데이터마이닝
(Data Mining)

한국방송통신대학교
정보통계학과 장영재 교수

4 강

나무모형

목차

1. 나무모형이란?

나무모형이란?

1) 나무모형의 소개

➢ 나무모형은 분석과정을 나무구조로 도형화하여 분류분석 혹은 회귀 분석을 수행하는 데이터마이닝의 대표적 분석기법

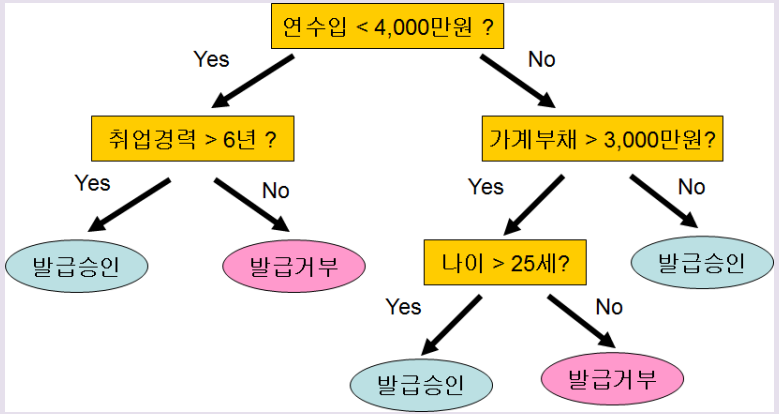
- 목표변수가 범주형인 경우에는 분류나무, 목표변수가 숫자형인 경우에는 회귀나무 모형을 이용

  - 나무구조에 의해 표현된 분류 및 회귀분석

나무모형이란?

1) 나무모형의 소개

〈그림1〉 신용카드 발급 승인관 관련한 나무모형



나무모형이란?

1) 나무모형의 소개

➢ 나무모형에 의한 의사결정 규칙은 데이터를 바탕으로 이루어짐

- 수집된 고객들의 데이터를 이용(데이터베이스에 기록된 많은 자사 신용카드 소지자들의 사용대금 및 상환 기록 등)하여 뿌리노드를 분할하며, 중간노드를 거쳐 최종노드에 다다르는 분리규칙을 도출할지 결정

- 분류나무 규칙을 미래 고객에게 적용하여 연체가 발생하는 규칙에 해당되는 신청자에게 승인거부라는 의사결정을 수행

나무모형이란?

2) 나무모형의 목적

➢ 나무모형은 분류 외에도 등급화, 세분화, 변수선택, 상호작용 탐색 등의 목적으로 사용 가능

- 분류: 카드발급, 은행대출 승인 등에 관한 의사결정
- 예측: 신용카드 신청자들이 카드를 발급받은 후의 신용상태라든지 월간 평균 신용카드 사용액 등을 예측
- 등급화: 신용카드 발급자들을 일정한 기준에 따라 몇 개 등급으로 구분
- 세분화: 군집분석의 결과 고객들을 여러 개의 군집으로 나눌 수 있고 군집결과를 목표변수로 사용하여 나무모형을 만들 수 있음 (고객세분화)
- 변수선택: 나무모형에서 뿌리노드 및 중간노드의 분할에 사용되는 변수 등 유용한 변수의 선택
- 상호작용탐색 : 입력변수 중 일부 변수들의 특별한 조합이 가지는 특별한 효과를 찾아내는 것

나무모형이란?

3) 상호작용의 발견

- 입력변수의 개수가 많은 경우 고려해야 하는 상호작용의 개수가 매우 많아져서 많은 시간과 노력이 소요됨

  - 타이태닉 승선자의 생존율 분석(본교재 참고)의 경우 좌석 등급(“1등석 〉 2등석 〉 3등석 〉 승무원”)과 성별(“여성〉남성”) 및 연령(“어린이〉성인”) 모두 생존과 사망을 구분 짓는 매우 중요한 역할을 하는 변수
  - 상호작용이 없는 로지스틱 회귀분석이라는 통계모형을 이용하여 분석해 보면 세 변수 모두 유의한 변수임을 확인할 수 있음

나무모형이란?

