Ian Lynch - Project Two

Ian Lynch

3/16/2021

#Set Seed  
set.seed(2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

require(mlbench)

## Loading required package: mlbench

library(e1071)  
library(klaR)

## Loading required package: MASS

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(nnet)  
library(rpart)  
library(MASS)  
library(randomForest)

## randomForest 4.6-14

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

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(expss)

##   
## Use 'expss\_output\_viewer()' to display tables in the RStudio Viewer.  
## To return to the console output, use 'expss\_output\_default()'.

##   
## Attaching package: 'expss'

## The following objects are masked from 'package:dplyr':  
##   
## between, compute, contains, first, last, na\_if, recode, vars

library(writexl)

#Prep Dataset  
data(BreastCancer)  
BreastCancer <- na.omit(BreastCancer)   
BreastCancer$Id <- NULL

#SVM tuning  
#library(e1071)  
mysvm <- svm(Class ~., BreastCancer)  
mysvm.pred <- predict(mysvm, BreastCancer)  
table(mysvm.pred, BreastCancer$Class)

##   
## mysvm.pred benign malignant  
## benign 431 8  
## malignant 13 231

SVMPredict <- as.data.frame(mysvm.pred)  
colnames(SVMPredict) <- "SVM Tuning"  
#mysvm.pred benign malignant  
# benign 431 8  
# malignant 13 231

#NaiveBayes  
#library(klaR)  
mynb <- NaiveBayes(Class ~., BreastCancer)  
mynb.pred <- predict(mynb,BreastCancer)

table(mynb.pred$class,BreastCancer$Class)

##   
## benign malignant  
## benign 431 3  
## malignant 13 236

NBPredict <- as.data.frame(mynb.pred$class)  
colnames(NBPredict) <- "Naive Bayes"  
# benign malignant  
# benign 431 3  
# malignant 13 236

#Neural Net  
#library(nnet)  
mynnet <- nnet(Class ~ ., BreastCancer, size=1)

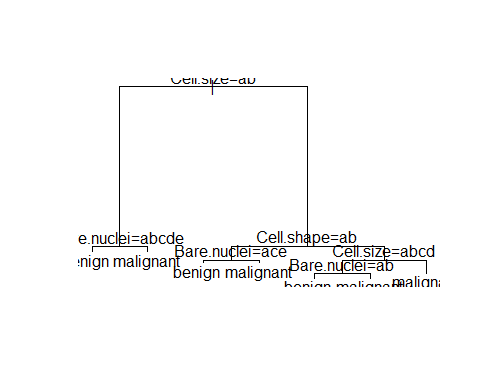
## # weights: 83  
## initial value 442.961096   
## iter 10 value 49.156890  
## iter 20 value 27.965406  
## iter 30 value 18.940677  
## iter 40 value 18.775892  
## iter 50 value 18.691843  
## iter 60 value 18.671388  
## iter 70 value 18.665875  
## iter 80 value 18.660203  
## iter 90 value 18.659652  
## iter 100 value 18.659286  
## final value 18.659286   
## stopped after 100 iterations

mynnet.pred <- predict(mynnet,BreastCancer,type="class")  
table(mynnet.pred,BreastCancer$Class)

##   
## mynnet.pred benign malignant  
## benign 442 1  
## malignant 2 238

NNPredict <- as.data.frame(mynnet.pred)  
  
colnames(NNPredict) <- "Neural Net"  
#mynnet.pred benign malignant  
# benign 438 10  
# malignant 6 229

#Decision Tree  
#library(rpart)  
mytree <- rpart(Class ~ ., BreastCancer)  
plot(mytree); text(mytree)



summary(mytree)

