13강. 나무모형(1)

- 나무모형이란
- 나무모형 분할 방법

1. 나무모형이란?

1) 나무모형의 소개

- 분석과정을 나무구조로 도형화하여 분류분석 혹은 회귀분석을 수행하는 최신 분석기법
- 반응변수가 범주형인 경우의 나무모형: 분류나무, 의사결정나무
- 반응변수가 숫자형인 경우의 나무모형: 회귀나무
- 분류분석과 회귀분석의 과정이 나무구조에 의해 표현되므로 분석자가 그 과정을 쉽게 이해하고 설명할 수 있는 장점
- 통계학을 전공하지 않은 일반인도 분석결과를 쉽게 해석할 수 있다는 점에서 실무에서의 활용도 높음

1. 나무모형이란?

1) 나무모형의 소개

예. 신용카드회사에 신용카드 발급 요청 서류 접수

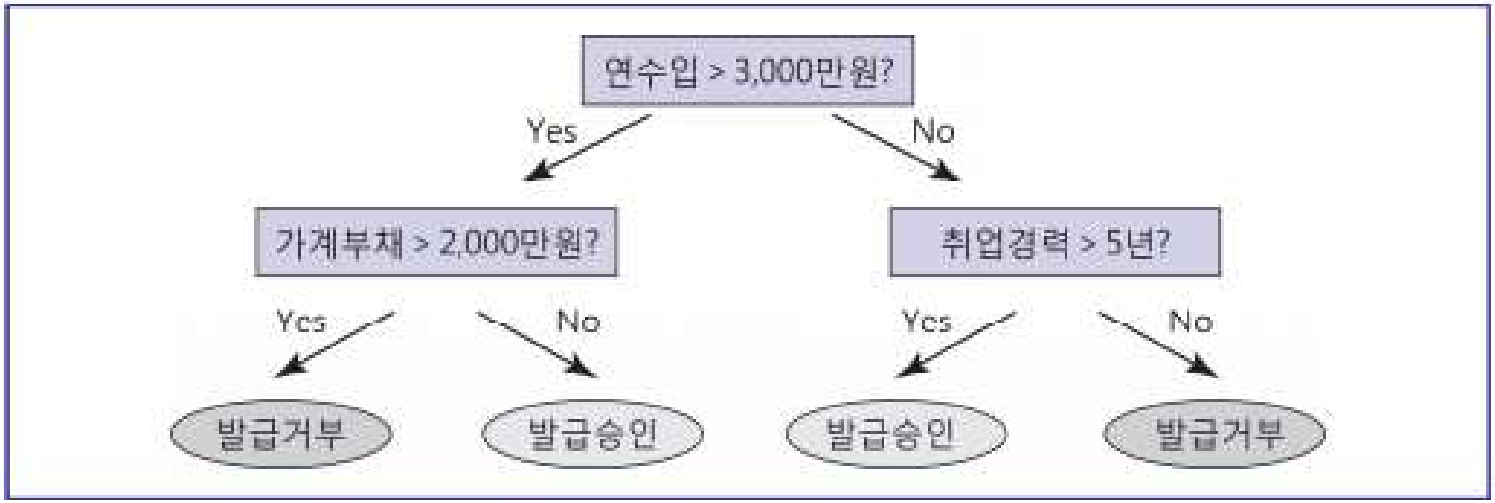
- 신청자들에 대한 연령, 소득수준, 주거형태 등의 인구 사회적 속성과 금융거래 기록 및 신용도 등을 이용하여 발급승인 혹은 발급거부

- 발급이 거부된 고객에게 거부이유를 설명해야 할 때 복잡한 통계모형을 설명하여 이해시키기란 너무 어려우므로 단순하면서도 이해가 쉬운 분류규칙을 필요로 함

1. 나무모형이란?

