Project Two

Ashlesha Tiwari

2023-03-19

Github link: <https://github.com/AshleshaTiwari/ProjectTwo_BreastCancerData/blob/main/ProjectTwo_Tiwari_Ashlesha.Rmd>

require(mlbench)

## Loading required package: mlbench

data(BreastCancer)  
ls(BreastCancer)

## [1] "Bare.nuclei" "Bl.cromatin" "Cell.shape" "Cell.size"   
## [5] "Cl.thickness" "Class" "Epith.c.size" "Id"   
## [9] "Marg.adhesion" "Mitoses" "Normal.nucleoli"

BreastCancer <- na.omit(BreastCancer)   
BreastCancer$Id <- NULL   
str(BreastCancer)

## 'data.frame': 683 obs. of 10 variables:  
## $ Cl.thickness : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 5 5 3 6 4 8 1 2 2 4 ...  
## $ Cell.size : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 8 1 10 1 1 1 2 ...  
## $ Cell.shape : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 4 1 8 1 10 1 2 1 1 ...  
## $ Marg.adhesion : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 5 1 1 3 8 1 1 1 1 ...  
## $ Epith.c.size : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 2 7 2 3 2 7 2 2 2 2 ...  
## $ Bare.nuclei : Factor w/ 10 levels "1","2","3","4",..: 1 10 2 4 1 10 10 1 1 1 ...  
## $ Bl.cromatin : Factor w/ 10 levels "1","2","3","4",..: 3 3 3 3 3 9 3 3 1 2 ...  
## $ Normal.nucleoli: Factor w/ 10 levels "1","2","3","4",..: 1 2 1 7 1 7 1 1 1 1 ...  
## $ Mitoses : Factor w/ 9 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 5 1 ...  
## $ Class : Factor w/ 2 levels "benign","malignant": 1 1 1 1 1 2 1 1 1 1 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:16] 24 41 140 146 159 165 236 250 276 293 ...  
## ..- attr(\*, "names")= chr [1:16] "24" "41" "140" "146" ...

set.seed(100)  
train.index <- sample(row.names(BreastCancer), 0.6\*dim(BreastCancer)[1])   
test.index <- setdiff(row.names(BreastCancer), train.index)   
train.df <- BreastCancer[train.index,]  
test.df <- BreastCancer[test.index,]

library(e1071)  
  
mysvm <- svm(Class ~ ., train.df)  
mysvm.pred <- predict(mysvm, test.df)  
table(mysvm.pred,test.df$Class)

##   
## mysvm.pred benign malignant  
## benign 169 5  
## malignant 6 94

Naive Bayes

library(klaR)

## Loading required package: MASS

mynb <- NaiveBayes(Class ~ ., train.df)  
mynb.pred <- predict(mynb,test.df)  
head(mynb.pred$class)

## 3 5 6 8 10 17   
## benign benign malignant benign benign benign   
## Levels: benign malignant

table(mynb.pred$class,test.df$Class)

##   
## benign malignant  
## benign 169 3  
## malignant 6 96

str(mysvm.pred)

