기계학습 개론

- Qualitative Classification (질적 분류)

정보통신공학과

Prof. Jinkyu Kang

MYONGJI UNIVERSITY

이번주 수업의 목차

- 결정트리
  - 원리
  - 노드에서의 질문
  - 학습 알고리즘
  - 특성

MYONGJI
UNIVERSITY

들어가는 말

- 세상에는 참으로 많은 데이터가 있다.
  - 계량 데이터
    - 점수, 매출액, GDP, 속도, 마찰계수, 토끼 개체수 등
    - 거리 개념 있다. 5는 31보다 크다. 5는 10보다 7에 가깝다.
  - 비계량 데이터
    - 직업, 행정 구역, 혈액형, 성씨, PC 브랜드 등
    - 거리 개념 없다. ‘O형은 B형보다 A형에 가깝다’는 성립 안한다.

MYONGJI
UNIVERSITY

들어가는 말

- 비계량 데이터의 분류를 다룸
  - 질적 분류기
    - 결정 트리
    - 스트링 인식기

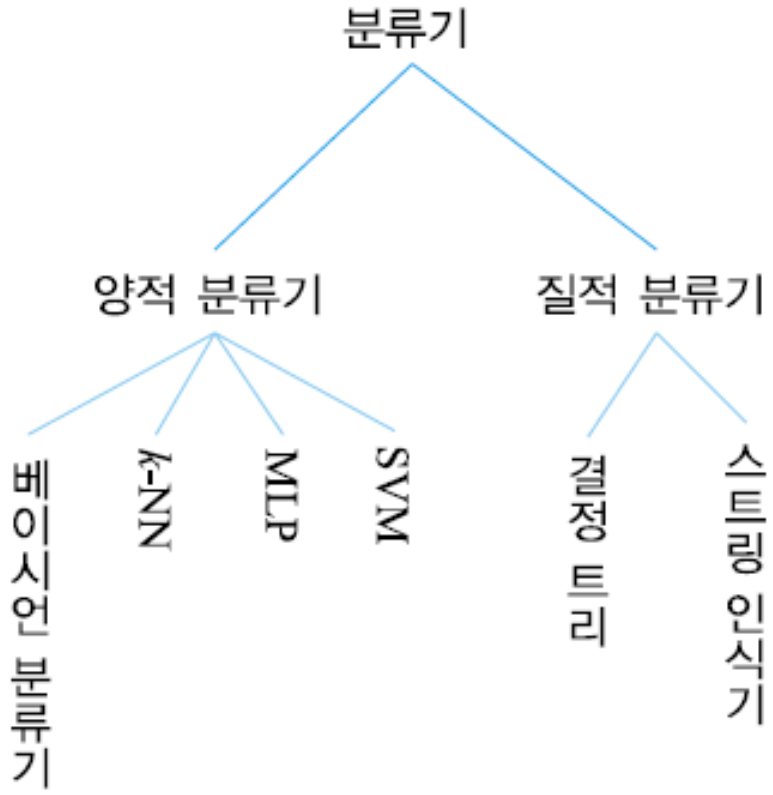


그림 6.1 양적 분류기와 질적 분류기

결정트리 I 원리

- 결정 트리의 원리
  - 스무고개와 개념이 비슷
  - 최적 기준에 따라 자동으로 질문을 만들어야 함

- 몇 가지 고려 사항

1. 노드에서 몇 개의 가지로 나눌 것인가?
2. 각 노드의 질문을 어떻게 만들 것인가?
3. 언제 멈출 것인가?
4. 잎 노드를 어느 부류에 할당할 것인가?

