의사 결정 나무

의사 결정 나무

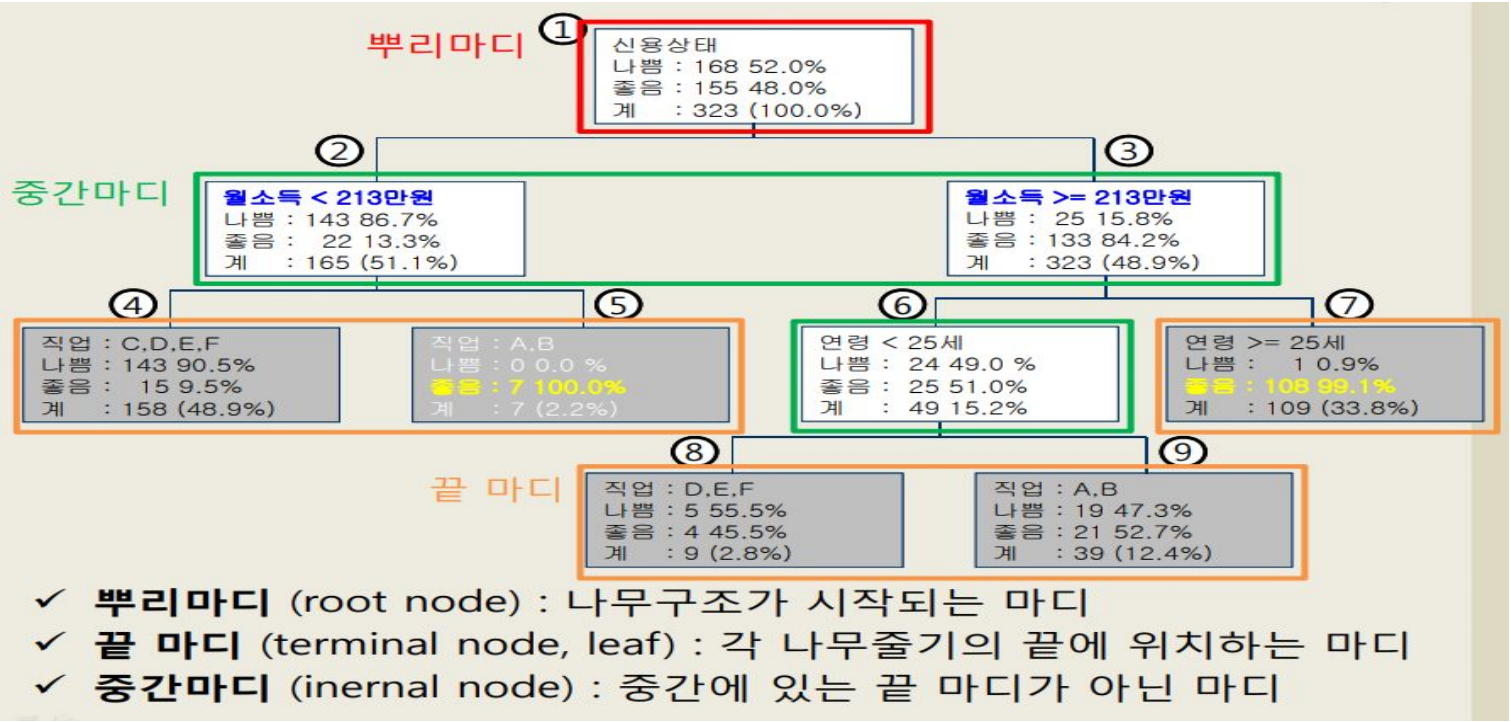
의사결정나무분석은 탐색과 모형화라는 두 가지 특징을 모두 가지고 있다.

즉, 의사결정나무분석은 판별분석, 회귀분석 등과 같은 모수적(parameter) 모형을 분석하기 위해 사전에 이상치(outlier)를 검색하거나 분석에 필요한 변수 또는 모형에 포함되어야 할 상호작용의 효과를 찾아내기 위해서 사용될 수도 있고,

의사결정나무 자체가 분류 또는 예측모형으로 사용될 수도 있다.

