R 교육 세미나

ToBig's 8기 류호성

Decision Tree

의사 결정 나무

Contents

Decision Tree

분류 / 예측을 하는 지도학습 방법론

의사결정 규칙(Decision Rule) 에 따라서 Tree를 생성한 후에 분류/예측하는 방법론

분류트리 / 회귀트리

https://ko.wikipedia.org/wiki/%EA%B2%B0%EC%A0%95_%ED%8A%B8%EB%A6%AC_%ED%95%99%EC%8A%B5%EB%B2%95

Unit 01 | Decision tree Intro

감기 유무를 맞추자!

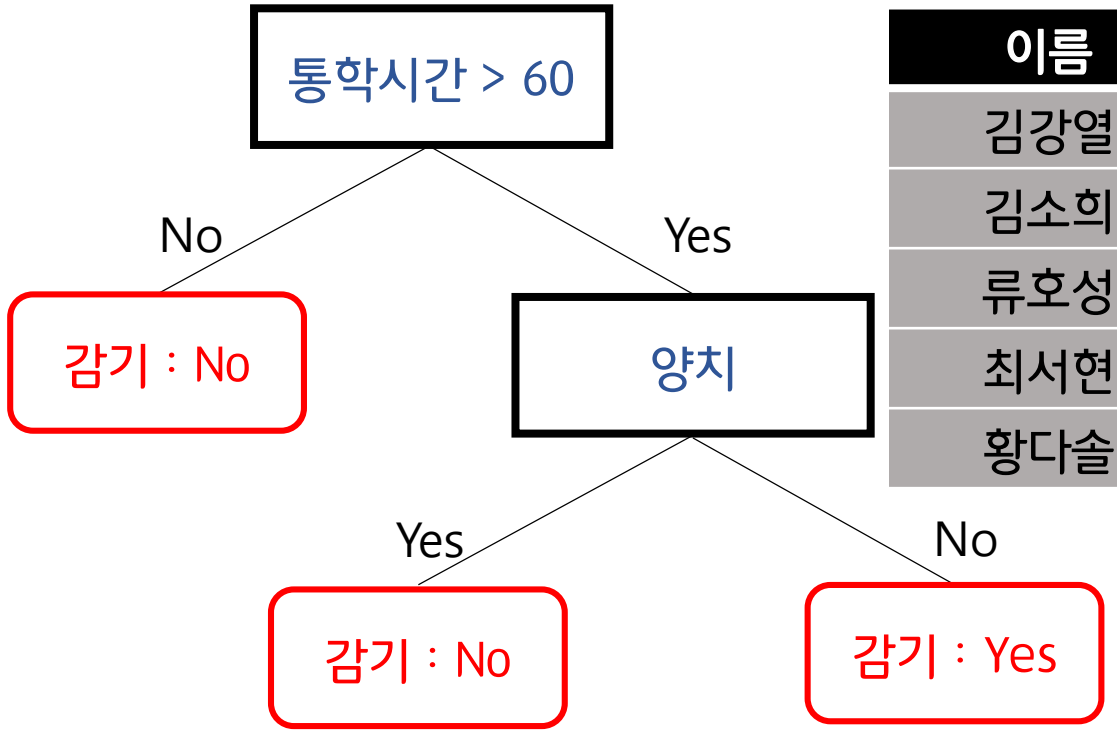
타겟변수
(종속변수)

- 예시 데이터!

이름	통학시간(분)	히트텍	양치	감기
김강열	40	X	X	No
김소희	30	O	O	No
류호성	120	O	X	Yes
최서현	90	X	X	Yes
황다솔	100	O	O	No