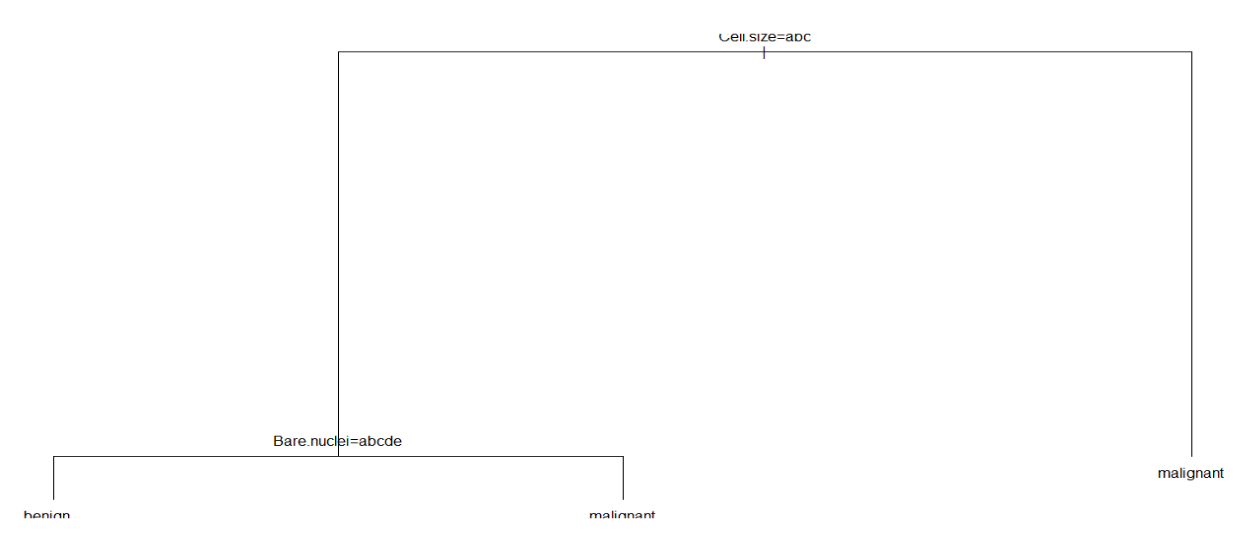
## Factor w/ 2 levels "benign","malignant": 1 1 2 1 1 1 1 2 1 1 ...  
## - attr(\*, "names")= chr [1:274] "3" "5" "6" "8" ...

str(mynb.pred)

## List of 2  
## $ class : Factor w/ 2 levels "benign","malignant": 1 1 2 1 1 1 1 2 1 1 ...  
## ..- attr(\*, "names")= chr [1:274] "3" "5" "6" "8" ...  
## $ posterior: num [1:274, 1:2] 1.00 1.00 2.07e-14 1.00 1.00 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : chr [1:274] "3" "5" "6" "8" ...  
## .. ..$ : chr [1:2] "benign" "malignant"

Decision Tree

library(MASS)  
library(rpart)  
mytree <- rpart(Class ~ ., train.df)  
plot(mytree); text(mytree)



summary(mytree)

## Call:  
## rpart(formula = Class ~ ., data = train.df)  
## n= 409   
##   
## CP nsplit rel error xerror xstd  
## 1 0.80714286 0 1.0000000 1.0000000 0.06854101  
## 2 0.08571429 1 0.1928571 0.2642857 0.04143646  
## 3 0.01000000 2 0.1071429 0.1714286 0.03395051  
##   
## Variable importance  
## Cell.size Cell.shape Bare.nuclei Normal.nucleoli Epith.c.size   
## 21 17 17 15 14   
## Bl.cromatin Cl.thickness   
## 13 1   
##   
## Node number 1: 409 observations, complexity param=0.8071429  
## predicted class=benign expected loss=0.3422983 P(node) =1  
## class counts: 269 140  
## probabilities: 0.658 0.342   
## left son=2 (282 obs) right son=3 (127 obs)  
## Primary splits:  
## Cell.size splits as LLLRRRRRRR, improve=133.7650, (0 missing)  
## Cell.shape splits as LLLRRRRRRR, improve=128.5963, (0 missing)  
## Normal.nucleoli splits as LLRRRRRRRR, improve=126.8815, (0 missing)  
## Bare.nuclei splits as LLRRRRRRRR, improve=120.8560, (0 missing)  
## Epith.c.size splits as LLRRRRRRRR, improve=113.6033, (0 missing)  
## Surrogate splits:  
## Cell.shape splits as LLLRRRRRRR, agree=0.944, adj=0.819, (0 split)  
## Normal.nucleoli splits as LLRLRRRRRR, agree=0.905, adj=0.693, (0 split)  
## Epith.c.size splits as LLLRRRRRRR, agree=0.900, adj=0.677, (0 split)  
## Bare.nuclei splits as LLLRRRRRRR, agree=0.900, adj=0.677, (0 split)  
## Bl.cromatin splits as LLLRRRRRRR, agree=0.892, adj=0.654, (0 split)  
##   
## Node number 2: 282 observations, complexity param=0.08571429  
## predicted class=benign expected loss=0.07092199 P(node) =0.6894866  
## class counts: 262 20  
## probabilities: 0.929 0.071   
## left son=4 (268 obs) right son=5 (14 obs)  
## Primary splits:  
## Bare.nuclei splits as LLLLLRRR-R, improve=21.67165, (0 missing)  
## Normal.nucleoli splits as LLRRRRRL-R, improve=18.78326, (0 missing)  
## Cl.thickness splits as LLLLLLRRRR, improve=16.38074, (0 missing)  
## Marg.adhesion splits as LLLRRRRRRR, improve=15.32522, (0 missing)  
## Bl.cromatin splits as LLLRR-R---, improve=11.36783, (0 missing)  
## Surrogate splits:  
## Cl.thickness splits as LLLLLLLLRR, agree=0.968, adj=0.357, (0 split)  
## Normal.nucleoli splits as LLLLLRRL-L, agree=0.965, adj=0.286, (0 split)  
## Cell.shape splits as LLLLRRRRRR, agree=0.957, adj=0.143, (0 split)  
## Marg.adhesion splits as LLLLLLLRRR, agree=0.957, adj=0.143, (0 split)  
## Epith.c.size splits as LLLLRRRRRR, agree=0.954, adj=0.071, (0 split)  
##   
## Node number 3: 127 observations  
## predicted class=malignant expected loss=0.05511811 P(node) =0.3105134  
## class counts: 7 120  
## probabilities: 0.055 0.945   
##   
## Node number 4: 268 observations  
## predicted class=benign expected loss=0.0261194 P(node) =0.6552567  
## class counts: 261 7  
## probabilities: 0.974 0.026   
##   
## Node number 5: 14 observations  
## predicted class=malignant expected loss=0.07142857 P(node) =0.03422983  
## class counts: 1 13  
## probabilities: 0.071 0.929

