2021년도 2학기 바이오빅데이터와데이터마이닝
중간고사 보고서

이화여자대학교 컴퓨터공학과
1871056 한지수

1번 a) 문제 풀이

Apriori 원리에 따라 데이터베이스로부터 각 원소 항목의 지지도를 구하고자 합니다.
k=1 조건부터 시작하여 1-itemset → 2-itemset → 3-itemset 순으로 빈발항목집합을 구합니다. Count는 편의상 분자만 표기하였습니다.

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}	3	50
{c,d}	2	33

⇒

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

Minsup(=support threshold 50)를 만족하는
3원소 집합은 1개이므로 4원소 집합으로
확장 불가. 알고리즘은 여기서 중단

Support threshold 를 고려하여 이를 만족하는 상위 빈발항목집합은 {b,d}, {a,b,c}이며, 이를 바탕으로 연관 규칙을 생성하게 된다.

∴ Apriori 원리를 통해 추출된 빈발항목집합: $\{a\}, \{b\}, \{c\}, \{d\}, \{a, b\}, \{a, c\}, \{b, c\}, \{b, d\}, \{a, b, c\}$

1871056 한지수

1번 b) 문제 풀이

상위빈발항목집합 {b,d}, {a,b,c}를 바탕으로 연관 규칙을 구한다.

연관 규칙	Confidence
$\{b\} \rightarrow \{d\}$	$c(\{b\} \rightarrow \{d\}) = \frac{\sigma(\{b,d\})}{\sigma(\{b\})} = \frac{3}{5} = 0.75$
$\{a\} \rightarrow \{b,c\}$	$c(\{a\} \rightarrow \{b,c\}) = \frac{\sigma(\{a,b,c\})}{\sigma(\{a\})} = \frac{3}{4} = 0.75$
$\{a,b\} \rightarrow \{c\}$	$c(\{a,b\} \rightarrow \{c\}) = \frac{\sigma(\{a,b,c\})}{\sigma(\{a,b\})} = \frac{3}{4} = 0.75$
$\{a,c\} \rightarrow \{b\}$	$c(\{a,c\} \rightarrow \{b\}) = \frac{\sigma(\{a,b,c\})}{\sigma(\{a,b\})} = \frac{3}{4} = 0.75$
$\{b\} \rightarrow \{a,c\}$	$c(\{b\} \rightarrow \{a,c\}) = \frac{\sigma(\{a,b,c\})}{\sigma(\{b\})} = \frac{3}{5} = 0.6$
$\{b,c\} \rightarrow \{a\}$	$c(\{b,c\} \rightarrow \{a\}) = \frac{\sigma(\{a,b,c\})}{\sigma(\{b,c\})} = \frac{3}{4} = 0.75$
$\{c\} \rightarrow \{a,b\}$	$c(\{c\} \rightarrow \{a,b\}) = \frac{\sigma(\{a,b,c\})}{\sigma(\{c\})} = \frac{3}{4} = 0.75$

오른쪽의 표를 참고하여 $Confidence(\%) \geq 60 (= confidence\ threshold)$를 만족하는 연관규칙은 다음과 같습니다.

∴추출된 연관 규칙

- $\{b\} \rightarrow \{d\}$
- $\{a\} \rightarrow \{b,c\}$
- $\{a,b\} \rightarrow \{c\}$
- $\{a,c\} \rightarrow \{b\}$
- $\{b\} \rightarrow \{a,c\}$
- $\{b,c\} \rightarrow \{a\}$
- $\{c\} \rightarrow \{a,b\}$

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 각 분기에 비교할 3가지 속성은 다음과 같습니다.

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

이 때 최선의 분할이 되는 최선의 속성을 고르기 위해서는 셋 중에서 최선의 information gain을 주는 것을 골라야 하며,
Marital Status는 속성 특성으로 다중 분할이 가능한 점을 고려하면 다음과 같이 6가지 비교가 필요합니다.

- Refund의 이진 분할 속성에서 나오는 Gini(Children)
- Marital Status 의 Multi-way split({Single},{Married}, {Divorced})의 Gini
- Marital Status 의 Two-way split의 Gini
    - {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

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

2. Marital Status 의 ({Single},{Married}, {Divorced})의 Gini

		Cheat		Gini(t)
		Yes	No	
Marital Status	{Single}	2	2	0.5
	{Married}	0	4	0
	{Divorced}	1	1	0.5
Gini = <u>0.3</u>				

3. Marital Status 의 ({Single, Married}, {Divorced})의 Gini

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

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

		Cheat		Gini(t)
		Yes	No	
Marital Status	{S, D}	3	3	0.5
	{M}	0	4	0
Gini = <u>0.3</u>				

5. Marital Status 의 ({Married, Divorced}, {Single}) 의 Gini

		Cheat		Gini(t)
		Yes	No	
Marital Status	{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		**<u>0.300</u>**		0.343		0.375		0.400		0.420	

1871056 한지수

2번 문제 풀이

첫 번째 분기 속성 고르기 결론

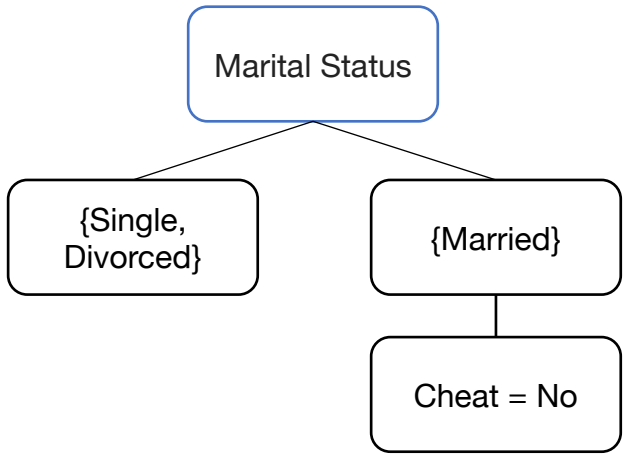
가장 낮은 Gini Index를 가지는 경우의 수는 다음과 같습니다.

