13기 정규세션

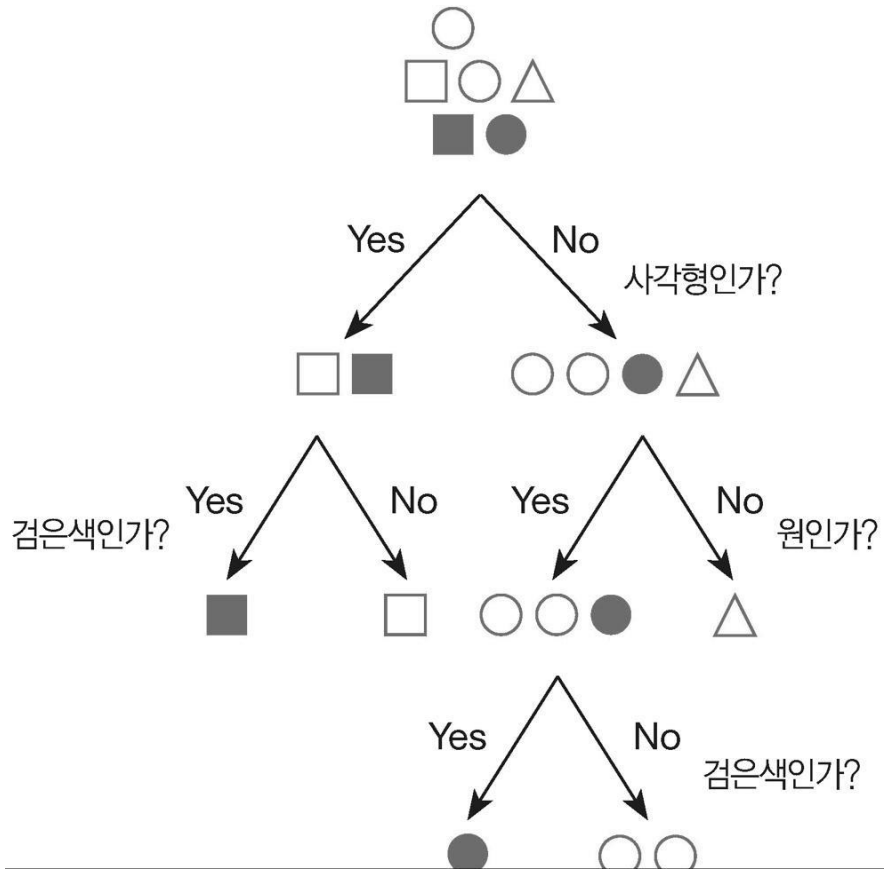
ToBig's 12기 윤기오

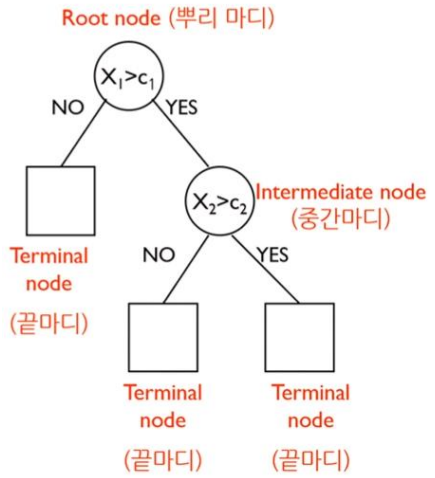
의사결정나무

Decision Tree

Contents

Unit 01 | 의사결정나무란?

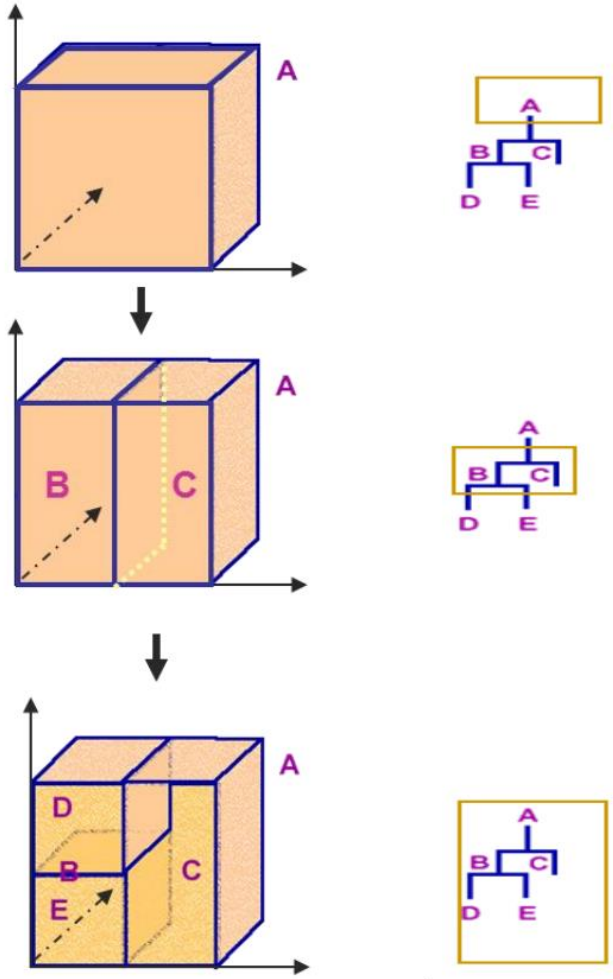




의사결정나무

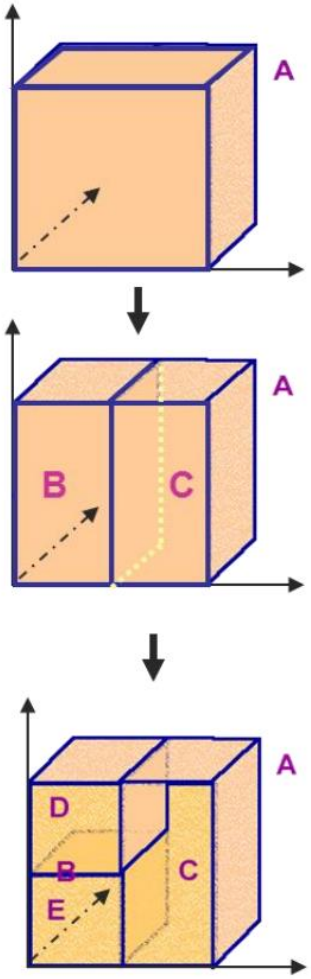
의사결정 규칙을 나무구조로 나타내어 전체 데이터를 소집단으로 **분류**하거나 **예측**하는 분석 방법

Unit 01 | 의사결정나무란?



