library(MASS)

data(biopsy)

str(biopsy)

biopsy$ID = NULL

names(biopsy) = c("thick", "u.size", "u.shape", "adhsn", "s.size", "nucl", "chrom", "n.nuc", "mit", "class")

names(biopsy)

biopsy.v2 = na.omit(biopsy)

library(reshape2)

library(ggplot2)

biop.m = melt(biopsy.v2, id.var="class")

ggplot(data=biop.m, aes(x=class, y=value)) + geom\_boxplot() +

facet\_wrap(~variable,ncol = 3)

library(corrplot)

bc = cor(biopsy.v2[ ,1:9]) #create an object of the features

corrplot.mixed(bc)

set.seed(123) #random number generator

ind = sample(2, nrow(biopsy.v2), replace=TRUE, prob=c(0.7, 0.3))

train = biopsy.v2[ind==1,] #the training data set

test = biopsy.v2[ind==2,] #the test data set

str(test) #confirm it worked

table(train$class)

table(test$class)

full.fit = glm(class~., family=binomial, data=train)

summary(full.fit)

confint(full.fit)

exp(coef(full.fit))

library(car)

vif(full.fit)

train$probs = predict(full.fit, type="response")

train$probs[1:5] #inspect the first 5 predicted probabilities

contrasts(train$class)

train$predict = rep("benign", 474)

train$predict[train$probs>0.5]="malignant"

table(train$predict, train$class)

mean(train$predict==train$class)

test$prob = predict(full.fit, newdata=test, type="response")

test$predict = rep("benign", 209)

test$predict[test$prob>0.5]="malignant"

table(test$predict, test$class)

mean(test$predict==test$class)

library(bestglm)

train$y=rep(0,474)

train$y[train$class=="malignant"]=1

head(train[ ,13])

biopsy.cv = train[ ,-10:-12]

head(biopsy.cv)

bestglm(Xy = biopsy.cv, IC="CV", CVArgs=list(Method="HTF", K=10, REP=1), family=binomial)

reduce.fit = glm(class~thick+u.size+nucl, family=binomial, data=train)

train$cv.probs = predict(reduce.fit, type="response")

train$cv.predict = rep("benign", 474)

train$cv.predict[train$cv.probs>0.5]="malignant"

table(train$cv.predict, train$class)

test$cv.probs = predict(reduce.fit, newdata=test, type="response")

test$predict = rep("benign", 209)

test$predict[test$cv.probs>0.5]="malignant"

table(test$predict, test$class)

bestglm(Xy= biopsy.cv, IC="BIC", family=binomial)

bic.fit=glm(class~thick+adhsn+nucl+n.nuc, family=binomial, data=train)

test$bic.probs = predict(bic.fit, newdata=test, type="response")

test$bic.predict = rep("benign", 209)

test$bic.predict[test$bic.probs>0.5]="malignant"

table(test$bic.predict, test$class)

lda.train = train[ ,-11:-15]

lda.train[1:3,]

lda.test = test[ ,-11:-15]

lda.test[1:3,]

lda.fit = lda(class~., data=lda.train)

lda.fit

plot(lda.fit, type="both")

lda.predict = predict(lda.fit)

train$lda = lda.predict$class

table(train$lda, train$class)

lda.test = predict(lda.fit, newdata = test)

test$lda = lda.test$class

table(test$lda, test$class)

mean(test$lda==test$class)

qda.fit = qda(class~., data=lda.train)

qda.fit

qda.predict = predict(qda.fit)

train$qda = qda.predict$class

table(train$qda, train$class)

qda.test = predict(qda.fit, newdata=test)

test$qda = qda.test$class

table(test$qda, test$class)

library(ROCR)

bad.fit = glm(class~thick, family=binomial, data=test)

test$bad.probs = predict(bad.fit, type="response") #save probabilities

pred.full = prediction(test$prob, test$class)

perf.full = performance(pred.full, "tpr", "fpr")

plot(perf.full, main="ROC", col=1)

pred.bic = prediction(test$bic.probs, test$class)

perf.bic = performance(pred.bic, "tpr", "fpr")

plot(perf.bic, col=2, add=TRUE)

pred.bad = prediction(test$bad, test$class)

perf.bad = performance(pred.bad, "tpr", "fpr")

plot(perf.bad, col=3, add=TRUE)

legend(0.6, 0.6, c("FULL", "BIC", "BAD"),1:3)

auc.full = performance(pred.full, "auc")

auc.full

auc.bic = performance(pred.bic, "auc")

auc.bic

auc.bad = performance(pred.bad, "auc")

auc.bad