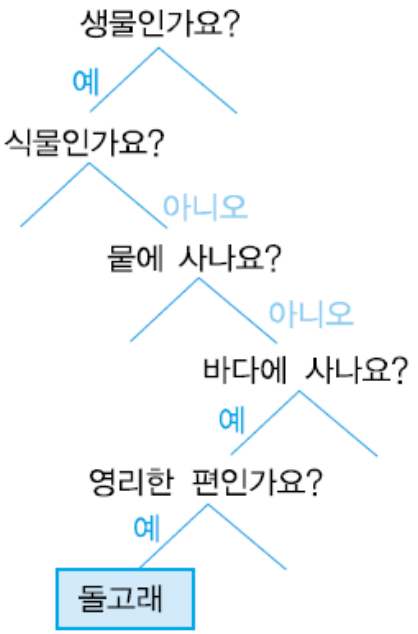
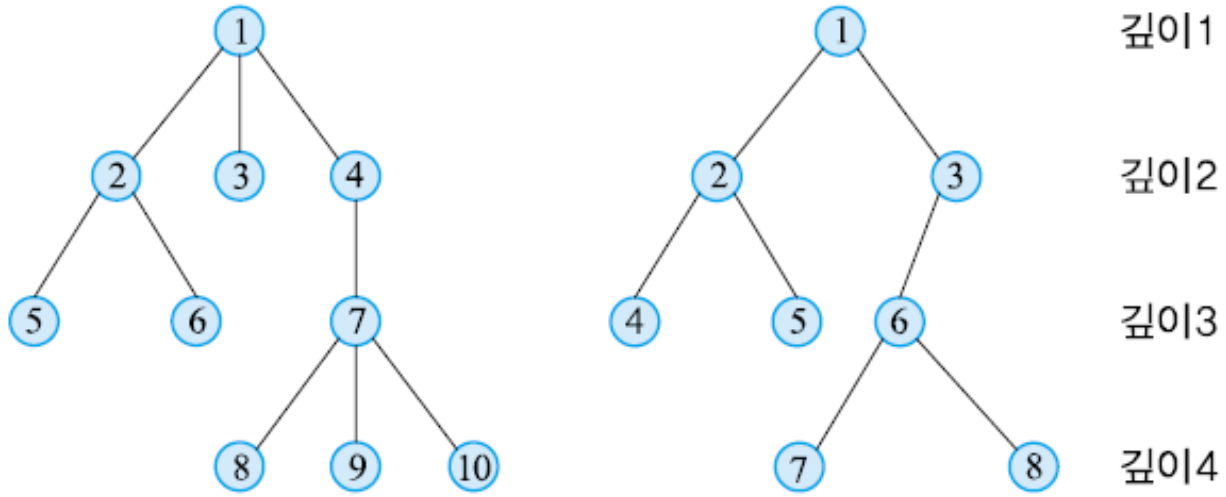


그림 6.2 스무고개 놀이

결정트리 I 원리

- 결정 트리의 표현
  - 트리 또는 이진 트리 사용



(a) 트리 (b) 이진 트리

그림 6.3 트리와 이진 트리

결정트리 | 노드에서의 질문

- 결정 트리의 노드
  - 노드의 분기

$$\left.\begin{aligned} X_{Tleft} \cup X_{Tright} &= X_T \\ X_{Tleft} \cap X_{Tright} &= \varnothing \end{aligned}\right\} \quad (6.1)$$

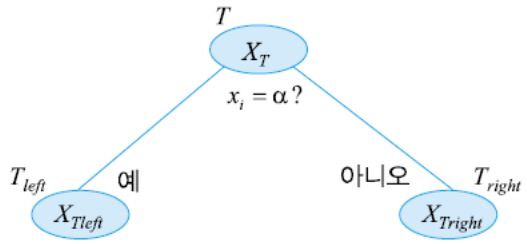


그림 6.5 노드의 분기

  - 질문 x_i=α? 어떻게 만들 것인가?
    - d 개의 특징이 있고 그들이 평균 n 개의 값을 가진다면 dn 개의 후보 질문
    - 그들 중 어느 것을 취해야 가장 유리한가?

MYONGJI UNIVERSITY

결정트리 | 노드에서의 질문

- 유리한 정도의 판단 기준은?
  - X_{Tleft}와 X_{Tright}가 동질일 수록 좋다.

