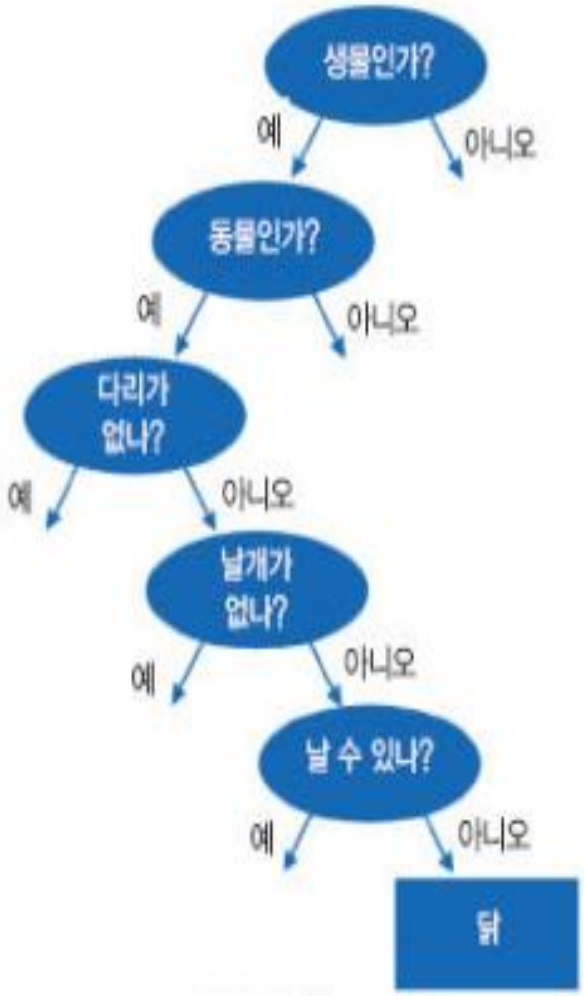
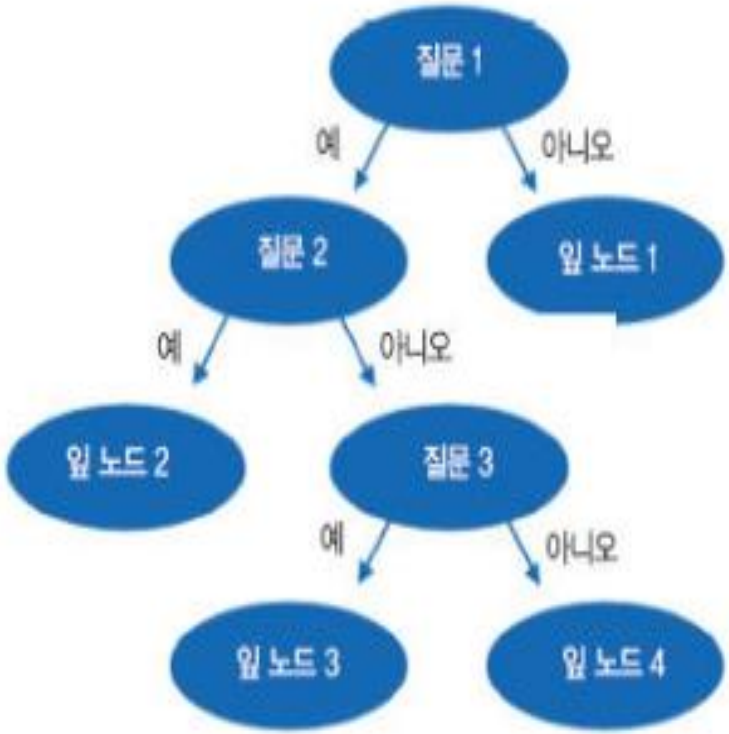
의사결정트리 알고리즘

