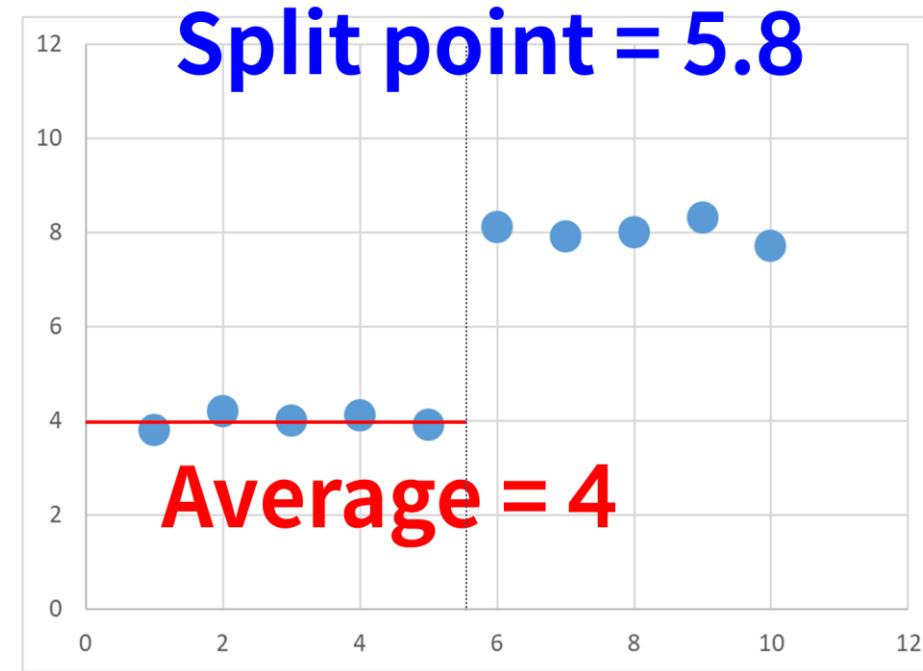
2 회귀나무 모형 추정

회귀모형에서 불순도를 측정하는 과정

- 불순도(impurity) 측정

- Sum of the squared error (SSE: $\sum_{i=1}^n (y_i - \hat{y})^2$)
- 예시

X	Y
1	3.8
2	4.2
3	4.0
4	4.1
5	3.9



$$SSE(\text{Left Child}) = (3.8 - 4.0)^2 + (4.2 - 4.0)^2 + \dots + (3.9 - 4.0)^2 = 0.1$$