- 불순도 측정 기준
  - 엔트로피 $$im(T) = -\sum_{i=1}^{M} P(\omega_i \mid T) \log_2 P(\omega_i \mid T) \quad (6.2)$$
  - 지니 불순도 $$im(T) = 1 - \sum_{i=1}^{M} P(\omega_i \mid T)^2 = \sum_{i \neq j} P(\omega_i \mid T) P(\omega_j \mid T) \quad (6.3)$$
  - 오분류 불순도 $$im(T) = 1 - \max_i P(\omega_i \mid T) \quad (6.4)$$
  - 노드 T에서 ω_i가 발생할 확률은

$$P(\omega_i \mid T) = \frac{X_T\text{에서 } \omega_i\text{에 속한 샘플의 수}}{|X_T|} \quad (6.5)$$

MYONGJI UNIVERSITY

결정트리 | 노드에서의 질문

- 예제 1) 불순도 측정

샘플	특징 벡터			부류
	x_1 (직업)	x_2 (선호 품목)	x_3 (몸무게)	
1	3	1	50.6	ω_2
2	2	3	72.8	ω_1
3	3	5	88.7	ω_3
4	2	2	102.2	ω_2
5	5	5	92.3	ω_2
6	3	4	65.3	ω_2
7	2	3	67.8	ω_1
8	7	1	47.8	ω_3
9	2	3	45.6	ω_1

노드 T의 샘플 집합 X_T가 아래와 같다고 하자.

$$X_T = \{(\mathbf{x}_1,\omega_2),\ (\mathbf{x}_2,\omega_1),\ (\mathbf{x}_3,\omega_3),\ (\mathbf{x}_4,\omega_2),\ (\mathbf{x}_5,\omega_2),\ (\mathbf{x}_6,\omega_2),\ (\mathbf{x}_7,\omega_1),\ (\mathbf{x}_8,\omega_3),\ (\mathbf{x}_9,\omega_1)\}$$

$$P(\omega_1 \mid T) = 3/9,\ P(\omega_2 \mid T) = 4/9,\ P(\omega_3 \mid T) = 2/9$$

엔트로피 불순도: $im(T) = -(\frac{3}{9}\log_2\frac{3}{9} + \frac{4}{9}\log_2\frac{4}{9} + \frac{2}{9}\log_2\frac{2}{9}) = 1.5305$

지니 불순도: $im(T) = 1 - (\frac{3^2}{9^2} + \frac{4^2}{9^2} + \frac{2^2}{9^2}) = 0.642$

오분류 불순도: $im(T) = 1 - \frac{4}{9} = 0.556$

MYONGJI UNIVERSITY

결정트리 | 노드에서의 질문

- 노드에서 질문 선택
  - 불순도 감소량이 최대인 질문을 취함
    - 불순도 감소량

$$\Delta im(T) = im(T) - \frac{|X_{Tleft}|}{|X_T|} im(T_{left}) - \frac{|X_{Tright}|}{|X_T|} im(T_{right}) \qquad (6.6)$$

MYONGJI UNIVERSITY

결정트리 | 노드에서의 질문

- 노드에서 질문 생성
  - 비계량인 경우 $x_i=\alpha$?
  - 계량인 경우 $x_i<\alpha$?
    - 이산
      - 이산 값에 따라 α를 결정
    - 연속
      - 실수 범위를 구간화 하여 α 결정
      - 또는 샘플의 값 분포를 보고 두 값의 가운데를 α로 결정

결정트리 | 노드에서의 질문

- 예제 2) 후보 질문 생성

직업 (x_1): [1,7]의 정수 (1 = 디자이너, 2 = 스포츠맨, 3 = 교수, 4 = 의사, 5 = 공무원, 6 = NGO, 7 = 무직)

선호 품목 (x_2): [1,5]의 정수 (1 = 의류, 2 = 전자 제품, 3 = 스포츠 용품, 4 = 책, 5 = 음식)

몸무게 (x_3): 실수

x_1에 의한 후보 질문: x_1=1?, x_1=2?, x_1=3?, x_1=4?, x_1=5?, x_1=6?, x_1=7?

x_2에 의한 후보 질문: x_2=1?, x_2=2?, x_2=3?, x_2=4?, x_2=5?

