Project 2 – compare classifiers code

Hannah Furumo

3/10/2021

dim(BreastCancer)

## [1] 699 11

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

set.seed(2)  
ind <- sample(2, nrow(BreastCancer), replace = TRUE, prob=c(0.8, 0.2))

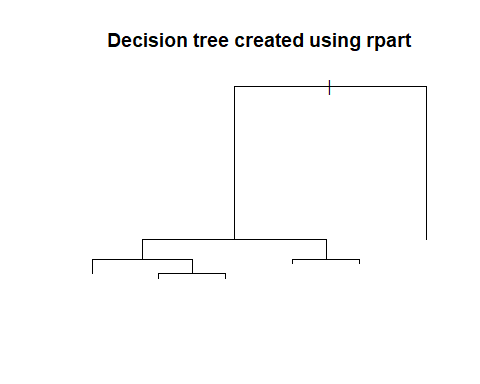
# create model using recursive partitioning on the training data set  
require(rpart)

## Loading required package: rpart

## Warning: package 'rpart' was built under R version 4.0.3

x.rp <- rpart(Class ~ ., data=BreastCancer[ind == 1,])  
# predict classes for the evaluation data set  
x.rp.pred <- predict(x.rp, type="class", newdata=BreastCancer[ind == 2,])  
# score the evaluation data set (extract the probabilities)  
x.rp.prob <- predict(x.rp, type="prob", newdata=BreastCancer[ind == 2,])

plot(x.rp, main="Decision tree created using rpart")



require(party)

## Loading required package: party

## Warning: package 'party' was built under R version 4.0.4

## Loading required package: grid

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 4.0.3

## Loading required package: modeltools

## Warning: package 'modeltools' was built under R version 4.0.3

## Loading required package: stats4

## Loading required package: strucchange

## Warning: package 'strucchange' was built under R version 4.0.4

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.3

##   
## Attaching package: 'zoo'

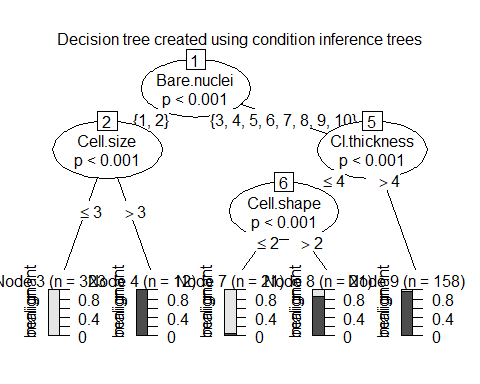
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Loading required package: sandwich

## Warning: package 'sandwich' was built under R version 4.0.4

x.ct <- ctree(Class ~ ., data=BreastCancer[ind == 1,])  
x.ct.pred <- predict(x.ct, newdata=BreastCancer[ind == 2,])  
x.ct.prob <- 1- unlist(treeresponse(x.ct, BreastCancer[ind == 2,]), use.names=F)[seq(1,nrow(BreastCancer[ind == 2,])\*2,2)]

plot(x.ct, main="Decision tree created using condition inference trees")



# create model using random forest and bagging ensemble using conditional inference trees  
x.cf <- cforest(Class ~ ., data=BreastCancer[ind == 1,], control = cforest\_unbiased(mtry = ncol(BreastCancer)-2))  
x.cf.pred <- predict(x.cf, newdata=BreastCancer[ind == 2,])  
x.cf.prob <- 1- unlist(treeresponse(x.cf, BreastCancer[ind == 2,]), use.names=F)[seq(1,nrow(BreastCancer[ind == 2,])\*2,2)]

# create model using bagging (bootstrap aggregating)  
require(ipred)

## Loading required package: ipred

## Warning: package 'ipred' was built under R version 4.0.4

x.ip <- bagging(Class ~ ., data=BreastCancer[ind == 1,])  
x.ip.prob <- predict(x.ip, type="prob", newdata=BreastCancer[ind == 2,])

# create model using svm (support vector machine)  
require(e1071)

## Loading required package: e1071

## Warning: package 'e1071' was built under R version 4.0.3

# svm requires tuning  
x.svm.tune <- tune(svm, Class~., data = BreastCancer[ind == 1,],  
 ranges = list(gamma = 2^(-8:1), cost = 2^(0:4)),  
 tunecontrol = tune.control(sampling = "fix"))  
# display the tuning results (in text format)  
x.svm.tune

##   
## Parameter tuning of 'svm':  
##   
## - sampling method: fixed training/validation set   
##   
## - best parameters:  
## gamma cost  
## 0.0625 1  
##   
## - best performance: 0.02234637

# If the tuning results are on the margin of the parameters (e.g., gamma = 2^-8),   
# then widen the parameters.  
# I manually copied the cost and gamma from console messages above to parameters below.  
x.svm <- svm(Class~., data = BreastCancer[ind == 1,], cost=4, gamma=0.0625, probability = TRUE)  
x.svm.prob <- predict(x.svm, type="prob", newdata=BreastCancer[ind == 2,], probability = TRUE)

require(ROCR)

## Loading required package: ROCR

## Warning: package 'ROCR' was built under R version 4.0.4

png(filename="roc\_curve\_5\_models.png", width=700, height=700)

# create an ROCR prediction object from rpart() probabilities  
x.rp.prob.rocr <- prediction(x.rp.prob[,2], BreastCancer[ind == 2,'Class'])  
# prepare an ROCR performance object for ROC curve (tpr=true positive rate, fpr=false positive rate)  
x.rp.perf <- performance(x.rp.prob.rocr, "tpr","fpr")  
# plot it  
plot(x.rp.perf, col=2, main="ROC curves comparing classification performance of five machine learning models")  
# Draw a legend.  
legend(0.6, 0.6, c('rpart', 'ctree', 'cforest','bagging','svm'), 2:6)  
# ctree  
x.ct.prob.rocr <- prediction(x.ct.prob, BreastCancer[ind == 2,'Class'])  
x.ct.perf <- performance(x.ct.prob.rocr, "tpr","fpr")  
# add=TRUE draws on the existing chart   
  
# ctree  
x.ct.prob.rocr <- prediction(x.ct.prob, BreastCancer[ind == 2,'Class'])  
x.ct.perf <- performance(x.ct.prob.rocr, "tpr","fpr")  
  
# cforest  
x.cf.prob.rocr <- prediction(x.cf.prob, BreastCancer[ind == 2,'Class'])  
x.cf.perf <- performance(x.cf.prob.rocr, "tpr","fpr")  
  
# bagging  
x.ip.prob.rocr <- prediction(x.ip.prob[,2], BreastCancer[ind == 2,'Class'])  
x.ip.perf <- performance(x.ip.prob.rocr, "tpr","fpr")  
  
# svm  
x.svm.prob.rocr <- prediction(attr(x.svm.prob, "probabilities")[,2], BreastCancer[ind == 2,'Class'])  
x.svm.perf <- performance(x.svm.prob.rocr, "tpr","fpr")  
  
plot(x.ct.perf, col=3, add=TRUE)  
plot(x.cf.perf, col=4, add=TRUE)  
plot(x.ip.perf, col=5, add=TRUE)  
plot(x.svm.perf, col=6, add=TRUE)