3) 상호작용의 발견

〈그림2〉 타이태닉 데이터 로지스틱 회귀분석 결과

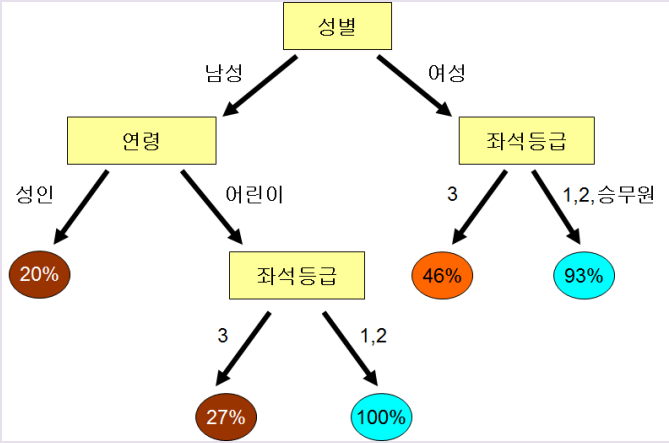
```
> setwd("C:/data/")
> titanic = read.csv("titanic.csv",header=T)
> summary(titanic)
    Class         Age           Sex         Survived
 Crew  :885   Adult:2092   Female: 470   No :1490
 First :325   Child: 109   Male  :1731   Yes: 711
 Second:285
 Third :706
> fit.logit = glm(Survived ~ Class + Age+ Sex, family = binomial, data = titanic)
> library(car)
> Anova(fit.logit)
Analysis of Deviance Table (Type II tests)

Response: Survived
      LR Chisq Df Pr(>Chisq)
Class   119.03  3  < 2.2e-16 ***
Age      18.85  1  1.413e-05 ***
Sex     352.91  1  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

나무모형이란?

3) 상호작용의 발견

〈그림3〉 타이태닉 데이터 의사결정나무



나무모형이란?

3) 상호작용의 발견

1. 성별 생존율은 좌석등급에 따라 영향
2. 좌석등급에 따른 생존율은 승객의 성별에 따라 영향
3. 승객의 연령에 따른 생존율은 성별에 따라 영향
4. 좌석등급에 따른 생존율은 연령과 성별에 의해 영향
   (연령과 좌석등급과 성별과의 3차 상호작용)

➢ 상호작용을 로지스틱 회귀분석으로 확인(상호작용 변수 추가)

나무모형이란?

3) 상호작용의 발견

〈그림4〉 타이태닉 데이터 로지스틱 회귀분석 상호작용 포함 결과

```
> fit.logit1 = glm(Survived ~ Class + Age + Sex + Class:Sex + Age:Sex + Class:Age:Sex,
+ family = binomial(link = "logit"), data = titanic)
> Anova(fit.logit1)
Analysis of Deviance Table (Type II tests)

Response: Survived
              LR Chisq Df Pr(>Chisq)
Class           120.73  3  < 2.2e-16 ***
Age              20.34  1  6.486e-06 ***
Sex             352.91  1  < 2.2e-16 ***
Class:Sex        57.29  3  2.233e-12 ***
Age:Sex           8.64  1   0.003295 **
Class:Age:Sex    37.26  4  1.590e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

✓ 나무모형이 아니었더라면, 현실적으로 로지스틱 회귀분석에서 3차 상호작용이 있다 하더라도 그 구체적 내용을 파악하는 것은 쉽지 않은 작업

나무모형이란?

4) 나무모형의 역사

- 나무모형은 최초에 상호작용을 탐색하는 방법으로 개발
  - AID (automatic interaction detection) 라는 방법을 개발한 선퀴스트와 모건 (Sonquist and Morgan, 1964)이 효시
- AID를 개량한 THAID (Morgan and Messenger, 1973) 라는 알고리즘이 개발되었으며, 카스 (Kass, 1980)에 의해서 CHAID (chi-squared automatic interaction detection) 라는 알고리즘이 완성
- 브라이먼, 프리드먼, 올센과 스톤 (Breiman, Friedman, Olshen, and Stone, 1984)은 CART (classification and regression trees) 라는 나무모형 알고리즘을 완성
  - 나무모형의 학술적 연구를 진흥시키는 결정적 역할

나무모형이란?

