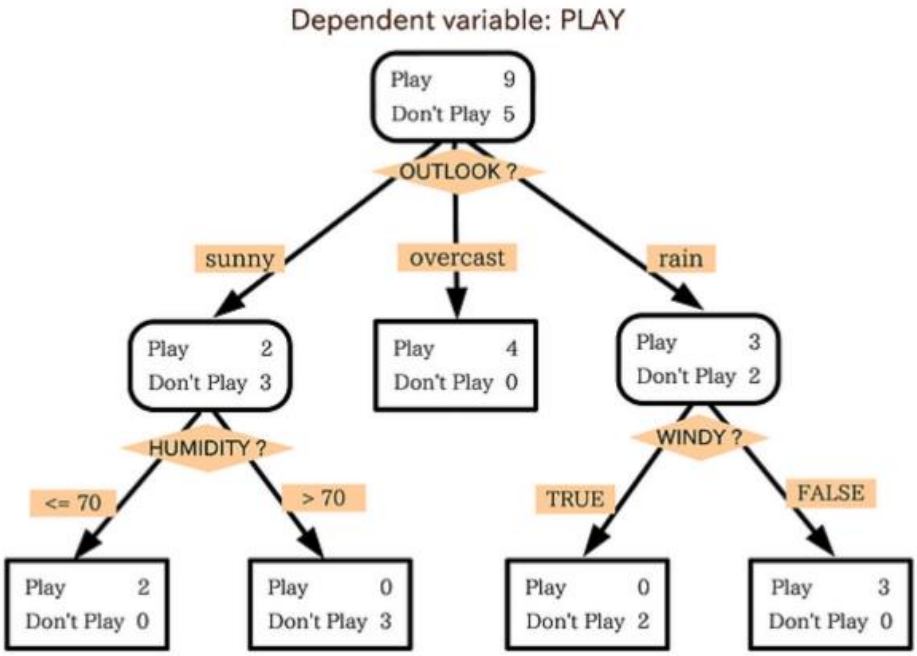
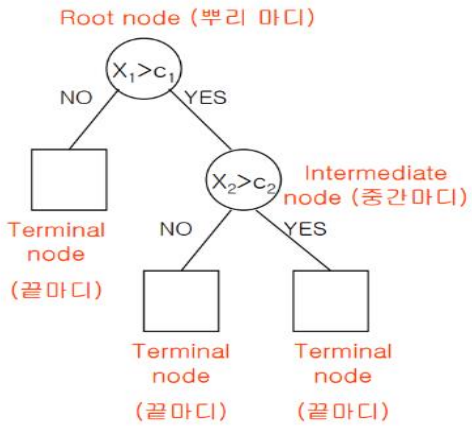
# 의사결정 나무(Decision Tree)

## 의사결정 나무란?

-한번에 하나씩 설명변수(x)를 사용하여 예측 가능한 규칙들의 집합을 생성하는 알고리즘.




## Decision Tree의 구조



초기지점인 root node로부터 시작해서 분기를 시작하여

분기를 거듭할수록 각 분류에 해당하는 데이터 개수는 줄어든다.

모든 Terminal Node안에 존재하는 데이터 개수를 합치면 Root Node의 데이터 개수와 같다.(모든 Terminal Node들은 교집합이 없음)

## CART method(Classification And Regression Tree)

-의사결정 나무를 생성하는 알고리즘 중 가장 대중적인 방법론

-이름 자체에서 유추할 수 있는 것처럼 분류(이산형)와 예측(연속형) 모두 가능

Regression 척도(회귀나무)

RSS (Residual Sum of Square)을 최소로 만든다.

분기를 위해 데이터를 나눌 때 각 범위의 데이터 값과 그 범위의 평균 차이의 제곱이 가장 작아질 때 분기한다.

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Classification 척도(분류 나무)

구분 뒤 각 영역의 순도(homogeneity)의 증가,불확실성(uncertainty) 혹은 불순도의 감소 ➔ 즉, 정보획득을 분기의 목표로 둠

순도를 계산하는 방법

1. 지니계수
2. Cross-entropy(불확실성)

$$Entropy(A) = \sum_{i=1}^{d} R_i \left( - \sum_{k=1}^{m} p_k \log_2 (p_k) \right)$$

0~1사이의 값을 가지며 0에 가까울수록 좋다.