Unit 01 | Decision tree Intro

다른 데이터(train data)로
이미 만들어진 Tree Model

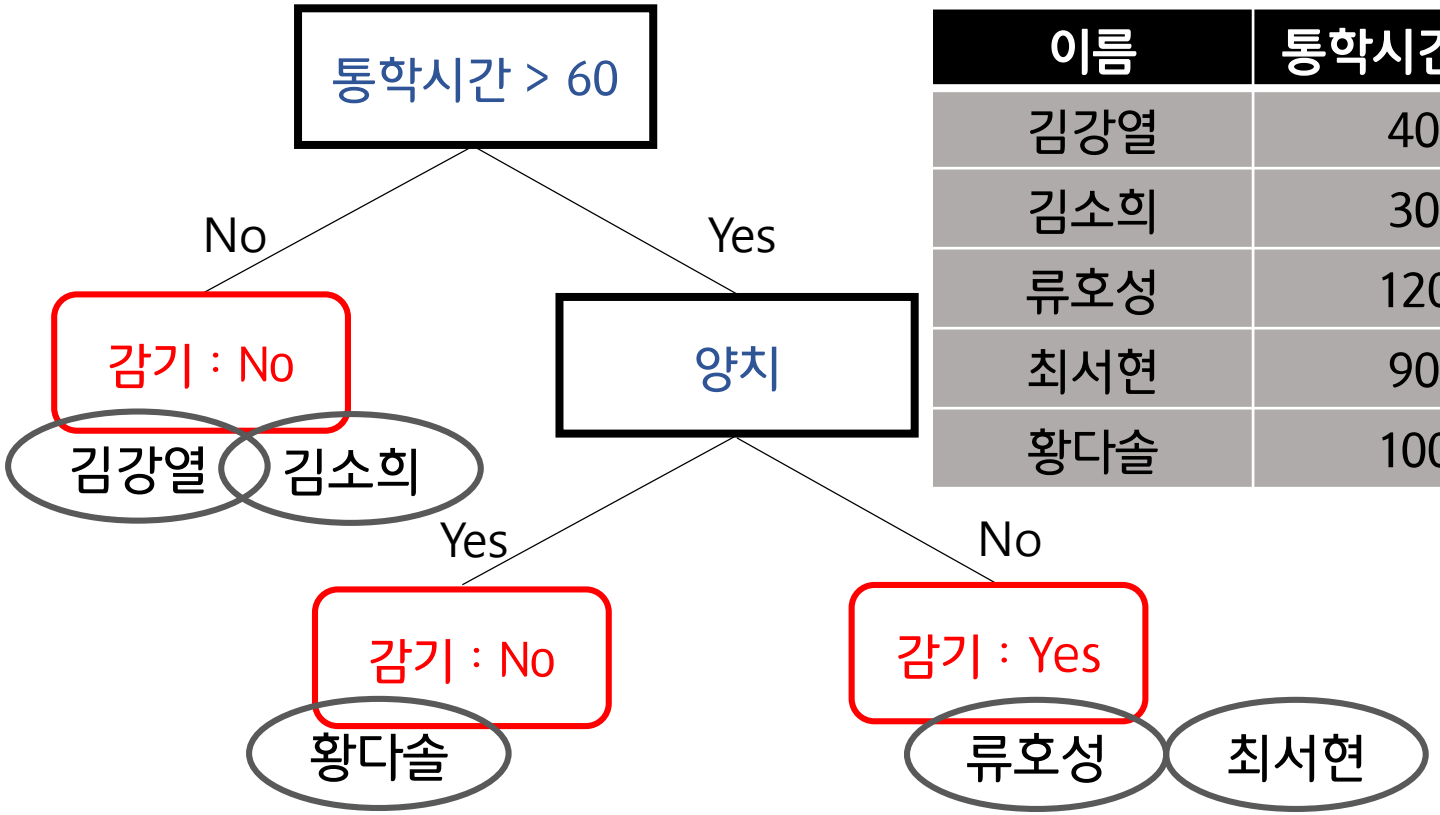


Test data

이름	통학시간(분)	히트텍	양치	감기
김강열	40	X	X	No
김소희	30	O	O	No
류호성	120	O	X	Yes
최서현	90	X	X	Yes
황다솔	100	O	O	No

Unit 01 | Decision tree Intro

다른 데이터(train data)로
이미 만들어진 Tree Model

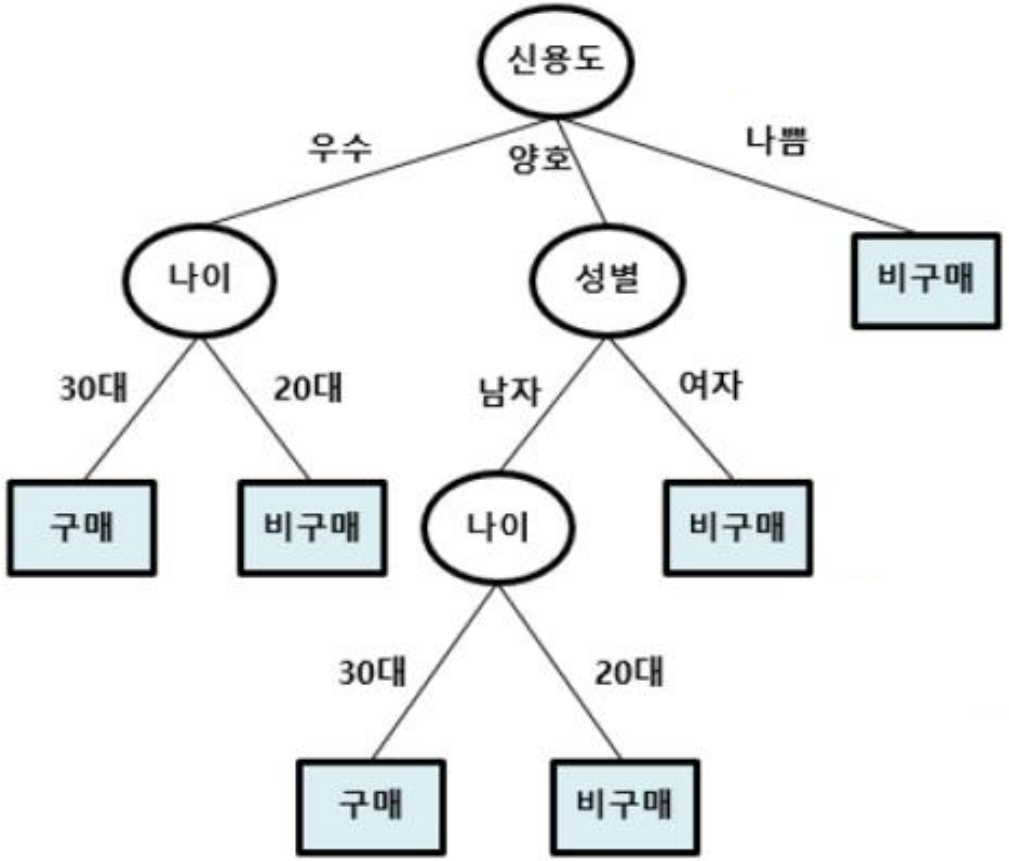


Test data

이름	통학시간(분)	히트텍	양치	감기
김강열	40	X	X	No
김소희	30	O	O	No
류호성	120	O	X	Yes
최서현	90	X	X	Yes
황다솔	100	O	O	No

Accuracy : 1
성능이 좋은 Tree

Unit 02 | Decision tree Intro



뿌리마디(root node) : 시작되는 마디로 전체 자료 포함
부모마디(parent node) : 주어진 마디의 상위마디
자식마디(child node) : 하나의 마디로부터 분리되어 나간 마디들
중간마디(internal node) : 부모마디와 자식마디가 모두 있는 마디
끝 마디 (terminal node) : 자식마디가 없는 마디
가지 (branch) : 뿌리마디로부터 끝마디까지 연결된 마디들
깊이(depth) : 뿌리마디로부터 끝마디까지의 중간마디들의 수

Unit 02 | Rule of Tree

Tree 를 만드는 기준은 어떻게 될까?