의사결정트리 알고리즘은
데이터 사이에 존재하는 패턴을
예측 가능한 규칙들의 조합으로
나타내는 알고리즘입니다

예를 들면
질문을 던져서 대상을 좁혀나가는
스무고개 놀이와 비슷한 개념입니다


생물인가?
예
아니오
동물인가?
예
아니오
다리가 없나?
예
아니오
날개가 없나?
예
아니오
날 수 있나?
예
아니오
닭

(a) 스무고개


질문 1
예
아니오
질문 2
잎 노드 1
예
아니오
잎 노드 2
질문 3
예
아니오
잎 노드 3
잎 노드 4

(b) 결정트리

그런데 이때 질문을 할때
가장 중요한 질문들을
처음에 해야
빨리 정답을 맞출수 있습니다

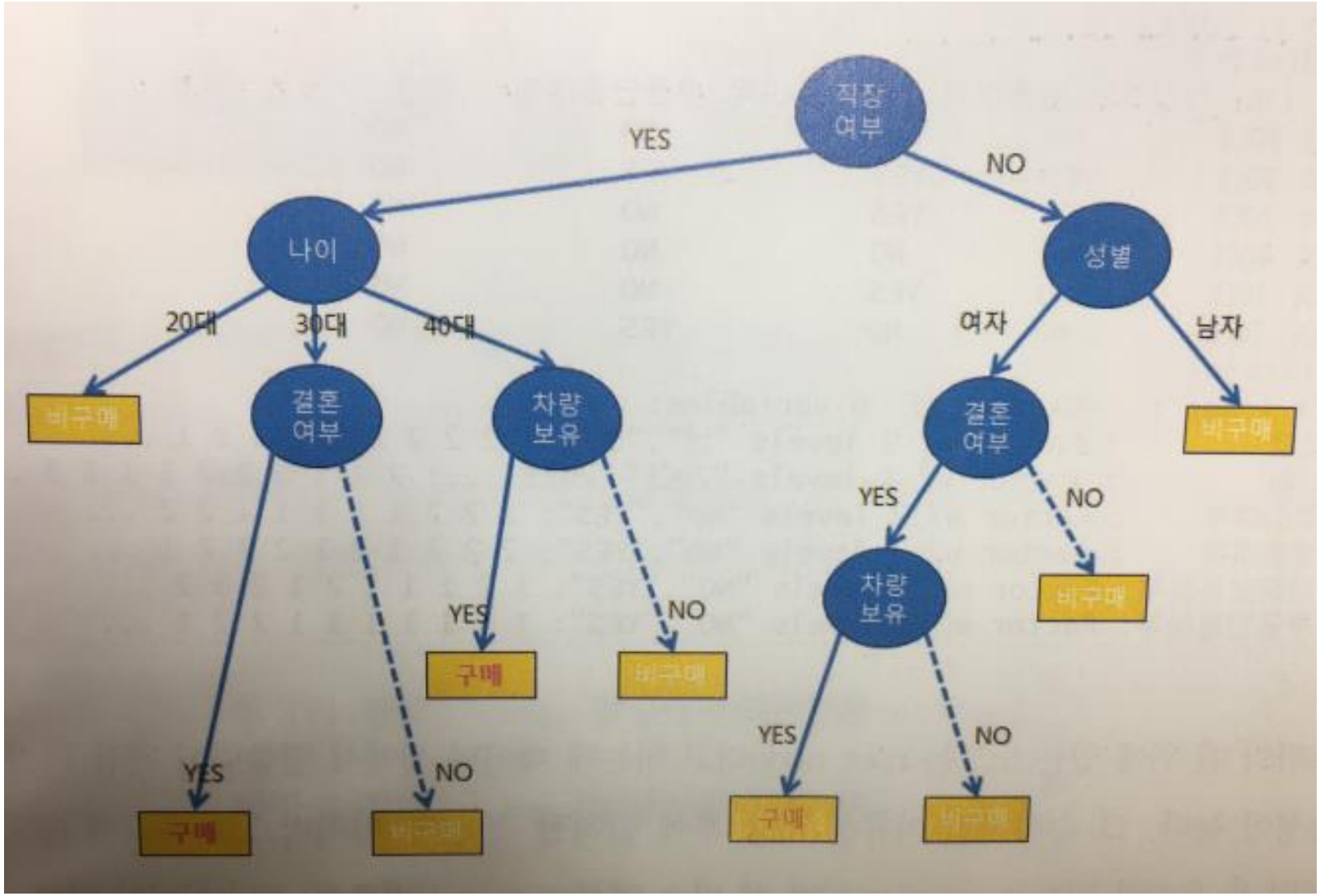
어떤 질문이 중헌디?