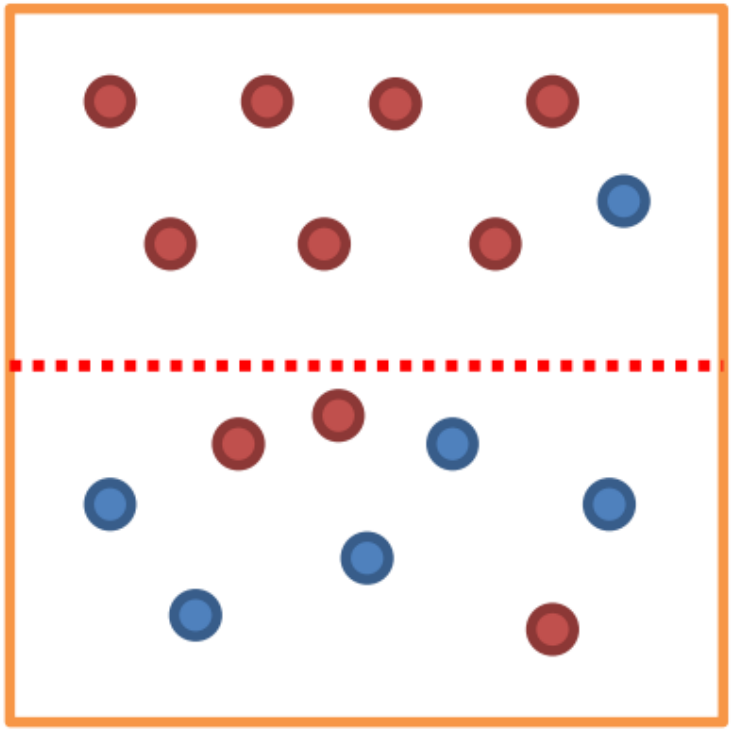
설명변수(X)가 3개짜리인 다변량 데이터에 의사결정나무를 적용한 모습

Unit 01 | 의사결정나무란?



그렇다면 무슨 기준으로 데이터를 나눌까? 의사결정나무를 적용한 모습

Unit 02 | 의사결정나무 알고리즘



순도(homogeneity) / 불순도(impurity)

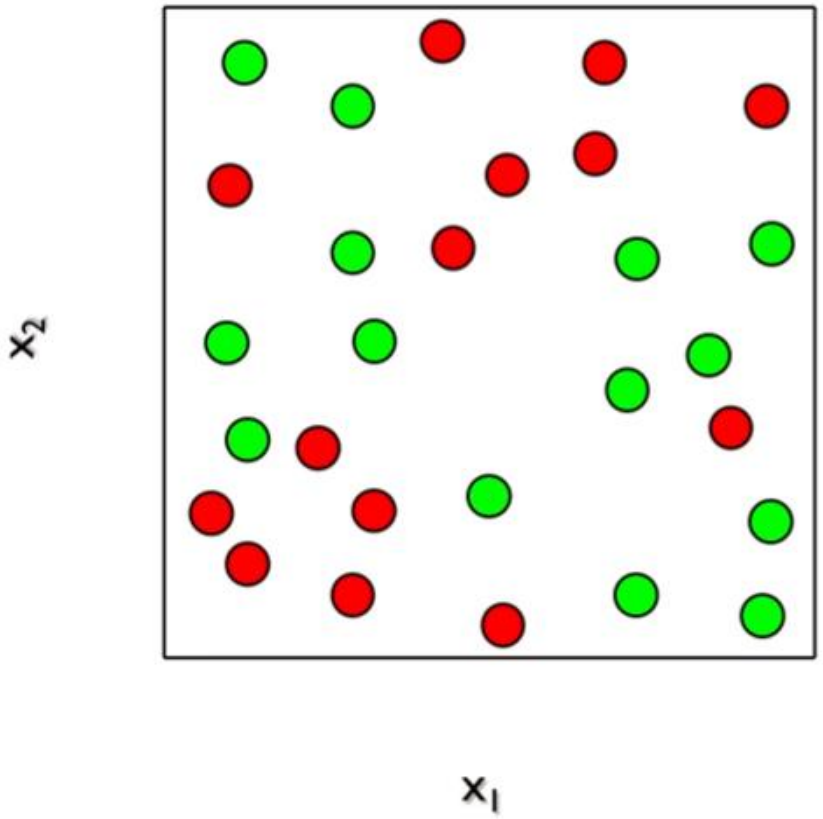
구분된 각 영역의 순도가 증가, 불순도가 감소하는 방향으로 학습된다.

순도가 높다 = 각 영역의 데이터들이 동질하다

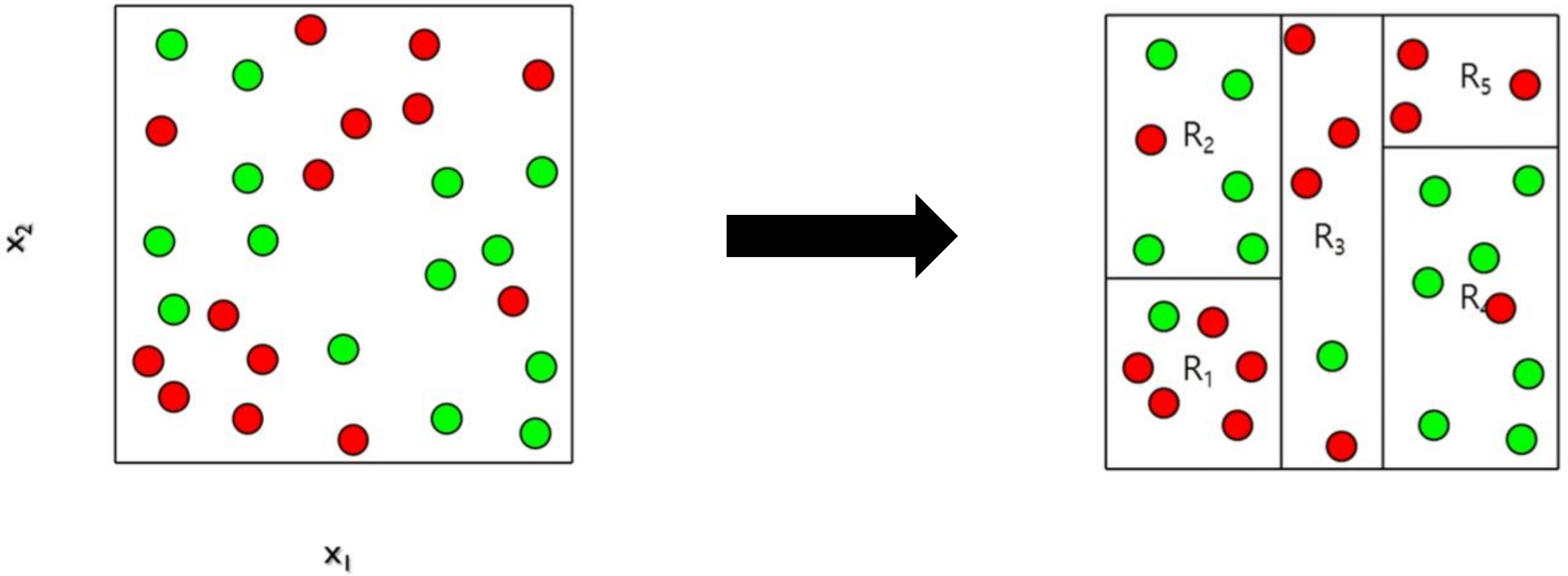
정보이론에서는 이를 정보획득(Information Gain) 이라고 한다.