2. Marital Status 의 ({Single},{Married}, {Divorced})
4. Marital Status 의 ({Single, Divorced}, {Married})
6. Taxable Income ({≤97},{>97})

이 중에서 어떤 경우를 분기 속성으로 고르는지에 따라 새로운 의사결정트리가 만들어집니다.

본 문제에서는 4번 속성을 고르기로 결정하였으며, 이에 따라 만들어지는 결정 트리는 다음과 같습니다.

Marital Status이 Married인 경우 Cheat=No로 바로 분류 가능하며, 이에 따라 남은 노드인 {Single, Divorced}인 경우에 대하여 분기 속성을 고르도록 하겠습니다.



1871056 한지수

2번 문제 풀이

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

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

		Cheat		Gini(t)
		Yes	No	
Refund	Yes	0	2	0
	No	3	1	0.375
Gini = <u>0.25</u>				

2. Marital Status 의 ({Single}, {Divorced})의 Gini

		Cheat		Gini(t)
		Yes	No	
Marital Status	{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 (K)	70		85		90		95		125		220			
Split Position (K)	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		**<u>0.25</u>**		0.4		0.4	

1871056 한지수

2번 문제 풀이

두 번째 분기 속성 고르기 결론

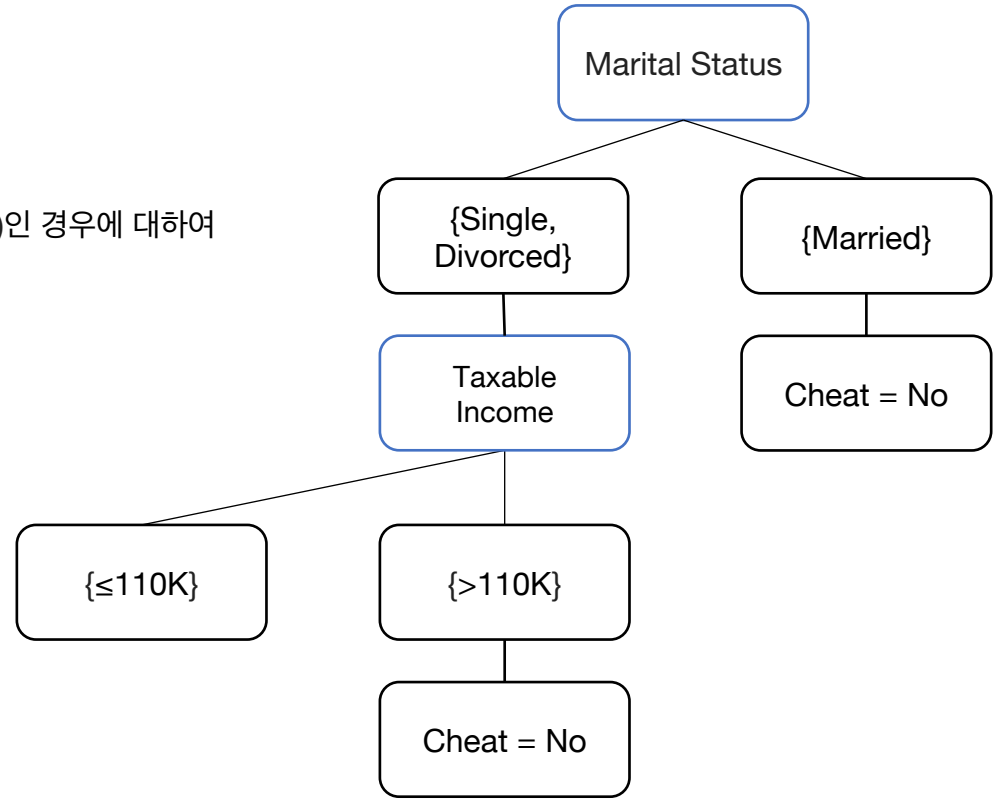
가장 낮은 Gini Index 0.25를 가지는 경우의 수는 다음과 같습니다.

1. Refund의 Gini
3. Taxable Income ({≤110K},{>110K})

이 중에서 어떤 경우를 분기 속성으로 고르는지에 따라 새로운 의사결정트리가 만들어집니다.

본 문제에서는 3번 속성을 고르기로 결정하였으며, 이에 따라 만들어지는 결정트리는 다음과 같습니다.

Taxable Income이 {>110}) 인 경우 Cheat=No로 바로 분류 가능하며, 이에 따라 남은 노드인 ({≤110})인 경우에 대하여 분기 속성을 고르도록 하겠습니다.



2번 문제 풀이

세 번째 분기 속성 고르기 Marital Status가 {Single, Divorced}이며, Taxable Income이 {≤110K} 에 해당되는 instance인 3,5,8,10에 대하여 속성을 고르고자 합니다.

