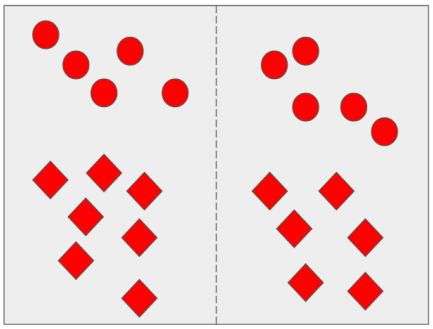
의사결정나무(지니지수, 엔트로피 등)

가. 다음의 그림을 보고 문제 1~2 번에 답하시오



indicator1

문제 1. 위 그림에서 indicator1로 데이터를 분류했을 때 왼쪽노드(Left) 노드의 지니계수를 구하는 다음의 계산식을 완성하시오.

$$1-(\frac{(\ \ )}{12})^2-(\frac{5}{(\ \ )})^2$$

문제 2. 위 그림에서 indicator1로 데이터를 분류했을 때 오른쪽노드(Right) 노드의 지니계수를 구하는 다음의 계산식을 완성하시오.

$$1-(\frac{(\ \ )}{(\ \ )})^2-(\frac{5}{(\ \ )})^2$$

나. 다음 아래의 도표는 "Good"과 "Bad" 두 상태를 가지는 데이터가 의사결정나무 모델에 의해 Left와 Right 노드로 분리된 상태를 정리한 것이다. 문제 3~5번 문제에 답하시오.

	Good	Bad	Total
Left	5	15	20
Right	10	5	15
Total	15	20	35

문제 3. Left 노드의 지니지수를 구하는 다음의 계산식을 완성하시오.

$$1-(\frac{(\ \ )}{20})^2-(\frac{15}{20})^2$$

문제 4. Right 노드의 지니지수를 구하는 다음의 계산식을 완성하시오.

$$1-(\frac{(\ \ )}{(\ \ )})^2-(\frac{(\ \ )}{(\ \ )})^2$$

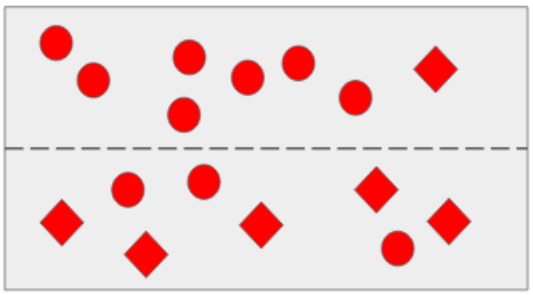
문제 5. Left 노드와 Right 노드 전체의 지니지수를 구하는 다음의 계산식을 완성하시오.

$$(1-(\frac{(\ \ )}{20})^2-(\frac{15}{20})^2)\times\frac{(\ \ )}{35}+(1-(\frac{(\ \ )}{(\ \ )})^2-(\frac{(\ \ )}{(\ \ )})^2)\times\frac{(\ \ )}{35}$$

문제 6. Left 노드와 Right 노드 전체의 지니지수를 구하는 다음의 계산식을 완성하시오.

$$2\times(\frac{5}{20}\times\frac{15}{20}\times\frac{(\ \ )}{35}+\frac{10}{15}\times\frac{5}{15}\times\frac{(\ \ )}{35})$$

다. 다음의 그림을 보고문제 6~9에 답하시오.