4) 나무모형의 역사

- 비슷한 시기 저명한 컴퓨터과학자인 퀸란 (Quinlan, 1983)에 의해 C4.5라는 나무모형이 개발
  - CART와 C4.5는 비모수적 방법 (이후 모수적 접근방법으로 나무모형의 효율성과 강건성을 추구한 방법들이 출현)
- FACT (Loh and Vanichsetakul, 1988)라는 나무모형은 모수적 접근방법을 채택한 최초의 나무모형이지만, 나무모형의 가지치기를 허용하지 않은 단점이 있어서, 후에 퀘스트(Quest) (Loh and Shih, 1997)의 개발이 뒤따름
  - 퀘스트 방법은 FACT와는 달리 이진분할 (binary split)을 실시하는 방법이었고, 나무모형의 가지치기를 채택
  - 2001년에는 크루즈(Cruise) (Kim and Loh, 2001)라는 다중 분할 나무모형이 개발(상호작용을 미리 탐색하여 명확화 가능)

나무모형이란?

4) 나무모형의 장단점

➢ 나무모형의 장점을 요약하면

① 입력변수의 형태에 관계없이 적용 가능

② 이해 및 해석이 용이

③ 상호작용을 쉽게 포착

④ 결측치의 처리가 용이

- 분할변수에 결측치가 있는 경우에는 서로게이트 (surrogate)라는 대리변수를 사용하는 처리방법이 있다.

⑤ 나무모형이 구축되고 나면 외부데이터에 대한 분류 및 예측이쉽고 빠르게 이루어짐

나무모형이란?

4) 나무모형의 장단점

➢ 나무모형의 단점을 요약하면

① 나무구조의 단순성과 분리점의 경직성 때문에 분류 성과 및 예측성과가 다른 모형보다 떨어질 수도 있음

- 연속형 변수인 경우 분리점 경계에 있는 값은 잘못 예측될 가능성 커짐

② 나무구조는 불안정적일 수 있음

- 관찰치의 수가 적은 경우에 데이터에 약간의 변화가 가해지면 나무구조는 변형이 될 수 있음 (∵나무모형의 분할방법이 데이터에 크게 의존)

2. 분류나무모형의 분할방법

2. 분류나무모형의 분할방법

1) CART 방법

➢ CART 방법에서는 순수도를 높게 하는 것과 동등하게 불순도(impurity)를 낮게 하는 방법으로 진행

- 지니지수는 불순도(impurity)를 측정하는 하나의 함수

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

- CART (Classification And Regression Trees)는 좌우 2개 가지로만 분할하는 이진분할 방법

2. 분류나무모형의 분할방법

1) CART 방법

- CART 방법에서는 순수도를 높게 하는 것과 동등하게 불순도(impurity)를 낮게 하는 방법으로 진행

  - CART는 모든 가능한 분할규칙중 분할개선도 (Goodness of split)를 최대화 시키는 분할규칙을 선택

$$\text{분할개선도} = \text{지니지수}(t) - \frac{N_{t_1}}{N_t}\text{지니지수}(t_1) - \frac{N_{t_2}}{N_t}\text{지니지수}(t_2)$$

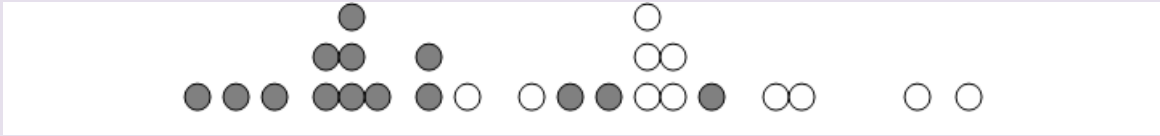
단, 노드 t_1 과 t_2는 노드 t의 자식노드들이며, N_t는 노드에 속한 관찰치의 개수

2. 분류나무모형의 분할방법

1) CART 방법

➢ 연속형 입력변수의 예

- 목표변수가 흑, 백 2개 집단으로 구성되어 있을 때,

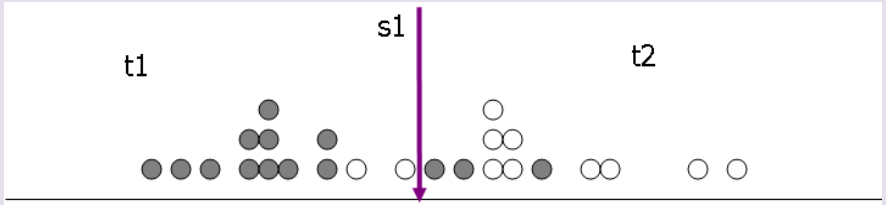


$$\text{지니지수} = 1 - \left(\frac{11}{25}\right)^2 - \left(\frac{14}{25}\right)^2 = 0.492$$