1) 나무모형의 소개

- 나무모형의 예

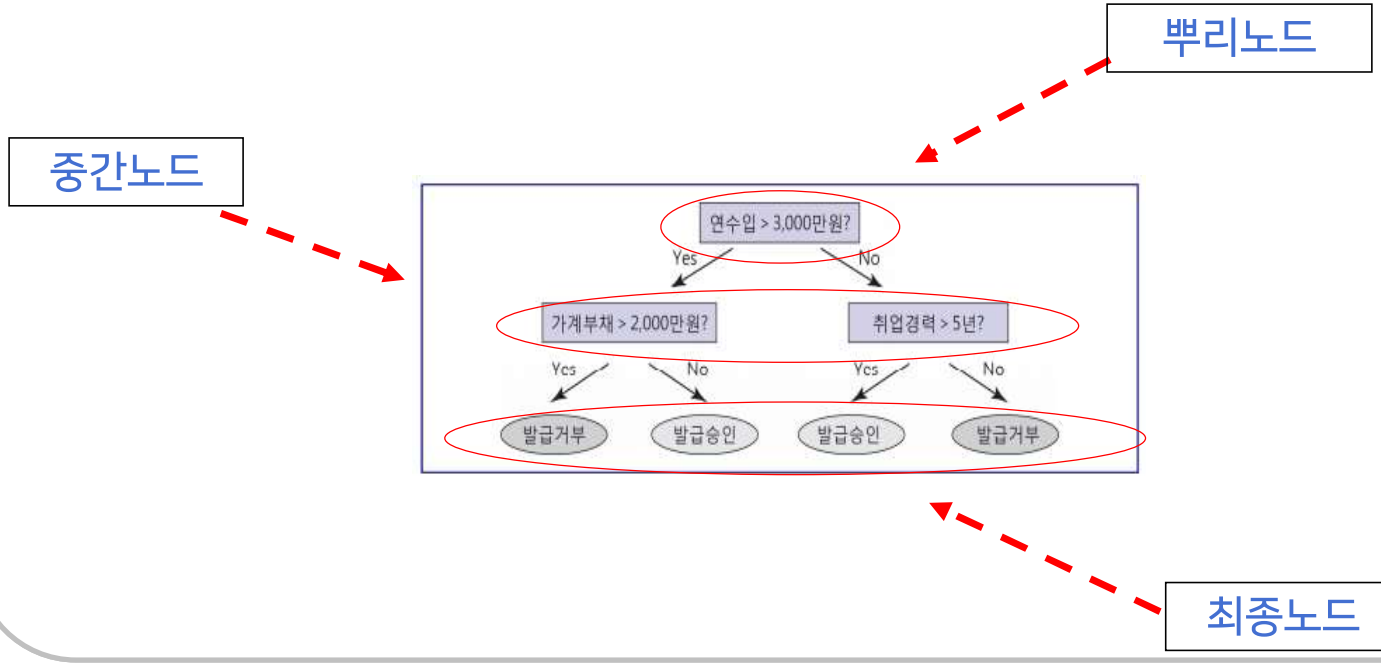


- 신용카드 발급여부에 대한 규칙이 매우 일목요연하게 표현되어 있음

1. 나무모형이란?

1) 나무모형의 소개

- 노드라 불리는 몸통과 노드와 노드를 연결해주는 줄기로 구성.
  노드에서는 규칙을, 줄기에서는 규칙에 대한 답을 제공



1. 나무모형이란?

1) 나무모형의 소개

- **분류규칙 생성과정**
  - 신용카드 회사는 자기회사에 수집된 고객들의 데이터를 활용
  - 데이터에 기록된 신용카드 소지자들의 사용대금 및 상환 기록을 참고
  - 대부분의 고객은 연체없이 좋은 신용도를 유지하게 되지만,
  - 일부 고객들은 3개월 이상 대금을 갚지 못하여 회사에 손실을 입히는 고객도 있음
  - 연체를 발생시키는 고객들을 잘 구분하여 분류해 주는 나무모형 분류규칙을 개발
  - 분류나무 규칙을 미래 고객에게 적용하여 연체가 발생하는 규칙에 해당되는 신청자에게 승인거부라는 의사결정

1. 나무모형이란?