문제 6. 전체 테두리 안의 데이터에 대한 엔트로피를 구하는 다음의 수식을 완성하시오.

$$-\frac{(\quad)}{16}\log_2\frac{(\quad)}{16}-\frac{(\quad)}{16}\log_2\frac{(\quad)}{16}$$

문제 7. 데이터를 위와 같이 점선으로 분류했을 경우, 위쪽 부분의 엔트로피를 구하는 수식을 작성하시오.

$$-\frac{7}{8}\log_2\frac{(\quad)}{(\quad)}-\frac{(\quad)}{(\quad)}\log_2\frac{(\quad)}{(\quad)}$$

문제 8. 데이터를 위와 같이 빨간 점선으로 분류했을 경우, 아래쪽 부분의 엔트로피를 구하는 다음의 수식을 완성하시오.

$$-\frac{(\quad)}{(\quad)}\log_2\frac{(\quad)}{(\quad)}-\frac{(\quad)}{(\quad)}\log_2\frac{(\quad)}{(\quad)}$$

문제 9. 데이터를 위와 같이 빨간 점선으로 분류했을 경우, 전체 엔트로피를 구하는 수식을 작성하시오.(계산식만 작성하시오.)

$$0.5\times(-\frac{(\quad)}{(\quad)}\log_2\frac{(\quad)}{(\quad)}-\frac{(\quad)}{(\quad)}\log_2\frac{(\quad)}{(\quad)})+0.5\times(-\frac{(\quad)}{(\quad)}\log_2\frac{(\quad)}{(\quad)}-\frac{(\quad)}{(\quad)}\log_2\frac{(\quad)}{(\quad)})$$

라. 다음 도표는 온도, 습도, 바람에 따라 테니스 강습을 했는지 안 했는지에 대한 데이터이다. 이 도표를 보고 문제 10번에 답하시오.

Temperature	Humidity	Windy	Class
Hot	High	False	N
Hot	High	True	N
Hot	High	False	P
Mild	High	False	P
Cold	Normal	False	P
Cold	Normal	True	N
Cold	Normal	True	P
Mild	High	False	N
Cold	Normal	False	N
Mild	Normal	False	P
Mild	Normal	True	P
Mild	High	True	P
Hot	Normal	False	N
Mild	High	True	P

문제 10. Windy를 기준으로 Class 데이터를 분리하면 전체 지니지수가 얼마인가?(계산식만 작성하시오.)

문제 11. 의사결정나무의 알고리즘으로, 불순도의 측도로 엔트로피 지수를 사용하며 각 마디에서 다지분리(multiful split)가 가능하며 범주형 입력변수에 대해서는 범주의 수만큼 분리가 일어나는 알고리즘은?

① SVM

② C5.0

③ CART

④ CHIAD

문제12. 의사결정나무의 알고리즘으로, 불순도의 측도로 카이제곱 통계량을 사용하며 가지치기를 하지 않고 적당한 크기에서 나무모형의 성장을 중지하는 알고리즘은?

① ID3

② C5.0

③ CART

④ CHIAD

문제 13. 의사결정나무의 알고리즘으로, 불순도의 측도로 지니지수를 사용하며 이진분리(binary split)를 수행하는 알고리즘은?

① ID3

② C5.0

③ CART

④ CHIAD

문제 14. 다음 중 의사결정나무에 대한 활용사례로 부적절한 것은?

① 시장세분화

② 고객속성에 따른 대출한도액 예측

③ 상품추천을 통한 교차판매

④ 신용도에 따른 고객 분류

문제 15. 다음 중 의사결정나무 모형의 특징으로 가장 부적절한 것은?

① 비모수적 방법이다.

② 설명이 용이하다.

③ 잡음데이터에 민감하다.

④ 계산이 단순하고 빠르다.

문제 16. 다음 중 의사결정나무의 활용분야로 가장 적절한 것은?

① 텍스트 분석

② 교호작용의 파악

③ 교차판매 예측

④ 장바구니 분석

답안

1. 7, 12
2. 6, 11, 11
3. 5
4. 5, 15, 10, 15
5. 5, 20, 5, 15, 10, 15, 15
6. 20, 15
   6.(오타) 10, 10, 6, 6
7. 7, 8, 1, 8, 1, 8
8. 3, 8, 3, 8, 5, 8, 5, 8
9. 7, 8, 7, 8, 1, 8, 1, 8, 3, 8, 3, 8, 5, 8, 5, 8
10. 생략
11. 2
12. 4
13. 3
14. 3
15. 3
16. 2