2 회귀나무 모형 추정

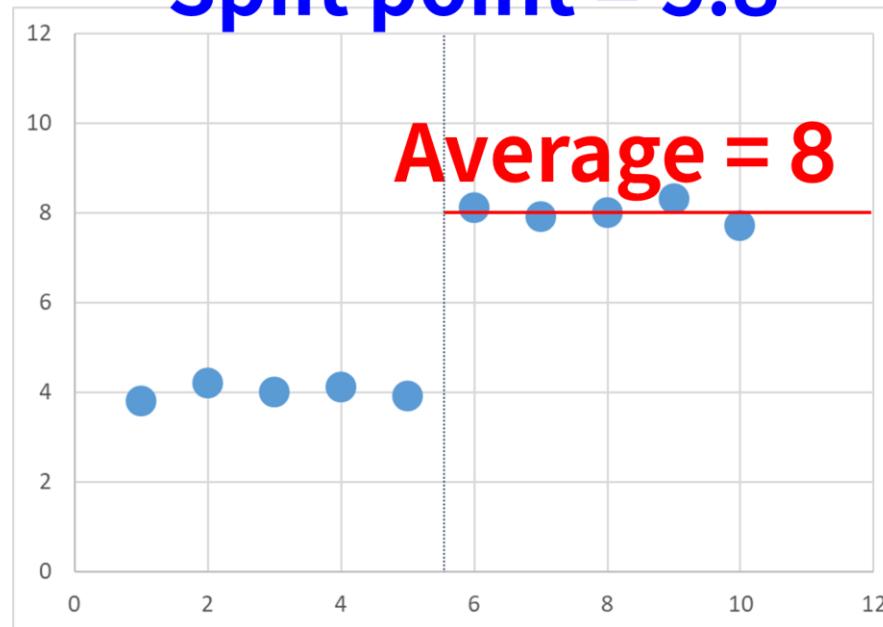
회귀모형에서 불순도를 측정하는 과정

- 불순도(impurity) 측정

- Sum of the squared error (SSE: $\sum_{i=1}^n (y_i - \hat{y})^2$)
- 예시

X	Y
6	8.2
7	7.8
8	8.0
9	8.4
10	7.6

Split point = 5.8

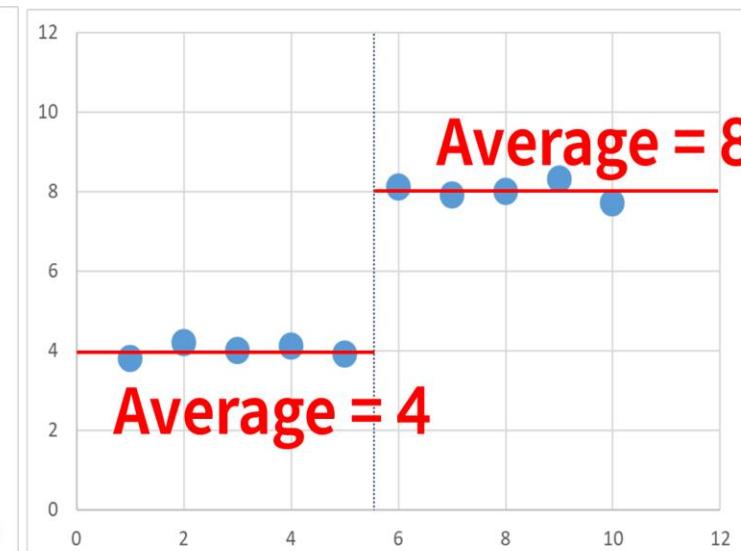
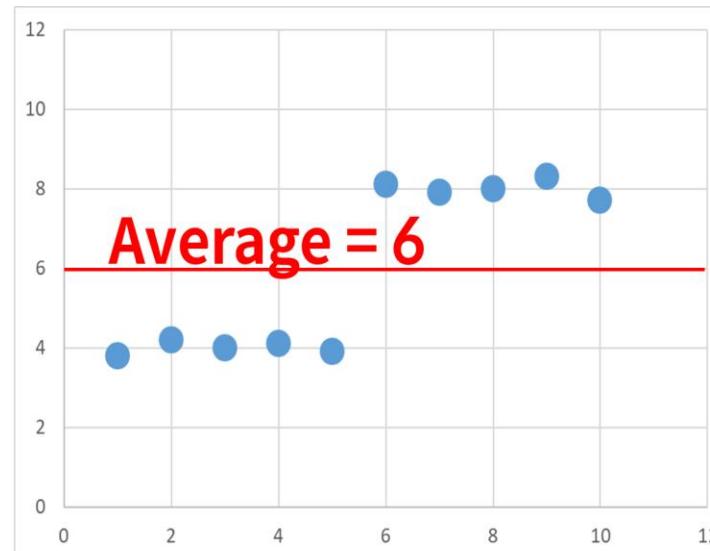


$$SSE(\text{Right Child}) = (8.2 - 8.0)^2 + (7.8 - 8.0)^2 + \dots + (7.6 - 8.0)^2 = 0.4$$