2) 나무모형의 목적

1. 분류: 카드 발급 여부 결정 혹은 대출 승인 여부 의사결정
2. 예측: 월간 평균 신용카드 사용액 예측
3. 등급화: 고객을 일정한 기준에 따라 몇 개 등급으로 나눔.
   수입이 많고 신용도가 높은 고객은 높은 등급을 부여하여 많은 액수의
   신용한도액을 허용
4. 세분화: 군집분석의 결과 고객들을 여러 개의 군집으로 분류.
   군집결과를 반응변수로 사용하여 나무모형. 신규고객에 대한 고객세분화 가능
5. 변수선택: 나무모형에서 사용되는 변수는 매우 유용한 변수
6. 상호작용탐색 : 설명변수 중 일부 변수들의 특별한 조합이 가지는 특별한 효과를
   찾아내는 것 의미. 추후 분석에서는 통계모형에 상호작용을 포함

1. 나무모형이란?

3) 나무모형의 장점

1. 설명변수의 형태에 관계없이 적용 가능하다.
   ① 설명변수가 명목형, 순서형, 숫자형 여부에 영향 받지 않고 나무모형의 방법을 적용할 수 있다.
   ② 이는 나무모형이 기타 통계모형과 구별되는 큰 장점이다.

2. 이해 및 해석이 용이한 장점이 있다.
   ① 나무구조를 따라서 뿌리노드에서 최종노드까지 따라가면 이해할 수 있고 해석할 수 있다.
   ② 추가로 나무구조로부터 중요한 변수에 대한 아이디어도 얻을 수 있다.

1. 나무모형이란?

3) 나무모형의 장점

3. 상호작용을 쉽게 찾아낸다.

   나무구조를 해석하다 보면 2개 이상 변수간의 상호작용이 발견되는 경우가 많다.

4. 결측치의 처리가 용이하다.

   매 단계 변수 1개만 사용하므로 결측치의 영향이 적으며 분할변수에
   결측치가 있는 경우에는 surrogate 라는 대리변수를 사용하는 처리방법이 있다.

5. 나무모형이 구축되고 나면 외부데이터에 대한 분류 및 예측이
   쉽고 빠르게 이루어진다.

1. 나무모형이란?