어떤 컬럼이 중헌디 ?

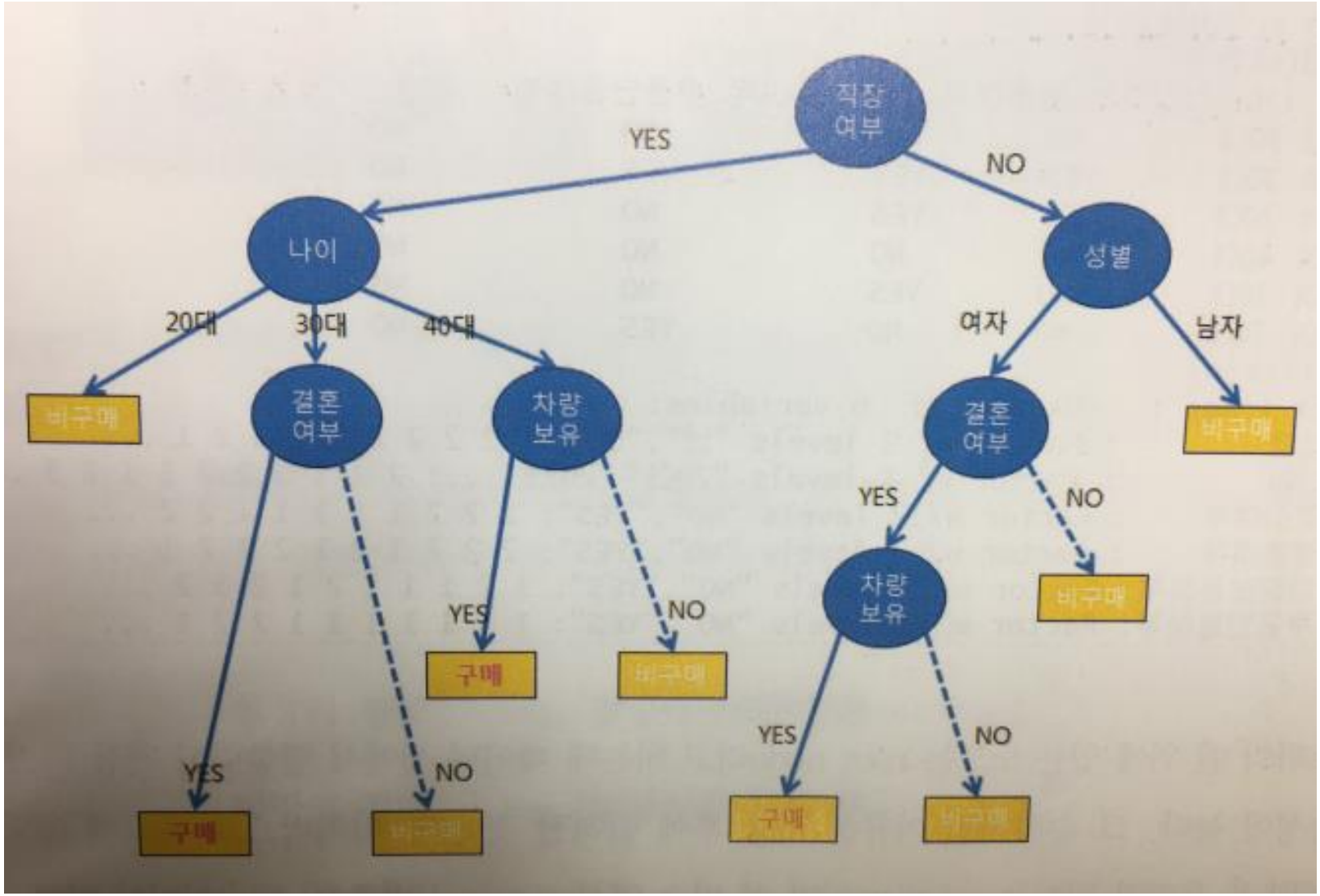
라벨

고객번호	성별	나이	직업	결혼여부	차소유	구매여부
1	남	30대	NO	YES	NO	NO
2	여	20대	YES	YES	YES	NO
3	여	20대	YES	YES	NO	NO
4	여	40대	NO	NO	NO	NO
5	여	30대	NO	YES	NO	NO
6	여	30대	NO	NO	YES	NO
7	여	20대	NO	YES	NO	NO
8	여	20대	NO	YES	YES	YES
9	여	30대	YES	YES	NO	YES
10	남	40대	YES	NO	YES	NO
11	남	20대	NO	NO	NO	NO
12	남	30대	NO	YES	YES	NO
13	남	20대	YES	NO	NO	NO
14	여	30대	YES	YES	NO	YES
15	남	30대	YES	YES	YES	YES

직장이 있는겨 ?


직장 여부
YES
NO
나이
성별
20대
30대
40대
여자
남자
비구매
결혼 여부
차량 보유
결혼 여부
비구매
YES
NO
