데이터마이닝
(Data Mining)

한국방송통신대학교
정보통계학과 장영재 교수

7 강 / 앙상블모형
: R을 이용한 실습

목차

1.배깅과 부스팅관련 R 함수

배깅과 부스팅관련 R 함수

1) bagging 함수

➢ 함수의 구조

bagging(formula, data, mfinal = 100, control,...)

➢ 기능

훈련데이터를 이용하여 배깅 앙상블을 수행한다. R의 bagging 오브젝트를 생성.
rpart 패키지를 필요로 함

➢ 옵션

- formula : R에서 사용하는 모형 관련 공식. 옵션 data의 data frame에 존재하는 변수이름만 사용가능함
- data : 훈련데이터에 해당하는 data frame 이름
- mfinal : 배깅앙상블의 분류기 개수. 디폴트는 100개
- Control : rpart.control과 같은 역할. rpart.control에서 설명

배깅과 부스팅관련 R 함수

2) predict.bagging 함수

- 함수의 구조

  predict.bagging(object, newdata, newmfinal=length(object$trees), ...)

- 기능

  생성된 배깅 앙상블모형 오브젝트에 새로운 데이터 newdata를 적용하여 예측

- 옵션
  - object : 배깅 오브젝트 이름
  - newdata : 예측의 대상인 data frame
  - newmfinal : 예측에 사용할 배깅 오브젝트내 분류기의 개수. 디폴트는 배깅 오브벡트의 분류기 개수

배깅과 부스팅관련 R 함수

3) errorevol 함수

- 함수의 구조

  errorevol(object, newdata)

- 기능

  R의 배깅 오브젝트를 대상으로 분류기 개수의 증가에 따라 오분류율의 변화를 출력

- 옵션
  - object : 배깅 오브젝트 이름
  - newdata : 예측의 대상인 data frame

배깅과 부스팅관련 R 함수

4) plot.errorevol 함수

- 함수의 구조

  plot.errorevol(x, y = NULL, ...)

- 기능

  errorevol 오브젝트의 오분류율을 그림으로 출력

- 옵션
  - x : errorevol 오브젝트 이름
  - y : 비교를 위한 또 다른 errorevol 오브젝트 이름. 드폴트는 없음

배깅과 부스팅관련 R 함수

5) boosting 함수

- 함수의 구조

  boosting(formula, data, boos = TRUE, mfinal = 100, coeflearn = 'Breiman', control,...)

- 기능

  훈련데이터를 이용하여 부스팅 앙상블 수행. R의 boosting 오브젝트 생성. rpart 패키지를 필요로 함

- 옵션
  - formula : R에서 사용하는 모형 관련 공식. 옵션 data의 data frame에 존재하는 변수이름만 사용가능함
  - data : 훈련데이터에 해당하는 data frame 이름
  - boos : 부스팅의 방식 선택. TRUE이면 표본추출에 의한 분류기 생성 방식을 사용, FALSE이면 가중치 반영된 분류기 생성 방식을 사용. 디폴트는 TRUE
  - mfinal : 부스팅 앙상블의 분류기 개수. 디폴트는 100개

배깅과 부스팅관련 R 함수

5) boosting 함수

➢ 옵션

- coeflearn : 분류기의 중요도 α_b의 정의.

  만약 'Breiman' 이면 $\alpha_b = \frac{1}{2}\log\frac{1-Err_b}{Err_b}$ 공식 사용,

  'Freund'이면 $\alpha_b = \log\frac{1-Err_b}{Err_b}$ 공식 사용,

  'Zhu' 이면 $\alpha_b = \log\frac{1-Err_b}{Err_b} + \log(nclass-1)$ 공식 사용. 디폴트는 'Breiman'

- control : rpart.control 과 같은 역할. rpart.control에서 설명

배깅과 부스팅관련 R 함수

6) predict.boosting 함수

- 함수의 구조

  predict.boosting(object, newdata, newmfinal=length(object$trees), ...)