분류 트리 :
종속변수(타겟변수)가 범주형 변수일 때,
만들어지는 모델(트리)

회귀 트리 :
종속변수(타겟변수)가 연속형 변수일 때,
만들어지는 모델(트리)

두 트리의 만드는 방식은 같은데 split(growing) 기준과 예측하는 방식이 다르다

Unit 02 | Rule of Tree

Tree 형성	Split Rule (Growing Rule) : Tree의 분리 규칙 / 성장 규칙 - 부모 마디로부터 자식 마디를 생성하는 기준을 정한다 - 목표 변수의 분포를 잘 분할하는 기준으로 Split 한다 Stop Rule : 정지 규칙 - 분리가 더 이상 일어나지 않도록 하는 기준을 정한다 Pruning - Tree가 지나치게 많은 마디를 가지고 있을 경우, Overfitting의 문제가 발생할 수 있다. - 부적절한 마디를 잘라내 모형을 단순화한다.
타당성 평가	이익도표(gain chart)나 위험도표(risk chart) 또는 Cross validation을 이용해서 Tree 평가
해석 및 예측	생성한 Tree 에 새로운 데이터(test data) 대입해보고 분류 및 예측을 한다.

Unit 02 | Rule of Tree

Tree 형성 (분류 Tree일 경우)

Split Rule (Growing Rule) : Tree의 분리 규칙 / 성장 규칙

- 부모 마디로부터 자식 마디를 생성하는 기준을 정한다
- 목표 변수의 분포를 잘 분할하는 기준으로 Split 한다

불순도(Impurity) : 얼마나 다른 범주의 개체들이 포함되어 있는가를 의미한다

1. **카이제곱 통계량 : p-값이 가장 작은 분할 기준에 따라 자식 마디를 생성한다.**

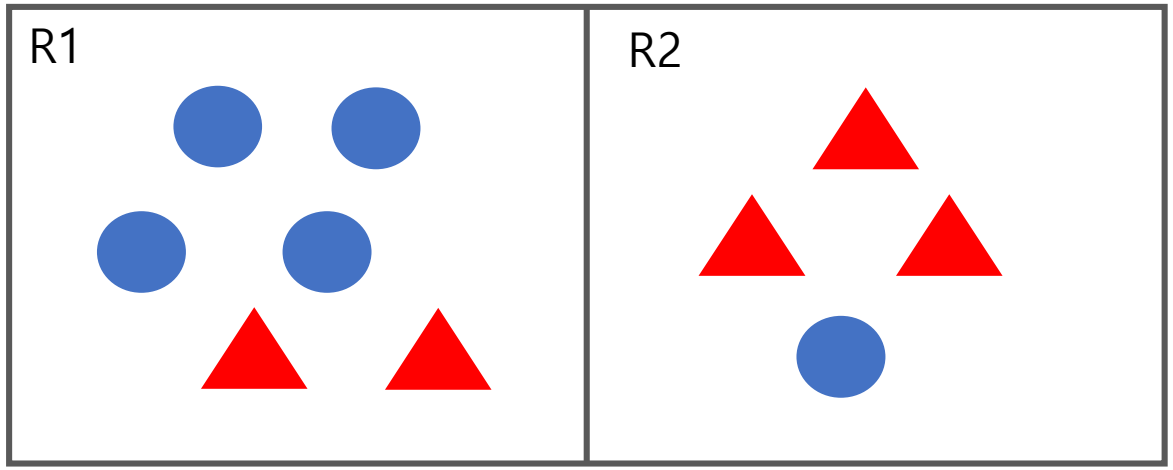
2. **지니 계수 : 지니 계수를 감소시켜주는 분할 기준에 따라 자식 마디를 생성한다.**

3. **엔트로피 계수 : 엔트로피 계수가 가장 작은 분할 기준에 따라 자식 마디를 생성한다.**

Unit 02 | Rule of Tree

지니 계수 (Gini index)

- 지니 계수를 가장 감소시켜주는 설명변수와 그 때의 최적 분리로 자식 마디를 생성한다.
- $I(A) = 1 - \sum_1^m {p_k}^2$, p_k :직사각형 A에서 K 집단에 속하는 관측치비율



