2021년도 2학기 바이오빅데이터와데이터마이닝
중간고사 보고서

이화여자대학교 컴퓨터공학과
1871056 한지수

1번 a) 문제 풀이

Apriori 원리에 따라 k=1 조건부터 시작하여 1-itemset → 2-itemset → 3-itemset 순으로 빈발항목집합을 구한다.

k=1	Count	Support(%)
{a}	4	66
{b}	5	83
{c}	4	66
{d}	3	50
{e}	2	33

⇒

k=2	Count	Support(%)
{a,b}	4	66
{a,c}	3	50
{a,d}	2	33
{b,c}	4	66
{b,d}	2	33
{c,d}	2	33

⇒

k=3	Count	Support(%)
{a,b,c}	3	50
{b,c,d}	2	33
{a,c,d}	1	16

(가지치기 끝 설명)

Support threshold 를 고려하여 이를 만족하는 상위 빈발항목집합은 {b,c}, {a,b,c}이다.

1871056 한지수

1번 b) 문제 풀이

빈발항목집합 {b,c}, {a,b,c}를 바탕으로 연관규칙을 구합니다.

연관규칙	Confidence
b ⟶ c	$c(b \to c) = \frac{\sigma(\{b,c\})}{\sigma(\{b\})} = \frac{4}{5} = 0.8$
a ⟶ bc	$c(a \to bc) = \frac{\sigma(\{a,b,c\})}{\sigma(\{a\})} = \frac{3}{4} = 0.75$
ab ⟶ c	$c(ab \to c) = \frac{\sigma(\{a,b,c\})}{\sigma(\{a,b\})} = \frac{3}{4} = 0.75$
ac ⟶ b	$c(ac \to b) = \frac{\sigma(\{a,b,c\})}{\sigma(\{a,b\})} = \frac{3}{4} = 0.75$
b ⟶ ac	$c(b \to ac) = \frac{\sigma(\{a,b,c\})}{\sigma(\{b\})} = \frac{3}{5} = 0.6$
bc ⟶ a	$c(bc \to a) = \frac{\sigma(\{a,b,c\})}{\sigma(\{b,c\})} = \frac{3}{4} = 0.75$
c ⟶ ab	$c(c \to ab) = \frac{\sigma(\{a,b,c\})}{\sigma(\{c\})} = \frac{3}{4} = 0.75$

minconf인 60를 만족하는 연관규칙은 다음과 같다.

- b ⟶ c
- a ⟶ bc
- ab ⟶ c
- ac ⟶ b
- b ⟶ ac
- bc ⟶ a
- c ⟶ ab

1871056 한지수

2번 문제 풀이

문제 풀이 방향

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Cheat 클라스를 예측하기 위한 결정트리를 위해 분기에 사용할 3가지 속성은 다음과 같습니다.

- Refund - 명목형 (이진: Yes, No) 속성
- Material Status - 명목형 (다중: Single, Married, Divorced) 속성
- Taxable Income (60K~220K) 연속형 속성

이 때 최선의 분할이 되도록 최선의 속성을 고르기 위해서는 셋 중에서 가장 최선의 information gain을 주는 것을 골라야 하며, Material Status는 속성 특성 덕분에 다중 분할이 가능한 점을 고려하면 다음과 같이 6가지 비교가 필요합니다.

- Refund의 이진 분할 속성에서 나오는 GIni(Children) 속성
- Material Status의 Multi-way split({Single},{Married}, {Divorced})의 Gini
- Material Status의 Two-way split의 GINI (3가지 경우)
    - {Single, Married}, {Divorced}
    - {Single, Divorced}, {Married}
    - {Married, Divorced}, {Single}
- Taxable Income의 GINI 중 split position을 고려한) 가장 낮은 Gini값

즉, 본 문제를 해결하기 위해 위 6가지 경우의 수 GINI 값 구하여 분기마다 최선의 속성을 구하고자 합니다.

(Gini값은 소수점 셋째자리에서 버림하여 표현됩니다.)

1871056 한지수

2번 문제 풀이

첫 번째 분기 속성 고르기

1. Refund의 Gini(Children) 속성

		Cheat		Gini(t)
		Yes	No	
Refund	Yes	0	3	0
	No	3	4	0.489
Gini = 0.342				

2. Material Status의 ({Single},{Married}, {Divorced})의 Gini

		Cheat		Gini(t)
		Yes	No	
Material Statue	{Single}	2	2	0.5
	{Married}	0	4	0
	{Divorced}	1	1	0.5
Gini = 0.3				

3. Material Status의 ({Single, Married}, {Divorced})의 Gini

		Cheat		Gini(t)
		Yes	No	
Material Statue	{S, M}	2	6	0.375
	{D}	1	1	0.5
Gini = 0.4				