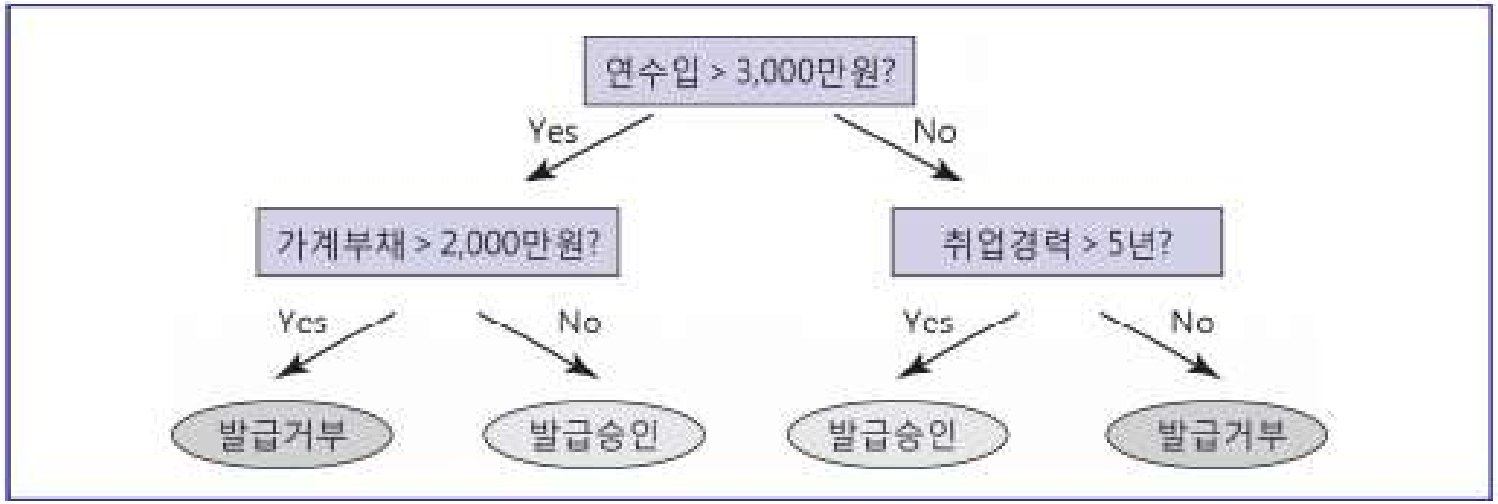
4) 나무모형의 단점

1. 나무구조의 단순성과 분리점의 경직성 때문에 분류 성과 및 예측성과가 다른 모형보다 떨어질 수도 있다.
   ① 연속형 변수인 경우 분리점 경계에 있는 값은 잘못 예측될 가능성이 커진다.

2. 나무구조는 불안정적일 수 있다.
   ① 관찰치의 수가 적은 경우에 데이터에 약간의 변화가 가해지면 나무구조는 변형이 될 수 있다.
   ② 이것은 나무모형의 분할방법이 데이터에 크게 의존하기 때문이다.

2. 나무모형 분할방법

- 분할규칙: ❖연수입 > 3000만원?
  - ❖중간노드: 가계부채 > 2000만원?
  - ❖취업경력 > 5년?



2. 나무모형 분할방법

- 나무모형에 분할규칙은 나무구조를 생성하는 과정 중에서 가장 중요한 틀을 잡아주는 일
- 분할방법의 기본적 아이디어
  - ✓ 나무구조의 매 단계마다 가급적 같은 그룹의 관찰치들이 같은 노드에 속하게 되도록 분할규칙을 찾는다.
  - ✓ 뿌리노드에서는 두 개 집단의 관찰치들이 섞여 있지만,
  - ✓ 분할이 진행됨에 따라 중간노드에서는 점점 더 동질적인 노드가 만들어지게 되고
  - ✓ 최종노드에서는 두 개 집단 중 한 집단이 압도적으로 많아지도록 함

2. 나무모형 분할방법

- **현재 가장 많이 사용되고 있는 세 가지 방법**

  1. CART 방법
  2. CHAID 방법
  3. QUEST 방법

2. 나무모형 분할방법

1) CART 방법

- 지니지수 (Gini Index)

$$\text{지니지수}(t) = 1 - \sum_{j=1}^{J} p(j \mid t)^2$$

  - ❖ 불순도 (impurity) 를 측정하는 지수
  - ❖ 여러 그룹의 관찰치들이 섞여 있으면 불순도가 높음
  - ❖ 한 그룹의 관찰치들만 있으면 불순도가 0 이 됨
  - ❖ 0~0.5 사이의 값을 가짐
  - ❖ *P(j/t) : t* 노드에서 *j* 그룹에 속한 관찰치들의 비율

2. 나무모형 분할방법

1) CART 방법

- 지니지수 (Gini Index)

$$\text{지니지수}(t) = 1 - \sum_{j=1}^{J} p(j \mid t)^2$$

예1) 어떤 노드 t에 그룹1의 관찰치가 10개, 그룹2의 관찰치가 10개 있는 경우

$$p(1 \mid t) = \frac{10}{20}, \text{ and } p(2 \mid t) = \frac{10}{20}.$$

$$\text{지니지수} = 1 - (0.5)^2 - (0.5)^2 = 0.5$$

예2) 어떤 노드 t에 그룹1의 관찰치가 20개, 그룹2의 관찰치가 없는 경우

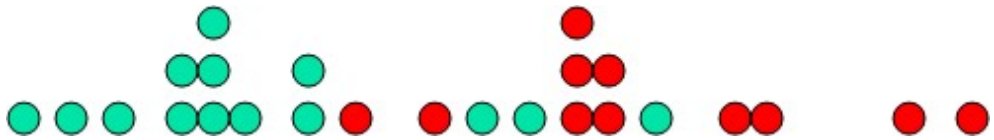
$$p(1 \mid t) = \frac{20}{20}, \text{ and } p(2 \mid t) = \frac{0}{20}.$$