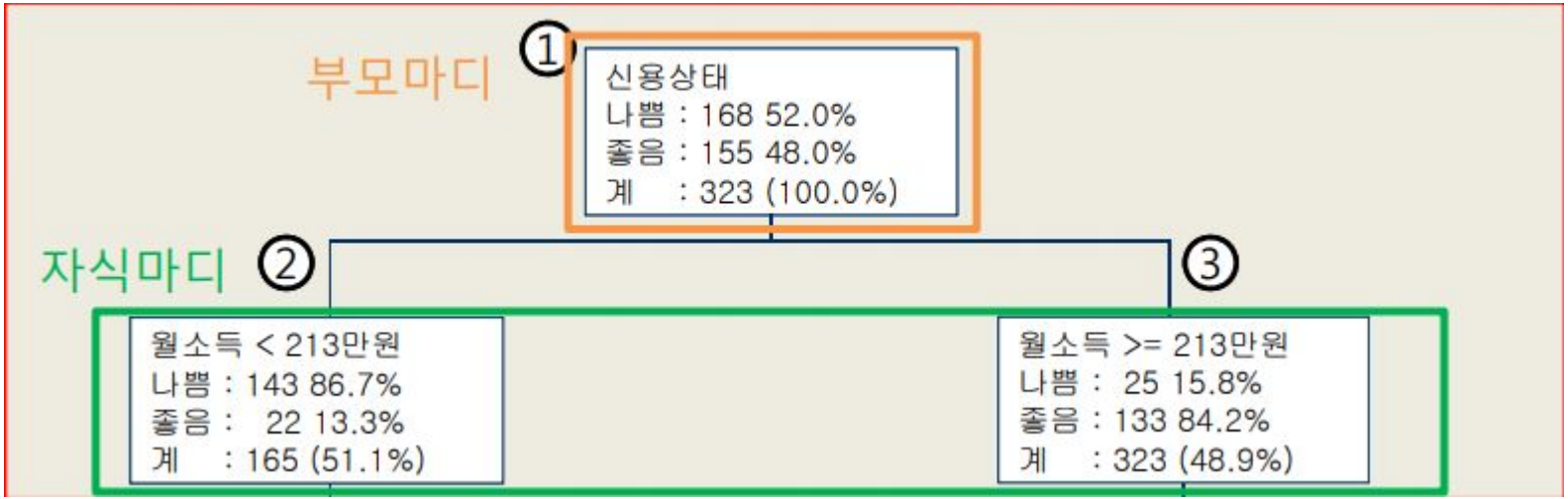
탐색 특징으로 이상치를 검색하거나 분석에 필요한 변수나 모형에 포함되어야 할 상호작용의 효과를 찾는데 사용됨.
모형화 특징으로 의사결정나무 자체가 분류 또는 예측모형으로 사용됨.

의사 결정 나무 구성 요소



- ✓ **뿌리마디** (root node) : 나무구조가 시작되는 마디
- ✓ **끝 마디** (terminal node, leaf) : 각 나무줄기의 끝에 위치하는 마디
- ✓ **중간마디** (inernal node) : 중간에 있는 끝 마디가 아닌 마디

의사 결정 나무 구성 요소



부모마디:자식마디의 상위 마디

자식마디:하나의 마디로 부터 분리된 마디

의사 결정 나무 구성 요소

가지:하나의 마디로 부터 끝 마디 까지 연결된 마디들

마디:가지를 이루고 있는 마디의 개수

의사 결정 나무의 분리 기준

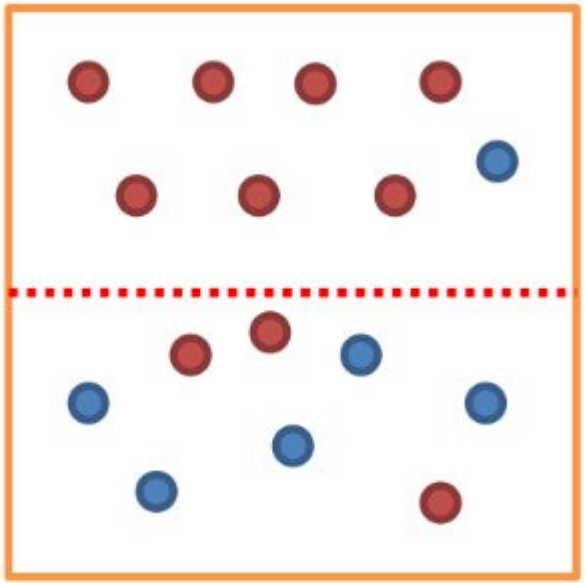
분리기준:어떤 입력변수를 입력하여 분리하는것이 의사결정나무 모델의 목표에 가장 잘 부합 하는지에 대한 기준

목표 변수의 분포를 구분하는 정도:순수도 또는 불순도

순도를 계산하는 방법

1.엔트로피:m개의 레코드중 A영역에 속하는 엔트로피는 다음과 같이 정의 됩니다.