4. Material Status의 ({Single, Divorced}, {Married}) 의 Gini

		Cheat		Gini(t)
		Yes	No	
Material Statue	{S, D}	3	3	0.5
	{M}	0	4	0
Gini = 0.3				

5. Material Status의 ({Married, Divorced}, {Single}) 의 Gini

		Cheat		Gini(t)
		Yes	No	
Material Statue	{M, D}	1	5	0.277
	{S}	2	2	0.5
Gini = 0.366				

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

6. Taxable Income의 Gini 중 split position을 고려한) 가장 낮은 Gini

Sorted Value	60		70		75		85		90		95		100		120		125		220	
Split Position	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
GINI	0.420		0.400		0.375		0.343		0.417		0.400		0.300		0.343		0.375		0.400		0.420	

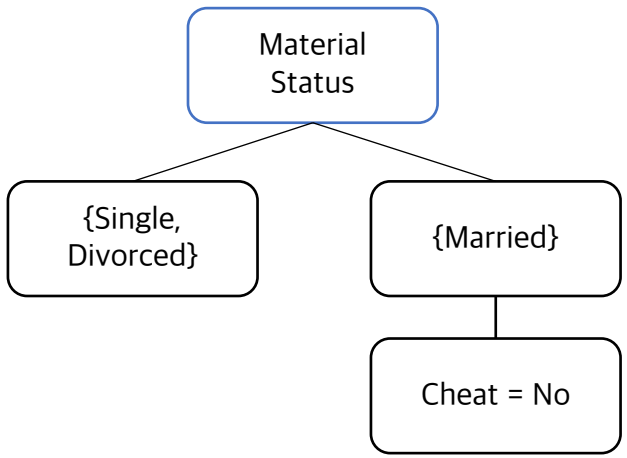
2번 문제 풀이

두 번째 분기 속성 고르기

⇒ 결론. 가장 낮은 Gini Index를 가지는 경우의 수는 다음과 같다.

2. Material Status의 ({Single},{Married}, {Divorced})
4. Material Status의 ({Single, Divorced}, {Married})
6. Taxable Income ({≤97},{>97})

본 문제에서는 4번을 고르기로 결정하였으며, 이에 따라 만들어지는 결정트리는 다음과 같다.



2번 문제 풀이

두 번째 분기 속성 고르기 Marital Status의 {Single, Divorced}인 해당되는 instance인 1,3,5,7,8,10에 대하여 속성을 고르고자 한다.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
3	No	Single	70K	No
5	No	Divorced	95K	Yes
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
10	No	Single	90K	Yes

1. Refund의 Gini(Children) 속성

		Cheat		Gini(t)
		Yes	No	
Refund	Yes	0	2	0
	No	3	1	0.375
Gini = 0.25				

2. Material Status의 ({Single}, {Divorced})의 Gini

		Cheat		Gini(t)
		Yes	No	
Material Statue	{Single}	2	2	0.5
	{Divorced}	1	1	0.5
Gini = 0.5				

3. Taxable Income의 Gini 중 split position을 고려한) 가장 낮은 Gini

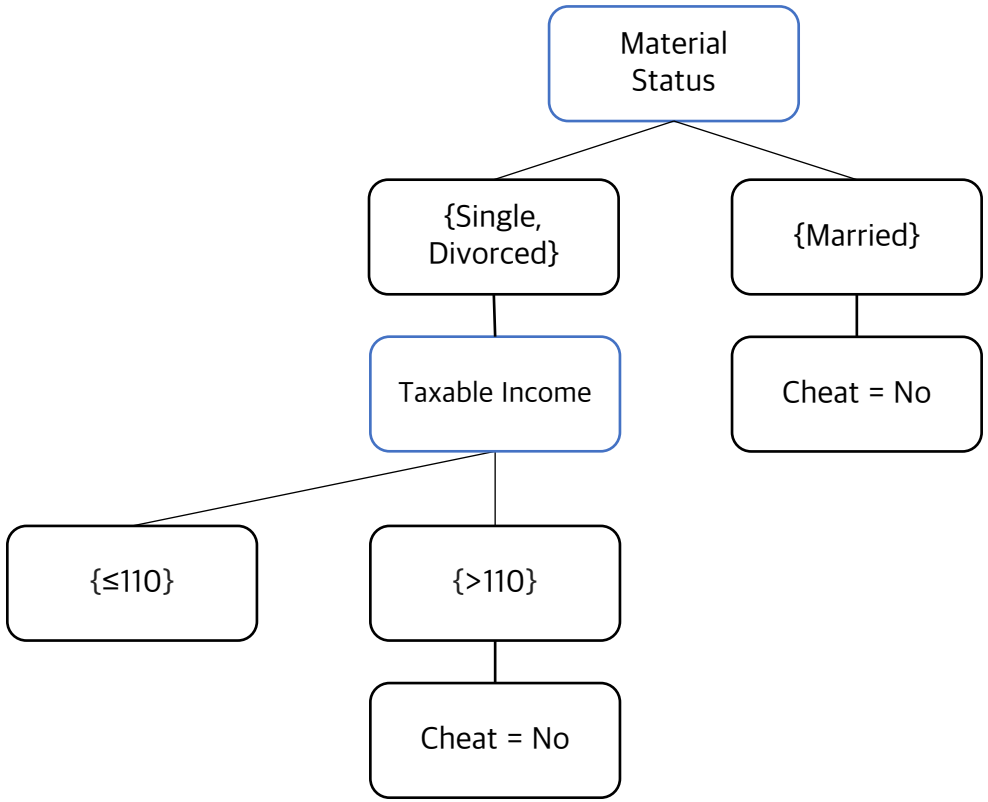
Sorted Value	70		85		90		95		125		220			
Split Position	65		77		87		92		110		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	1	2	2	1	3	0	3	0	3	0
No	0	3	1	2	1	2	1	2	1	2	2	1	2	1
GINI(t)	1	0.5	0	0.48	0.5	0.5	0.444	0.444	0.375	0	0.48	0	0.48	0
GINI	0.5		0.4		0.5		0.444		0.25		0.4		0.4	