2 회귀나무 모형 추정

회귀모형에서 불순도를 측정하는 과정

X	Y
1	3.8
2	4.2
3	4.0
4	4.1
5	3.9
6	8.2
7	7.8
8	8.0
9	8.4
10	7.6



$$SSE(\text{Parent}) = (3.8 - 6.0)^2 + (4.2 - 6.0)^2 + \dots + (7.6 - 6.0)^2 = 40.5$$

$$SSE(\text{Left Child}) = (3.8 - 4.0)^2 + (4.2 - 4.0)^2 + \dots + (3.9 - 4.0)^2 = 0.1$$

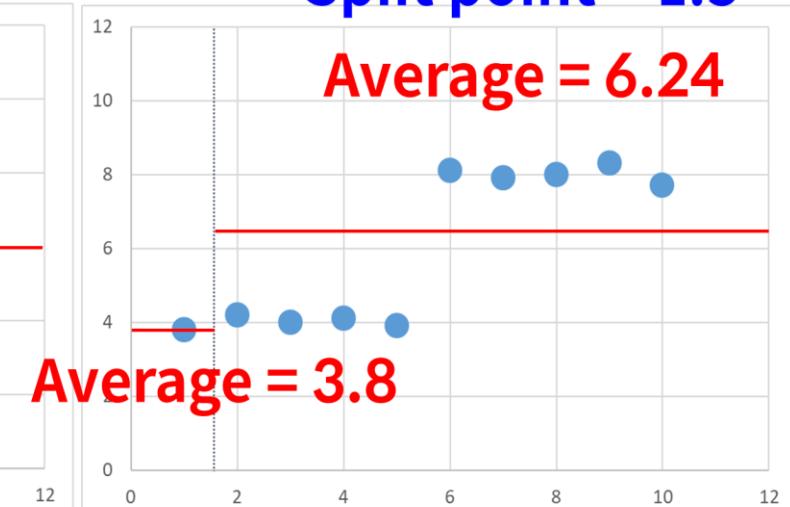
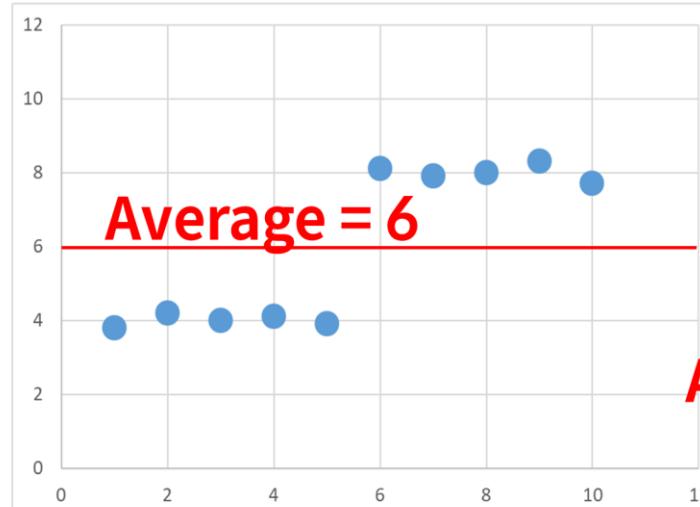
$$SSE(\text{Right Child}) = (8.2 - 8.0)^2 + (7.8 - 8.0)^2 + \dots + (7.6 - 8.0)^2 = 0.4$$

$$\text{Information Gain} = 40.5 - (0.1 + 0.4) = 40.0$$

2 회귀나무 모형 추정

분기 후보 1

X	Y
1	3.8
2	4.2
3	4.0
4	4.1
5	3.9
6	8.2
7	7.8
8	8.0
9	8.4
10	7.6



$$SSE(\text{Parent}) = (3.8 - 6.0)^2 + (4.2 - 6.0)^2 + \dots + (7.6 - 6.0)^2 = 40.5$$