YES
NO
구매
비구매
차량 보유
비구매
YES
NO
YES
NO
구매
구매
비구매

직장이 없다고 그럼 남자여? 여자여?



화장품 구매에 가장 중요한 컬럼이 뭔가?

라벨

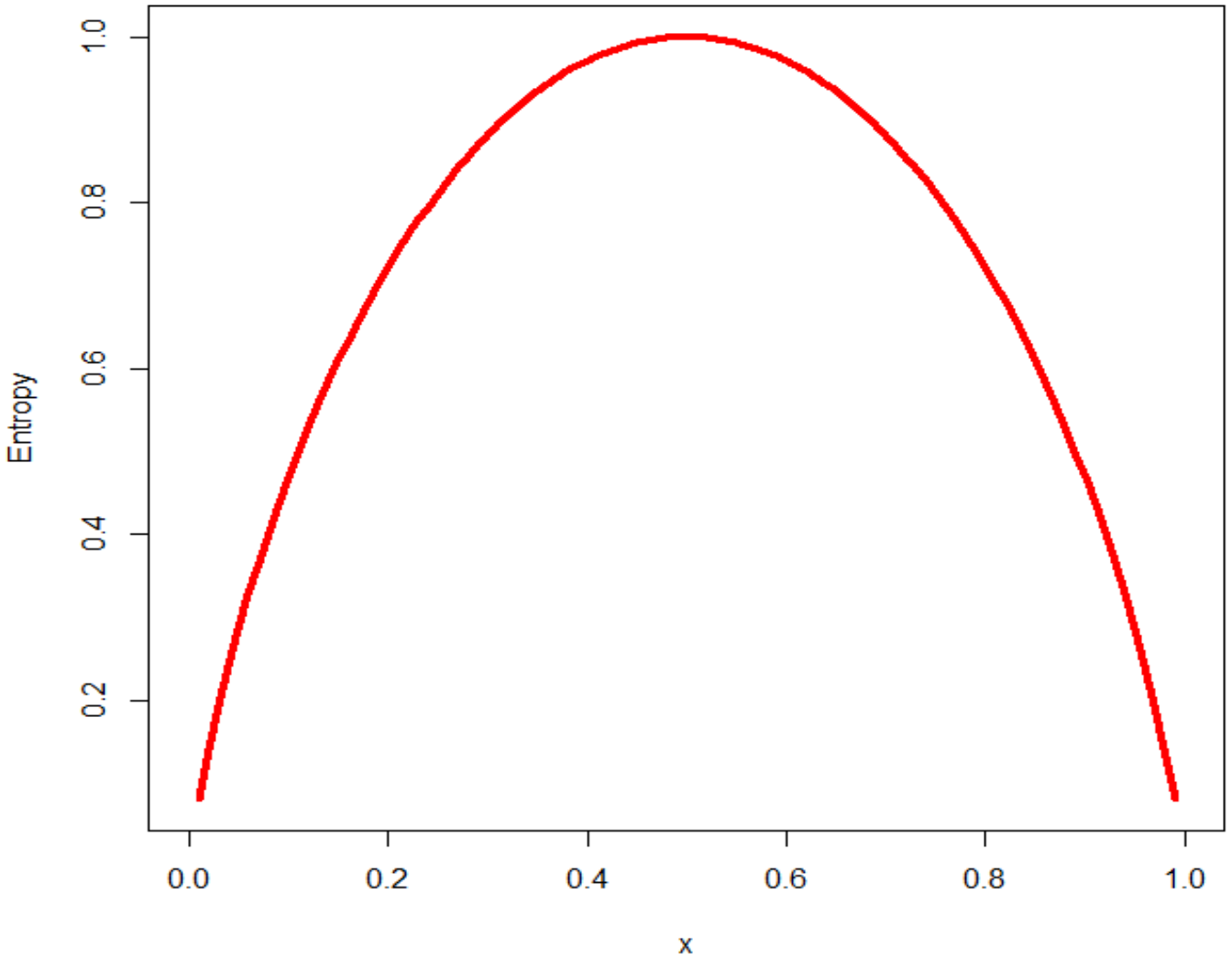
고객번호	성별	나이	직업	결혼여부	차소유	구매여부
1	남	30대	NO	YES	NO	NO
2	여	20대	YES	YES	YES	NO
3	여	20대	YES	YES	NO	NO
4	여	40대	NO	NO	NO	NO
5	여	30대	NO	YES	NO	NO
6	여	30대	NO	NO	YES	NO
7	여	20대	NO	YES	NO	NO
8	여	20대	NO	YES	YES	YES
9	여	30대	YES	YES	NO	YES
10	남	40대	YES	NO	YES	NO
11	남	20대	NO	NO	NO	NO
12	남	30대	NO	YES	YES	NO
13	남	20대	YES	NO	NO	NO
14	여	30대	YES	YES	NO	YES
15	남	30대	YES	YES	YES	YES