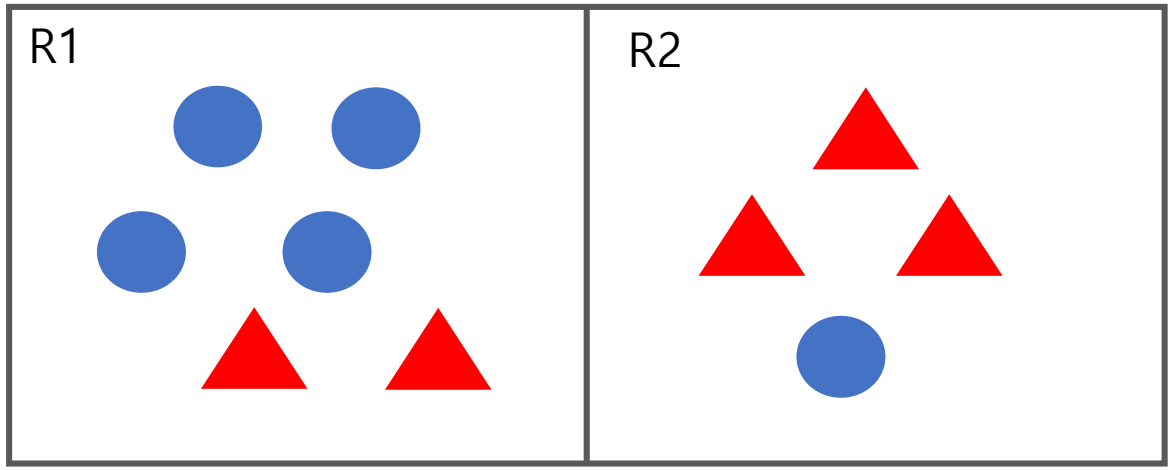
$$I(R1) = 1 - \left\{ \left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2 \right\} = 0.444$$

$$I(R2) = 1 - \left\{ \left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right\} = 0.375$$

$$G = \frac{6}{10} \times I(R1) + \frac{4}{10} \times I(R2) = \frac{6}{10} \times 0.444 + \frac{4}{10} \times 0.375 = 0.4164$$

엔트로피 계수 (Entropy index)

- 엔트로피 계수를 가장 감소시켜주는 설명변수와 그 때의 최적 분리로 자식 마디를 생성한다.
- $\text{entropy}(A) = -\sum_1^m p_k \log_2(p_k)$, p_k :직사각형 A에서 K 집단에 속하는 관측치비율



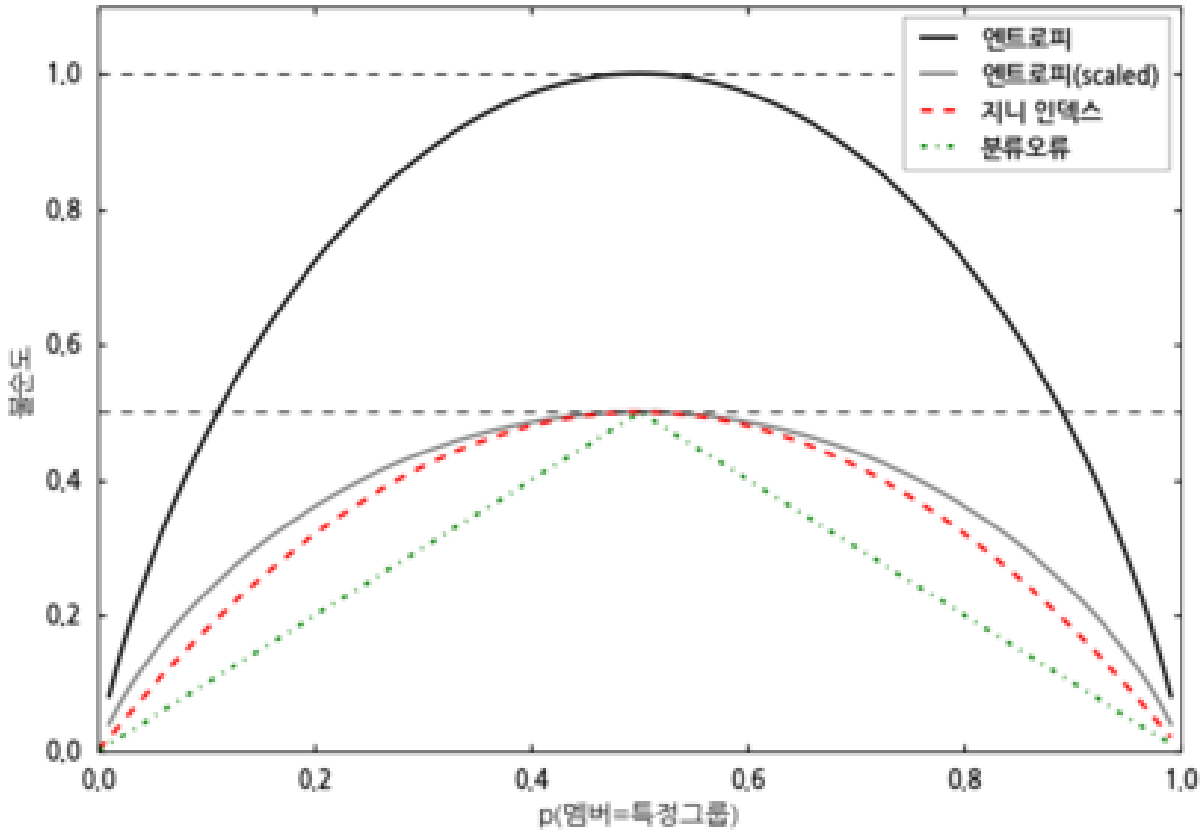
$$I(R1) = -\left\{\frac{4}{6} \times log_2\left(\frac{4}{6}\right) + \frac{2}{6} \times log_2\left(\frac{2}{6}\right)\right\}$$
$$= 0.9182$$

$$I(R2) = -\left\{\frac{3}{4} \times log_2\left(\frac{3}{4}\right) + \frac{1}{4} \times log_2\left(\frac{1}{4}\right)\right\}$$
$$= 0.8113$$

$$G = \frac{6}{10} \times I(R1) + \frac{4}{10} \times I(R2) = \frac{6}{10} \times 0.9182 + \frac{4}{10} \times 0.8113 = 0.8754$$