$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2 (p_k)$$



만약 전체 부분의 엔트로피를 계산한다고 하면 다음과 같다.

-10/16 * log2(10/16) - 6/16 * log2(6/16) ~0.95

순도를 계산하는 방법

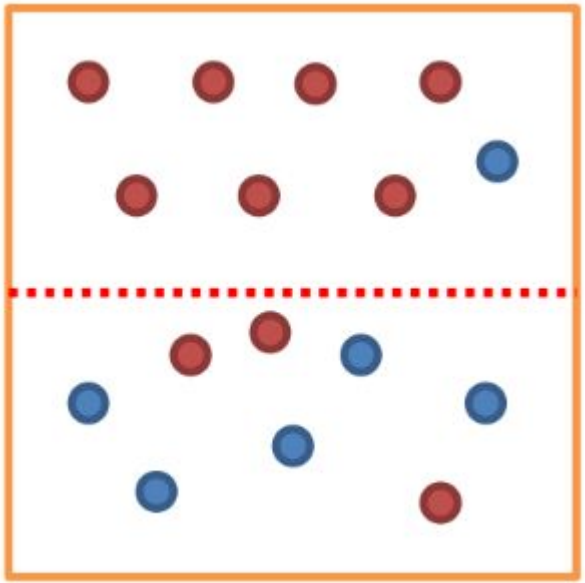
빨간선을 기준으로 엔트로피를 계산하는 방법

전체 넓이를 1이라고 했을때

0.5 *(-⅞*log2(7/8) -⅛*log2(1/8))+0.5*(-⅜*log2(3/8)-5/8log2(5/8)) ~ 0.75

엔트로피 감소를 통해 영역의 순도가 증가하고 불순도의 감소를 만들고

이를 최대한의 정도로 만든다.



순도를 계산하는 방법

2. 지니계수(Gini Index)

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

의사결정나무 분석모델

소득,주택크기,구입여부로 의사결정 나무모델 진행

구입여부를 Y로 소득 및 주택 크기를 x로 진행

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

의사결정나무 분석모델

1.변수를 하나의 기준으로 정렬한다.

첫번째와 나머지 부분을 분리하고 전과의 엔트로피를 비교한다.

나누기전 엔트로피 = **-12/24*log2(12/24)-12*24*log(12/24) = 1**

나누고난 후 엔트로피 **1/24 * (log2(1)) + 23/24 (-12/23*log2(12/23))**

-11/23*log(11/23) ~ 0.96

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

의사결정나무 분석모델

이후 분기 지점을 두번째 레코드로 두고 처음 두 개 레코드와 나머지 **22**개 레코드 간의 엔트로피를 계산한 뒤 정보획득을 알아봅니다. 이렇게 순차적으로 계산한 뒤,

이번엔 다른 변수인 소득을 기준으로 정렬하고 다시 같은 작업을 반복합니다. 모든 경우의 수 가운데 정보획득이 가장 큰 변수와 그 지점을 택해 첫번째 분기를 하게 됩니다.

이후 또 같은 작업을 반복해 두번째, 세번째… 이렇게 분기를 계속 해 나가는 과정이 바로 의사결정나무의 학습입니다.

그렇다면 **1**회 분기를 위해 계산해야 하는 경우의 수는 총 몇 번일까요**?** 개체가 **n**개, 변수가 **d**개라고 할 때 경우의 수는 **d(n−1)**개가 됩니다. 분기를 하지 않는 경우를 제외하고 모든 개체와 변수를 고려해 보는 것입니다.

가지치기

가지치기 **(pruning) :** 적절하지 않은 마디를 제거하여, 적당한 크기의 부나무**(subtree)** 구조를 가지도록 하는 규칙

과적합을 막기 위해 사용

CC(T)=Err(T)+α×L(T)

CC(T)=의사결정나무의 비용 복잡도(=오류가 적으면서 terminal node 수가 적은 단순한 모델일 수록 작은 값)

ERR(T)=검증데이터에 대한 오분류율

L(T)=terminal node의 수(구조의 복잡도)

Alpha=ERR(T)와 L(T)를 결합하는 가중치(사용자에 의해 부여됨, 보통 0.01~0.1의 값을 씀)