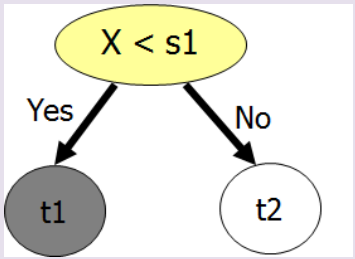
2. 분류나무모형의 분할방법

1) CART 방법

- 분할점이 s1으로 설정되었다고 가정하면,



위 분할점 s1에 의한 나무모형은



2. 분류나무모형의 분할방법

1) CART 방법

$$\text{지니지수}(t_1) = 1 - \left(\frac{11}{13}\right)^2 - \left(\frac{2}{13}\right)^2 = 0.26$$

$$\text{지니지수}(t_2) = 1 - \left(\frac{9}{12}\right)^2 - \left(\frac{3}{12}\right)^2 = 0.375$$

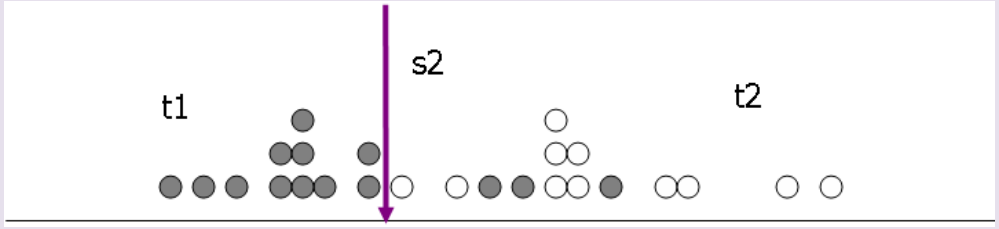
$$\text{가중평균} = 0.26 \times \frac{13}{25} + 0.375 \times \frac{12}{25} = 0.315$$

- 분할점 s1에 의해 지니지수는 0.492에서 0.315로 감소하는 효과가 발생. 즉, s1의 분할개선도는 0.177 (=0.492 - 0.315)

2. 분류나무모형의 분할방법

1) CART 방법

- 다음과 같은 분할점 s2에 의해 분할되는 경우



- 분할점 s2에 의해 지니지수는 0.492에서 0.189로 감소하는 효과가 발생
  즉, s2의 분할개선도는 0.303 (=0.492 - 0.189)

  → s2가 더 효율적으로 분할했음

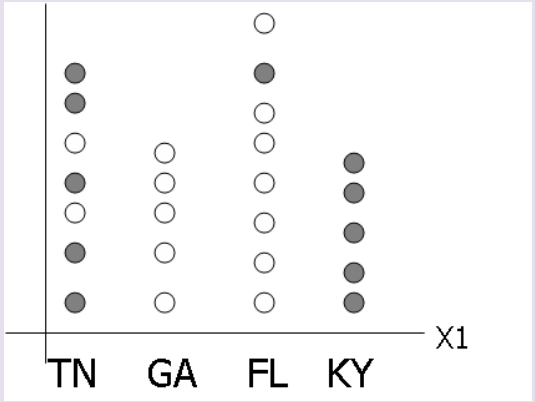
✓ 만약 변수가 2개 이상이 되면각 변수마다 최적의 분할점이 존재하며 이들 중 분할개선도를 최대화 해주는 것을 선택

2. 분류나무모형의 분할방법

1) CART 방법

➢ 범주형 입력변수의 예

- 아래와 같은 입력변수가 범주형이며 목표변수는 2개 집단으로 구성된 데이터의 경우(범주의 개수는 TN, GA, FL, KY 등의 네 개)



2. 분류나무모형의 분할방법

1) CART 방법

➢ 범주형 입력변수의 예

- CART에서는 모든 가능한 부분집합으로 분할을 고려

좌측분할	우측분할	지니지수
TN	GA, FL, KY	0.434
GA	TN, FL, KY	0.396
FL	TN, GA, KY	0.399
KY	TN, GA, FL	0.300
TN, GA	FL, KY	0.492
TN, FL	GA, KY	0.488
TN, KY	GA, FL	0.207