$$SSE(\text{Left Child}) = (3.8 - 3.8)^2 = 0$$

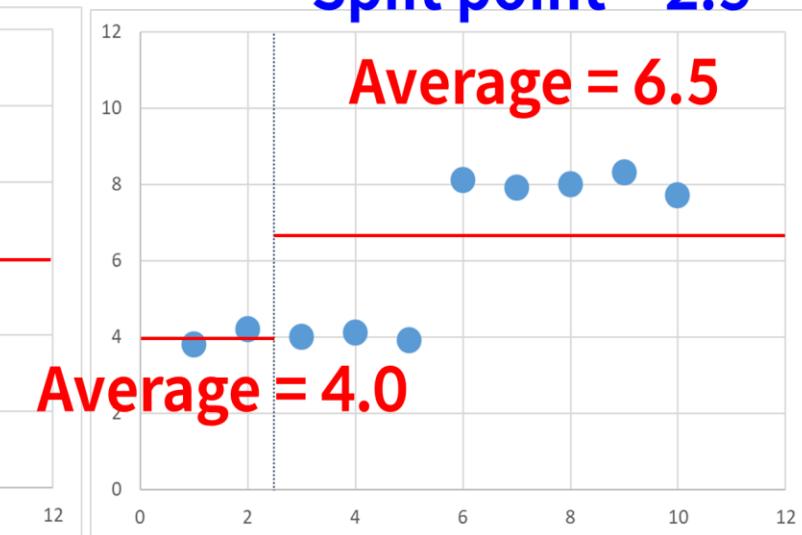
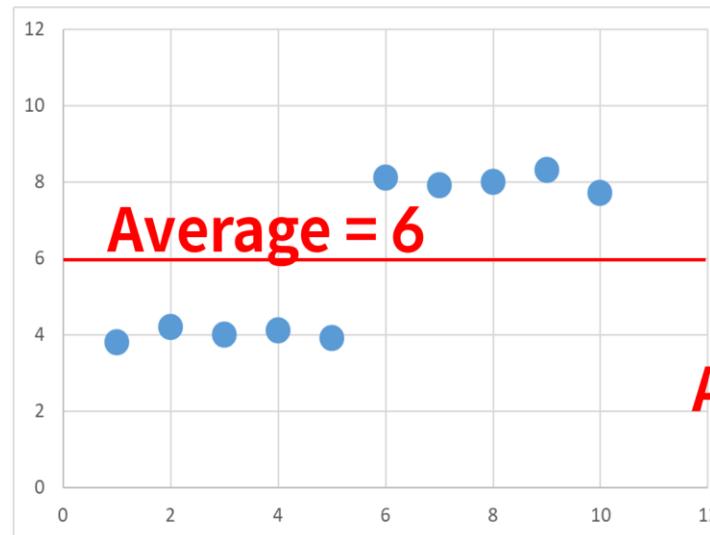
$$SSE(\text{Right Child}) = (4.2 - 6.24)^2 + (4.0 - 6.24)^2 + \dots + (7.6 - 6.24)^2 = 34.92$$

$$\text{Information Gain} = 40.5 - (0 + 34.92) = 5.08$$

2 회귀나무 모형 추정

분기 후보 2

X	Y
1	3.8
2	4.2
3	4.0
4	4.1
5	3.9
6	8.2
7	7.8
8	8.0
9	8.4
10	7.6



$$SSE(\text{Parent}) = (3.8 - 6.0)^2 + (4.2 - 6.0)^2 + \dots + (7.6 - 6.0)^2 = 40.5$$