그렇다면 '순도' 혹은 '불순도' 는 어떻게 계산할까?

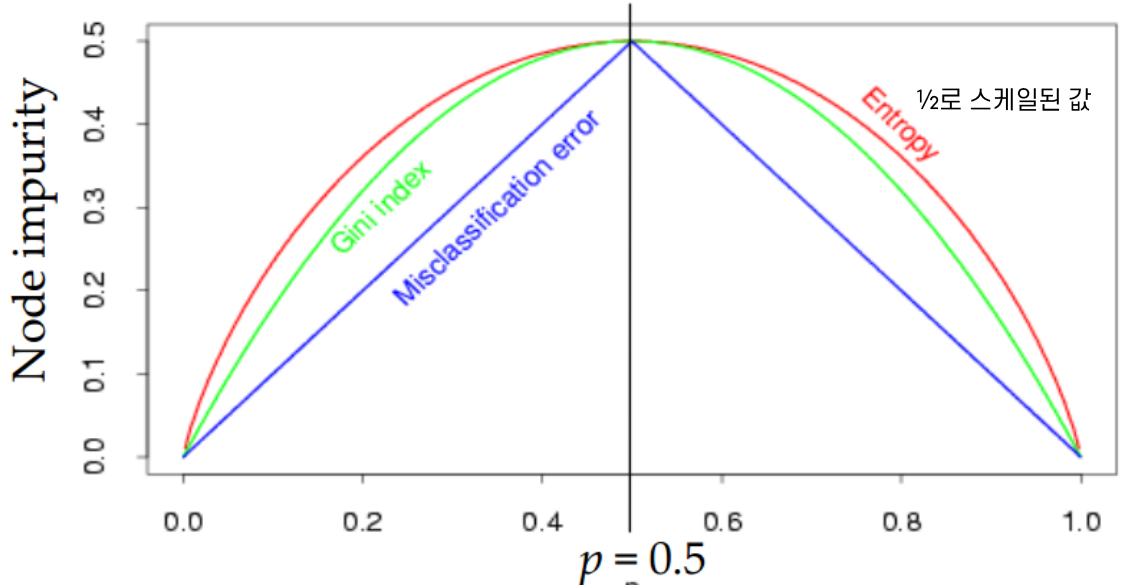
Unit 02 | 의사결정나무 알고리즘



Unit 02 | 의사결정나무 알고리즘



Unit 02 | 의사결정나무 알고리즘



Q1. 불순도(Impurity)를 측정하는 지표는?

A. Entropy, Gini Index, Misclassification error 등

Q2. 어떤 기준으로 노드를 놓아야 하며, 어떤 노드를 가장 위에 놓아야 할까?

A. ID3 & CART 알고리즘

Unit 02 | 의사결정나무 알고리즘

ID3 *Entropy 지수를 활용한 알고리즘*

ID3

- Entropy를 도입하여 Decision Tree의 가지를 나눠보자!
- Information Gain = 전체 Entropy – 속성별 Entropy
- Information Gain이 높을수록 명확한 정보를 얻을 수 있음

Class가 3개인 Feature “A”에 대한 Information Gain은?

$$Gain(A) = Info(D) - Info_A(D_i)$$

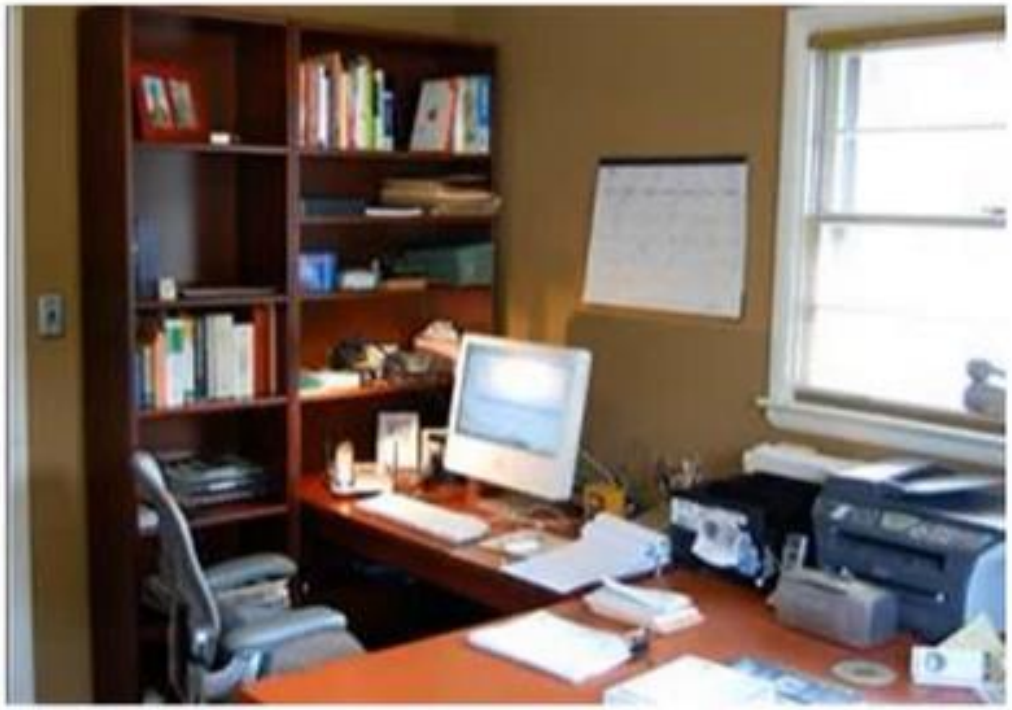
$$Info(D) = Entropy_{label}$$

$$Info_A(D_i) = -\sum_{j=1}^{3} \frac{|D_j|}{|D|} * Entropy_{label_j}$$

Unit 02 | 의사결정나무 알고리즘

엔트로피(Entropy) *불순도(Impurity) 를 측정하는 지표 1 (ID3)*

High Entropy (messy)



Low Entropy (Clean)

엔트로피(Entropy) *불순도(Impurity) 를 측정하는 지표 1 (ID3)*

엔트로피란?

- 무질서도를 정량화해서 표현한 값

- 어떤 집합의 Entropy가 높을수록(무질서할수록) 그 집단의 특징을 찾는 것이 어렵다.