종합해보면 {TN,KY}와 {GA,FL}의 두 개 집합으로 분할하는 것이 지니지수를 최소화 시키는 분할집합

2. 분류나무모형의 분할방법

1) CART 방법

➢ 연속형과 범주형 입력변수가 혼재되어 있는 경우

- 연속형과 범주형 입력변수가 혼재되어 있는 경우에도 각 입력변수가 가지는 분할후보점 및 분할후보 부분집합을 모두 비교하여 가장 분할개선도를 최대화 해주는 것을 선택

➢ 반복적 분할

- 분할점 혹은 분할집합을 결정하는 절차를 자식노드 마다 반복적으로 적용해주면 데이터가 재귀적 (recursive)으로 분할되면서 나무구조가 완성

2. 분류나무모형의 분할방법

1) CART 방법

➢ 불순도 함수의 종류

- CART 나무모형에서는 지니지수를 불순도 함수로 사용하고 있지만, 불순도 함수로는 다음과 같은 다양한 종류가 존재

- 지니지수(t) $= 1 - \sum_{j}^{J} p(j|t)^2$
- 엔트로피(t) $= -\sum_{j}^{J} p(j|t) \times \log_2 p(j|t)$
- 디비언스(t) $= -2\sum_{j}^{J} n_j \log(p(j|t))$
- 분류오분류율(t) $= 1 - \max\{p(1|t), p(2|t), \ldots, p(J|t)\}$

2. 분류나무모형의 분할방법

2) C4.5 방법

➢ C4.5 나무모형에서는 엔트로피를 불순도 함수로 사용

- 연속형 입력변수의 경우에는 CART와 동일한 방법으로 분할후보점을 탐색하되, 엔트로피 불순도 함수를 사용한다는 점만 차이

- 범주형 입력변수의 경우에는 각 범주가 하나의 가지를 차지하도록 분할 (3개의 범주가 있는 범주형 입력변수라면 해당 노드는 3개의 자식노드로 분할)

  - C4.5에서는 불순도의 감소량을 정보이익 (information gain)이라고 칭하며 다음과 같이 정의

$$\text{정보이익} = \text{엔트로피}(t) - \sum_{j}^{K} \frac{N_{t_i}}{N_t} \text{엔트로피}(t_i)$$

2. 분류나무모형의 분할방법

2) C4.5 방법

- 정보이익은 분할 가지의 수가 많을수록 증가하는 경향이 있음.
  따라서 범주의 개수가 많은 범주형 입력변수일수록 정보이익이
  커지게 되므로, 그 변수가 분할 변수로 선택되게 되는 불합리한 측면

- 이 문제를 극복하기 위해 C4.5에서는 정보이익비율 (gain ratio)을
  가지 분할의 기준으로 삼음

정보이익비율 = 정보이익 / 내재정보

✓ 단, 내재정보(intrinsic information)란

$$-\sum_{j}^{K} \frac{N_{t_i}}{N_t} \log_2 \left(\frac{N_{t_i}}{N_t}\right)$$

2. 분류나무모형의 분할방법

3) CHAID 방법

➢ CHAID (Chi-squared Automatic Interaction Detection) 방법은 분할표 검정에서 사용되는 카이제곱 검정방법을 사용하여 분할점 혹은 분할집합을 결정하는 방법

- 연속형 입력변수에 대하여는 범주화를 수행하고 CHAID 방법의 분할규칙을 따름

1) 입력변수의 범주값 중에서 목표변수와의 카이제곱 검정의 결과가 유의하게 다르지 않은 범주들은 서로 통합

2) 통합된 입력변수의 범주값과 목표변수의 소속집단 정보를 이용하여 분할표를 생성. 분할표에 대한 카이제곱 검정의 결과가 유의하다면 이 변수는 그룹을 나누는데 유의한 변수

3) 모든 입력변수에 2)의 과정을 반복하여 가장 유의한 입력변수를 선택. 선택된 변수는 1)에서 결정된 범주별로 노드가 분리. 경우에 따라서 2개 가지로 분리될 수도 있지만 3개 이상의 가지로 분리될 수도 있음

2. 분류나무모형의 분할방법

4) QUEST 방법

➢ 분할후보점이 많은 변수를 더 선호하는 CART의 특성(변수선택 편향)을 수정하기 위해 QUEST (Quick Unbiased Efficient Statistical Tree) 방법이 제안