$$SSE(\text{Left Child}) = (3.8 - 4.0)^2 + (4.2 - 4.0)^2 = 0.08$$

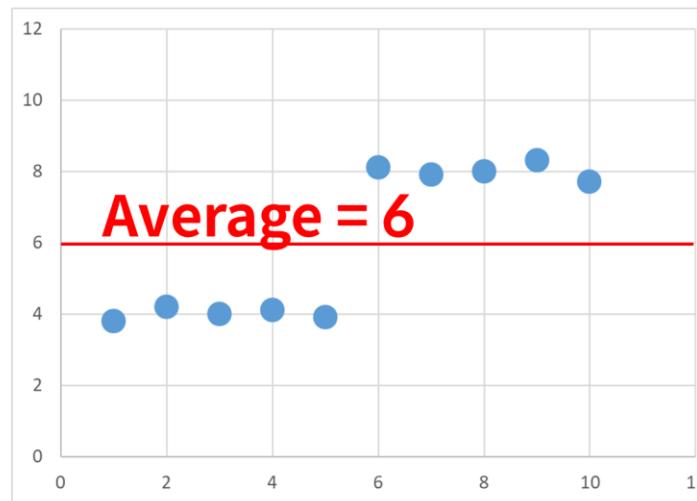
$$SSE(\text{Right Child}) = (4.0 - 6.5)^2 + (4.1 - 6.5)^2 + \dots + (7.6 - 6.5)^2 = 30.22$$

$$\text{Information Gain} = 40.5 - (0.08 + 30.22) = 9.7$$

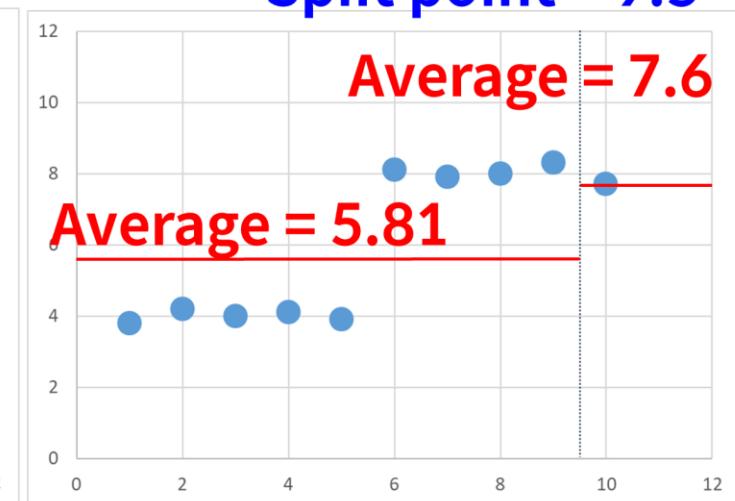
2 회귀나무 모형 추정

분기 후보 9

X	Y
1	3.8
2	4.2
3	4.0
4	4.1
5	3.9
6	8.2
7	7.8
8	8.0
9	8.4
10	7.6



Split point = 9.5



$$SSE(\text{Parent}) = (3.8 - 6.0)^2 + (4.2 - 6.0)^2 + \dots + (7.6 - 6.0)^2 = 40.5$$

$$SSE(\text{Left Child}) = (3.8 - 5.81)^2 + (4.2 - 5.81)^2 + \dots + (8.4 - 5.81)^2 = 30.45$$