- 우리의 목적 : Entropy를 감소시키는 방향으로 분류하기

  엔트로피 감소 = 불순도 감소 = 순도 증가 = 정보 획득

Unit 02 | 의사결정나무 알고리즘

엔트로피(Entropy) *불순도(Impurity) 를 측정하는 지표 1 (ID3)*

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

m개의 레코드가 속하는 A영역에 대한 엔트로피

Pk = A영역에 속하는 레코드 가운데 k 범주에 속하는 레코드의 비율

A 영역에 속하는 모든 레코드가 동일한 범주에 속할 경우(=불확실성 최소 =순도 최대) 엔트로피는 0
범주가 둘 뿐이고 반반씩 섞여 있을 경우 (=불확실성 최대 =순도 최소) 엔트로피는 1

Unit 02 | 의사결정나무 알고리즘

ID3 *Entropy 지수를 활용한 알고리즘*

엔트로피(Entropy) *불순도(Impurity) 를 측정하는 지표 1 (ID3)*

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

age	*income*	*student*	*credit_rating*	*Class: buys_computer*
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

Buys_computer에 대한 Entropy를 구해보기
Class : No 5개 / Yes 9개

Unit 02 | 의사결정나무 알고리즘

ID3 *Entropy 지수를 활용한 알고리즘*

$$Gain(A) = Info(D) - Info_A(D_i)$$

$$Info(D) = Entropy_{label}$$

$$Info_A(D_i) = -\sum_{j=1}^{3} \frac{|D_j|}{|D|} * Entropy_{label_j}$$

D = 주어진 데이터들의 집합
|D| = 주어진 데이터들의 집합의 데이터 갯수

Unit 02 | 의사결정나무 알고리즘

img	cartoon	winter	> 1	Family winter photo
	No	Yes	Yes	Yes
	No	Yes	No	No
	Yes	No	Yes	No
	Yes	Yes	Yes	No
	No	Yes	No	No
	No	No	Yes	No
	Yes	No	Yes	No
	yes	yes	no	no

ID3 *Entropy 지수를 활용한 알고리즘*

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

$$Gain(A) = Info(D) - Info_A(D_i)$$

전체 8개 사진
-> 겨울 가족 사진 Yes 1개
-> 겨울 가족 사진 No 7개

Info(D)

Unit 02 | 의사결정나무 알고리즘

img	cartoon	winter	> 1	Family winter photo
	No	Yes	Yes	Yes
	No	Yes	No	No
	Yes	No	Yes	No
	Yes	Yes	Yes	No
	No	Yes	No	No
	No	No	Yes	No
	Yes	No	Yes	No
	yes	yes	no	no

ID3 *Entropy 지수를 활용한 알고리즘*

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

$$Info_A(D_i) = -\sum_{j=1}^{3} \frac{|D_j|}{|D|} * Entropy_{label_j}$$

Gain(Cartoon)

Unit 02 | 의사결정나무 알고리즘

img	cartoon	winter	> 1	Family winter photo
	No	Yes	Yes	Yes
	No	Yes	No	No
	Yes	No	Yes	No
	Yes	Yes	Yes	No
	No	Yes	No	No
	No	No	Yes	No
	Yes	No	Yes	No
	yes	yes	no	no

ID3 *Entropy 지수를 활용한 알고리즘*

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

$$Info_A(D_i) = -\sum_{j=1}^{3} \frac{|D_j|}{|D|} * Entropy_{label_j}$$

Gain(Winter)

Unit 02 | 의사결정나무 알고리즘

img	cartoon	winter	> 1	Family winter photo
	No	Yes	Yes	Yes
	No	Yes	No	No
	Yes	No	Yes	No
	Yes	Yes	Yes	No
	No	Yes	No	No
	No	No	Yes	No
	Yes	No	Yes	No
	yes	yes	no	no

ID3 *Entropy 지수를 활용한 알고리즘*

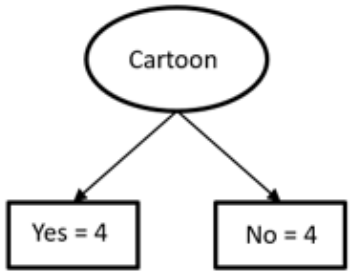
$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2 (p_k)$$

$$Info_A(D_i) = \sum_{j=1}^{3} \frac{|D_j|}{|D|} * Entropy_{label_j}$$

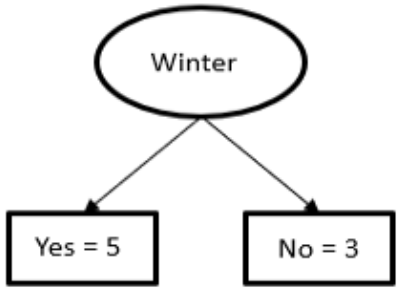
Gain(>1)

Unit 02 | 의사결정나무 알고리즘

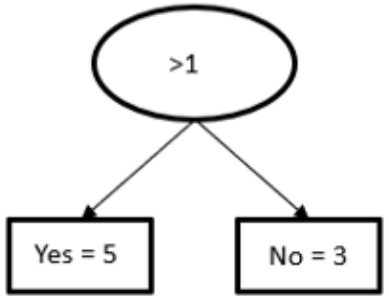
ID3 *Entropy 지수를 활용한 알고리즘*



0.138



0.093



0.093

가장 Information Gain이 높은 [카툰]으로 branch split을 했을 때 정보량 획득이 가장 크다.

-> 최초로 [카툰] 으로 데이터를 나눈다.

Unit 02 | 의사결정나무 알고리즘

CART *Gini Index를 활용한 알고리즘*

CART

- Gini Index를 도입하여 Decision Tree의 가지를 나눠보자!
- 데이터를 split 했을 때 불순한 정도
- 데이터의 대상 속성을 얼마나 잘못 분류할지 계산
- Binary split을 전제로 분석함
- Feature의 데이터 분류 개수가 k개일 때 $2^{k-1}-1$ 개 만큼의 split 생성

Class가 3개인 Feature “A”에 대한 Gini Index는?

$$Gini(A) = \sum_{j=1}^{2} \frac{|D_j|}{|D|} * Gini(D_i)$$

$$Gini(D_i) = 1 - \sum_{j=1}^{3} P_j^{\,p^2+(1-p)^2}$$

지니 지수(Gini Index) *불순도(Impurity) 를 측정하는 지표 2 (CART)*

지니 지수란?

- 데이터의 통계적 분산정도를 정량화해서 표현한 값

- 어떤 집합의 Gini Index가 높을수록 그 집단의 데이터가 분산 되어있다.

- Entropy와 크게 다르지 않다!