## Call:  
## rpart(formula = Class ~ ., data = BreastCancer)  
## n= 683   
##   
## CP nsplit rel error xerror xstd  
## 1 0.79079498 0 1.00000000 1.0000000 0.05215335  
## 2 0.05439331 1 0.20920502 0.2552301 0.03118544  
## 3 0.02510460 2 0.15481172 0.1799163 0.02655926  
## 4 0.01255230 3 0.12970711 0.1589958 0.02506475  
## 5 0.01000000 6 0.09205021 0.1506276 0.02443403  
##   
## Variable importance  
## Cell.size Cell.shape Bare.nuclei Epith.c.size Bl.cromatin   
## 21 18 16 15 14   
## Normal.nucleoli Cl.thickness   
## 14 1   
##   
## Node number 1: 683 observations, complexity param=0.790795  
## predicted class=benign expected loss=0.3499268 P(node) =1  
## class counts: 444 239  
## probabilities: 0.650 0.350   
## left son=2 (418 obs) right son=3 (265 obs)  
## Primary splits:  
## Cell.size splits as LLRRRRRRRR, improve=222.3221, (0 missing)  
## Cell.shape splits as LLLRRRRRRR, improve=216.4111, (0 missing)  
## Bare.nuclei splits as LLRRRRRRRR, improve=203.7284, (0 missing)  
## Bl.cromatin splits as LLLRRRRRRR, improve=196.3903, (0 missing)  
## Epith.c.size splits as LLRRRRRRRR, improve=193.1310, (0 missing)  
## Surrogate splits:  
## Cell.shape splits as LLLRRRRRRR, agree=0.917, adj=0.785, (0 split)  
## Epith.c.size splits as LLRRRRRRRR, agree=0.900, adj=0.743, (0 split)  
## Bare.nuclei splits as LLRRRRRRRR, agree=0.880, adj=0.691, (0 split)  
## Normal.nucleoli splits as LLRRRRRRRR, agree=0.877, adj=0.683, (0 split)  
## Bl.cromatin splits as LLLRRRRRRR, agree=0.876, adj=0.679, (0 split)  
##   
## Node number 2: 418 observations, complexity param=0.0251046  
## predicted class=benign expected loss=0.02870813 P(node) =0.6120059  
## class counts: 406 12  
## probabilities: 0.971 0.029   
## left son=4 (410 obs) right son=5 (8 obs)  
## Primary splits:  
## Bare.nuclei splits as LLLLLRRR-R, improve=11.68296, (0 missing)  
## Normal.nucleoli splits as LLLR-RRL-R, improve=11.68296, (0 missing)  
## Cl.thickness splits as LLLLLLRRRR, improve=10.32214, (0 missing)  
## Bl.cromatin splits as LLLLR-R---, improve= 8.53307, (0 missing)  
## Epith.c.size splits as LLLRRRRRRR, improve= 4.63208, (0 missing)  
## Surrogate splits:  
## Cl.thickness splits as LLLLLLLLRR, agree=0.988, adj=0.375, (0 split)  
## Normal.nucleoli splits as LLLR-RRL-L, agree=0.988, adj=0.375, (0 split)  
## Mitoses splits as LLRLL-LL-, agree=0.983, adj=0.125, (0 split)  
##   
## Node number 3: 265 observations, complexity param=0.05439331  
## predicted class=malignant expected loss=0.1433962 P(node) =0.3879941  
## class counts: 38 227  
## probabilities: 0.143 0.857   
## left son=6 (23 obs) right son=7 (242 obs)  
## Primary splits:  
## Cell.shape splits as LLRRRRRRRR, improve=20.58158, (0 missing)  
## Cell.size splits as LLLRRRRRRR, improve=18.27650, (0 missing)  
## Bare.nuclei splits as LRRRRRRRRR, improve=16.81493, (0 missing)  
## Bl.cromatin splits as LLRRRRRRRR, improve=13.91034, (0 missing)  
## Marg.adhesion splits as LLRRRRRRRR, improve=11.17148, (0 missing)  
## Surrogate splits:  
## Bl.cromatin splits as LRRRRRRRRR, agree=0.932, adj=0.217, (0 split)  
##   
## Node number 4: 410 observations  
## predicted class=benign expected loss=0.01219512 P(node) =0.6002928  
## class counts: 405 5  
## probabilities: 0.988 0.012   
##   
## Node number 5: 8 observations  
## predicted class=malignant expected loss=0.125 P(node) =0.01171303  
## class counts: 1 7  
## probabilities: 0.125 0.875   
##   
## Node