직업이 중요한가 결혼여부가 중요한가?

그 질문의 우선순위를 정할때 필요한게
엔트로피(entropy) 입니다

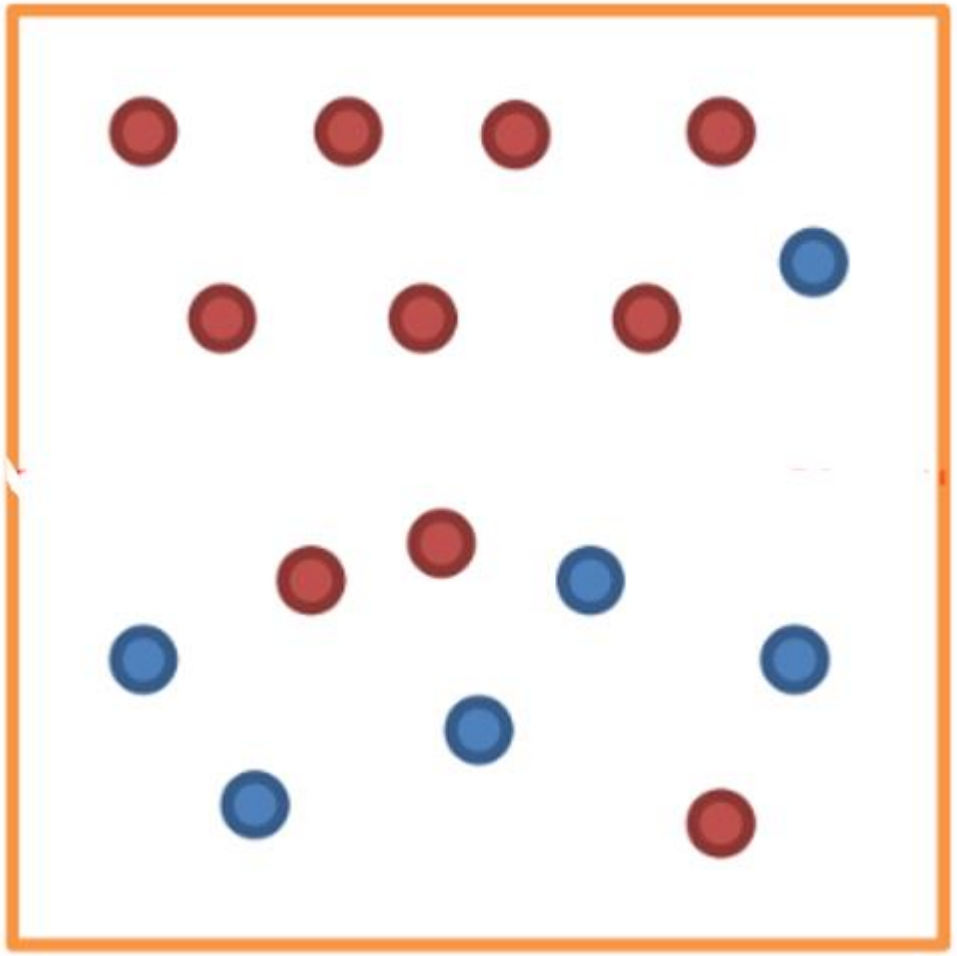
엔트로피는 불확실성의 정도를 말합니다

x 축이 확률이고 y 축이 엔트로피입니다



아래의 박스의 엔트로피를 구해보면 ?

(불순도)