- 우리의 목적 : Gini Index를 감소시키는 방향으로 분류하기

  지니 지수 감소 = 불순도 감소 = 순도 증가 = 정보 획득

Unit 02 | 의사결정나무 알고리즘

CART *Gini Index를 활용한 알고리즘*

age	income	student	credit_rating	Class: buys_computer
youth	high	no	fair	no
youth	high	no	excellent	no
middle_aged	high	no	fair	yes
senior	medium	no	fair	yes
senior	low	yes	fair	yes
senior	low	yes	excellent	no
middle_aged	low	yes	excellent	yes
youth	medium	no	fair	no
youth	low	yes	fair	yes
senior	medium	yes	fair	yes
youth	medium	yes	excellent	yes
middle_aged	medium	no	excellent	yes
middle_aged	high	yes	fair	yes
senior	medium	no	excellent	no

A,B,C가 있을 때
A | B C
B | A C
C | A B
다 해보아야 한다.

Age $\text{Gini}_{age}(D)$

Credit $\text{Gini}_{credit}(D)$

Income $\text{Gini}_{income}(D)$

Student $\text{Gini}_{student}(D)$

이 중 가장 작은 Gini Index 값을 가지는 변수가 최초 split이 됨

Unit 02 | 의사결정나무 알고리즘

Age에 대한 Gini Index를 먼저 구해보자

	RID	age	income	student	credit_rating	class_buys_computer
0	1	youth	high	no	fair	no
1	2	youth	high	no	excellent	no
7	8	youth	medium	no	fair	no
8	9	youth	low	yes	fair	yes
10	11	youth	medium	yes	excellent	yes

AGE : youth, senior, middle_aged

	RID	age	income	student	credit_rating	class_buys_computer
2	3	middle_aged	high	no	fair	yes
3	4	senior	medium	no	fair	yes
4	5	senior	low	yes	fair	yes
5	6	senior	low	yes	excellent	no
6	7	middle_aged	low	yes	excellent	yes
9	10	senior	medium	yes	fair	yes
11	12	middle_aged	medium	no	excellent	yes
12	13	middle_aged	high	yes	fair	yes
13	14	senior	medium	no	excellent	no

CART *Gini Index를 활용한 알고리즘*

$$Gini(A) = \sum_{j=1}^{2} \frac{|D_j|}{|D|} * Gini(D_i)$$

$$Gini(D_i) = 1 - \sum_{j=1}^{3} P_j^2 \quad (p^2 + (1-p)^2)$$

Unit 02 | 의사결정나무 알고리즘

CART *Gini Index를 활용한 알고리즘*

Gini Index

$Min(Gini_{age_i}) = 0.357$ - > middle_aged

$Min(Gini_{income_i}) = 0.443$

$Min(Gini_{credit}) = 0.429$

$Min(Gini_{student}) = 0.367$

Age

cart는 분류된 변수를 다시 쓰게 된다. (여기서 다시 경우의 수 구함)

Middle_aged

	age	income	student	credit_rating	class_buys_computer
2	middle_aged	high	no	fair	yes
6	middle_aged	low	yes	excellent	yes
11	middle_aged	medium	no	excellent	yes
12	middle_aged	high	yes	fair	yes

Youth,senior

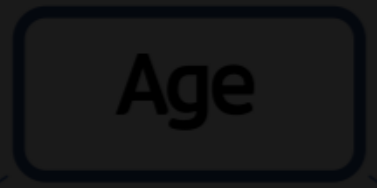
	age	income	student	credit_rating	class_buys_computer
0	youth	high	no	fair	no
1	youth	high	no	excellent	no
3	senior	medium	no	fair	yes
4	senior	low	yes	fair	yes
5	senior	low	yes	excellent	no
7	youth	medium	no	fair	no
8	youth	low	yes	fair	yes
9	senior	medium	yes	fair	yes
10	youth	medium	yes	excellent	yes
13	senior	medium	no	excellent	no

Unit 02 | 의사결정나무 알고리즘

CART *Gini Index를 활용한 알고리즘*

Gini Index

$$Min(Gini_{age_i}) = 0.357$$
$$Min(Gini_{income_i}) = 0.443$$
$$Min(Gini_{credit}) = 0.429$$
$$Min(Gini_{student}) = 0.367$$



Middle_aged

Youth,senior

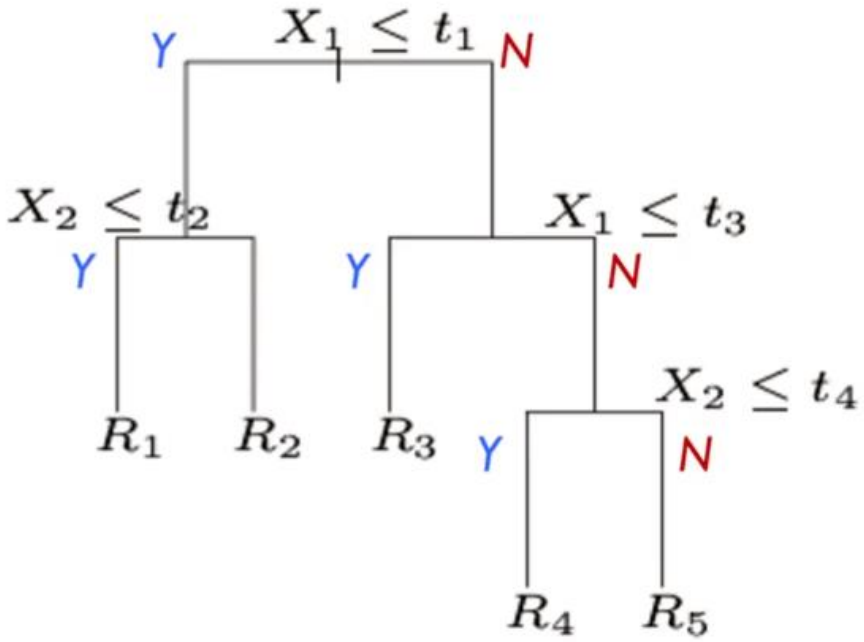
Q. 연속형 변수에는 의사결정나무를 활용할 수 없을까?

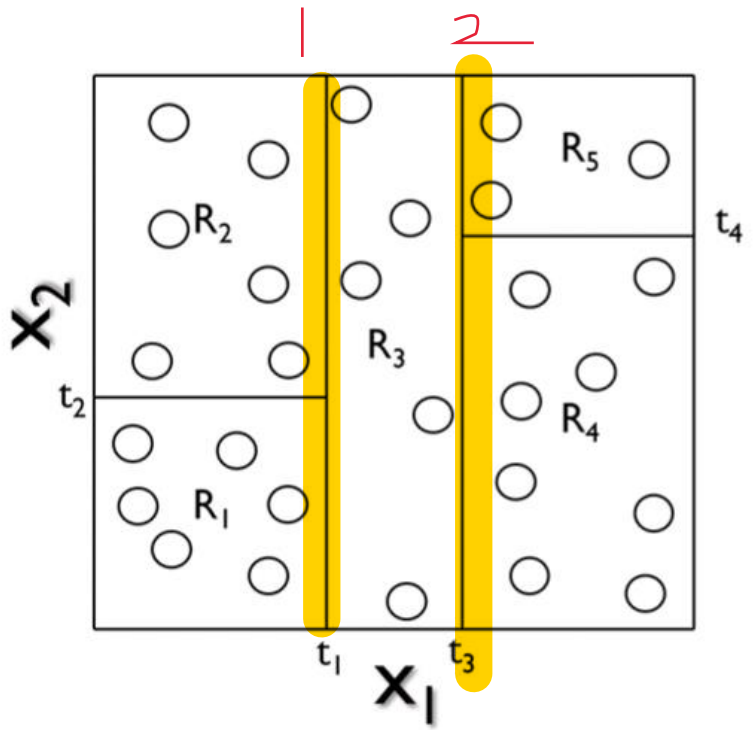
	age	income	student	credit_rating	class_buys_computer
2	middle_aged	high	no	fair	yes
6	middle_aged	low	yes	excellent	yes
11	middle_aged	medium	no	excellent	yes
12	middle_aged	high	yes	fair	yes