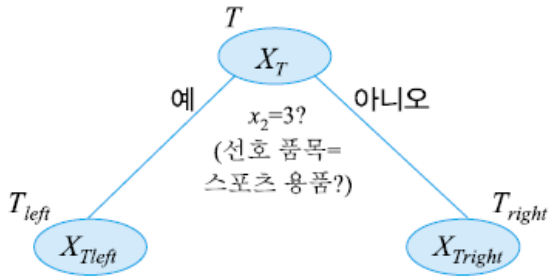
표에서 x_3의 값의 분포를 조사하면,

45.6, 47.8, 50.6, 65.3, 67.8, 72.8, 88.7, 92.3, 102.2

x_3에 의한 후보 질문: x_3<46.7?, x_3<49.2?, x_3<57.95?, x_3<66.55?, x_3<70.3?, x_3<80.75?, x_3<90.5?, x_3<97.25?

결정트리 | 노드에서의 질문

- 예제 3) 불순도 감소량



X_T={($\mathbf{x}_1$,ω_2), ($\mathbf{x}_2$,ω_1), ($\mathbf{x}_3$,ω_3), ($\mathbf{x}_4$,ω_2), ($\mathbf{x}_5$,ω_2), ($\mathbf{x}_6$,ω_2), ($\mathbf{x}_7$,ω_1), ($\mathbf{x}_8$,ω_3), ($\mathbf{x}_9$,ω_1)}}

X_{Tleft}={($\mathbf{x}_2$,ω_1), ($\mathbf{x}_7$,ω_1), ($\mathbf{x}_9$,ω_1)}

X_{Tright}={($\mathbf{x}_1$,ω_2), ($\mathbf{x}_3$,ω_3), ($\mathbf{x}_4$,ω_2), ($\mathbf{x}_5$,ω_2), ($\mathbf{x}_6$,ω_2), ($\mathbf{x}_8$,ω_3)}

그림 6.6 '선호 품목=스포츠 용품?' 이라는 질문에 따른 분기 결과

(6.6)의 불순도 감소량 $\Delta im(T) = 0.642 - \frac{3}{9} * 0.0 - \frac{6}{9} * 0.444 = 0.346$

MYONGJI UNIVERSITY

결정트리 I
Python을 이용한 불순도 감소량 계산

- x1==1? : 0.0
- x1==2? : 0.20864197530864198
- x1==3? : 0.0864197530864198
- x1==4? : 0.0
- x1==5? : 0.058641975308642014
- x1==6? : 0.0
- x1==7? : 0.11419753086419748

- x2==1? : 0.05467372134038789
- x2==2? : 0.058641975308642014
- x2==3? : 0.345679012345679
- x2==4? : 0.058641975308642014
- x2==5? : 0.05467372134038789

- x3<46.7? : 0.086419753
- x3<49.2? : 0.086419753
- x3<57.95? : 0.012345679
- x3<66.55? : 0.008641975
- x3<70.3? : 0.008641975
- x3<80.75? : 0.08641975
- x3<90.5? : 0.13403880
- x3<97.25? : 0.05864197

MYONGJI UNIVERSITY

결정트리 | 학습 알고리즘

- 결정 트리 학습 알고리즘
  - 언제 멈출 것인가?
    - 과적합 vs. 설익은 수렴
  - 잎 노드의 부류 할당

$$\left.\begin{array}{l} T\text{의 부류를 } \omega_k\text{로 한다.} \\ \text{이때 } k = \underset{i}{\operatorname{argmax}}\ P(\omega_i \mid T) \end{array}\right\}$$