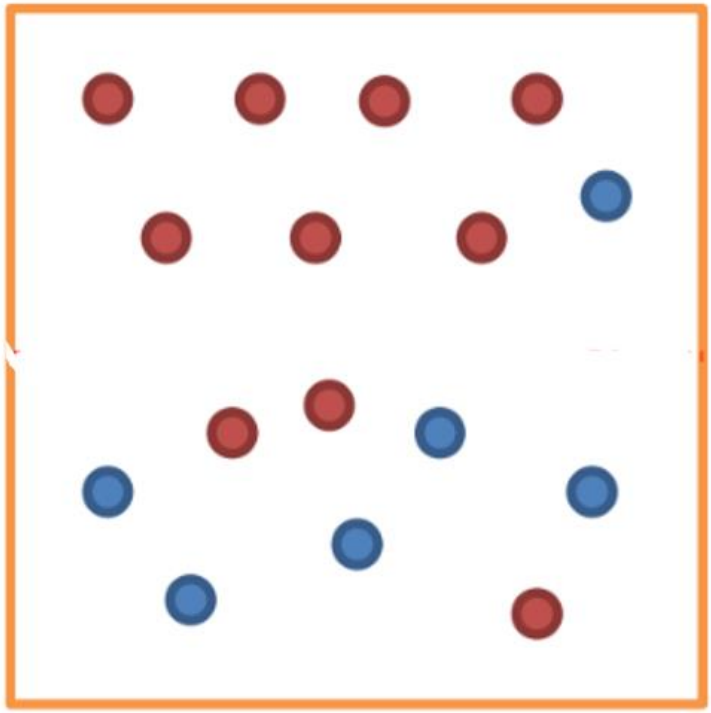


엔트로피 공식에 대입해서

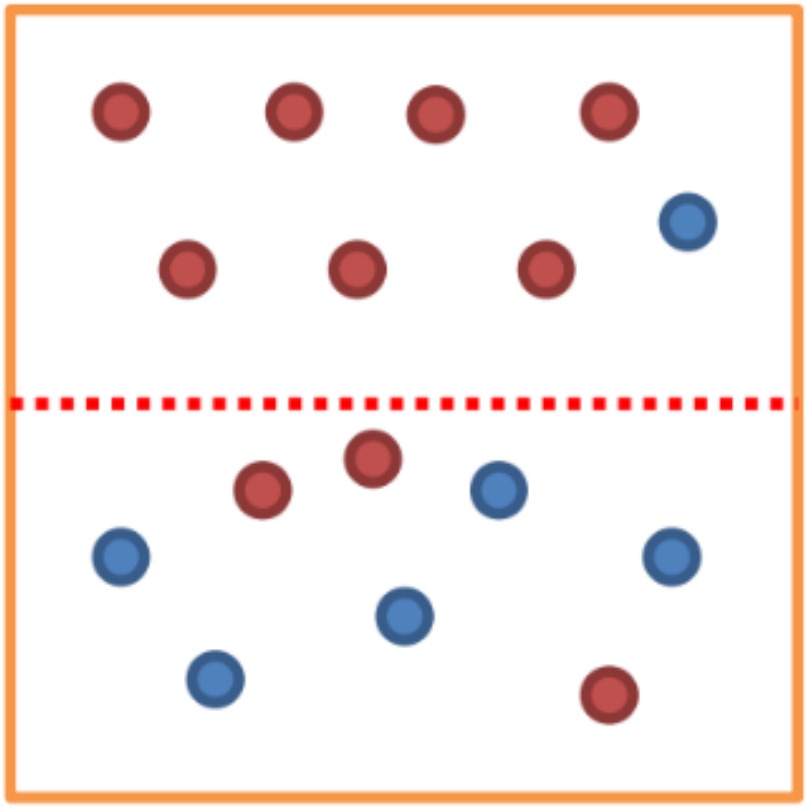
$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2 (p_k)$$

전체 16개 가운데 빨간색 동그라미(범주(k)=1)는 10개, 파란색(범주(k)=2)은 6개이군요. 그럼 A 영역의 엔트로피는 다음과 같습니다.

$$Entropy(A) = -\frac{10}{16}\log_2\left(\frac{10}{16}\right) - \frac{6}{16}\log_2\left(\frac{6}{16}\right) \approx 0.95$$