	age	income	student	credit_rating	class_buys_computer
0	youth	high	no	fair	no
1	youth	high	no	excellent	no
3	senior	medium	no	fair	yes
4	senior	low	yes	fair	yes
5	senior	low	yes	excellent	no
7	youth	medium	no	fair	no
8	youth	low	yes	fair	yes
9	senior	medium	yes	fair	yes
10	youth	medium	yes	excellent	yes
13	senior	medium	no	excellent	no

Unit 02 | 의사결정나무 알고리즘

연속형 변수





연속형 변수

Q. 연속형 변수를 split하는 방법

1. 전체 데이터를 모두 기준점으로 한다.

2. 중위수, 사분위수를 기준점으로 한다.

3. Label의 class가 바뀌는 수를 기준점으로 한다.

Unit 02 | 의사결정나무 알고리즘

연속형 변수

Step 1. Split할 연속형 변수를 sorting한다.

	ID	STREAM	SLOPE	ELEVATION	VEGETATION
0	1	False	steep	3900	chapparal
1	2	True	moderate	300	riparian
2	3	True	steep	1500	riparian
3	4	False	steep	1200	chapparal
4	5	False	flat	4450	conifer
5	6	True	steep	5000	conifer
6	7	True	steep	3000	chapparal

300~5000

ELEVATION
300
1200
1500
3000
3900
4450
5000

Unit 02 | 의사결정나무 알고리즘

연속형 변수

Step 2. Label의 class가 바뀌는 지점을 찾는다.

	ID	STREAM	SLOPE	ELEVATION	VEGETATION
1	2	True	moderate	300	riparian
3	4	False	steep	1200	chapparal
2	3	True	steep	1500	riparian
6	7	True	steep	3000	chapparal
0	1	False	steep	3900	chapparal
4	5	False	flat	4450	conifer
5	6	True	steep	5000	conifer

(1)
(2)
(3)
(4)

Step 3. 경계의 평균값으로 기준값을 잡는다.

	ID	STREAM	SLOPE	ELEVATION	VEGETATION
1	2	True	moderate	300	riparian
3	4	False	steep	1200	chapparal
2	3	True	steep	1500	riparian
6	7	True	steep	3000	chapparal
0	1	False	steep	3900	chapparal
4	5	False	flat	4450	conifer
5	6	True	steep	5000	conifer

750
1,350
2250
4175

Unit 02 | 의사결정나무 알고리즘

연속형 변수

Step 4. 구간별 경계값을 기준으로 Entropy 또는 Gini를 산출한다.

$$\text{Gain}(elec_{750}) = Info(D) - Info_{elec_{750}}(D)$$
$$\text{Gain}(elec_{1350}) = Info(D) - Info_{elec_{1350}}(D)$$
$$\text{Gain}(elec_{2250}) = Info(D) - Info_{elec_{2250}}(D)$$
$$\text{Gain}(elec_{4175}) = Info(D) - Info_{elec_{4175}}(D)$$

$$Max(\text{Gain}(elec))$$

Unit 02 | 의사결정나무 알고리즘

연속형 변수

Step 5. 최종 Split point 선택!