1. Refund의 Gini

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

2. Marital Status 의 ({Single}, {Divorced})의 Gini

		Cheat		Gini(t)
		Yes	No	
Marital Status	{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 (K)	70		85		90		95			
Split Position (K)	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

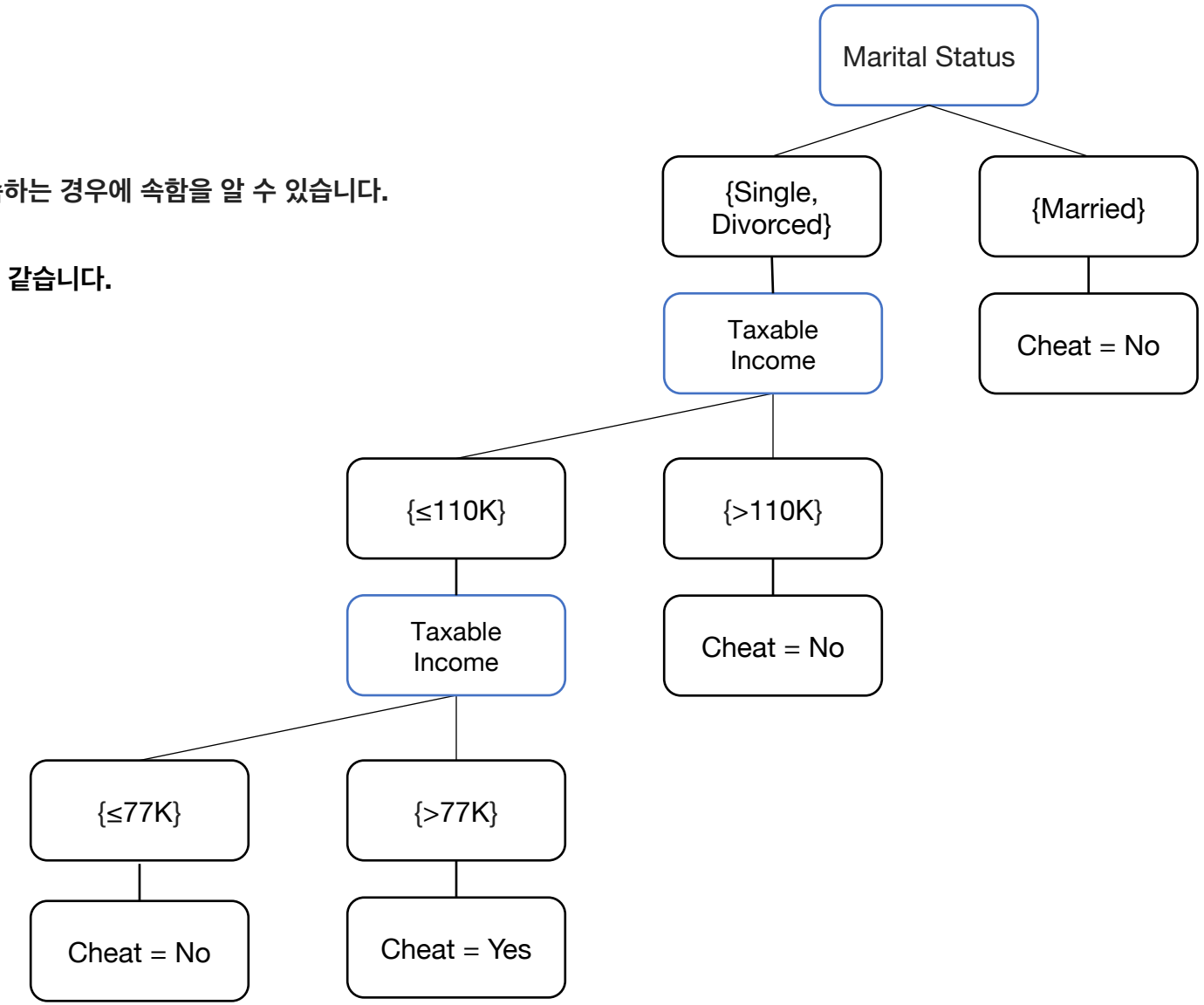
1871056 한지수

2번 문제 풀이

세 번째 분기 속성 고르기 결론 및 최종 결론

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

불순도 척도인 Gini index가 0인 점으로 보아 모든 레코드가 하나의 클래스에 속하는 경우에 속함을 알 수 있습니다.
이에 따라 만들어지는 결정 트리는 다음과 같습니다.
3번 조건을 적용하면서 모든 데이터가 분류 완료되며, 완성된 결정 트리는 다음과 같습니다.



1871056 한지수

3번 문제 풀이

breast-cancer-wisconsin.data 데이터셋 소개

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%)

breast-cancer-wisconsin.names

- 699 instance (단, 결측치는 17개 객체에 존재하며 “?”, NA로 명시되어있다. 모든 속성 데이터를 가지고 있는 객체는 682개이다.)
- 10개 속성과 class attribute로 구성되어있다.
    1. Sample code number (단순 id 숫자이므로 결정트리 속성에 제외)
    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 를 통해 Class(11번째 속성)을 분류하고자 합니다.