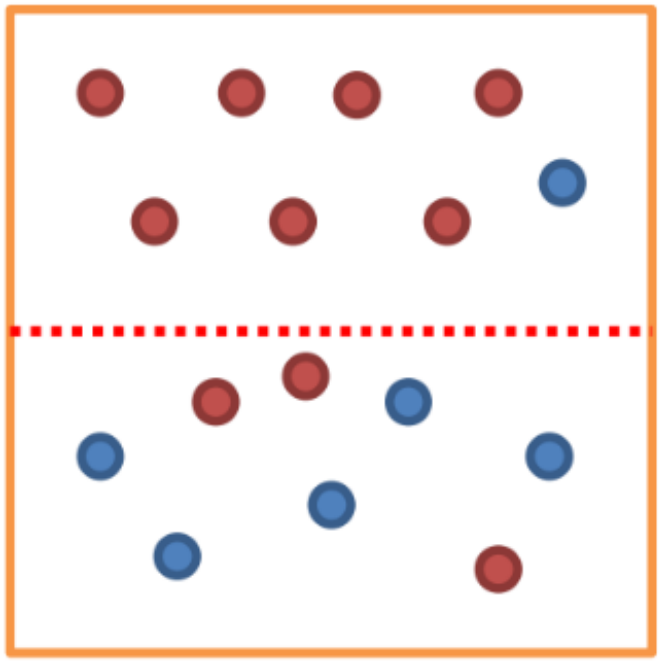


이번에는 아래와 같이 분할한 상태에서
의 불순도를 각각 구해보겠습니다



$$Entropy(A) = \sum_{i=1}^{d} R_i \left(- \sum_{k=1}^{m} p_k \log_2 (p_k) \right)$$

$$Entropy(A) = 0.5 \times \left(-\frac{7}{8} \log_2 (\frac{7}{8}) - \frac{1}{8} \log_2 (\frac{1}{8}) \right) + 0.5 \times \left(-\frac{3}{8} \log_2 (\frac{3}{8}) - \frac{5}{8} \log_2 (\frac{5}{8}) \right) \approx 0.75$$



