의사결정 나무에서 연속형 변수 경계값 나누는 방법

[표 6.11] 데이터

x	100	120	160	180	186	190	210	250	270	300
y	1	1	1	-1	-1	-1	-1	1	1	1

[풀이]

x	100	120	160	180	186	190	210	250	270	300
y	1	1	1	-1	-1	-1	-1	1	1	1
중간값	110	140	170	183	188	200	230	260	285	
	≤ >	≤ >	≤ >	≤ >	≤ >	≤ >	≤ >	≤ >	≤ >	
1분류	1 5	2 4	3 3	3 3	3 3	3 3	3 3	4 2	5 1	
-1분류	0 4	0 4	0 4	1 3	2 2	3 1	4 0	4 0	4 0	
엔트로피	0.892	0.8	0.690	0.925	0.971	0.925	0.690	0.8	0.892	

(예) 중간값이 110인 경우:

$$\text{엔트로피} = \frac{1}{10}\times\{-1\times\log_2(1)-0\times\log_2(0)\}+\frac{9}{10}\times\left\{-\frac{5}{9}log_2(\frac{5}{9})-\frac{4}{9}log_2(\frac{4}{9})\right\}=0.892$$

위 데이터에서 엔트로피가 최소가 되는 가지분할은 c=**170 또는 230**이다. 두 경우 분류기의 정확도는 각각 70%, 30%이다.

배깅 방법을 적용하기 위하여 10개의 붓스트랩 표본을 추출한 결과와 엔트로피가 최소가 되는 각각의 분류기가 [표 6.12]와 같다.

$$\text{엔트로피계수} = -\sum_{i=1}^{K} p_i \log_2 p_i$$

ex) 집단 2개 $\text{엔트로피계수} = -p\log_2 p-(1-p)\log_2(1-p)$

$$I(A)=\frac{n_{1\cdot}}{n_{\cdot\cdot}}I(A_1)+\frac{n_{2\cdot}}{n_{\cdot\cdot}}I(A_2)+\cdots+\frac{n_{a\cdot}}{n_{\cdot\cdot}}I(A_a)$$

p_i

2. 한 백화점의 어느 상품매장을 방문하는 사람 10 명에 대해 상품 구매여부와 나이를 조사해보니 다음과 같다. 구매하는 사람을 Y 집단, 구매하지 않는 사람을 N 집단으 로 표시하고 나이를 오름차순으로 정렬하였다. 의사결정나무 모형을 적용하기 위해 나이를 두 개의 집단으로 나누고자 한다. 어떠한 경계값으로 나누어야 좋은가? (지니계수 이용하여 문제 풀기)

지니계수 $= 1 - \sum_{i=1}^{K} p_i^2$

구매여부	N	N	N	Y	N	Y	N	Y	Y	Y
나이	25	27	31	33	35	41	43	49	51	55

26 29 32 34 38 42 46 50 53

< >

Y	0 5	0 2	0 3	1 3	1 4	2 4	2 5	3 5	4 5
N	1 4	5 3	5 2	4 2	4 1	3 1	3 0	2 0	1 0

$$\frac{1}{10}\left(1-\left(\frac{0}{1}^2+\frac{1}{1}^2\right)\right)+\frac{9}{10}\left(1-\left(\left(\frac{5}{9}\right)^2+\left(\frac{4}{9}\right)^2\right)\right)$$

$$=\frac{9}{10}\left(1-\frac{41}{81}\right)=\frac{9}{10}\cdot\frac{40}{81}=\frac{4}{9}=0.44$$