Unit 02 | Rule of Tree

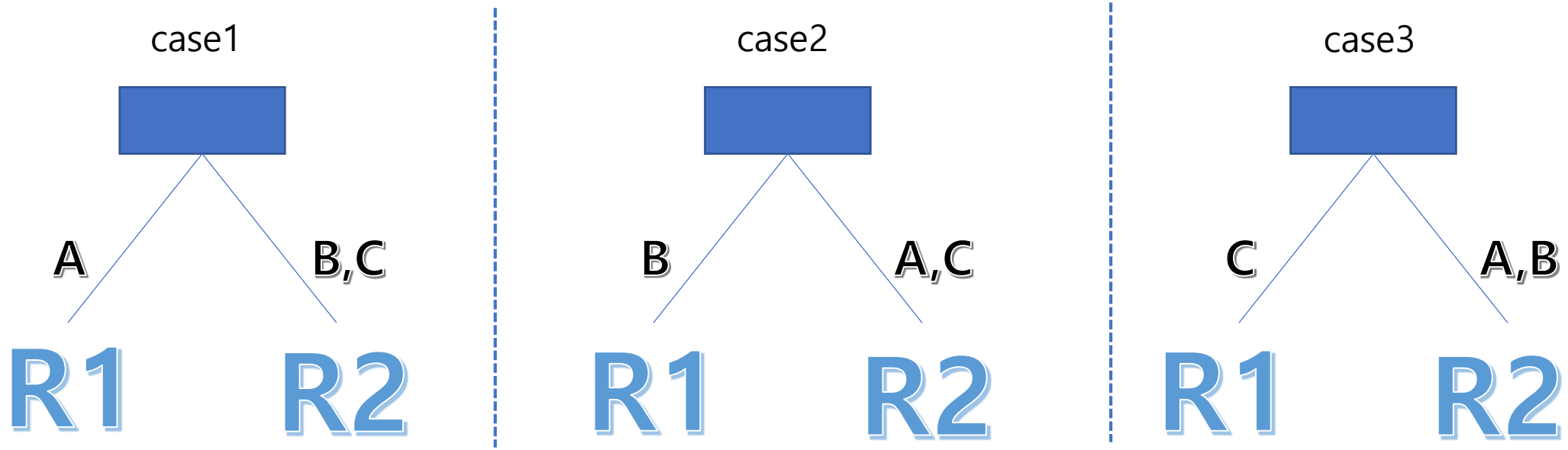
불순도 다이어그램



두 집단의 분류 확률이 같은 경우 (p_k = 0.5)
→ 어느 집단으로 분류될지가 모호함
→ 불순도 최대

Unit 02 | Rule of Tree

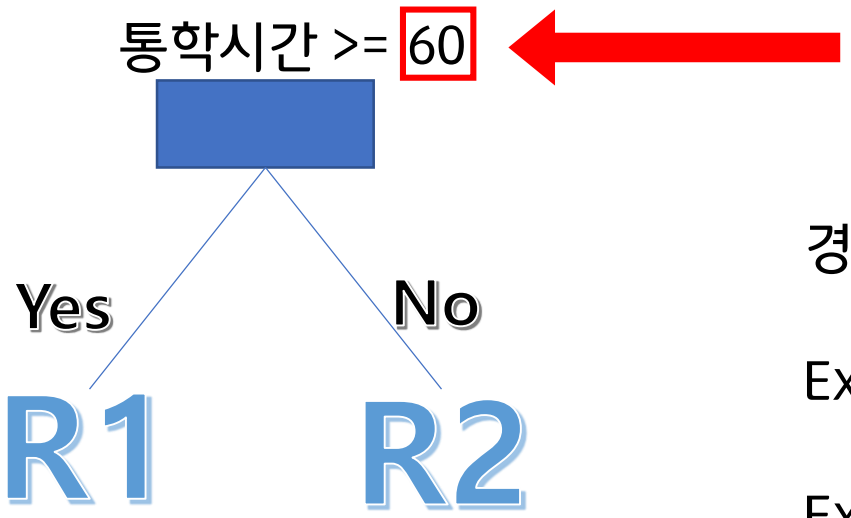
설명변수 : 범주형 (“A”,”B”,”C”)일 경우



Case 별로 지니 계수 혹은 엔트로피 계수를 비교해 가장 작은 값을 갖는 Split rule(case)을 선택한다

Unit 02 | Rule of Tree

설명변수 : 연속형일 경우 - 경계값의 기준을 찾는다.



지니 계수 혹은 엔트로피 계수를 비교해
가장 작은 값을 갖도록 하는 경계값을 잘 찾아야 된다.

경계값 후보

Ex) 1사분위수 / 2사분위수 / 3사분위수

Ex) seq(from=1사분위수,to=3사분위수,length.out=10)

Ex) seq(from=min(x),to=max(x),length.out=20)

Unit 02 | Rule of Tree

모든 설명 변수에 대한 Split 기준 중에 불순도 (지니 계수 / 엔트로피 계수)가
가장 작은 값을 갖는 설명 변수와 그 때의 기준으로 Split을 한다!

칠판 주목!