1871056 한지수

3번 문제 풀이

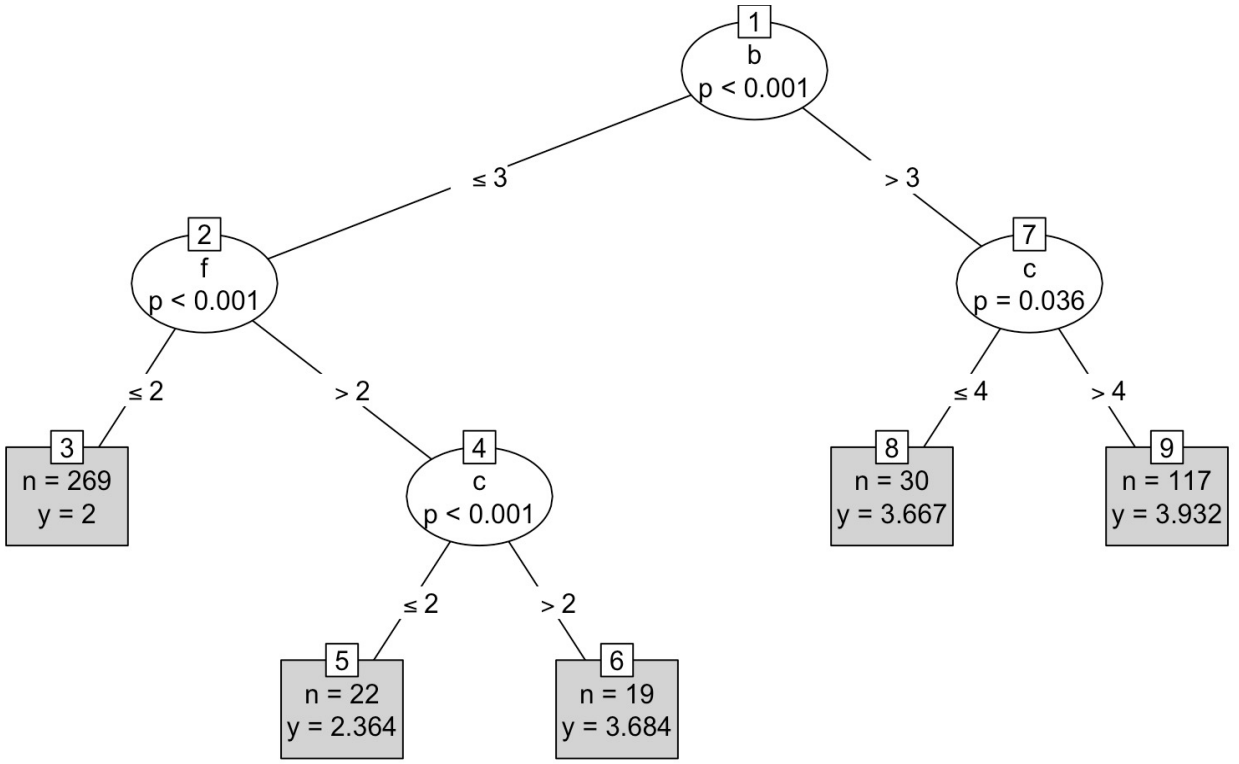
midterm.R 코드 파이프라인 및 결정트리 결과

1. Dataset Preperation
    a. 데이터 셋의 column name 수정 (2-11행: a-i, class)
    b. 결측치 처리: “?”, NA 가진 instance
2. Split training set and test set (train: test = 0.7:0.3)
3. Training set의 Decision Tree: ctree() "conditional inference tree", Decision Tree인 bio_ctree 생성
4. Classification with test set and check accuracy

⇒ bio_ctree: 노드 1,2,4,7에서 4개의 분할 생성

Train dataset = 480 instances, Test dataset = 202 instances

Test set accuracy: 72.4444%



생성된 bio_ctree 결정트리

➢ 본 문제에는 2-10행 중 3,4,5,7 행(Column b,c,d,f) 속성이 결정트리 분류 속성에 들어갔습니다.

- 난수 생성 등의 컴퓨터에 세팅된 여러 변수로 인해 코드 실행시 다른 결정 트리 결과가 나올 수 있습니다.
- Gini index는 소수 셋째자리에서 버림 처리되었습니다.
- ctree 개념은 p-test에 의한 significance를 기준으로 분기가 됩니다. 현재 비교자하는 gini index와 달라 비교 결과가 상이하게 나올 가능성이 존재합니다.

1871056 한지수

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	248	47	274	21	276	19	292	3	292	3	292	3	293	2	294	1	295	0	295	0
Class 4	2	171	8	165	9	144	50	123	72	101	92	81	107	66	124	49	126	47	173	0
GINI(t)	0.015	0.338	0.067	0.207	0.055	0.200	0.249	0.046	0.317	0.056	0.364	0.068	0.391	0.057	0.417	0.039	0.419	0	0.466	1
GINI	0.166		0.117		**0.113**		0.194		0.259		0.311		0.343		0.376		0.377		0.466	

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	229	66	266	29	284	11	288	7	290	5	292	3	294	1	295	0	295	0	295	0
Class 4	0	173	7	166	25	148	48	125	73	100	97	76	115	58	133	40	137	36	173	0
GINI(t)	0	0.399	0.049	0.253	0.148	0.128	0.244	0.100	0.321	0.090	0.374	0.073	0.404	0.033	0.428	0	0.433	0	0.466	1
GINI	0.204		0.134		0.141		0.204		0.269		0.323		0.357		0.391		0.399		0.466	

1871056 한지수

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	28	267	266	29	283	12	288	7	291	4	292	3	293	2	294	1	294	1	295	0
Class 4	0	173	13	160	41	132	75	98	100	73	131	42	138	35	150	23	151	22	173	0
GINI(t)	0	0.477	0.088	0.259	0.221	0.152	0.327	0.124	0.380	0.098	0.427	0.124	0.435	0.102	0.447	0.079	0.448	0.083	0.466	1
GINI	0.448		0.157		0.200		0.282		0.334		0.398		0.409		0.428		0.430		0.466	