number 6: 23 observations, complexity param=0.0125523  
## predicted class=benign expected loss=0.2173913 P(node) =0.03367496  
## class counts: 18 5  
## probabilities: 0.783 0.217   
## left son=12 (16 obs) right son=13 (7 obs)  
## Primary splits:  
## Bare.nuclei splits as LRLRL----R, improve=4.968944, (0 missing)  
## Bl.cromatin splits as LLLRR-RR--, improve=4.968944, (0 missing)  
## Cl.thickness splits as LLLLRRRRRR, improve=3.381643, (0 missing)  
## Epith.c.size splits as LLRRRRRRRR, improve=1.992754, (0 missing)  
## Cell.shape splits as LRRRRRRRRR, improve=1.397516, (0 missing)  
## Surrogate splits:  
## Bl.cromatin splits as LLLRR-RR--, agree=0.913, adj=0.714, (0 split)  
## Cl.thickness splits as LLLLLRRRRR, agree=0.870, adj=0.571, (0 split)  
## Mitoses splits as LRLR----R, agree=0.870, adj=0.571, (0 split)  
## Marg.adhesion splits as LLLLLLLRRR, agree=0.826, adj=0.429, (0 split)  
## Normal.nucleoli splits as LLRRLL-L--, agree=0.826, adj=0.429, (0 split)  
##   
## Node number 7: 242 observations, complexity param=0.0125523  
## predicted class=malignant expected loss=0.08264463 P(node) =0.3543192  
## class counts: 20 222  
## probabilities: 0.083 0.917   
## left son=14 (68 obs) right son=15 (174 obs)  
## Primary splits:  
## Cell.size splits as LLLLRRRRRR, improve=5.297663, (0 missing)  
## Bare.nuclei splits as LLRRRRRRRR, improve=4.093695, (0 missing)  
## Cell.shape splits as LLLLRRRRRR, improve=2.958548, (0 missing)  
## Bl.cromatin splits as LLLLRLRRRR, improve=2.838336, (0 missing)  
## Marg.adhesion splits as LLLLLRRRRR, improve=2.754821, (0 missing)  
## Surrogate splits:  
## Cell.shape splits as LLLLRRRRRR, agree=0.789, adj=0.250, (0 split)  
## Epith.c.size splits as LLRRRRRRRR, agree=0.777, adj=0.206, (0 split)  
## Marg.adhesion splits as LRRRRRRRRR, agree=0.744, adj=0.088, (0 split)  
## Bl.cromatin splits as LLRRRRRRRR, agree=0.736, adj=0.059, (0 split)  
## Bare.nuclei splits as RRRRRRLRRR, agree=0.723, adj=0.015, (0 split)  
##   
## Node number 12: 16 observations  
## predicted class=benign expected loss=0 P(node) =0.02342606  
## class counts: 16 0  
## probabilities: 1.000 0.000   
##   
## Node number 13: 7 observations  
## predicted class=malignant expected loss=0.2857143 P(node) =0.0102489  
## class counts: 2 5  
## probabilities: 0.286 0.714   
##   
## Node number 14: 68 observations, complexity param=0.0125523  
## predicted class=malignant expected loss=0.25 P(node) =0.09956076  
## class counts: 17 51  
## probabilities: 0.250 0.750   
## left son=28 (14 obs) right son=29 (54 obs)  
## Primary splits:  
## Bare.nuclei splits as LLRRR-RRRR, improve=7.600529, (0 missing)  
## Cl.thickness splits as LLLLLLRRRR, improve=3.558824, (0 missing)  
## Normal.nucleoli splits as LLRRRLLLRR, improve=2.951389, (0 missing)  
## Marg.adhesion splits as LLLLLRRRRR, improve=2.615385, (0 missing)  
## Bl.cromatin splits as LLLLRLLR-R, improve=1.640351, (0 missing)  
##   
## Node number 15: 174 observations  
## predicted class=malignant expected loss=0.01724138 P(node) =0.2547584  
## class counts: 3 171  
## probabilities: 0.017 0.983   
##   
## Node number 28: 14 observations  
## predicted class=benign expected loss=0.2857143 P(node) =0.0204978  
## class counts: 10 4  
## probabilities: 0.714 0.286   
##   
## Node number 29: 54 observations  
## predicted class=malignant expected loss=0.1296296 P(node) =0.07906296  
## class counts: 7 47  
## probabilities: 0.130 0.870