Tree 형성 (분류 Tree일 경우)

Stop Rule : 정지 규칙

- 분리가 더 이상 일어나지 않도록 하는 기준을 정한다

1. **분리를 더 이상 하더라도 불순도가 줄어들지 않을 경우**
2. **자식 마디에 남아 있는 sample 수가 적은 경우(일반적으로 전체 데이터의 5%)**
3. **분석자가 미리 정해 놓은 깊이에 도달했을 경우**

Unit 02 | Rule of Tree

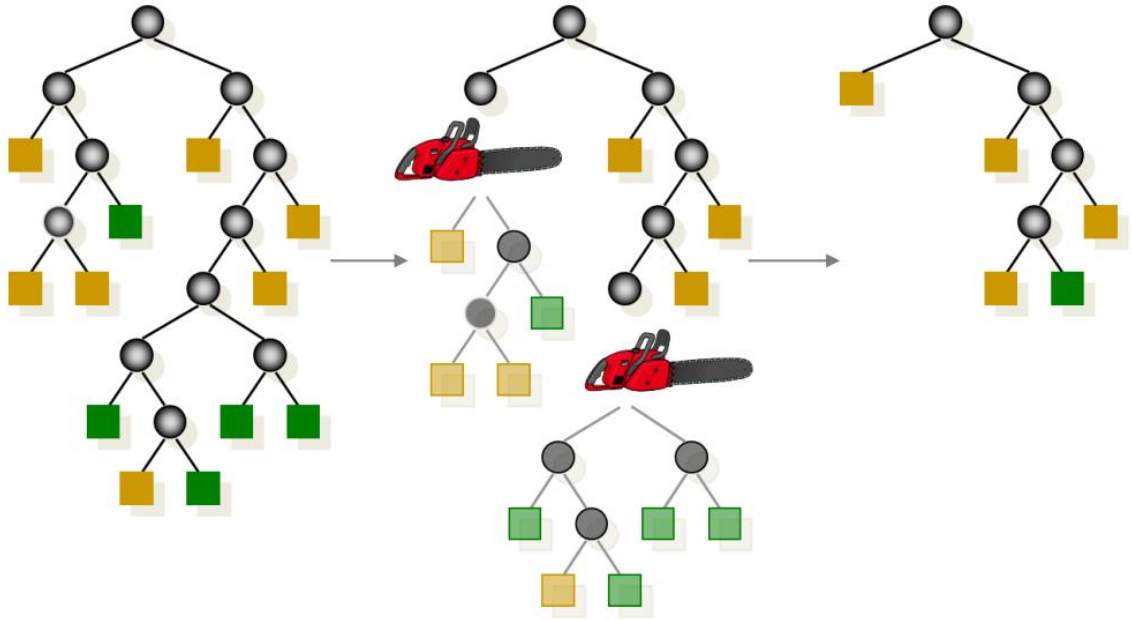
Tree 형성 (분류 Tree일 경우)

Pruning

- Tree가 지나치게 많은 마디를 가지고 있을 경우, Overfitting의 문제가 발생할 수 있다.
  (Train data에만 좋은 모델!
  Test data에는 합당하지 않는 경우가 발생한다)

- 이를 해결하기 위해 부적절한 마디를 잘라내 모형을 단순화한다.

Pre- pruning/ Post- pruning



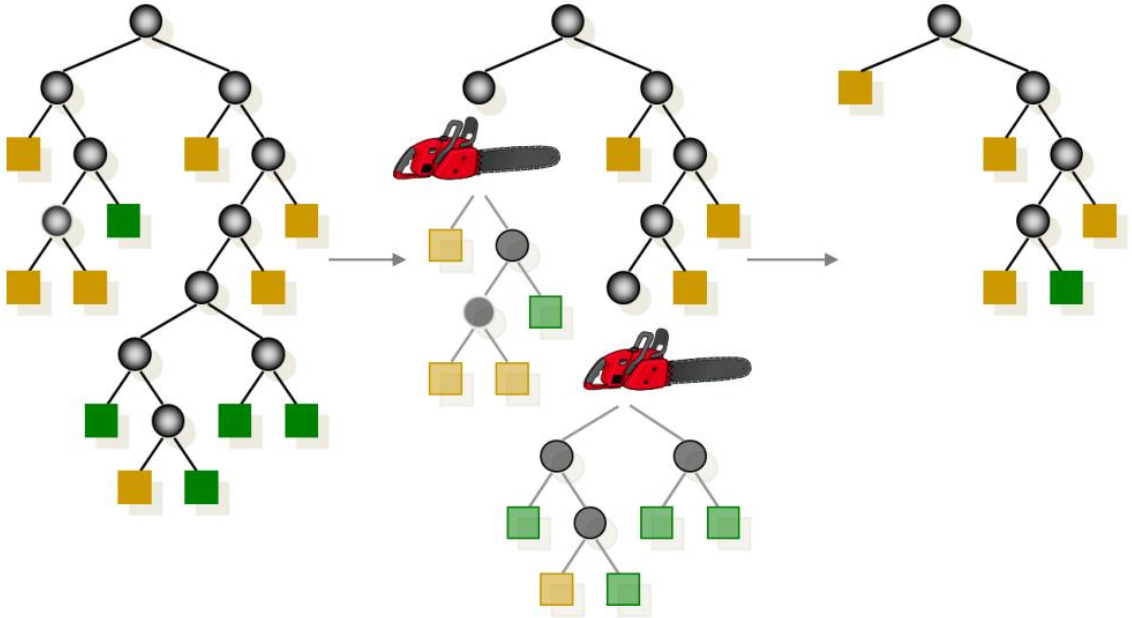
Unit 02 | Rule of Tree

Post-Pruning

Tree를 먼저 완성시킨 후에, Pruning을 하는 방식

1. **비용 복잡도(cost-complexity) 에 의거한 pruning**

2. **Pessimistic pruning**



Unit 02 | Rule of Tree

비용 복잡도(cost-complexity)에 의거한 Pruning

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

CC(T) : Tree의 비용 복잡성(= 오류가 적으면서 끝 마디 수가 적은 단순한 모델일수록 작은 값)