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	261	34	271	24	279	16	284	11	291	4	291	4	291	4	293	2	293	2	295	0
Class 4	14	159	22	151	35	138	42	131	58	115	62	111	68	105	78	95	84	89	173	0
GINI(t)	0.096	0.290	0.138	0.236	0.198	0.186	0.224	0.142	0.277	0.064	0.289	0.067	0.307	0.070	0.332	0.040	0.346	0.042	0.466	1
GINI	0.176		0.174		0.194		0.199		0.223		0.234		0.252		0.271		0.287		0.466	

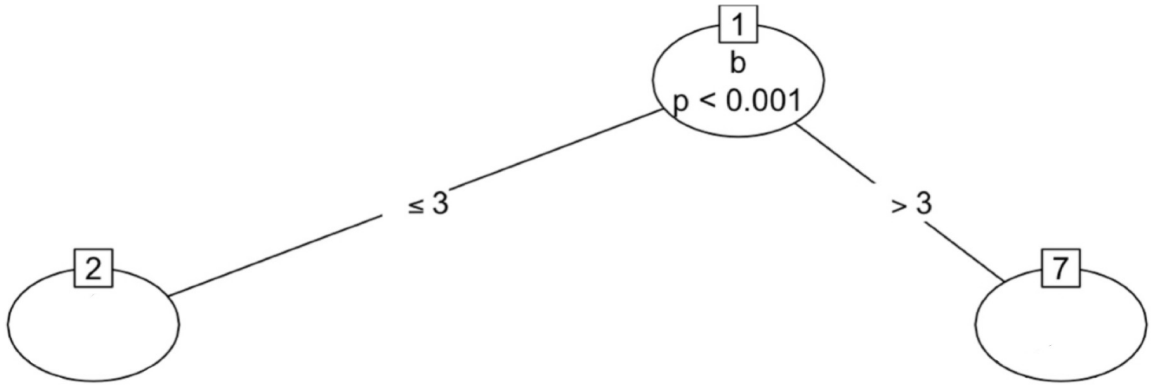
1871056 한지수

3번 문제 풀이

첫 번째 분기 속성 고르기 결론

가장 낮은 Gini Index 0.113을 가지는 경우의 수는 다음과 같습니다.
- Column b의 ({≤3},{>3})

가장 낮은 Gini Index에 따라 만들어지는 결정 트리는 다음과 같습니다.
노드 1에서 Column b의 특성에 따라 노드 2,7로 나뉨을 볼 수 있습니다.
다음 분기는 Column b가 {≤3}인 경우에 대하여 분기 속성을 고르도록 하겠습니다.



〔첫번째 분기 속성을 바탕으로 분기된 결정트리〕

1871056 한지수

3번 문제 풀이

두 번째 분기 속성 고르기

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

Split Position	1		2		3	
Split Position	<=	>	<=	>	<=	>
Class 2	248	47	274	21	287	8
Class 4	2	171	7	166	29	144
GINI(t)	0.015	0.338	0.048	0.199	0.166	0.099
GINI	0.166		0.108		0.144	

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	229	58	265	22	283	4	286	1	286	1	287	0	287	0	287	0	287	0	287	0
Class 4	0	29	5	24	12	17	19	10	26	3	28	1	28	1	29	0	29	0	29	0
GINI(t)	0	0.444	0.036	0.499	0.078	0.308	0.116	0.165	0.152	0.375	0.161	0	0.161	0	0.166	0	0.166	0	0.166	0
GINI	0.122		0.103		0.093		0.118		0.155		0.161		0.161		0.166		0.166		0.166	

1871056 한지수

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	244	43	265	22	280	7	283	4	284	3	286	1	286	1	286	1	286	1	287	0
Class 4	11	18	14	15	17	12	23	6	25	4	26	3	26	3	27	2	27	2	29	0
GINI(t)	0.082	0.416	0.095	0.482	0.107	0.465	0.139	0.48	0.148	0.489	0.152	0.375	0.152	0.375	0.157	0.444	0.157	0.444	0.166	1
GINI	0.146		0.140		0.129		0.149		0.156		0.155		0.155		0.160		0.160		0.166	

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	259	28	269	18	276	11	280	7	282	4	285	2	285	2	285	2	285	2	287	0
Class 4	1	28	2	27	6	23	8	21	11	18	12	17	14	15	15	14	16	13	29	0
GINI(t)	0.007	0.5	0.014	0.48	0.041	0.437	0.054	0.375	0.072	0.297	0.077	0.188	0.089	0.207	0.095	0.218	0.100	0.231	0.166	1
GINI	0.094		**0.080**		0.084		0.082		0.087		0.084		0.095		0.101		0.106		0.166	