재귀적 분기와 가지치기

| 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 |

한 변수를 기준으로 정렬한 후 가능 한 모든 분기점에 대해서

Cross entropy를 구해 분기 전과 비교한다.

| 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 |

Lot size변수로 정렬 후

1번 || 나머지23개로 분기 설정

Cross-entropy 계산

1~2번 || 나머지22개로 분기 설정

Cross-entropy 계산

.

(반복)

.

1~23번 || 나머지 1개로 분기 설정

Cross-entropy 계산

나머지 변수인 income으로 동일과정을 반복한후

정보획득이 가장 큰 (분기 후 Cross-entropy가 가장 크게 줄어드는) 분기를 선택

1회 분기를 위해 소용되는 시간은

T(N)=D(N-1) (D는 변수의 개수,N은 분기 전 데이터 개수)

가지치기(pruning)

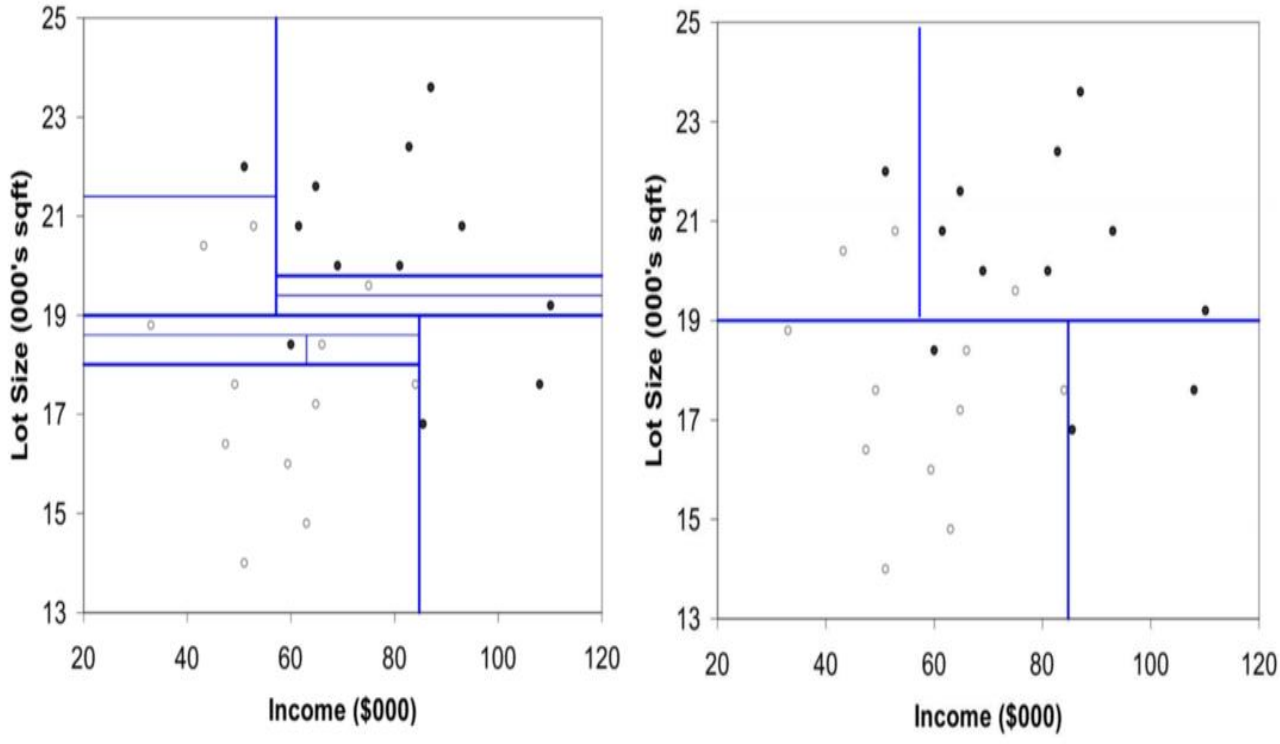
분기를 계속해서 거듭하다 보면 모든 terminal node의 불순도가 0이 되는데