2번 문제 풀이

두 번째 분기 속성 고르기

⇒ 결론. 가장 낮은 Gini Index 0.25를 가지는 경우의 수는 다음과 같다.
1. Refund의 Gini(Children)
3. Taxable Income ({≤110},{>110})

본 문제에서는 3번 조건을 고르기로 결정하였으며, 이에 따라 만들어지는 결정트리는 다음과 같다.



2번 문제 풀이

세 번째 분기 속성 고르기 Marital Status의 {Single, Divorced}인 해당되는 instance인 1,3,5,7,8,10에 대하여 속성을 고르고자 한다.

1. Refund의 Gini(Children) 속성

		Cheat		Gini(t)
		Yes	No	
Refund	Yes	0	0	1
	No	1	3	0.375
Gini = 0.375				

2. Material Status의 ({Single}, {Divorced})의 Gini

		Cheat		Gini(t)
		Yes	No	
Material Statue	{Single}	2	1	0.444
	{Divorced}	1	0	0
Gini = 0.333				

3. Taxable Income의 Gini 중 split position을 고려한) 가장 낮은 Gini

Sorted Value	70		85		90		95			
Split Position	65		77		87		92		100	
	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	2	0	3	1	2	2	1	3	0
No	0	2	1	0	1	0	1	0	1	0
GINI(t)	1	0.5	0	0	0.5	0	0.444	0	0.375	0
GINI	0.5		0		0.25		0.333		0/375	