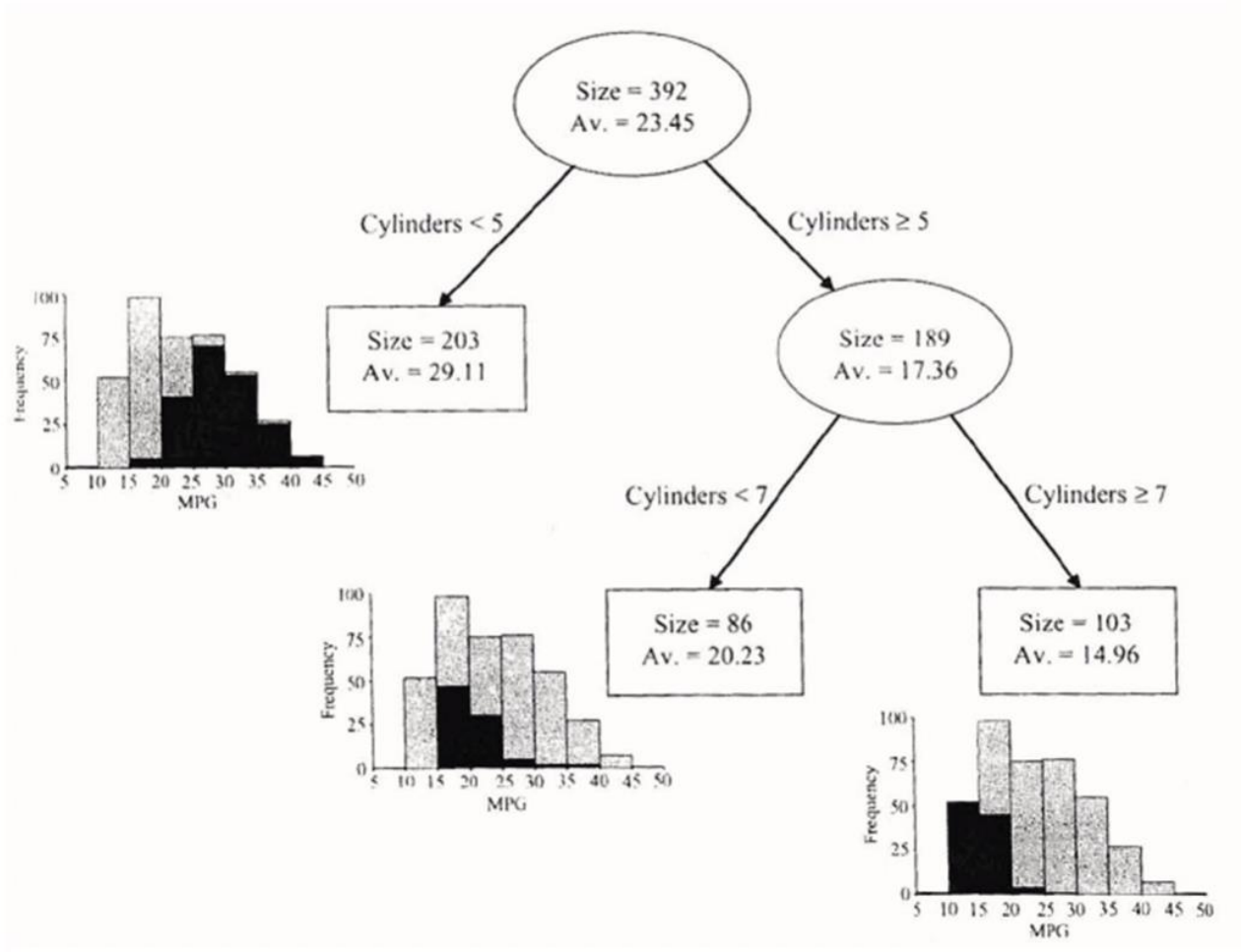
Stream 0.3

Slope 0.5

Elevation
750: 0.3
1350: 0.18
2250: 0.59
4175: 0.86

Unit 02 | 의사결정나무 알고리즘

연속형 변수



가지치기란?

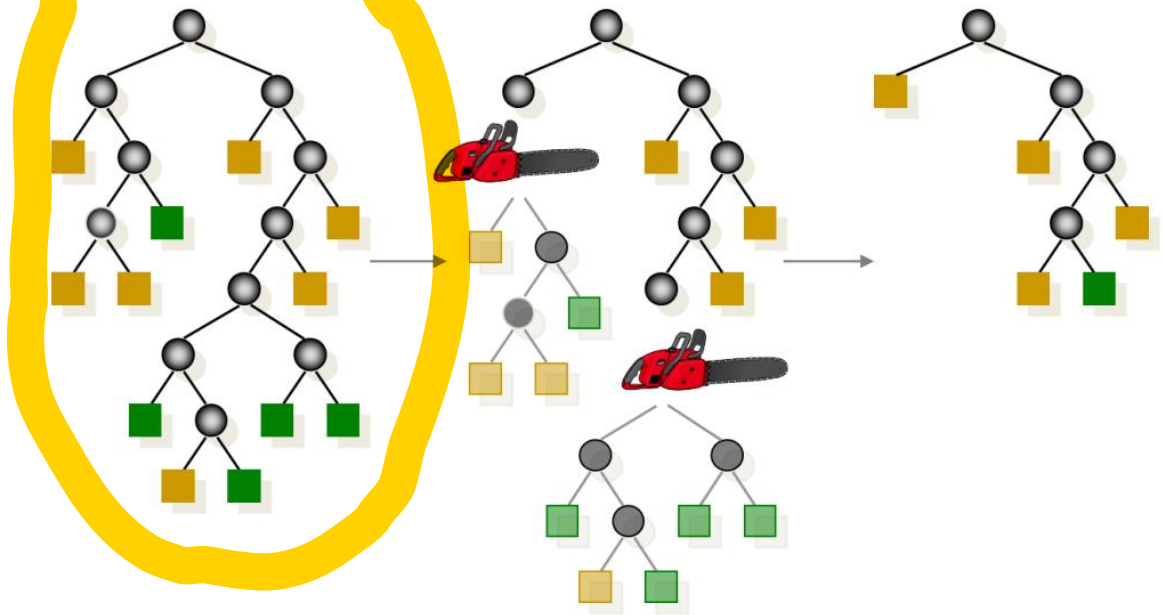
- 모든 terminal node (마지막 node) 의 순도가 100%인 상태를 Full tree 라고 함
- 이 경우 분기가 너무 많아 과적합(overfitting) 위험이 발생 & 모델이 너무 복잡해짐
- 분기가 증가할 경우 처음에는 오분류율이 감소하나 일정 수준 이상이 되면 일반화 능력이 떨어져 오히려 증가함
- 이를 방지하기 위해 적절한 수준에서 terminal node를 결합해주는 것

Unit 03 | 가지치기

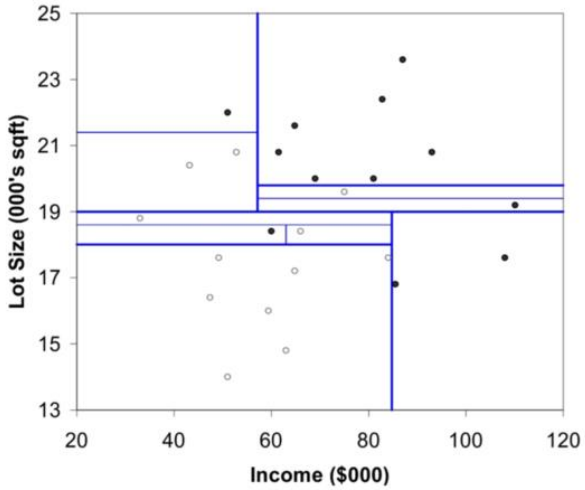
가지치기

Full tree
(terminal node의 불순도가 0)

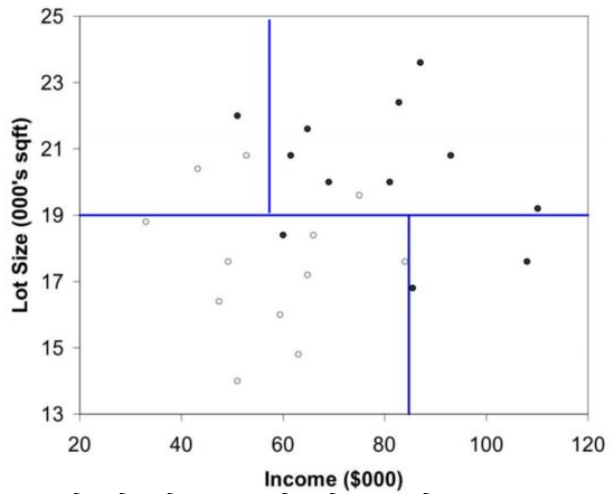
가지치기 후



Full tree
(terminal node의 불순도가 0)



가지치기 후



일반화 능력이좋다 .

가지치기의 종류

1. Pre-pruning (사전 가지치기)
   - 트리의 최대 depth나 분기점의 최소 개수를 미리 지정

2. Post-Pruning (사후 가지치기 또는 가지치기)
   - 트리를 만든 후 데이터포인트가 적은 노드를 삭제 or 병합

가지치기의 비용함수

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

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)

Unit 04 | 정리

장점

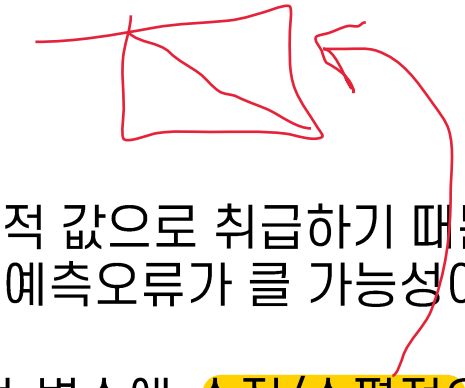
1. 결과를 해석하고 이해하기 용이하다.