- 기능

  생성된 부스팅 앙상블모형 오브젝트에 새로운 데이터 newdata를 적용하여 예측

- 옵션
  - object : 부스팅 오브젝트 이름
  - newdata : 예측의 대상인 data frame
  - newmfinal : 예측에 사용할 부스팅 오브젝트내 분류기의 개수. 디폴트는 부스팅 오브벡트의 분류기 개수

배깅과 부스팅관련 R 함수

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

배깅과 부스팅관련 R 함수

7) Predict 함수

- 함수의 구조

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

- 기능

생성된 나무모형 오브젝트에 새로운 데이터 newdata를 적용하여 예측

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

배깅과 부스팅관련 R 함수

7) Predict 함수

➢ 옵션

- type : 분류나무인 경우 “vector”이면 예측된 집단이 숫자로 출력되고, “prob”이면 집단별 예측 확률이 출력되며. “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

2. 랜덤포레스트관련R 함수

2. 랜덤포레스트관련 R 함수

1) Random Forest함수

➢ 함수의 구조

randomForest(formula, data, ntree=500, mtry, replace=TRUE, classwt=NULL, nodesize, maxnodes=NULL,
importance=FALSE, keep.forest=!is.null(y) && is.null(xtest), keep.inbag=FALSE, ...)

➢ 기능

훈련데이터를 이용하여 랜덤포레스트 앙상블을 수행. R의 랜덤포레스트 오브젝트를 생성

➢ 옵션

- formula : R에서 사용하는 모형 관련 공식. 옵션 data의 data frame에 존재하는 변수이름만 사용가능함
- data : 훈련데이터에 해당하는 data frame 이름
- ntree : 랜덤포레스트 앙상블의 분류기 개수. 디폴트는 500개

랜덤포레스트관련 R 함수

1) Random Forest함수

➢ 옵션

- mtry : 분류나무 중간노드마다 랜덤하게 선택되는 변수들의 개수 설정. 디폴트는 분류나무인 경우$\sqrt{p}$, 회귀나무인 경우 $\frac{p}{3}$
- replace : 관찰치를 랜덤추출할 때 TRUE이면 복원추출, FALSE는 비복원 추출. 복원추출이면 붓스트랩이라 함. 디폴트는 TRUE
- classwt : 집단에 대한 사전확률. 디폴트는 균등확률
- nodesize : 최종노드의 최소 데이터 수. 디폴트는 분류나무이면 1, 회귀나무이면 5
- maxnodes : 앙상블 내 나무모형이 가질 수 있는 최대 최종노드의 수. 디폴트는 제한없음
- importance : 입력변수의 중요도 계산 여부. 디폴트는 FALSE
- keep.forest : 앙상블 내 분류기의 정보 저장여부. 디폴트는 TRUE
- keep.inbag : 훈련데이터 관찰값이 붓스트랩 데이터에 포함되었는지여부를 저장한 ×의 행렬. 디폴트는 FALSE

2. 랜덤포레스트관련 R 함수

2) Importance 함수

➢ 함수의 구조

importance(x, type, class=NULL, scale=TRUE, ...)

Arguments

➢ 기능

생성된 랜덤포레스트 오브젝트를 이용하여 입력변수의 중요도를 계산

➢ 옵션

- x : 랜덤포레스트 오브젝트 이름
- type : '1' 혹은 '2' 선택. '1'은 정분류율의 평균감소값을 이용하여 계산, '2'는 불순도의 평균감소값을 이용하여 계산
- class : 분류의 문제에서 중요도를 계산할 특정 집단을 지정함. 디폴트는 없음
- scale : 중요도 계산에서 표준오차로 나누기 여부. 디폴트는 TRUE

2. 랜덤포레스트관련 R 함수

3) predict 함수

- 함수의 구조

  predict(object, newdata, type="response", predict.all=FALSE, ...)

- 기능

  생성된 랜덤포레스트 오브젝트에 새로운 데이터 newdata를 적용하여 예측

- 옵션
  - object : 랜덤포레스트 오브젝트 이름
  - newdata : 예측의 대상인 data frame
  - type : 예측값의 형태 지정. ‘response’, ‘prob.’, 혹은 ‘votes’ 를 선택가능
    ‘response‘는 예측집단, ’prob’는 집단별 확률, ‘votes’는 집단별 분류기의 투표수를 출력함.
    디폴트는 ‘response’
  - predict.all : 각 분류기의 예측결과 저장 여부, 디폴트는 FALSE

2. 랜덤포레스트관련 R 함수

4) Plot 함수

➢ 함수의 구조

plot(x, type="l", main, ...)

➢ 기능

랜덤포레스트 오브젝트의 오분류율 혹은 MSE를 계산

➢ 옵션

- x : 랜덤포레스트 오브젝트 이름
- type : plot내 선의 종류
- main : plot의 제목

강의를 마쳤습니다.

다음시간에는...