mytree.pred <- predict(mytree,BreastCancer,type="class")  
  
DTPredict <- as.data.frame(mytree.pred)  
  
colnames(DTPredict) <- "Decision Tree"

#Quadratic Discriminant Analysis  
#library(MASS)  
BreastCancerqda <- lapply(BreastCancer,as.numeric)  
BreastCancerqda$Class <- factor(BreastCancerqda$Class, labels = c("benign", "malignant"))  
  
myqda <- qda(Class ~ ., BreastCancerqda)  
myqda.pred <- predict(myqda, BreastCancerqda)  
table(myqda.pred$class,BreastCancerqda$Class)

##   
## benign malignant  
## benign 422 6  
## malignant 22 233

QDAPredict <- as.data.frame(myqda.pred$class)   
  
colnames(QDAPredict) <- "Quad Discrim Analysis"  
# benign malignant  
# benign 422 6  
# malignant 22 233

#Regularised Discriminant Analysis  
#library(klaR)  
myrda <- rda(Class ~ ., BreastCancer)  
myrda.pred <- predict(myrda, BreastCancer)  
table(myrda.pred$class, BreastCancer$Class)

##   
## benign malignant  
## benign 433 2  
## malignant 11 237

RDAPredict <- as.data.frame(myrda.pred$class)  
colnames(RDAPredict) <- "Reg Discrim Analysis"  
# benign malignant  
# benign 433 2  
# malignant 11 237

#Regularised Discriminant Analysis  
#library(klaR)  
myrda <- rda(Class ~ ., BreastCancer)  
myrda.pred <- predict(myrda, BreastCancer)  
table(myrda.pred$class, BreastCancer$Class)

##   
## benign malignant  
## benign 432 2  
## malignant 12 237

RDAPredict <- as.data.frame(myrda.pred$class)  
colnames(RDAPredict) <- "Reg Discrim Analysis"  
# benign malignant  
# benign 432 2  
# malignant 12 237

#Random Forests  
#library(randomForest)  
myrf <- randomForest(Class ~ .,BreastCancer)  
myrf.pred <- predict(myrf, BreastCancer)  
table(myrf.pred, BreastCancer$Class)

##   
## myrf.pred benign malignant  
## benign 444 0  
## malignant 0 239

RFPredict <- as.data.frame(myrf.pred)  
colnames(RFPredict) <- "Random Forests"  
#myrf.pred benign malignant  
# benign 444 0  
# malignant 0 239  
# 100% Accurate!!!

#Final Results  
#library(dplyr)  
#library(expss)  
  
Results <- as.data.frame(cbind(SVMPredict, NBPredict, NNPredict, DTPredict, QDAPredict,  
 RDAPredict,RFPredict))  
  
ActualResults <- dplyr::select(BreastCancer, Class)  
  
View(Results)  
View(ActualResults)  
  
MajorityPredict <- apply(Results,1,function(x) names(which.max(table(x))))  
  
PredictVsReal <- as.data.frame(cbind(ActualResults, MajorityPredict))  
View(PredictVsReal)  
  
Accuracy <- as.data.frame(ifelse (PredictVsReal$MajorityPredict == PredictVsReal$Class,   
 "True", "False"))  
colnames(Accuracy) <- "Accuracy"

#Majority Selection Validation in Excel  
#library(writexl)  
write\_xlsx(Results,"C:\\Users\\ianmc\\OneDrive\\Documents\\TBANLT 560\\Project Two\\Results.xlsx")  
  
MajorityDF <- as.data.frame(MajorityPredict)  
write\_xlsx(MajorityDF,"C:\\Users\\ianmc\\OneDrive\\Documents\\TBANLT 560\\Project Two\\Majority.xlsx")

#Ensemble Accuracy  
PredictVsReal <- cbind(PredictVsReal, Accuracy)  
View(PredictVsReal)  
  
SumAccuracy <- sum(with(PredictVsReal, Accuracy == "True"))  
SumObservations <- nrow(PredictVsReal)  
EnsembleAccuarcy <- SumAccuracy/SumObservations  
EnsembleAccuarcy

## [1] 0.9824305

#0.9795022  
#Ensemble 97.95% Accuracy!