2. 비모수적 모형이기 때문에 선형성, 정규성, 등분산성 등의 가정이 필요하지 않다.

3. 데이터를 가공할 필요가 거의 없다.

단점

데이터 형태가 이렇다면 어려움



1. 연속형 변수를 비연속적 값으로 취급하기 때문에 분리의 경계점 부근에서 예측오류가 클 가능성이 있다.

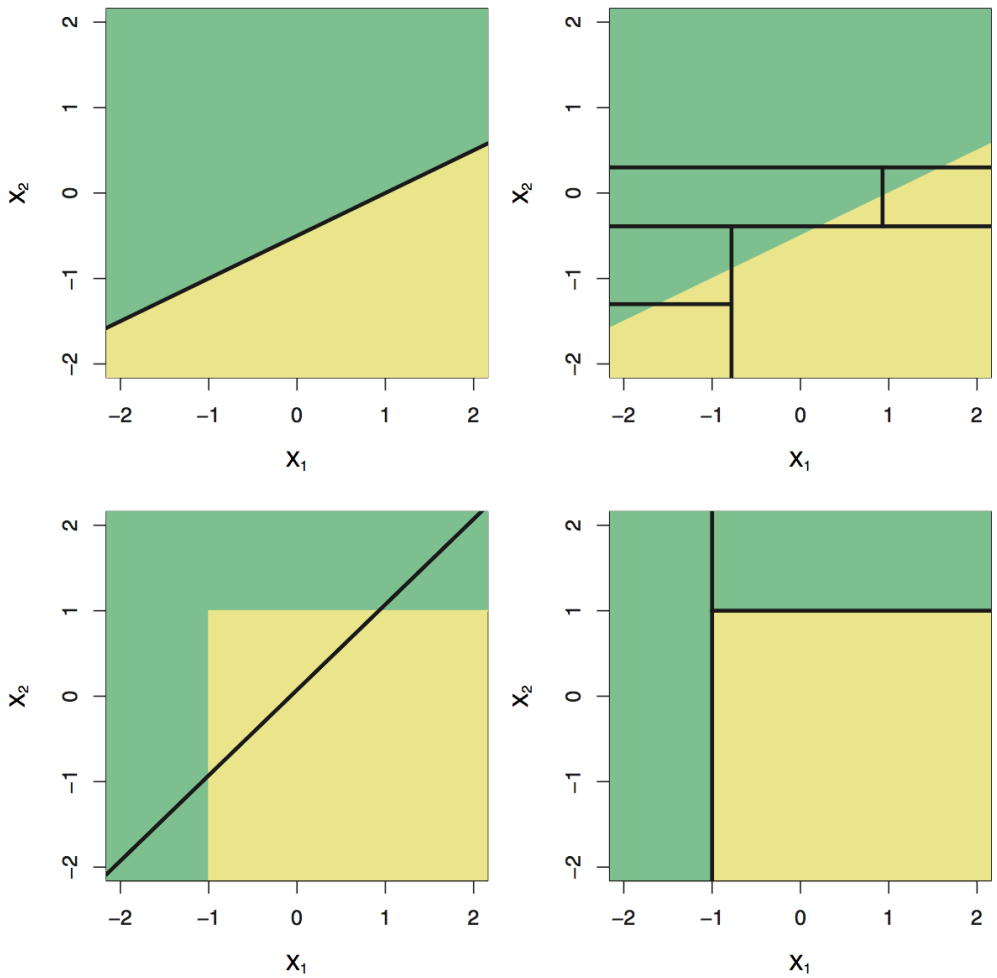
2. 데이터의 특성이 특정 변수에 수직/수평적으로 구분되지 못할 때 성능이 떨어지고 트리가 복잡해진다.

3. Overfitting 문제가 발생하기 쉽다.

4. 중간 단계에서 오류가 발생하면 다음 단계로 에러가 계속 전파된다.

5. 적은 개수의 노이즈에도 크게 영향을 받는다.

Unit 04 | 정리



단점 해결방안

앙상블!!

과제 1

[과제 1]

본인이 구현한 함수를 통해 다음 문제를 풀어주세요!

문제 1) 변수 'income'의 이진분류 결과를 보여주세요.

문제 2) 분류를 하는 데 가장 중요한 변수를 선정하고, 해당 변수의 Gini index를 제시해주세요.

문제3) 문제 2에서 제시한 feature로 DataFrame을 split한 후,
나눠진 2개의 DataFrame에서 각각 다음으로 중요한 변수를 선정하고 해당 변수의 Gini index를 제시해주세요.

주석 꼼꼼히 달아주세요!

과제 1

[과제 1]

주의사항

- 본인이 구현한 함수임을 증명하기 위해 주석 꼼꼼히 달아주세요.

- 이 데이터셋 뿐만 아니라 변수의 class가 더 많은 데이터에도 상관없이 적용 가능하도록 함수를 구현해 주세요.
  변수의 class가 3개를 넘는 경우 모든 이진분류 경우의 수를 따져보아야 합니다.
  Hint) itertools 라이브러리의 combinations 함수 & isin 함수 등이 활용될 수 있으며 이 밖에도 본인의 방법대로 마음껏 구현해주세요.

- 함수에 들어가는 변수나 flow 등은 본인이 변경해도 무관하며 결과만 똑같이 나오면 됩니다

과제 1

[과제 1]

get_gini(df, label) → $Gini(D_i) = 1 - \sum_{j=1}^{3} P_j$

get_attribute_gini_index(df, attribute, label) → $Gini(A) = \sum_{j=1}^{2} \frac{|D_j|}{|D|} * Gini(D_i)$

과제 2

[과제 2]

Entropy를 구하고, 각 변수에 대한 Gain을 구하는 함수를 구현하는 과제입니다.

DT_Assignment2.ipynb 파일에 있는 두가지 함수를 만들어 주시면 됩니다. 결과는 주어져 있습니다.

두번째 함수는 출력값이 꼭 주어진 형태와 일치할 필요는 없습니다. 봤을 때 각 변수에 대한 Gain을 알아볼 수 있도록 구성해주세요.

마찬가지로 주석 꼼꼼히 달라주세요!

참고자료

- https://ratsgo.github.io/machine%20learning/2017/03/26/tree/
- https://ratsgo.github.io/machine%20learning/2017/10/04/comparison/
- https://ai-times.tistory.com/161?category=126028
- https://dreamlog.tistory.com/576
- https://scikit-learn.org/stable/modules/tree.html
- https://yamalab.tistory.com/31
- Tobigs 11기 김유민 강의자료
- 고려대학교 김성범 교수님 강의

들어주셔서 감사합니다.