알고리즘 [6.1] 결정 트리 학습

입력: 훈련 집합 $X = \{(\mathbf{x}_1, t_1), \cdots, (\mathbf{x}_N, t_N)\}$

출력: 결정 트리 R

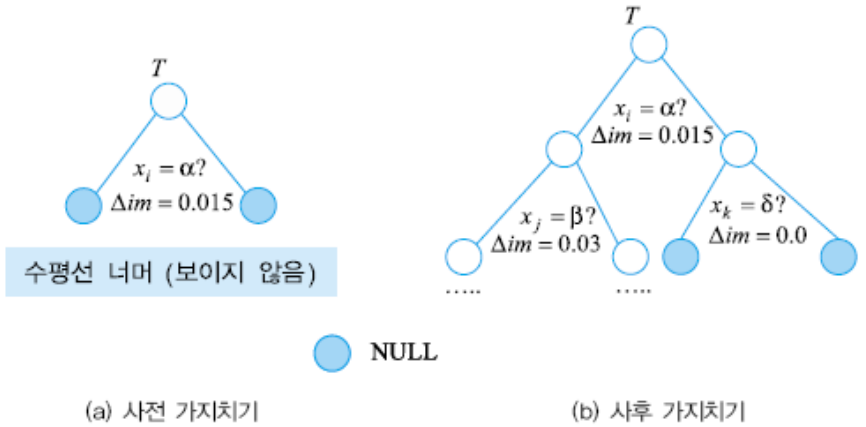
알고리즘:

```
1. 노드 하나를 생성하고 그것을 R이라 한다.  // 이것이 루트 노드이다.
2. T = R;
3. X_T = X;
4. split_node(T, X_T);   // 루트 노드를 시작점으로 하여 순환 함수를 호출한다.

5. split_node(T, X_T) {  // 순환 함수
6.     노드 T에서 후보 질문을 생성한다.
7.     모든 후보 질문의 불순도 감소량을 측정한다.  // (6.6) 또는 (6.7) 이용
8.     불순도 감소량이 최대인 질문 q를 선택한다.
9.     if (T가 멈춤 조건을 만족) {
10.        T에 부류를 할당한다.
11.        return;
12.    }
13.    else {
14.        q로 X_T를 X_Tleft와 X_Tright로 나눈다.
15.        새로운 노드 T_left와 T_right를 생성한다.
16.        split_node(T_left, X_Tleft);
17.        split_node(T_right, X_Tright);
18.    }
19. }
```

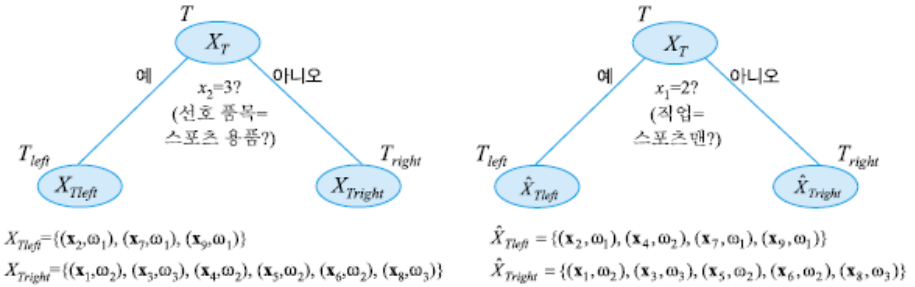
결정트리 | 특성

- 결정 트리의 특성
  - 특징 값에 대한 제약이 적다.
    - 계량, 비계량, 혼합 특징을 모두 다룰 수 있다.
    - 특징 전처리 불필요
  - 분류 결과가 '해석 가능'하다.
  - 인식 작업이 매우 빠르다.
  - 가지치기
    - 사전 가지치기
    - 사후 가지치기
  - 불안정성
  - 결정 트리 학습은 욕심 알고리즘
  - 손실 특징을 다루기 쉽다.
    - 대리 분기



(a) 사전 가지치기 (b) 사후 가지치기

그림 6.7 사전 가지치기와 사후 가지치기



(a) 주 분기 (b) 대리 분기

그림 6.8 주 분기와 대리 분기