$$SSE(\text{Right Child}) = (7.6 - 7.6)^2 = 0$$

$$\text{Information Gain} = 40.5 - (30.45 + 0) = 10.05$$

3 회귀나무 예시

토요타 코롤라 중고차 가격 예측 문제



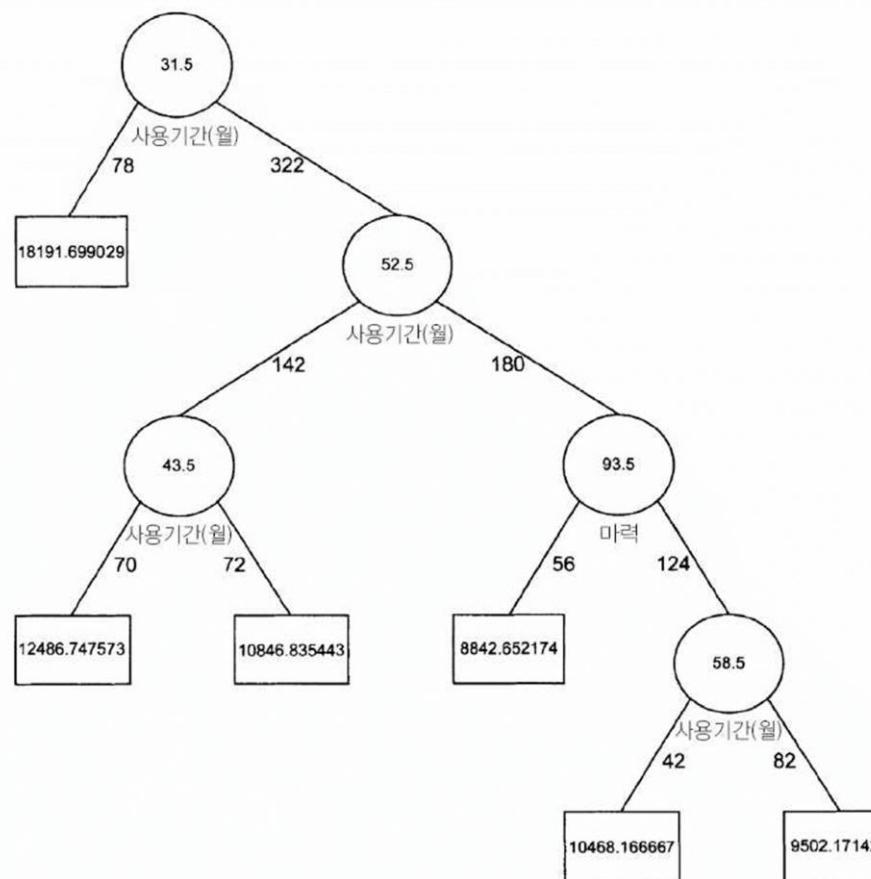
Dependent variable
(target)

Independent variables
(attributes, features)

Variable	Description
Price	Offer Price in EUROS
Age_08_04	Age in months as in August 2004
KM	Accumulated Kilometers on odometer
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)
HP	Horse Power
Met_Color	Metallic Color? (Yes=1, No=0)
Automatic	Automatic ((Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters
Doors	Number of doors
Quarterly_Tax	Quarterly road tax in EUROS
Weight	Weight in Kilograms

3 회귀나무 예시

토요타 코롤라 중고차 가격 예측 문제



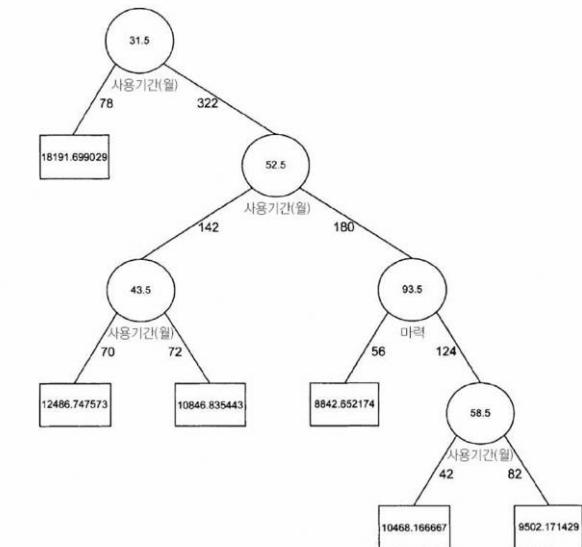
3 회귀나무 예시

의사결정나무의 장점

- 의사결정나무는 예측에 대한 설명을 제공할 수 있음
- 변수 선택 과정이 자동적으로 수행됨
- 특별한 통계적 가정을 요구하지 않음
- 결측치가 존재하는 상황에서도 모델 구축이 가능함

의사결정나무의 단점

- 한 번에 하나의 변수만 고려하므로 변수간 상호작용을 파악하기 어려움



sk4me@konkuk.ac.kr