Err(T) : 오분류율 (불순도)

L(T) : 끝마디의 수 (구조 복잡도)

Alpha : Err(T)와 L(T)를 결합하는 가중치(보통 0.01~0.1의 값을 쓴다)

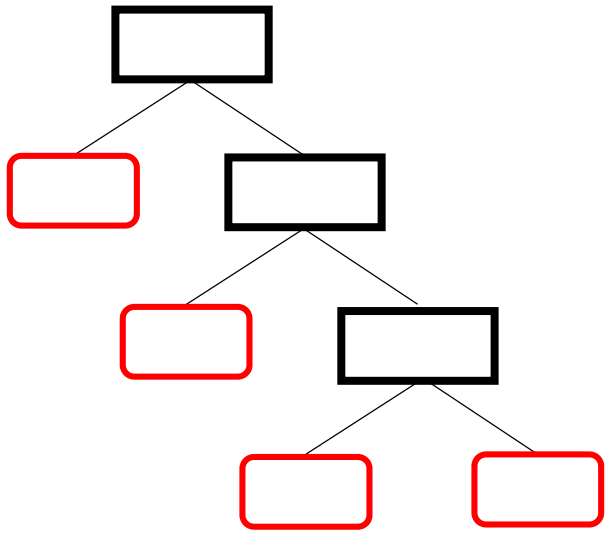
Pessimistic pruning

일반적으로 Split할 때,
오분류율 (불순도)이 작아지면서 Split이 된다.
그런데 Pessimistic pruning에서는 Split 할 때마다 오분류율 + (마디 수*0.5) 만큼의
오차가 더 있다고 가정을 한다.

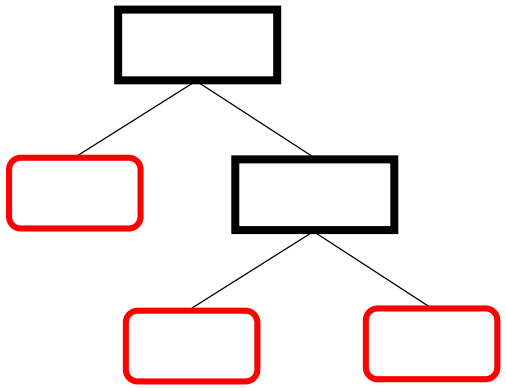
왜 Why?
→ 과적합을 피하기 위해서!

Unit 02 | Rule of Tree

Ex)



오분류율 : 0.45
끝 마디 수 : 4
Pessimistic error :
0.45+4*0.5 = 2.45



오분류율 : 0.66
끝 마디 수 : 3
Pessimistic error :
0.66+3*0.5 = 2.16

Pruning 한 Tree model의
Pessimistic error 값이 더 낮으므로,
Pruning을 한다!

Unit 02 | Rule of Tree

회귀 Tree(타겟변수 : 연속형) 일 때는?

분류 Tree일 때와 tree 형성 과정이 똑같은데 Split 기준과 예측 방법이 다르다!

분류 Tree의 불순도 :

카이제곱 통계량 / 지니 계수/ 엔트로피 계수

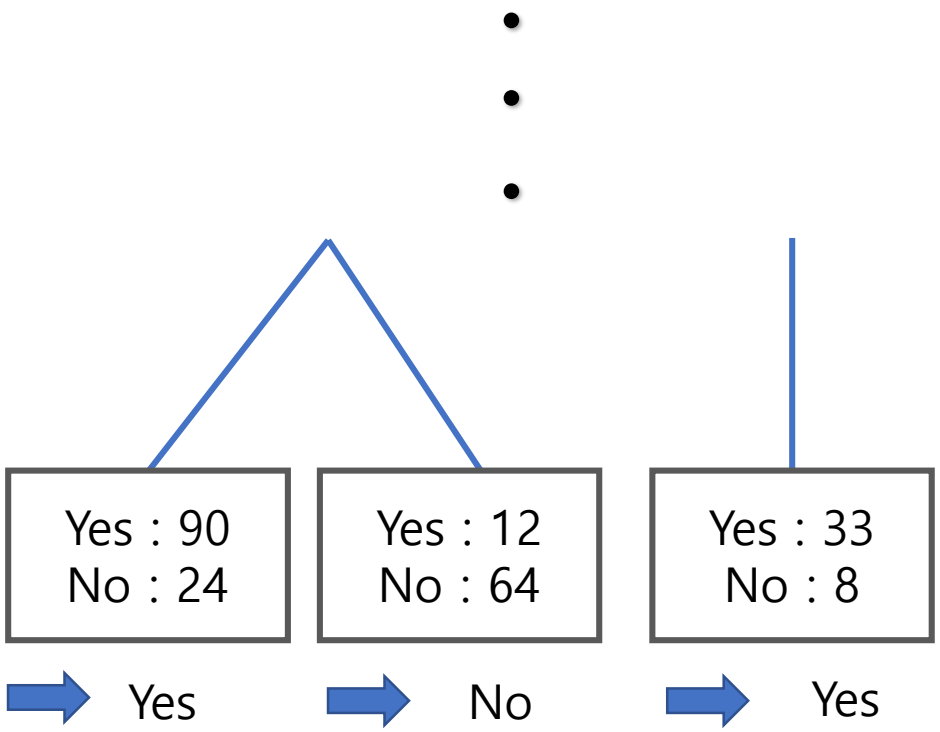
회귀 Tree의 불순도 :

F 통계량 / 분산의 감소량

→ 우리가 알고 있는 최소 제곱 추정량 (sse) 을 줄여 나가는 방식으로 split하면 된다.