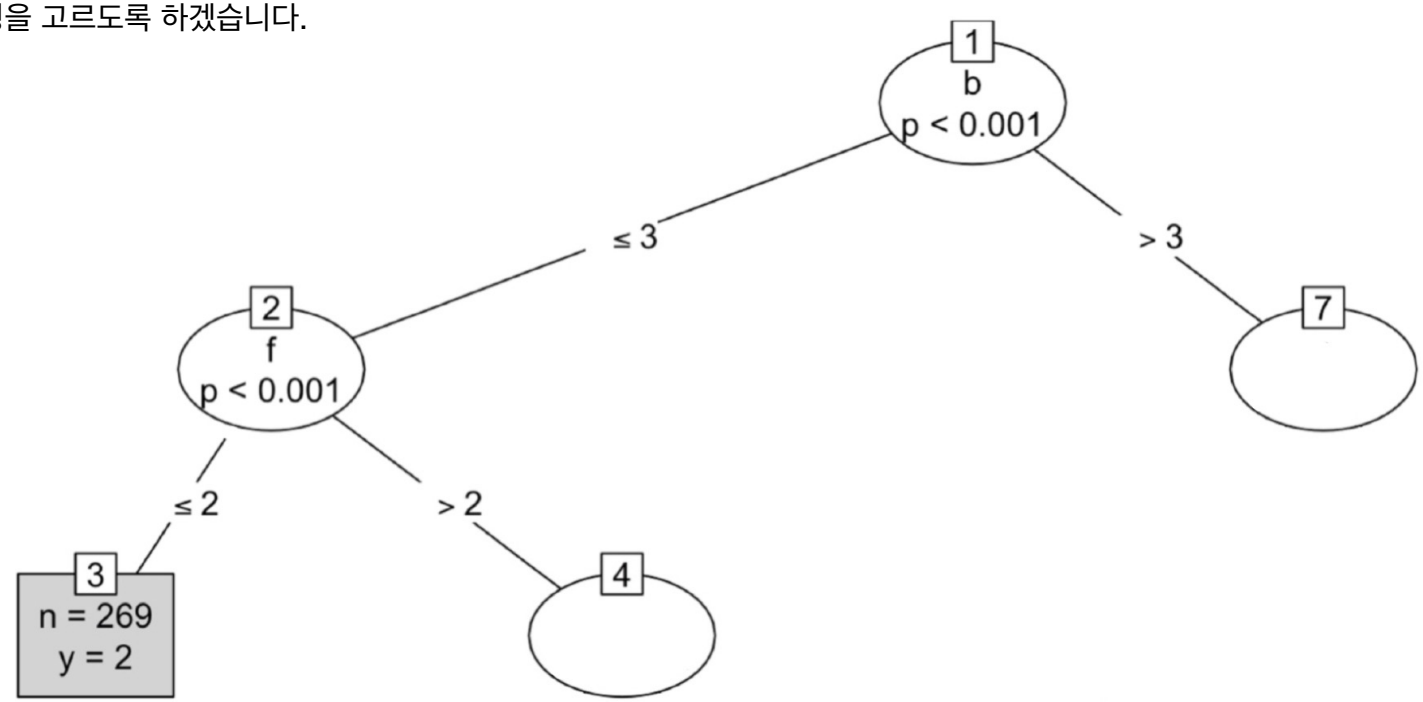
1871056 한지수

3번 문제 풀이

두 번째 분기 속성 고르기 결론

가장 낮은 Gini Index 0.080을 가지는 경우의 수는 다음과 같습니다.
- Column f의 ({≤2},{>2})

가장 낮은 Gini Index에 따라 만들어지는 결정 트리는 다음과 같습니다.
노드 2에서 Column f의 특성에 따라 단말 노드인 노드 3, 노드 4로 나뉨을 볼 수 있습니다.
Column b가 {≤3}이고, Column f가 {≤2}인 경우는 단말 노드로 결정되었음을 볼 수 있습니다.
다음 분기는 Column b가 {≤3}이고, Column f가 {>2} 인 경우에 대하여 분기 속성을 고르도록 하겠습니다.



〔두번째 분기 속성을 바탕으로 분기된 결정트리〕

1871056 한지수

3번 문제 풀이

세 번째 분기 속성 고르기

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

Split Position	1		2		3	
Split Position	<=	>	<=	>	<=	>
Class 2	12	6	14	4	18	0
Class 4	2	25	6	21	27	0
GINI(t)	0.244	0.312	0.42	0.268	0.48	1
GINI	0.291		0.336		0.48	

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	10	8	15	3	17	1	18	0	18	0	18	0	18	0	18	0	18	0	18	0
Class 4	0	27	3	24	12	15	17	10	24	3	26	1	26	1	27	0	27	0	27	0
GINI(t)	0	0.352	0.277	0.197	0.485	0.117	0.499	0	0.489	0	0.48347107	0	0.483	0	0.48	1	0.48	1	0.48	1
GINI	0.274		**0.229**		0.354		0.388		0.457		0.472727273		0.472		0.48		0.48		0.48	

1871056 한지수

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	12	6	13	5	14	4	15	3	16	2	17	1	17	1	17	1	17	1	18	0
Class 4	9	18	12	15	15	12	21	6	23	4	24	3	24	3	25	2	25	2	27	0
GINI(t)	0.489	0.375	0.499	0.375	0.499	0.375	0.486	0.444	0.483	0.444	0.485	0.375	0.485	0.375	0.481	0.444	0.481	0.444	0.48	1
GINI	0.428		0.444		0.455		0.477		0.478		0.475		0.475		0.479		0.479		0.48	

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

Split Position	3		4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2	7	11	11	7	16	2	16	2	16	2	16	2	16	2	18	0
Class 4	4	23	5	22	9	18	10	17	12	15	13	14	14	13	27	0
GINI(t)	0.462	0.437	0.429	0.366	0.460	0.18	0.473	0.188	0.489	0.207	0.494	0.21875	0.497	0.231	0.48	1
GINI	0.443		0.388		0.336		0.353		0.383		0.396		0.408		0.48	