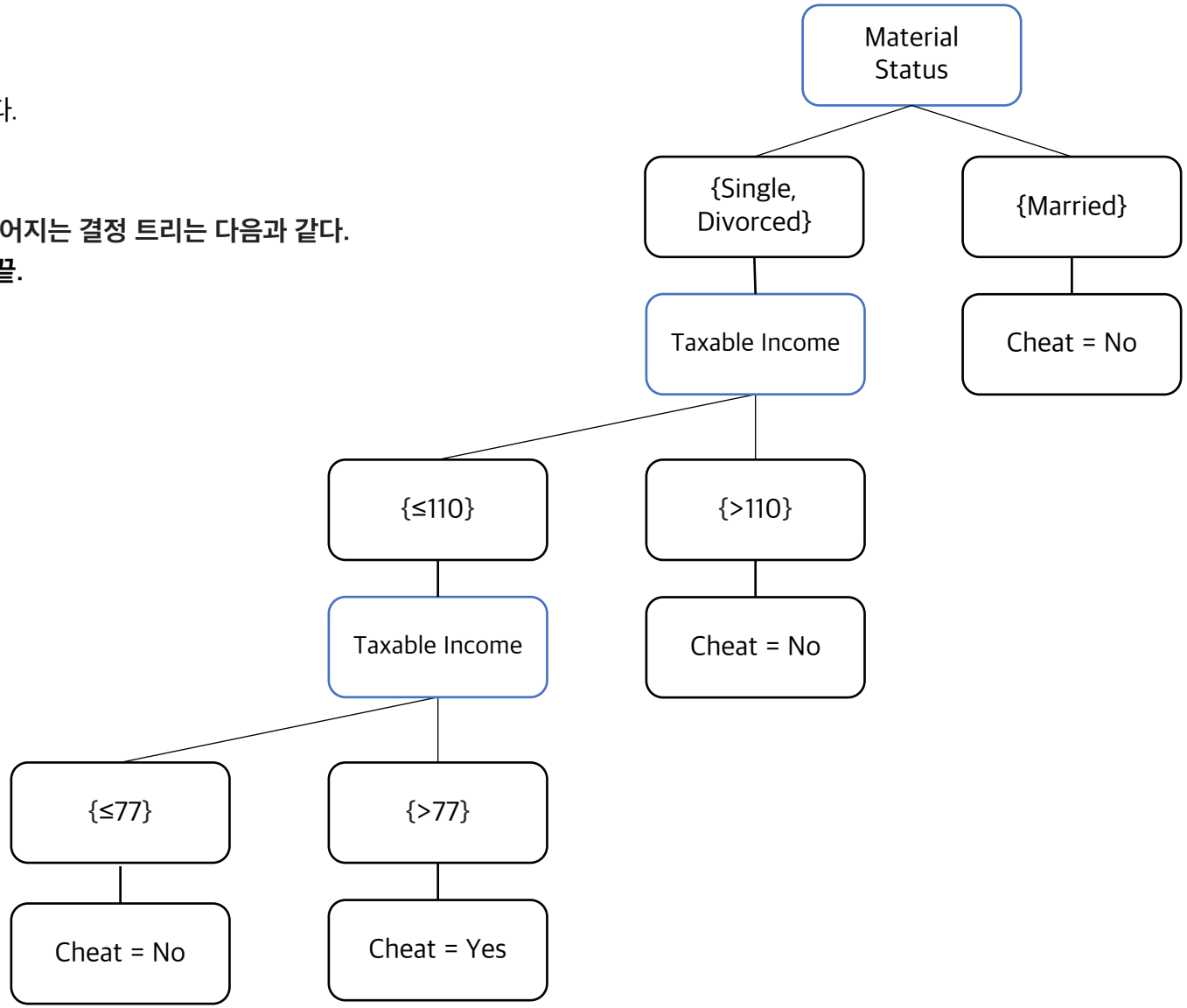
Tid	Refund	Marital Status	Taxable Income	Cheat
3	No	Single	70K	No
5	No	Divorced	95K	Yes
8	No	Single	85K	Yes
10	No	Single	90K	Yes

2번 문제 풀이

세 번째 분기 속성 고르기

⇒ 결론. 가장 낮은 Gini Index 0을 가지는 경우의 수는 다음과 같다.
3. Taxable Income ({≤77},{>77})

본 문제에서는 3번 조건을 고르기로 결정하였으며, 이에 따라 만들어지는 결정 트리는 다음과 같다.
3번 조건을 적용하면서 모든 데이터가 분류 완료되었으므로 분할 끝.



3번 문제 풀이

breast-cancer-wisconsin.data 데이터셋 소개 (breast-cancer-wisconsin.names 설명 참조)

5. Number of Instances: 699 (as of 15 July 1992)

6. Number of Attributes: 10 plus the class attribute

7. Attribute Information: (class attribute has been moved to last column)

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

8. Missing attribute values: 17

There are 16 instances in Groups 1 to 6 that contain a single missing (i.e., unavailable) attribute value, now denoted by "?".

9. Class distribution:

Benign: 458 (65.5%)
Malignant: 241 (34.5%)

- 699 instance (단, 결측치는 17개 객체에 존재하며 “?”로 명시되어있다. 모든 속성 데이터를 가지고 있는 객체는 682개이다.)
- 10개 속성과 class attribute로 구성되어있다.

1.	Sample code number (단순 id 숫자이므로 결정트리 속성에 제외하였다.)	
2.	Clump Thickness	1-10 연속형 속성
3.	Uniformity of Cell Size	1-10 연속형 속성
4.	Uniformity of Cell Shape	1-10 연속형 속성
5.	Marginal Adhesion	1-10 연속형 속성
6.	Single Epithelial Cell Size	1-10 연속형 속성
7.	Bare Nuclei	1-10 연속형 속성
8.	Bland Chromatin	1-10 연속형 속성
9.	Normal Nucleoli	1-10 연속형 속성
10.	Mitoses	1-10 연속형 속성
11.	Class	(2는 양성, 4는 음성으로 표기)

⇒ 결정트리를 통해 breast-cancer-wisconsine.data 의 2-10번 속성 (중 고른다는거 표현)을 통해 Class(11번째 속성)을 분류하고자 한다.