Unit 02 | Rule of Tree

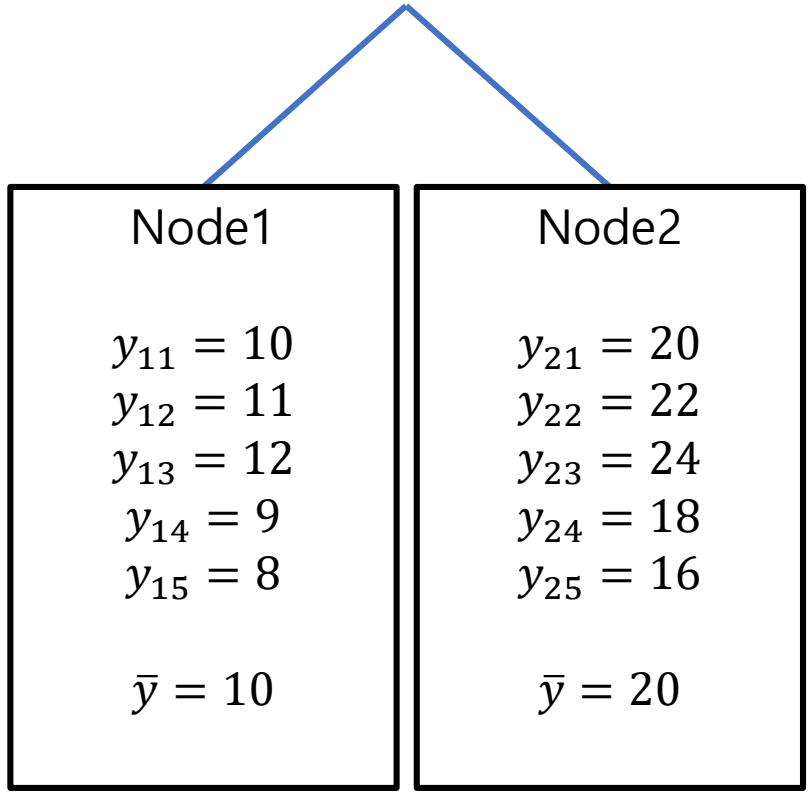
분류 Tree에서는 끝마디의 범주로 예측을 하면 된다.



그렇다면, 회귀 Tree에서는 어떻게 예측해야 되지..?

Unit 02 | Rule of Tree

회귀 Tree에서는 끝마디의 $\bar{y}$ 로 예측한다.



왼쪽 node 의 sse

: $(10-10)^2+(10-10)^2+\cdots+(10-10)^2=10$

오른쪽 node 의 sse

: $(20-20)^2+(22-20)^2+\cdots+(16-20)^2=40$

sse 의 합

: $\frac{5}{10}\times 10+\frac{5}{10}\times 40 = 25$

→ Sse 를 줄여 나가는 방식으로 Split

Unit 02 | Rule of Tree

Tree 장점	Tree 단점
- 해석의 용이성 → 사용자가 쉽게 이해 가능 - 상호작용 효과의 해석 → 두 개 이상의 변수가 종속변수에 어떻게 영향을 주는지 쉽게 알 수 있다. - 비모수적 모형 → 선형성,정규성,등분산성 등의 가정이 필요 없다. - 이상치에 민감하지 않다.	- 비연속성 → 분리의 경계점 근방에서는 예측오류가 클 가능성이 있다. - 축에 평행한 분할 → 한 노드에서 한 변수에 의해서만 분할이 일어난다. 그래서 직선으로만 분리 - 비안정성 → train data에 의존하는 경향이 크다.

Unit 03 | Tree 알고리즘

- Tree 패키지 (Binary Recursive Partitioning)
- Rpart 패키지 (Classification And Regression Tree)
  - 이 패키지들은 엔트로피, 지니계수를 기준으로 가지치기를 할 변수를 결정
  - 상대적으로 연산 속도는 빠르지만 과적합화의 위험성이 존재
  - 그래서 두 패키지를 사용할 경우에는 Pruning 과정을 거쳐서 의사결정나무를 최적화 하는 과정이 필요.
- Party 패키지(Unbiased recursive partitioning based on permutation tests)
  - p-test를 거친 Significance를 기준으로 가지치기를 할 변수를 결정
  - biased 될 위험이 없어 별도로 Pruning할 필요가 없다는 장점
  - 입력 변수의 레벨이 31개 까지로 제한되어 있다는 단점

R 코드에서 같이 봅시다!!

Unit 04 | ensemble

Tree는 해석이 용이하고, 의사결정 방법이 인간의 의사결정 방법과 유사하지만,
예측력이 많이 떨어지고, bias-variance 문제가 발생한다.

이를 해결하기 위해서 Tree를 여러 개 생성하고 결합하는
Ensemble (bagging, boosting, Random Forest) 이 나오게 됐다!
이로써 Tree의 예측력을 높인다.

Ensemble 방식들은 여러 개의 약한 학습기의 결합이다

자세한 건 **서현**이 강의에서 배우자!!

들어주셔서 감사합니다.

참고 링크

https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

http://contents.kocw.net/KOCW/document/2014/korea/choijonghu/5.pdf

https://m.blog.naver.com/PostView.nhn?blogId=2011topcit&logNo=220611261399&proxyReferer=https%3A%2F%2Fwww.google.co.kr%2F

http://datamining.dongguk.ac.kr/lectures/2011-2/dm/S1_tree2.pdf

http://www.dodomira.com/2016/05/29/564/