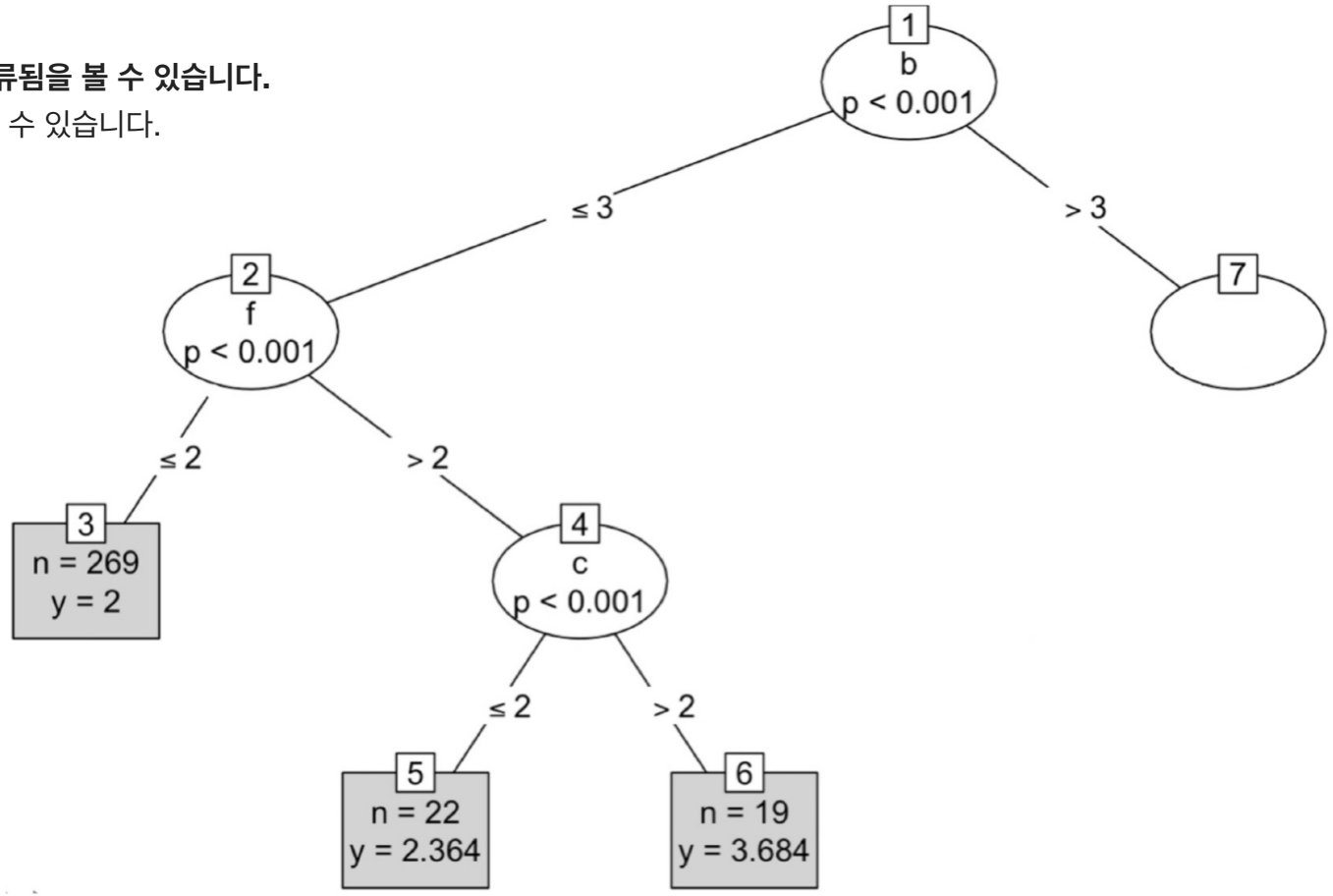
1871056 한지수

3번 문제 풀이

세 번째 분기 속성 고르기 결론

가장 낮은 Gini Index 0.229을 가지는 경우의 수는 다음과 같습니다.
- Column c의 ({≤2},{>2})

가장 낮은 Gini Index에 따라 만들어지는 결정 트리는 다음과 같습니다.
노드 4에서 Column c의 특성에 따라 둘로 나뉘며, 모두 단말 노드인 노드 5, 노드 6로 분류됨을 볼 수 있습니다.
Column b가 {≤3}인 경우에 대해서는 모두 단말 노드로 분류되며, 분류가 완료되었음을 알 수 있습니다.
다음 분기는 Column b가 {>3}인 경우에 대하여 분기 속성을 고르도록 하겠습니다.



〔세번째 분기 속성을 바탕으로 분기된 결정트리〕

1871056 한지수

3번 문제 풀이

네 번째 분기 속성 고르기

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

Split Position	4		5		6		7		8		9		10	
Split Position	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Class 2	5	3	5	3	5	3	6	2	7	1	8	0	8	0
Class 4	21	123	43	101	63	81	78	66	95	49	97	47	144	0
GINI(t)	0.310	0.046	0.186	0.056	0.136	0.068	0.132	0.057	0.127	0.039	0.140	0	0.099	1
GINI	**0.091**		0.097		0.099		0.098		0.098		0.097		0.099	

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	0	8	1	7	1	7	2	6	4	4	5	3	7	1	8	0	8	0	8	0
Class 4	0	144	2	142	13	131	29	115	47	97	69	75	87	57	104	40	108	36	144	0
GINI(t)	1	0.099	0.444	0.089	0.132	0.096	0.120	0.094	0.144	0.076	0.126	0.073	0.137	0.033	0.132	0	0.128	0	0.099	1
GINI	0.099		0.096		0.099		**0.099**		0.099		0.099		0.098		0.097		0.098		0.099	