$$\text{지니지수} = 1 - (1)^2 - (0)^2 = 0.$$

2. 나무모형 분할방법

1) CART 방법

(1) 연속형 변수의 예 : 1개 변수이며 2개 집단에 속한 연속형 자료



총 25개의 관찰치 중 적색 그룹은 11개, 녹색 그룹은 14개

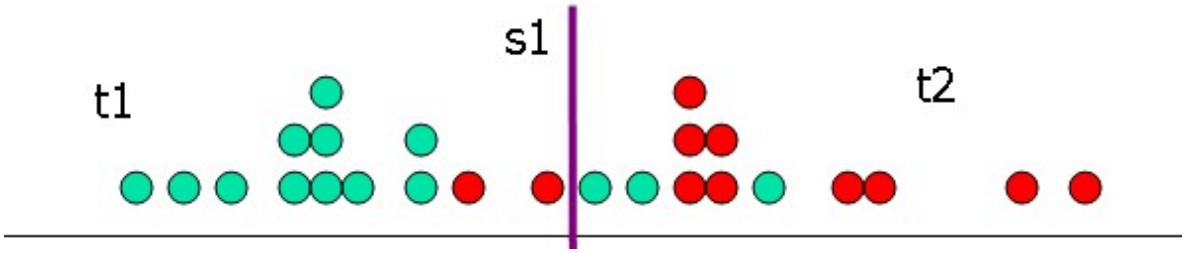
$$p(1|t)=\frac{11}{25},\ \text{and}\ p(2|t)=\frac{14}{25}.$$

$$\text{지니지수}=1-\left(\frac{11}{25}\right)^2-\left(\frac{14}{25}\right)^2=0.492$$

2. 나무모형 분할방법

1) CART 방법

- 분할점 s1

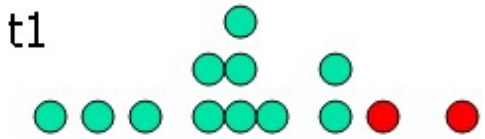


분할점 s1에 의해 현 노드가 t1 과 t2 의 두 개로 분할

2. 나무모형 분할방법

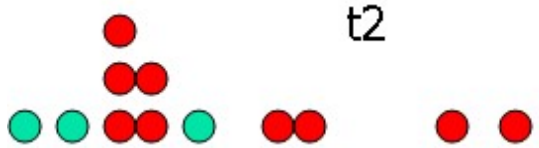
1) CART 방법

- 노드 t1



$$\text{지니지수} = 1 - \left(\frac{2}{13}\right)^2 - \left(\frac{11}{13}\right)^2 = 0.260$$

- 노드 t2



$$\text{지니지수} = 1 - \left(\frac{9}{12}\right)^2 - \left(\frac{3}{12}\right)^2 = 0.375$$