1871056 한지수

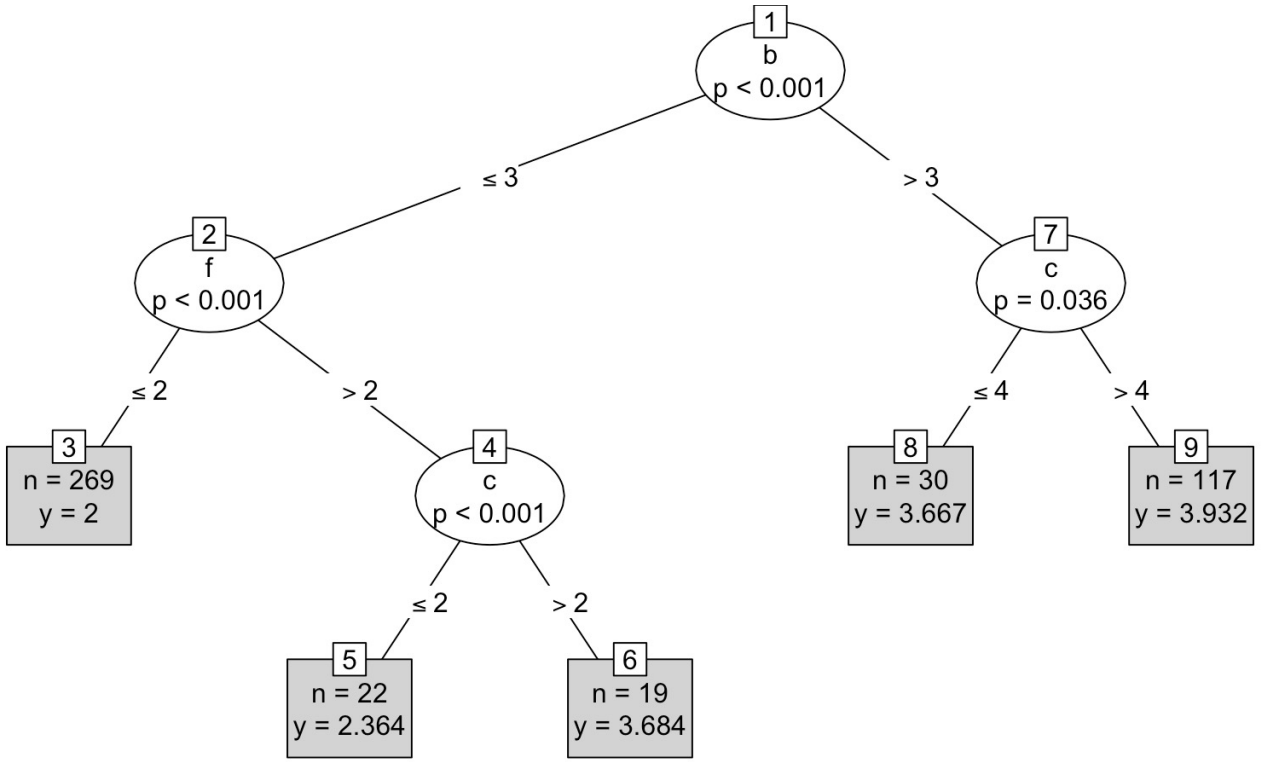
3번 문제 풀이

midterm.R 코드 파이프라인 및 결정트리 결과

1. prepare dataset
    a. 데이터 셋의 column name 수정
    b. “?”와 같은 데이터 값 수정
2. Split training set and test set
    1. Column b,c,d,f
3. Training set의 Decision Tree: ctree() "conditional inference tree", Decision Tree인 bio_ctree 생성
4. Classification with test set
5. Check accuracy