mytree.pred <- predict(mytree,test.df,type="class")  
table(mytree.pred,test.df$Class)

##   
## mytree.pred benign malignant  
## benign 170 9  
## malignant 5 90

Regularised Discriminant Analysis

library(klaR)  
myrda <- rda(Class ~ ., train.df)  
myrda.pred <- predict(myrda, test.df)  
table(myrda.pred$class,test.df$Class)

##   
## benign malignant  
## benign 170 3  
## malignant 5 96

Random Forest

library(randomForest)

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

myrf <- randomForest(Class ~ ., train.df)  
myrf.pred <- predict(myrf, test.df)  
head(myrf.pred)

## 3 5 6 8 10 17   
## benign benign malignant benign benign benign   
## Levels: benign malignant

table(myrf.pred, test.df$Class)

##   
## myrf.pred benign malignant  
## benign 169 2  
## malignant 6 97

Combining Classes

combine.classes<-data.frame(myrf.pred, myrda.pred$class,#myqda.pred,   
 mytree.pred,mysvm.pred, mynb.pred$class)  
  
combine.classes$myrf.pred<-ifelse(combine.classes$myrf.pred=="benign", 0, 1)  
combine.classes[,2]<-ifelse(combine.classes[,2]=="benign", 0, 1)  
combine.classes[,3]<-ifelse(combine.classes[,3]=="benign", 0, 1)  
combine.classes[,4]<-ifelse(combine.classes[,4]=="benign", 0, 1)  
combine.classes[,5]<-ifelse(combine.classes[,5]=="benign", 0, 1)  
str(combine.classes)

## 'data.frame': 274 obs. of 5 variables:  
## $ myrf.pred : num 0 0 1 0 0 0 0 1 0 0 ...  
## $ myrda.pred.class: num 0 0 1 0 0 0 0 1 0 0 ...  
## $ mytree.pred : num 0 0 1 0 0 0 0 1 0 0 ...  
## $ mysvm.pred : num 0 0 1 0 0 0 0 1 0 0 ...  
## $ mynb.pred.class : num 0 0 1 0 0 0 0 1 0 0 ...

combine.cl<-combine.classes[, -c(6,7)]  
majority.vote=rowSums(combine.classes[, -c(6,7)])  
head(majority.vote)

## 3 5 6 8 10 17   
## 0 0 5 0 0 0

combine.classes[,6]<-rowSums(combine.classes[,-c(6,7)])  
combine.classes[,7]<-ifelse(combine.classes[,6]>=4, "malignant", "benign")  
table(combine.classes[,7], test.df$Class)

##   
## benign malignant  
## benign 170 4  
## malignant 5 95