2. 나무모형 분할방법

1) CART 방법

- 노드 t1 과 노드 t2 의 지니지수 가중평균

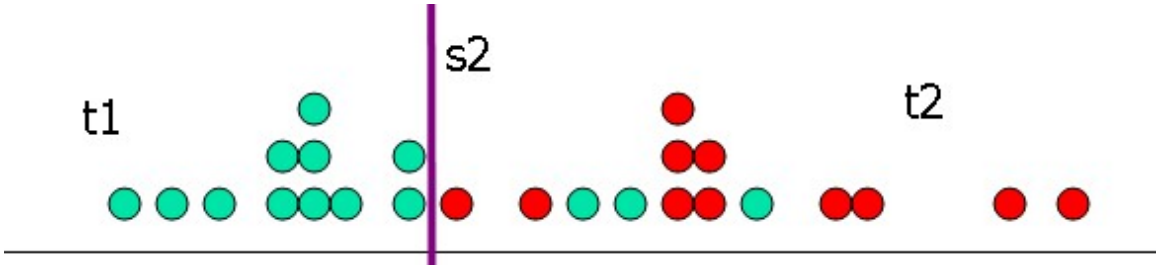
$$\text{가중평균} = 0.260 \times \frac{13}{25} + 0.375 \times \frac{12}{25} = 0.315$$

- 분할점 s1 에 의해 지니지수는 0.493에서 0.315로 감소함
- 감소도(s1) = 0.493 - 0.315 = 0.178

2. 나무모형 분할방법

1) CART 방법

- 분할점 s2

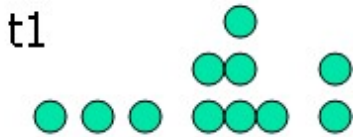


분할점 s2에 의해 현 노드가 t1 과 t2 의 두 개로 분할

2. 나무모형 분할방법

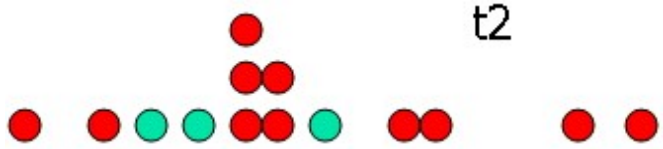
1) CART 방법

- 노드 t1



$$\text{지니지수} = 1 - \left(\frac{0}{11}\right)^2 - \left(\frac{11}{11}\right)^2 = 0.0$$

- 노드 t2



$$\text{지니지수} = 1 - \left(\frac{11}{14}\right)^2 - \left(\frac{3}{14}\right)^2 = 0.337$$

2. 나무모형 분할방법

1) CART 방법

- 노드 t1 과 노드 t2 의 지니지수 가중평균

$$\text{가중평균} = 0.0 \times \frac{11}{25} + 0.337 \times \frac{14}{25} = 0.189$$

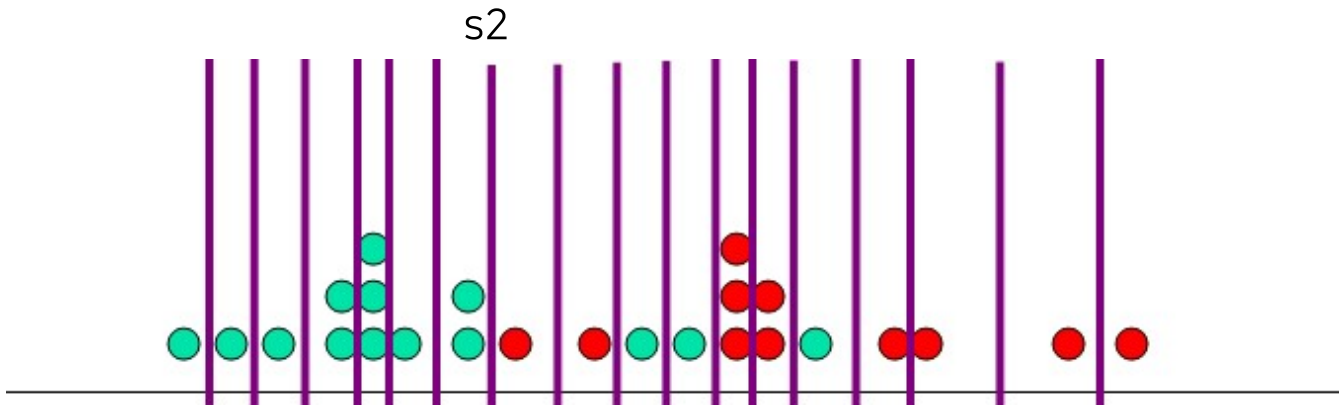
- 분할점 s2 에 의해 지니지수는 0.493에서 0.189로 감소함
- 감소도(s2)=0.493 - 0.189 = 0.303
- 감소도(s1)=0.493 - 0.315 = 0.178