⇒ bio_ctree: 노드 1,2,4,7에서 4개의 분할 생성

Test set accuracy: 60.4444%



생성된 bio_ctree 결정트리

3번 문제 풀이

첫 번째 분기 속성 고르기

1. Column b의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

2. Column c의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

3번 문제 풀이

첫 번째 분기 속성 고르기

3. Column e의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

4. Column f의 Gini 중 split position을 고려한) 가장 낮은 Gini

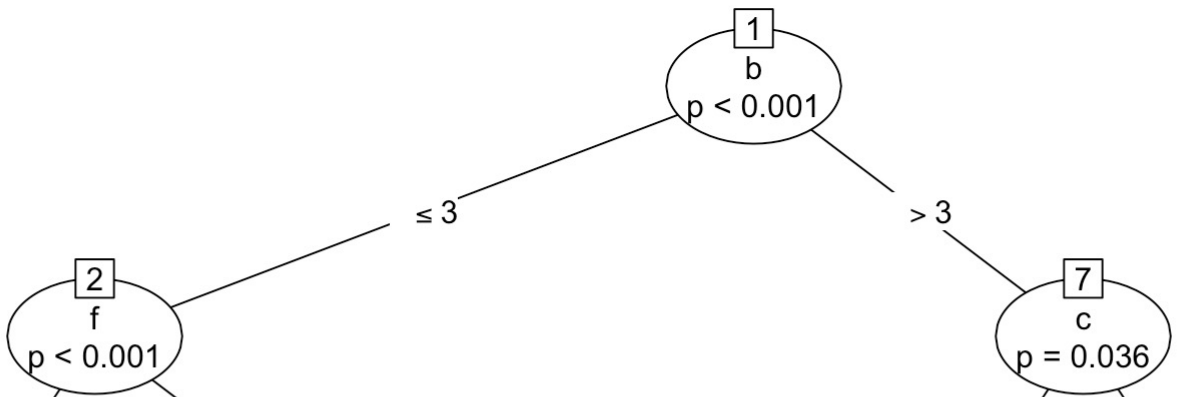
Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

3번 문제 풀이

첫 번째 분기 속성 고르기

⇒ 결론. 가장 낮은 Gini Index 0을 가지는 경우의 수는 다음과 같다.

본 문제에서는 3번 조건을 고르기로 결정하였으며, 이에 따라 만들어지는 결정 트리는 다음과 같다.



3번 문제 풀이

두 번째 분기 속성 고르기

1. Column b의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

2. Column c의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

3번 문제 풀이

두 번째 분기 속성 고르기

3. Column d의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

4. Column f의 Gini 중 split position을 고려한) 가장 낮은 Gini

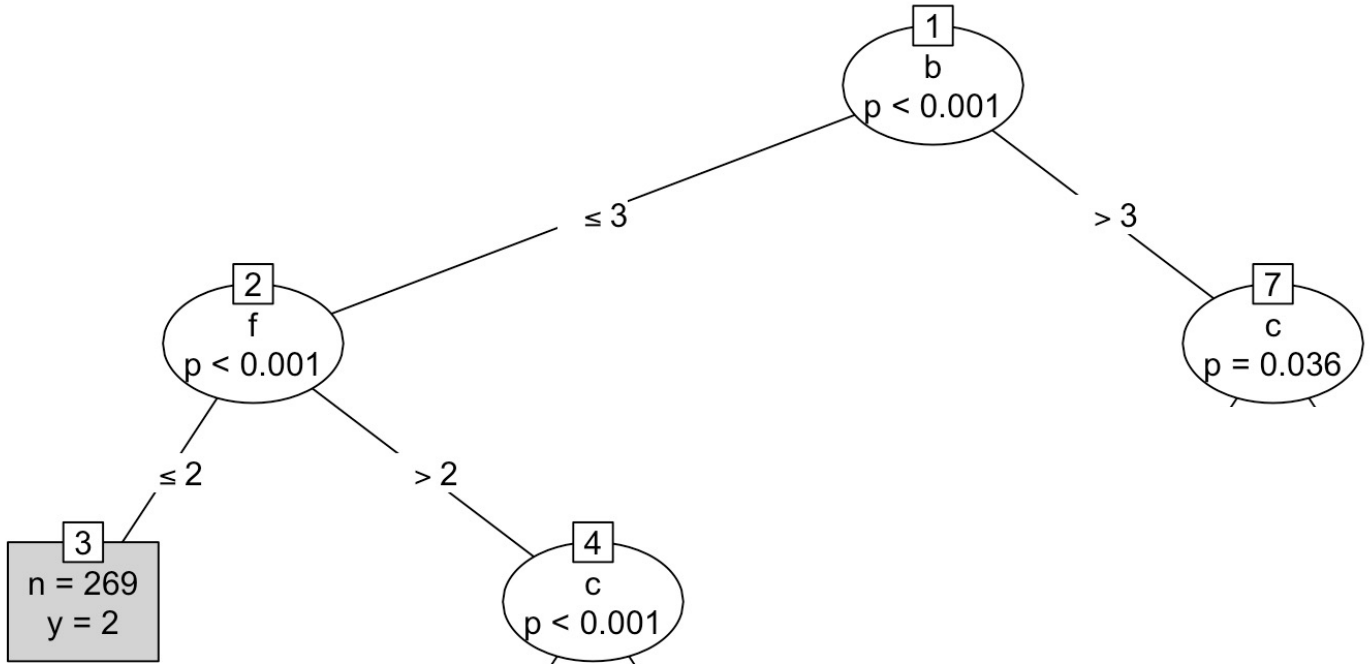
Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

3번 문제 풀이

두 번째 분기 속성 고르기

⇒ 결론. 가장 낮은 Gini Index 0을 가지는 경우의 수는 다음과 같다.

본 문제에서는 3번 조건을 고르기로 결정하였으며, 이에 따라 만들어지는 결정 트리는 다음과 같다.