- CART의 편향 현상에 대한 이유가 변수선택과 분할점 선택이 동시에 이루어지기 때문

  - 이를 분리하여 먼저 변수를 선택하고, 그리고 선택된 변수에 대해서 분할점을 선택하는 방법입력변수의 범주값 중에서 목표변수와의 카이제곱 검정의 결과가 유의하게 다르지 않은 범주들은 서로 통합

2. 분류나무모형의 분할방법

4) QUEST 방법

- 분할후보점이 많은 변수를 더 선호하는 CART의 특성(변수선택 편향)을 수정하기 위해 QUEST (Quick Unbiased Efficient Statistical Tree) 방법이 제안

1) 변수선택 : 일요인 분산분석 (Oneway ANOVA) 카이제곱 검정 이용

2) 분할점 선택 : 선택된 분할변수에 한하여 CART의 분할점 선택 방법, 2차판별분석 (Quadratic discriminant analysis), 범주형인 경우 Crimcoord 방법과 2차판별분석 방법을 혼용하여 분할점을 선택

3) 장점 : QUEST 방법은 연산속도가 매우 빠르며, CART의 약점을 보완한 방법으로 예측정확도도 우수한 방법

2. 분류나무모형의 분할방법

5) CRUISE 방법

- CRUISE는 QUEST 방법의 단점인 변수선택 부분을 개선하고, 변수간 상호작용을 좀 더 적극적으로 반영하고자 하는 방법

1) 변수선택 : 카이제곱 분할표 검정을 사용 (입력변수간 상호작용 존재여부에 대한 통계검정이 포함)

2) 분할점 선택 : 선택된 분할변수에 한하여 CART의 분할점 선택 방법, 1차판별분석 (Linear discriminant analysis) 방법, 범주형인 경우 Crimcoord 라는 방법과 1차판별분석 방법을 혼용하여 분할점을 선택 (선택된 변수에 대해 1차 판별분석을 수행하기 전에, 박스-콕스 변환 (Box-Cox transformation)을 우선 수행)

3) 장점 : 결측치를 처리하는 다양한 방법을 제시하며 다중분할 및 선형결합분할도 가능

3. 분류나무모형의 크기 선택

3. 분류나무모형의 크기 선택

1) 분할정지 방법

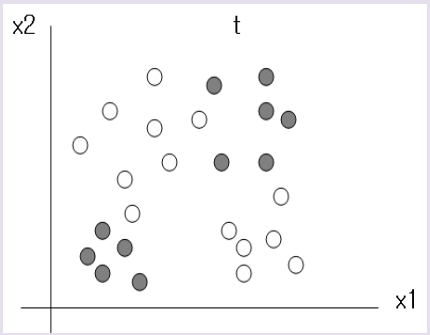
➢ 단계마다 분할이 꼭 필요한 것인지 통계적 유의성을 이용하여 평가

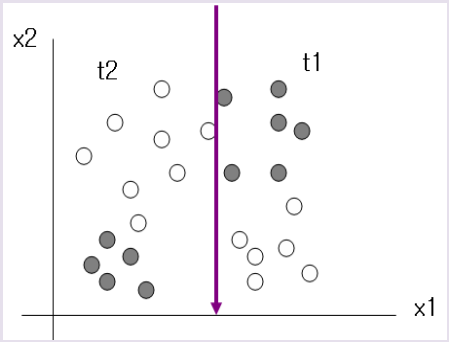
- 빠른 시간에 나무모형을 완성할 수 있다는 장점이 있지만, 다음절에 설명하는 가지치기 방법보다는 예측 정확도가 떨어지는 단점