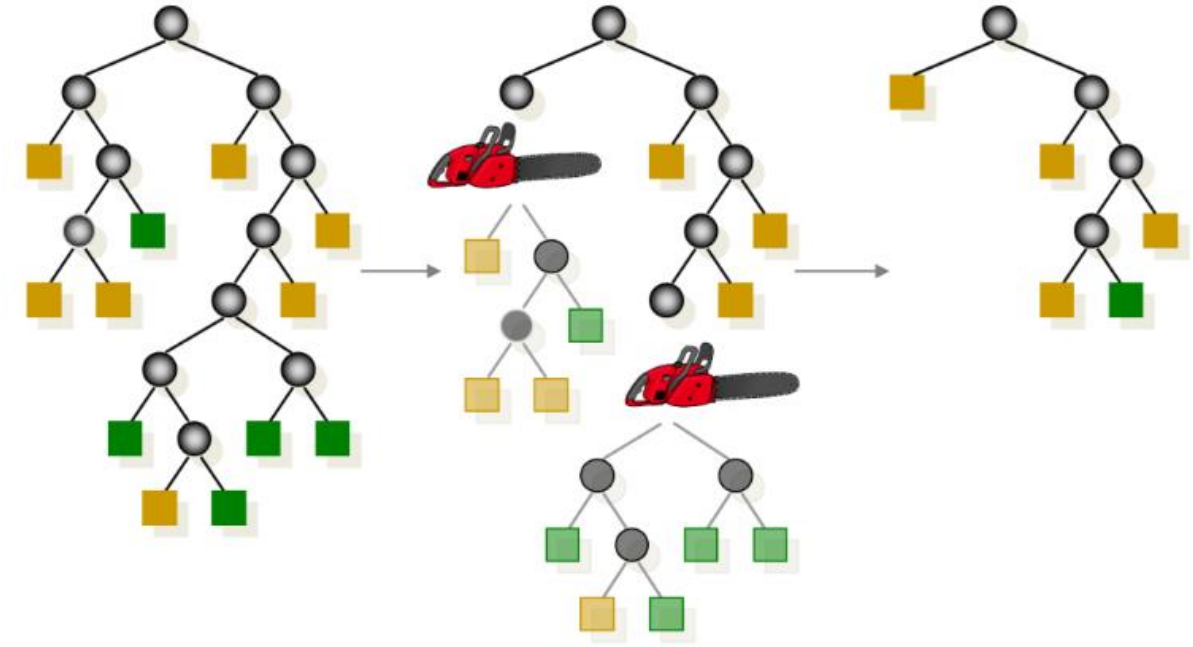
이는 over-fitting을 유발하게 된다.

Terminal node가 너무 많아지면(트리가 너무 깊어지면)

➔overfitting

➔새로운 데이터에 대한 설명력(generalization)감소





Terminal node를 합(merge)하는 방식으로 개수를 줄인다.

Pruning을 하는 방법

**Pruning을 위해서는 각 값에 대한 Cost Function을 정의하고 그 값을 가장 작게 하는 지점에서 가지치기**

CC(T) = Err(T) + a x L(T)

의사결정나무의 비용 복잡도 = 오분류율 + a * Terminal node개수

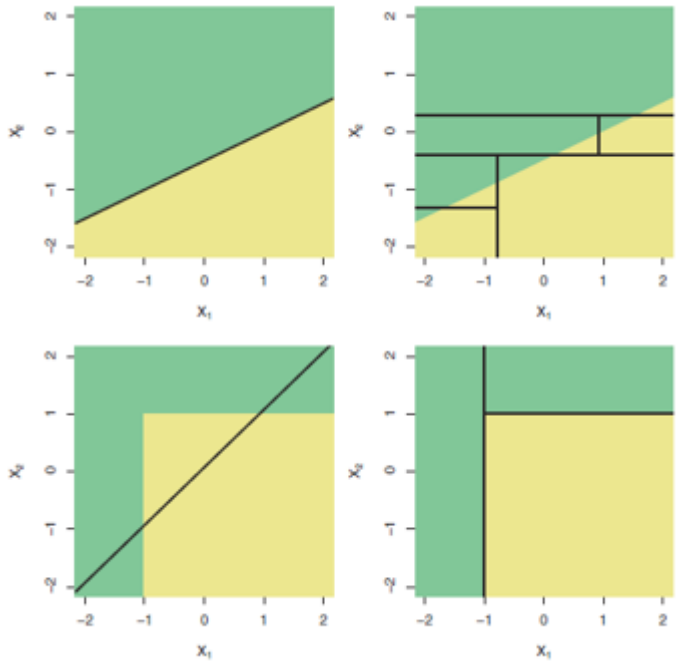
의사결정 나무의 장단점

장점:

구현 및 이해가 쉽다.

연속형 변수와 범주형 변수를 모두 처리할 수 있다.

단점:




데이터의 특성이 특정 변수(축)에 대해 수직/수평적으로 구분되지 못하면 분류율이 떨어진다.

약간의 차이로 분기 지점이 달라지면 전혀 다른 결과값을 만들어 낸다.