Assignment 4

Esther Chen

2022-11-03

Dataset

We are using a dataset of malignant and benign cell data which was collected by University of Wisconsin. The dataset consists of 699 rows of 9 variables of various qualities of the cells, and the response is whether the cell is benign or malignant.

Question 1

Create functions that calculate sensitivity, specificity, accuracy, and precision.

#TP / (TP + FN)

sensitivity = function(cm) {  
 return (cm[2,2]/(cm[2,2] + cm[2,1]))  
}  
#TN / (TN + FP)  
specificity = function(cm) {  
 return(cm[1,1]/(cm[1,2] +cm[1,1]))  
}  
#(TP + TN) / (TP + TN + FP + FN)  
accuracy <- function(cm) {  
 return (( cm[1,1] + cm[2,2]) / (cm[1,1] + cm[1,2] + cm[2,1] + cm[2,2]))  
}  
#TP / (TP + FP)  
precision = function(cm) {  
 return (cm[2,2] / (cm[1,2] + cm[2,2]))  
}

Question 2

Create training and testing set from data.

sample <- sample(c(TRUE, FALSE), nrow(cancer), replace = TRUE, prob=c(0.67, 0.33))  
  
#Predictors  
cancer.subset<- cancer[c('Class','Clump\_Thickness', 'Uniformity\_of\_Cell\_Size', 'Uniformity\_of\_Cell\_Shape', 'Marginal\_Adhesion', 'Single\_Epithelial\_Cell\_Size')]  
  
  
#Function to normalize data  
normalize <- function(x) {  
 return ((x - min(x)) / (max(x) - min(x)))   
}

#Normalized dataset  
cancer.subset.n <- as.data.frame(lapply(cancer.subset[2:6], normalize))

#Split into testing and training  
cancer.train <- cancer.subset.n[sample,]  
cancer.test <- cancer.subset.n[-sample,]

#Gather the labels  
train.labels <- cancer.subset[sample, 1, drop = TRUE]  
test.labels <- cancer.subset[-sample, 1, drop = TRUE]

Question 3

Perform a KNN prediction. Produce confusion matrix. Calculate sensitivity, specificity, accuracy, and precision.  
library(class)

#Determining the cluster size

nrow(cancer.train)  
  
sqrt(462)

#Train the KNN model  
knn.21 <- knn(train = cancer.train, test = cancer.test, cl= train.labels, k = 21)  
#Obtain the confusion matrix  
knn.cm <- table(knn.21, test.labels)  
#Calculate statistics  
sensitivity(knn.cm)  
specificity(knn.cm)  
accuracy(knn.cm)  
precision(knn.cm)

The sensitivity is .94, specificity is .97, accuracy .95, and precision .93.

Question 4

Perform a logistic regression. Use 0.7 as the threshold.

#Logistic regression model

logit <- glm(Class ~ Clump\_Thickness + Uniformity\_of\_Cell\_Size + Uniformity\_of\_Cell\_Shape + Marginal\_Adhesion + Single\_Epithelial\_Cell\_Size, data = cancer.subset, family = "binomial")  
#Obtain predictions  
preds <- predict(logit, type="response")  
  
summary(preds)  
#Obtain the confusion matrix  
log.cm <- table(cancer.subset$Class, preds>0.7)  
#Calculate statistics  
sensitivity(log.cm)  
specificity(log.cm)  
accuracy(log.cm)  
precision(log.cm)

The sensitivity is .90, specificity is .98, accuracy is .95, and precision .96.

Question 5

Calculate a decision tree.

library(rpart)

#Train the tree

cancer.tree = rpart(Class ~ Clump\_Thickness + Uniformity\_of\_Cell\_Size + Uniformity\_of\_Cell\_Shape + Marginal\_Adhesion + Single\_Epithelial\_Cell\_Size, data = cancer.subset[sample,], control = rpart.control(maxdepth=3), method="class")  
#Obtain predictions  
tree.pred = predict(cancer.tree, cancer.subset[-sample,], type ='class')  
  
tree.cm <- table(tree.pred, test.labels)  
  
sensitivity(tree.cm)  
specificity(tree.cm)  
accuracy(tree.cm)  
precision(tree.cm)

The sensitivity is .91, specificity .96, accuracy .94, and precision .93.

Question 6

Is there a method that is better than the others?

In the context of cancer cell detection, we are interested in specificity and sensitivity. A high specificity means less false positives, and a high sensitivity test is less likely to produce a false negative. Proper detection is important because there are costs associated with tests. However, with machine learning models the amount of resources expended is probably less than with traditional biological testing.

The KNN model we trained has a sensitivity of 94% and specificity of 97%. This means that 94% of malignant cells were correctly classed by the KNN, and 97% of benign cells are correctly classed.

The logistic regression model correctly classified 90% of malignant cells and 98% of benign cells.

The decision tree correctly classified 91% of malignant cells and 96% of benign cells.

I think that the KNN model generally has the best performance because it has the highest sensitivity and specificity, so if I had to select one model with the best performance I would go with the KNN. Though the logistic regression model has a slight edge in specificity, it has the lowest sensitivity of the three, meaning that it more often produces false positives.