3번 문제 풀이

세 번째 분기 속성 고르기

1. Column b의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

2. Column c의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

3번 문제 풀이

세 번째 분기 속성 고르기

3. Column d의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

4. Column f의 Gini 중 split position을 고려한) 가장 낮은 Gini

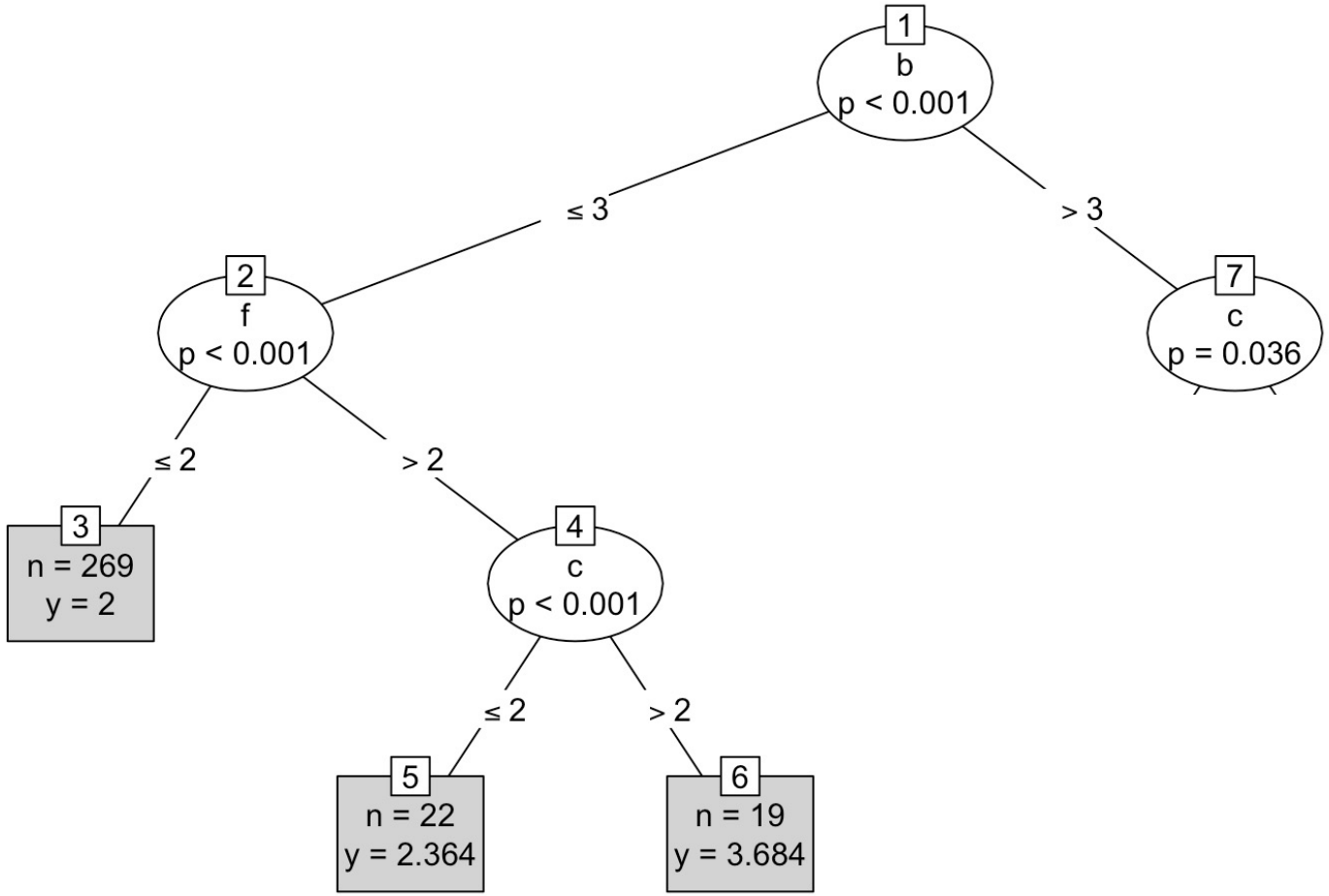
Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

3번 문제 풀이

세 번째 분기 속성 고르기

⇒ 결론. 가장 낮은 Gini Index 0을 가지는 경우의 수는 다음과 같다.

본 문제에서는 3번 조건을 고르기로 결정하였으며, 이에 따라 만들어지는 결정 트리는 다음과 같다.



