

5강 의사결정나무 II

통계·데이터과학과 장영재 교수



01 의사결정나무 관련 R 함수

02 R 사용 예제



1. 의사결정나무 관련 R 함수



01 의사결정나무 R 명령어 요약

1 rpart 함수

■ 함수의구조

`rpart(formula, data, weights, subset, na.action = na.rpart, method, parms, control, cost, ...)`

■ 기능

훈련데이터를이용하여의사결정나무를생성.R의part오브젝트를생성

■ 옵션

- formula : R에서 사용하는 모형 관련 공식. 옵션 data의 data frame에 존재하는 변수이름만 사용가능
- data : 훈련데이터에 해당하는 data frame 이름
- weights : 훈련데이터의 관측값에 대한 가중치 벡터. 디폴트는 균등가중치
- subset : 조건에 맞는 훈련데이터의 일부분만 사용하고자 할 때. 디폴트는 전체 사용



01 의사결정나무 R 명령어 요약

1 rpart 함수

■ 옵션

- `na.action = na.rpart` : 목표변수가 결측값이면 전체 관찰치를 삭제. 입력변수가 결측값인 경우에는 삭제하지 않음
- `method` : 의사결정나무의 종류를 지정함. "anova", "poisson", "class", "exp" 중에서 선택해야 함. "anova" 선택시 회귀나무, "poisson" 선택시 포아송 회귀나무, "class" 선택시 분류나무, "exp" 선택시 생존나무가 생성됨. 디폴트는 "class"
- `parms` : 분할을 위한 옵션. 회귀나무에는 해당사항 없음. 분류나무에는 사전확률을 주거나, 불순도 함수를 지정할 때 사용. 예를 들어 `parms = list(prior = c(0.65,0.35), split = "information")`인 경우, 집단1과 집단2의 사전확률은 65:35임. 불순도 함수는 엔트로피 함수 사용을 의미. 디폴트는 집단간 데이터비율 (proportional) 사전확률과 지니지수
- `control` : `rpart.control` 과 같은 역할. `rpart.control`에서 설명
- `cost` : 오분류 비용을 지정할 때 사용함. 디폴트는 균등비용



01 의사결정나무 R 명령어 요약

2 rpart.control 함수

■ 함수의구조

```
rpart.control(minsplit = 20, minbucket = round(minsplit/3), cp = 0.01,  
maxcompete = 4, maxsurrogate = 5, xval = 10, maxdepth = 30, ...)
```

■ 기능

의사결정나무를 생성할 때 분할 규칙 등을 설정

■ 옵션

- minsplit : 한 노드를 분할하기 위해 필요한 데이터의 개수. 이 값보다 적은 수의 관찰치가 있는 노드는 분할하지 않음. 디폴트는 20개
- minbucket : 최종노드에 포함되어야 하는 최소 데이터의 개수. minsplit이 지정되면, 자동으로 minsplit/3으로 지정됨. minbucket 보다 적은 수의 관찰치가 있는 노드는 존재하지 않게 됨



01 의사결정나무 R 명령어 요약

2 rpart.control 함수

■ 옵션

- `cp` : 비용복잡함수의 벌점모수. 노드를 분할할 때 분할 전과 비교하여 오분류율이 `cp` 값이상으로 향상되지 않으면 더 이상 분할하지 않음. 디폴트는 0.01
* 실제 적용 사례 : 만약 '`cp= 0.1`'이라면 벌점모수 0.1에 해당하는 비용복잡함수가 최소값이 될 때까지 계속 분할
- `maxcompete` : CART 분할방법에서 우수했던 분할후보점들을 `maxcompete` 개수만큼 출력함. 디폴트는 4개
- `maxsurrogate` : 결과물에 출력할 서로게이트 분할점의 개수 지정. 디폴트는 5개.
* 서로게이트-분할변수 값에 결측치가 있어서 분할이 어려울 경우, 이를 처리하기 위한 대체 분할변수를 찾아 분할을 실시하는 방법
- `xval` : 교차타당성의 fold 개수. 디폴트는 10
- `maxdepth` : 나무구조의 깊이 설정. 뿌리노드는 0. `maxdepth=5` 이면 나무구조는 뿌리노드로부터 5단계 아래로 내려감. 디폴트는 30



01 의사결정나무 R 명령어 요약

3 printcp 함수

■ 함수의구조

```
printcp(x, digits = getOption("digits") - 2)
```

■ 기능

R의 rpart 오브젝트를 대상으로 cp 값에 대한 나무구조 순서를 출력

■ 옵션

- x : rpart 오브젝트 이름
- digits : 출력할 숫자의 소수점 이하 자리수



01 의사결정나무 R 명령어 요약

4 prune 함수

- 함수의구조

`prune(tree, cp, ...)`

- 기능

의사결정나무의가지치기를실시

- 옵션

- tree : rpart 오브젝트 이름
- cp : 가지치기 할 비용복잡함수의 별점모수값



01 의사결정나무 R 명령어 요약

5 Plot 함수

■ 함수의구조

`plot(x, uniform = FALSE, branch = 1, compress = FALSE, nspace, margin = 0, ...)`

■ 기능

생성된 의사결정나무를 나무 그림으로 표현

■ 옵션

- `x` : `rpart` 오브젝트 이름
- `uniform` : 부모노드와 자식노드의 간격 크기. 만약 `FALSE` 이면 분할개선도에 비례하여 간격이 커지거나 작아짐. `TRUE`이면 균등한 간격을 유지. 디폴트는 `FALSE`
- `branch` : 중간노드를 잇는 가지의 형태. 만약 1이면 직교형이고 0이면 v자 형태. 0과 1사이 값이면 직교와 v자의 혼합형. 디폴트는 1