3. 분류나무모형의 크기 선택

2) 가지치기 방법

➢ 계속적으로 분할해 나가도록 허용한 뒤 적절하지 않은 마디를 제거하여 적당한 크기의 나무구조를 가지는 나무모형을 최종적인 예측모형으로 선택



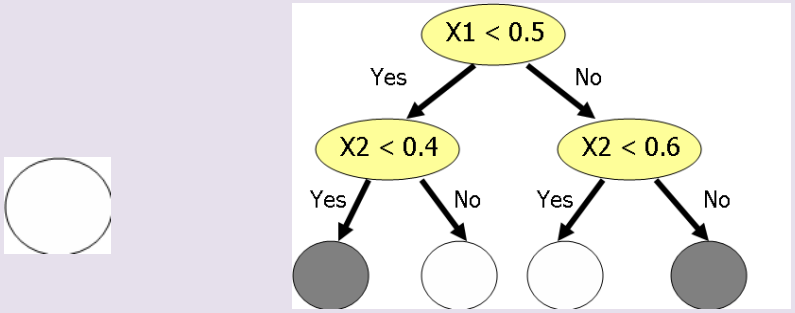


〈그림5〉 1회 분할에 그칠 경우

3. 분류나무모형의 크기 선택

2) 가지치기 방법

- 계속적으로 분할해 나가도록 허용한 뒤 적절하지 않은 마디를 제거하여 적당한 크기의 나무구조를 가지는 나무모형을 최종적인 예측모형으로 선택



〈그림6〉 분할정지 방법에 의한 나무(좌) 와 가지치기 방법에 의한 나무(우)

3. 분류나무모형의 크기 선택

2) 가지치기 이론

- CART 나무모형에서는 가지치기를 위해 비용복잡함수 (cost-complexity function)를 사용

$$\text{비용복잡함수}(\alpha) = \text{오분류율}(T) + \alpha \times |T|$$

- 단, T는 임의의 나무모형을 의미하고, |T|는 나무모형 T의 최종노드의 개수, α는 나무모형의 복잡도에 따른 벌점모수

3. 분류나무모형의 크기 선택

2) 가지치기 이론

- CART 나무모형에서는 가지치기를 위해 비용복잡함수 (cost-complexity function)를 사용

  - α 값이 증가함에 따라 구축되는 나무를 $T_0 \gg T_1 \gg T_2 \gg \cdots$라고 할 때, 10-fold 교차타당성 방법에 의하여 $T_0, T_1, T_2, \ldots$ 에 대한 오분류율을 계산하고, 이 중 가장 작은 오분류율을 보이는 T^*를 선택

$$Err(T^*) = \min\{Err(T_0), Err(T_1), Err(T_2), \ldots\}$$

3. 분류나무모형의 크기 선택

2) 가지치기 이론

- CART 나무모형에서는 가지치기를 위해 비용복잡함수 (cost-complexity function)를 사용

  - 1-s.e. 법칙은 위의 $T_0, T_1, T_2, \ldots$ 나무중에서 $Err(T^*) + s \cdot e(T^*)$ 의 값보다 작은 오분류율을 가지면서도 가장 크기가 작은 나무구조를 선택하는 방법

  ex) T^{**} 는 다음의 조건을 만족하는 나무구조 중 가장 작은 크기의 나무

  $$Err(T^*) < Err(T^{**}) \le Err(T^*) + s \cdot e(T^*)$$

  (0.5-s.e. 법칙이라면, $Err(T^*) + 0.5 \times s \cdot e(T^*)$를 사용)

4. 회귀나무모형의 분할방법

4. 회귀나무모형의 분할방법

1) CART 방법

- CART의 회귀나무는 분할을 위해 사용했던 불순도 함수 대신에 분산 함수를 사용하는 것이 차이점

$$\text{불순도}(t)=\frac{1}{N_t}\sum_{y_i \in t}(y_i-\bar{y}_t)^2$$

$$\text{분할개선도}=\text{불순도}(t)-\frac{N_{t_1}}{N_t}\text{불순도}(t_1)-\frac{N_{t_2}}{N_t}\text{불순도}(t_2)$$

4. 회귀나무모형의 분할방법

2) GUIDE 방법

- GUIDE (Generalized Unbiased Interaction Detection and Estimation) 방법은 CART의 변수선택에 대한 편향 현상을 수정하기 위해 개발된 방법

1) 변수선택 : 먼저 목표변수와 입력변수들을 이용하여 다중선형회귀분석을 수행하고 잔차의 분할표를 구성하여 카이제곱검정을 실시

2) 분할점 : CART의 분할점 선택 방법, 연속형인 입력변수인 경우 중위수 등

3) 장점 : GUIDE 방법은 분할변수 선택에 있어서 CART의 약점을 보완한 방법으로, 매우 다양한 회귀나무를 생성할 수 있으며, 예측정확도도 우수한 방법

강의를 마쳤습니다.

다음시간에는...