3번 문제 풀이

네 번째 분기 속성 고르기

1. Column b의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

2. Column c의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

3번 문제 풀이

네 번째 분기 속성 고르기

3. Column d의 Gini 중 split position을 고려한) 가장 낮은 Gini

Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

4. Column f의 Gini 중 split position을 고려한) 가장 낮은 Gini

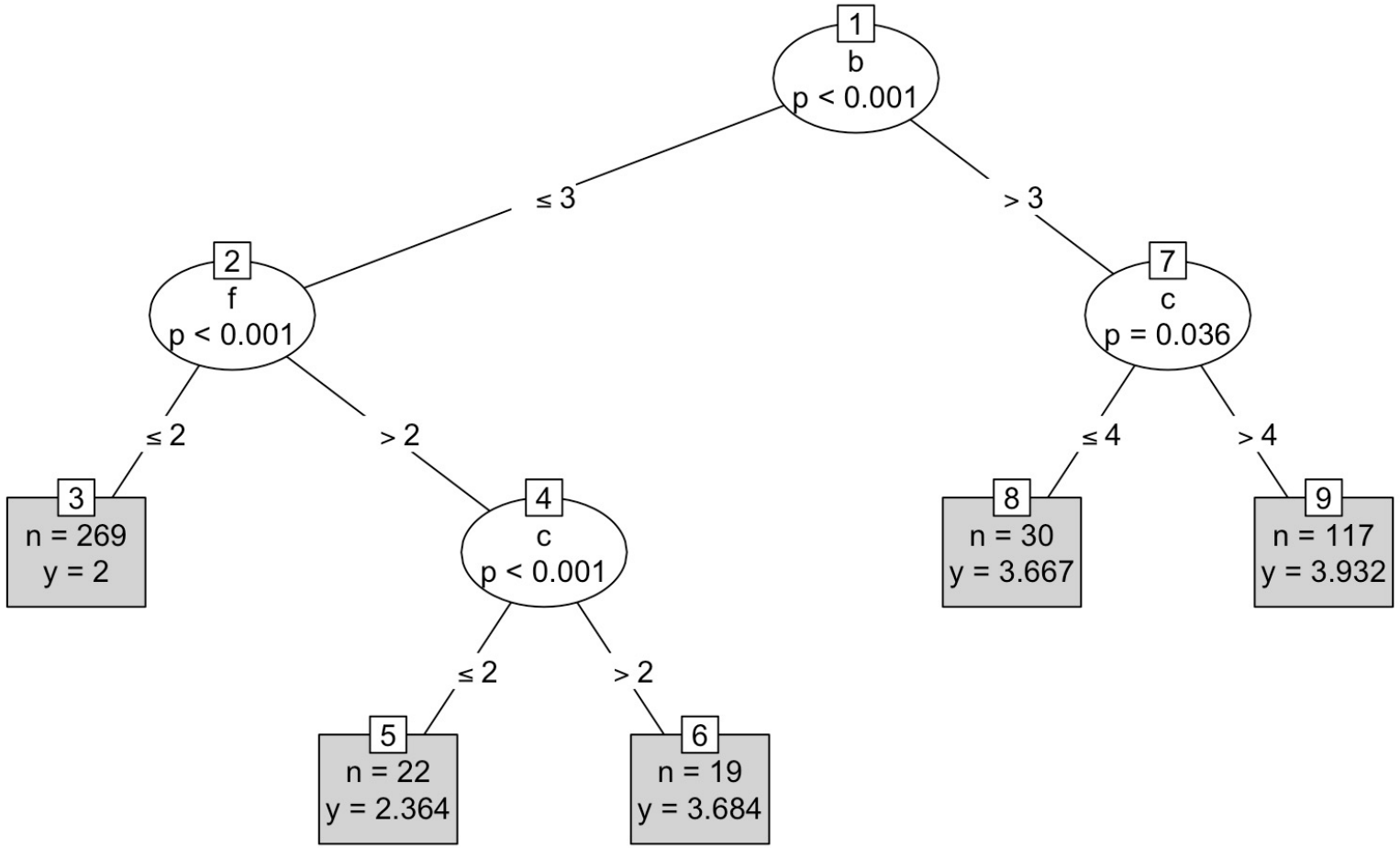
Split Position	1		2		3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2																				
Class 4																				
GINI(t)																				
GINI																				

3번 문제 풀이

네 번째 분기 속성 고르기

⇒ 결론. 가장 낮은 Gini Index 0을 가지는 경우의 수는 다음과 같다.

본 문제에서는 3번 조건을 고르기로 결정하였으며, 이에 따라 만들어지는 결정 트리는 다음과 같다.



3번 문제 풀이

최선의 분할 결정

노드 1,2,3,4,9,11,12에서 7개의 분할 생성

- 노드 1
- 노드 2
- 노드 4
- 노드 7

Sorted Value								
	<=	>	<=	>	<=	>	<=	>
GINI								