1871056 한지수

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	4	4	4	4	6	2	7	1	8	0	8	0	8	0	8	0	8	0	8	0
Class 4	20	124	24	120	38	106	57	87	71	73	79	65	88	56	103	41	105	39	144	0
GINI(t)	0.277	0.060	0.244	0.062	0.235	0.036	0.194	0.022	0.182	0	0.166	0	0.152	0	0.133	0	0.131	0	0.099	1
GINI	0.094		0.096		0.094		0.095		0.094		0.095		0.096		0.097		0.097		0.099	

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	2	6	2	6	3	5	4	4	6	2	6	2	6	2	8	0	8	0	8	0
Class 4	13	131	20	124	29	115	35	109	47	97	50	94	54	90	63	81	68	86	144	0
GINI(t)	0.231	0.083	0.165	0.088	0.169	0.079	0.184	0.068	0.200	0.039	0.191	0.040	0.18	0.042	0.199	0	0.188	0	0.099	1
GINI	0.098		0.099		0.098		0.098		0.095		0.096		0.096		0.093		0.099		0.099	

1871056 한지수

3번 문제 풀이

네 번째 분기 속성 고르기 결론 및 최종 결론

가장 낮은 Gini Index 0.091을 가지는 경우의 수는 다음과 같습니다.
- Column b의 ({≤1},{>1})

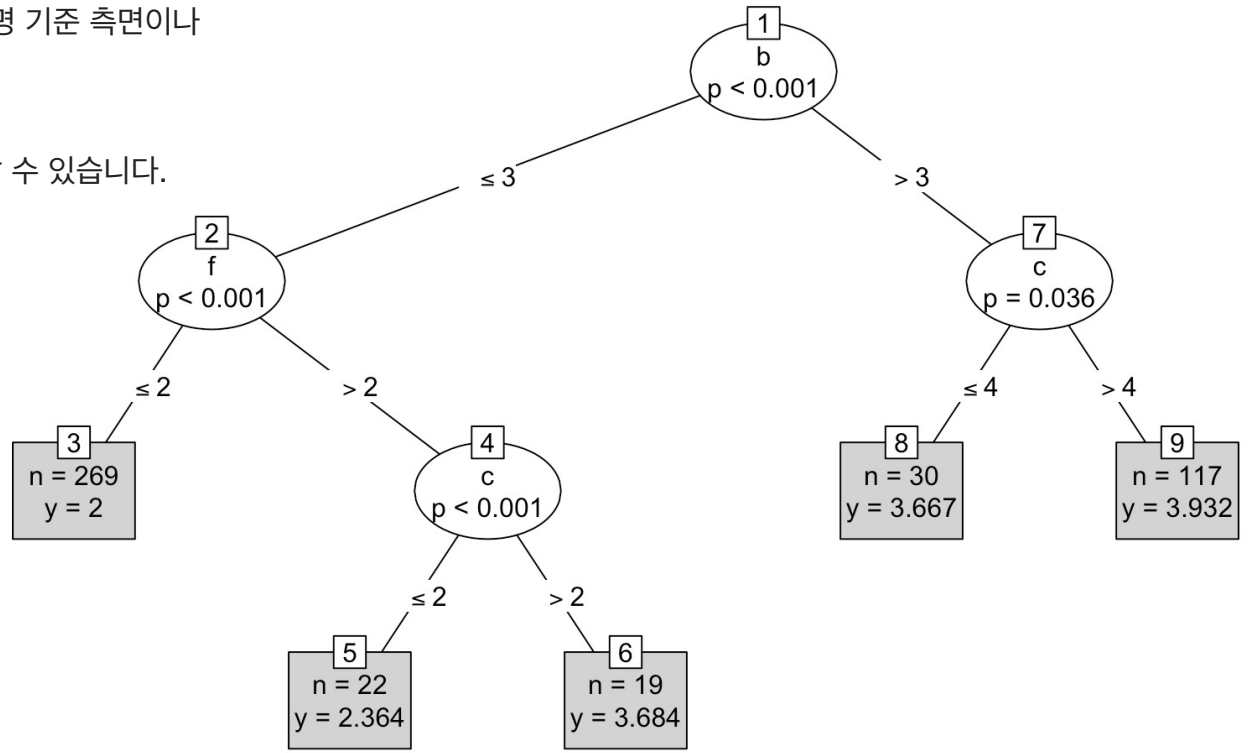
가장 낮은 Gini Index에 따라 만들어지는 ctree 결정 트리는 다음과 같습니다.

Column c의 ({≤4},{>4})가 선택된 이유는 간단합니다. 불순도 척도를 나타내는 다양한 방법 중 ctree는 p-test를 거친 significance를 기준으로 분류됩니다. Gini index에 따라 분류를 시도하여 비교해본다면 분명 기준 측면이나 여러가지 고려사항이 달라 다르게 결과가 나올 수 있다고 해석하였습니다.

노드 7에서 Column c의 특성에 따라 모두 단말 노드인 노드 8, 노드 9로 나뉨을 볼 수 있습니다.

Column b가 {>3}인 경우에 대해서도 끝까지 모두 단말 노드로 분류되며, 분류가 완료되었음을 알 수 있습니다.

이에 따라 모든 instance가 단말 노드로 분류되며, 결정 트리 분기가 완료되었습니다.



〔최종적으로 생성된 bio_ctree 결정트리〕

1871056 한지수