

# 7강 앙상블모형 II

통계·데이터과학과 장영재 교수



# 01 배깅과 부스팅 관련 R 함수

## 02 랜덤포레스트 관련 R 함수

### 03 R 사용 예제



# 1.배깅과 부스팅관련 R 함수



# 01 배깅과 부스팅 관련 R 함수

## 1 bagging 함수

### 함수의구조

`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 사용 가능함
- data : 훈련데이터에 해당하는 data frame 이름
- mfinal : 배깅앙상블의 분류기 개수. 디폴트는 100개
- control : rpart.control과 같은 역할



## 01 배깅과 부스팅 관련 R 함수

### 2 predict.bagging 함수

#### 함수의구조

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

#### 기능

생성된 배깅 앙상블모형 오브젝트에 새로운 데이터 `newdata`를 적용하여 예측

#### 옵션

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



## 01 배깅과 부스팅 관련 R 함수

### 3 importanceplot 함수

- 함수의구조

importanceplot(object, ...)

- 기능

R의 배깅 혹은 부스팅을 수행할 때 각 입력변수가 가지는 상대적 중요도를 표현

- 옵션

- object : 배깅 또는 부스팅 오브젝트 이름



## 01 배깅과 부스팅 관련 R 함수

### 4 errorevol 함수

#### 함수의구조

`errorevol(object, newdata)`

#### 기능

R의 배깅 또는 부스팅 오브젝트를 대상으로 분류기 개수의 증가에 따라 오분류율의 변화를 출력

#### 옵션

- `object`: 배깅 또는 부스팅 오브젝트 이름
- `newdata`: 예측의 대상인 data frame





## 01 배경과 부스팅 관련 R 함수

### 5 plot.errorevol 함수

#### ■ 함수의구조

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

#### ■ 기능

errorevol 오브젝트의 오분류율을 그림으로 출력

#### ■ 옵션

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





## 01 배경과 부스팅 관련 R 함수

### 6 boosting 함수

#### ■ 함수의구조

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

#### ■ 기능

훈련데이터를 이용하여 부스팅 앙상블 수행. R의 boosting 오브젝트 생성. rpart 패키지 필요

#### ■ 옵션

- `formula` : R에서 사용하는 모형 관련 공식. 옵션 `data`의 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



## 01 배경과 부스팅 관련 R 함수

### 6 boosting 함수

#### ■ 옵션

- `coflearn` : 분류기의 중요도  $\alpha_b$ 의 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



## 01 배깅과 부스팅 관련 R 함수

### 7 predict.boosting 함수

#### ■ 함수의구조

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

#### ■ 기능

생성된 부스팅 앙상블모형 오브젝트에 새로운 데이터 `newdata`를 적용하여 예측

#### ■ 옵션

- `object`: 부스팅 오브젝트 이름
- `newdata`: 예측의 대상인 data frame
- `newmfinal`: 예측에 사용할 부스팅 오브젝트 내 분류기의 개수. 디폴트는 부스팅 오브젝트의 분류기 개수



## 2. 랜덤포레스트 관련 R 함수



## 02 랜덤포레스트 관련 R 함수

### 1 randomForest함수

#### ■ 함수의구조

```
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 사용 가능함
- data : 훈련데이터에 해당하는 data frame 이름
- ntree : 랜덤포레스트 앙상블의 분류기 개수. 디폴트는 500개



## 02 램덤포레스트 관련 R 함수

### 1 randomForest함수

#### I 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 부트스트랩이라 함. 디폴트는 TRUE
- classwt : 집단에 대한 사전확률. 디폴트는 균등확률
- nodesize : 최종노드의 최소 데이터 수. 디폴트는 분류나무이면 1, 회귀나무이면 5
- maxnodes : 앙상블 내 의사결정나무가 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 부트스트랩 데이터에 포함되었는지 여부를 저장한 n by n 행렬. 디폴트는 FALSE



## 02 랜덤포레스트 관련 R 함수

### 2 importance 함수

#### ■ 함수의구조

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

#### ■ 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





## 02 랜덤포레스트 관련 R 함수

### 3 predict 함수

#### ■ 함수의구조

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

#### ■ 기능

생성된 랜덤포레스트 오브젝트에 새로운 데이터 `newdata`를 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 또는 예측값, 'prob'는 집단별 확률, 'votes'는 집단별 분류기의 투표수를 출력함. 디폴트는 'response'
- `predict.all` : 각 분류기의 예측결과 저장 여부, 디폴트는 FALSE



## 02 랜덤포레스트 관련 R 함수

### 4 plot 함수

#### ■ 함수의구조

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

#### ■ 기능

랜덤포레스트 오브젝트의 오분류율 혹은 MSE를 계산

#### ■ 옵션

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



## 02 랜덤포레스트 관련 R 함수

### 5 varImpPlot 함수

#### I 함수의구조

`varImpPlot(x, sort=TRUE, n.var=min(30, nrow(x$importance)), ...)`

#### I 기능

랜덤포레스트 변수 중요도 값들을 바차트 그래프로 표현

#### I 옵션

- `x` : 랜덤포레스트 오브젝트 이름
- `sort` : 수 중요도의 순서대로 정렬 여부
- `n.var` : 출력할 변수의 개수. 디폴트는 최대 30개



## 02 랜덤포레스트 관련 R 함수

### 6 partialPlot 함수

#### ■ 함수의구조

`partialPlot(x, pred.data, x.var, which.class, ...)`

#### ■ 기능

랜덤포레스트의 부분중속그림을 그래프로 표현(부분중속그림: 특정 변수 값 변동 시 전체 예측값에 미치는 영향의 시각화)

#### ■ 옵션

- `x` : 랜덤포레스트 오브젝트 이름
- `pred.data` : 부분중속그림을 그리기 위한 데이터
- `x.var` : 부분중속그림의 대상이 되는 변수명
- `which.class` : 분류의 경우 예측값을 계산할 변수명. 디폴트는 첫 번째 범주



# 3. R 사용 예제



다음시간 안내

# 8강. 신경망 모형 I