01 의사결정나무 R 명령어 요약

5 Plot 함수

I 옵션

- compress : 출력시 노드의 배치에 관한 사항. 만약 TRUE이면 더 압축된 의사결정 나무 그림을 출력해줌. 디폴트는 FALSE
- nspace : 중간노드와 최종노드 사이의 공간 여백값. 디폴트는 branch의 값과 동일
- margin : 나무 출력시 주변 여백값. 너무 작은 값을 사용하면 분할규칙이 잘리는 경우가 있음. 디폴트는 0



01 의사결정나무 R 명령어 요약

6 text 함수

■ 함수의구조

```
text(x, splits = TRUE, all = FALSE, digits = getOption("digits") - 3, use.n = FALSE, fancy = FALSE,  
fwidth = 0.8, fheight = 0.8, bg = par("bg"), col, ...)
```

■ 기능

plot에 의해 그려진 나무구조에 텍스트를 삽입

■ 옵션

- x : rpart 오브젝트 이름
- splits : 나무 출력시 분할규칙도 함께 출력하는지 여부. 디폴트는 TRUE.
- all : 최종노드만 분류집단명을 출력할지 여부. TRUE이면 중간노드에도 분류집단을 출력하게 됨. 디폴트는 FALSE



01 의사결정나무 R 명령어 요약

6 text 함수

■ 옵션

- digits : 분할규칙에 사용할 소수점 이하 자리 수
- use.n : 최종노드에 대한 정보 출력. TRUE 이면 정보를 출력하게 됨. 분류나무이면 각 집단별 관찰치 개수를 출력. 회귀나무이면 최종노드의 관찰치 개수를 출력함. 디폴트는 FALSE
- fancy : TRUE이면 중간노드는 타원으로, 최종노드는 직사각형으로 출력함. 디폴트는 FALSE
- fwidth : fancy 옵션이 TRUE일 때 사용하는 것으로, 타원과 직사각형의 넓이를 조절함. 디폴트는 0.8
- fheight : fancy 옵션이 TRUE일 때 사용하는 것으로, 타원과 직사각형의 높이를 조절함. 디폴트는 0.8
- bg : fancy 옵션이 TRUE일 때 사용하는 것으로, 나무의 배경색상
- col : 나무의 문자 색상



01 의사결정나무 R 명령어 요약

7 predict 함수

■ 함수의구조

```
predict(object, newdata, type = c("vector", "prob", "class", "matrix"), na.action = na.pass, ...)
```

■ 기능

생성된 의사결정나무 오브젝트에 새로운 데이터 newdata를 적용하여 예측

■ 옵션

- object : rpart 오브젝트 이름
- newdata : 예측의 대상인 data frame



01 의사결정나무 R 명령어 요약

7 predict 함수

I 옵션

- type : 분류나무인 경우
"vector"이면 예측된 집단이 숫자로 출력되고,
"prob"이면 집단별 예측 확률이 출력
"class"이면 예측된 집단이 factor 형태로 출력되고
"matrix"이면 위의 모든 것이 출력됨.
회귀나무인 경우, "vector" 혹은 "matrix"이면 예측값이 출력되고,
다른 옵션은 회귀나무와 관계없음
- na.action : newdata의 결측치에 대한 처리방법. na.omit은 결측치를 제외하는 방법이고 na.pass는 서로게이트 분할을 이용하는 방법임. 디폴트는 na.pass. 서로게이트란 결측치를 처리하기 위한 대체 분할법의 일종임



01 의사결정나무 R 명령어 요약

8 prp 함수

■ 함수의구조

`prp(object, type=2, extra="auto", digits=2, box.palette="auto", ...)`

■ 기능

생성된 의사결정나무 오브젝트를 나무구조의 형태로 출력

■ 옵션

- `object` : rpart 오브젝트 이름
- `type` : 의사결정나무 출력 형식 지정. '0'이면 중간노드를 그리지 않음
'1'이면 중간노드를 도형으로 표현하고 노드를 분할하는 조건을 노드 위쪽에 출력
'2'이면 중간노드를 그리면서 노드를 분할하는 조건을 노드 아래쪽에 출력
'3'이면 중간노드를 그리지 않으면서 왼쪽으로 분할하는 조건과 오른쪽으로 분할하는 조건을 모두 출력함
'4'이면 중간노드를 그리면서 왼쪽으로 분할하는 조건과 오른쪽으로 분할하는 조건을 모두 출력



01 의사결정나무 R 명령어 요약

8 prp 함수

I 옵션

- extra : 중간노드 내의 출력 형식을 지정
'0'인 경우: 분류나무이면 다수 범주를 출력하고 회귀나무이면 노드의 목표변수 평균을 출력
'1'인 경우: '0' 옵션의 결과에 추가하여 분류나무이면 범주별 관찰치 개수 출력, 회귀나무이면 노드의 총 관찰치 개수 출력
'2'인 경우: 분류나무인 경우 다수 범주의 관찰치 개수와 총 관찰치 개수 출력, 회귀나무는 해당사항 없음
- digits : 출력시 사용하는 소수점 자리수. 디폴트는 2
- box.palette : 노드별로 목표변수의 요약값을 색상으로 표현하는 색상표.
'auto'인 경우 모델의 행태에 따라 자동 선택. 'Grays'인 경우 회색으로 표현



2. R 사용 예제



다음시간 안내

6강. 앙상블 모형 I