불순도가 더 많이 감소, 더 효율적임

2. 나무모형 분할방법

1) CART 방법

- 모든 분할점

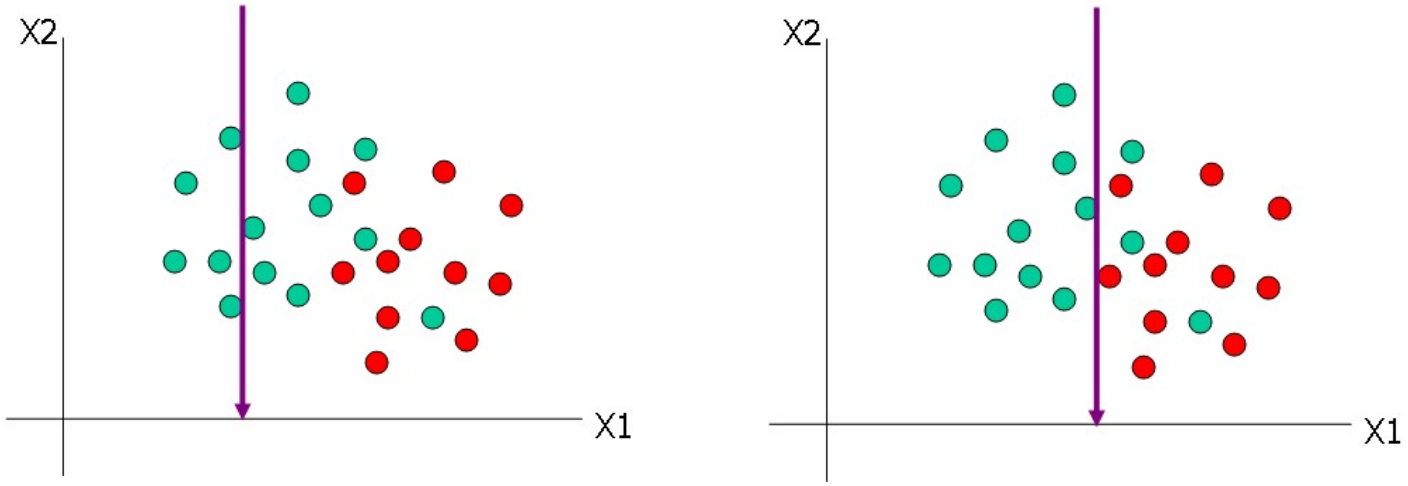


모든 분할점에 대해 지니지수를 계산해 보면 s2는 불순도를 최대로 감소시킴

2. 나무모형 분할방법

1) CART 방법

- 변수가 2개 이상인 경우

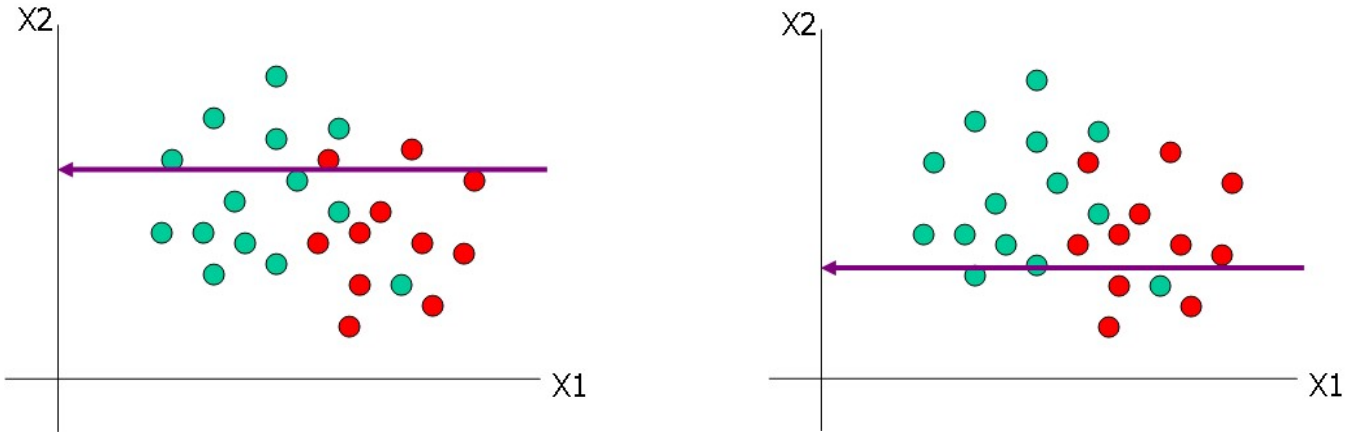


1. 우선 X1의 최적 분할점을 탐색하여 찾음

2. 나무모형 분할방법

1) CART 방법

- 변수가 2개 이상인 경우

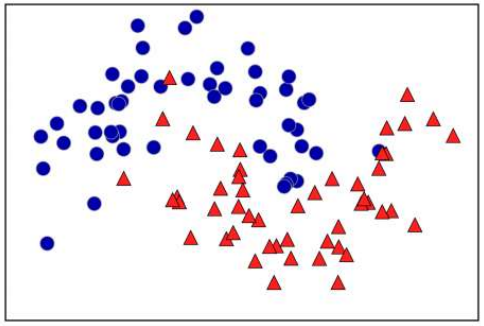


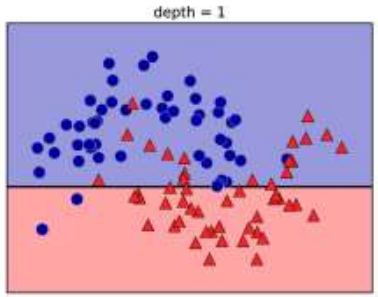