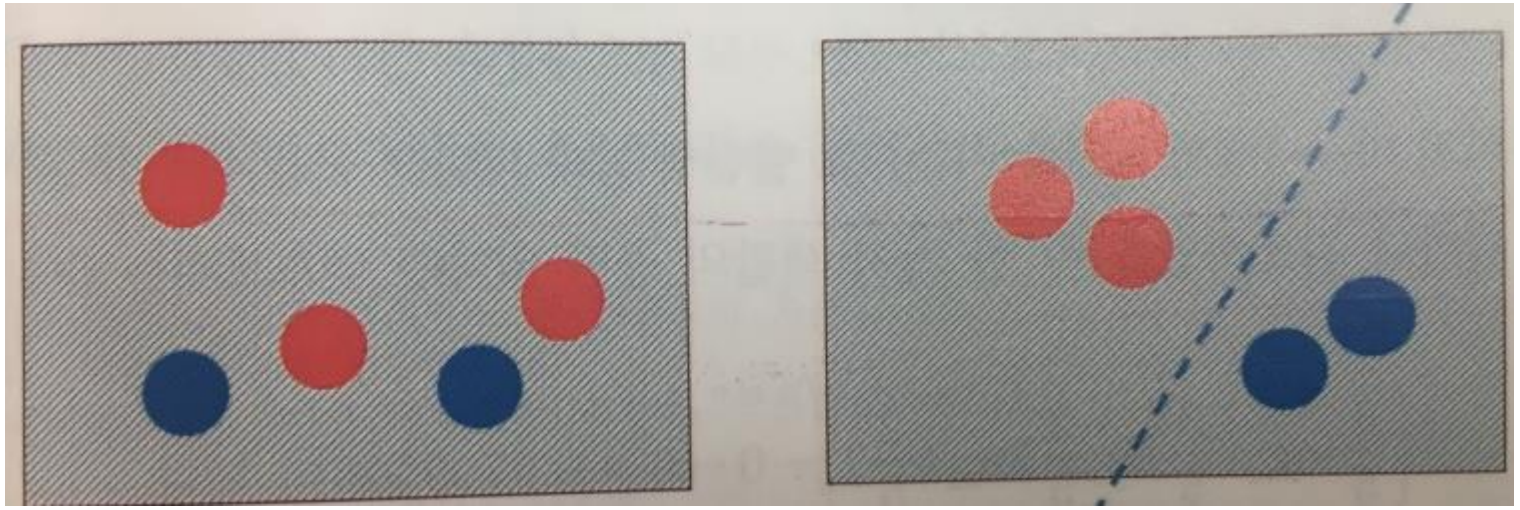
그래서 정보 획득량은 ?

분할전 엔트로피 - 분할후 엔트로피

이므로

0.95 − 0.75 = 0.2 입니다

문제1. 아래 그림을 보고 분할전과 분할후의 엔트로피를 각각 구해서 정보획득량을 출력하세요



문제2. 아래의 각각의 컬럼들의 정보획득량을 출력하세요

고객번호	성별	나이	직업	결혼여부	차소유	구매여부
1	남	30대	NO	YES	NO	NO
2	여	20대	YES	YES	YES	NO
3	여	20대	YES	YES	NO	NO
4	여	40대	NO	NO	NO	NO
5	여	30대	NO	YES	NO	NO
6	여	30대	NO	NO	YES	NO
7	여	20대	NO	YES	NO	NO
8	여	20대	NO	YES	YES	YES
9	여	30대	YES	YES	NO	YES
10	남	40대	YES	NO	YES	NO
11	남	20대	NO	NO	NO	NO
12	남	30대	NO	YES	YES	NO
13	남	20대	YES	NO	NO	NO
14	여	30대	YES	YES	NO	YES
15	남	30대	YES	YES	YES	YES

참고자료

1. 어서와 머신러닝은 처음이지

양지헌 선생님 지음

2. rstsgo’s blog

이기창 선생님

cafe.daum.net/oracleoracle

사랑하는 자여 네 영혼이 잘됨같이 네가 범사에 잘되고 강건하기를 내가 간구하노라

– 성경 요한삼서 1장 2절