2. X2의 최적 분할점을 탐색하여 찾음

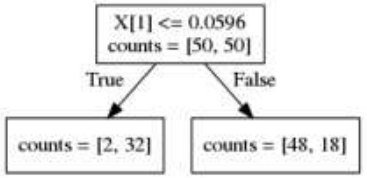
3. X1과 X2의 최적 분할점을 비교 – 불순도의 감소량을 비교함

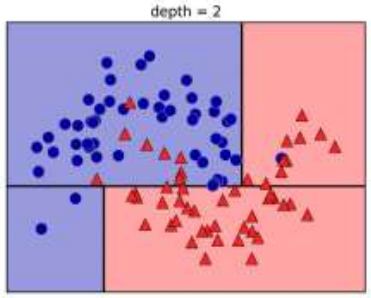
4. 불순도가 가장 많이 감소되는 최적분할점을 선택

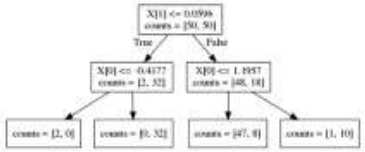
Building decision trees : recursive partitioning

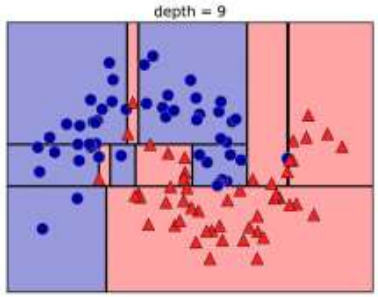


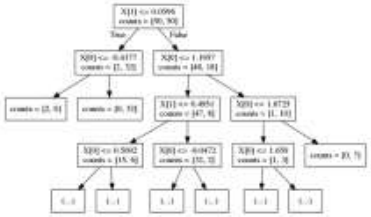












다음시간에는

14강 나무모형 (2)

수고했습니다.