Eric_Hirsch_621_Assignmrnt_2

Eric Hirsch

3/2/2022

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.5

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.5
```

```r
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.0.5

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1
## v purrr   0.3.4

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

Read the data

```
dfB <- read.csv("D:\\RStudio\\CUNY_621\\Assignment 2\\classification-output-data.csv")
dfB <- dfB %>%
  select(class, scored.class, scored.probability)
```

Homework 2

1. Review the raw confusion matrix -

```
dfB1 <- dfB %>%
  select(-scored.probability)

table(dfB1)
```

```
##      scored.class
## class   0   1
##     0 119   5
##     1  30  27
```

```
dfTable <- as.data.frame(table(dfB1))
```

The raw confusion matrix shows us the actual values as rows and the scored values as columns. The upper left-lower right diagonal contains correct predictions, while the opposite diagonal contains incorrect predictions.

We write a function to return accuracy (correct predictions over all predictions)

```
TableToDf <- function (df)
{
  df1 = as.data.frame(table(df))
  li <- vector(mode = "list", length = 4)
  names(li) <- c("TP", "TN", "FP", "FN")

  li$TP <- df1$Freq[4]
  li$TN <- df1$Freq[1]
  li$FP <- df1$Freq[3]
  li$FN <- df1$Freq[2]

  #if(is.na(li$TP)) {
  #  li$TP=1
```

```
  #}
  #  if(is.na(li$TN)) {
  #  li$TN=1
  #  }
  #  if(is.na(li$FP)) {
  #  li$FP=1
  #  }
  #  if(is.na(li$FN)) {
  #  li$FN=1
  #  }


  return(li)

}

x <- TableToDf(dfB1)
```

2. Accuracy Function

```
Accuracy <- function(df) {

li <- TableToDf(df)
Accuracy <- (li$TP + li$TN)/(li$TP + li$TN + li$FP + li$FN)

return (Accuracy)

}

print(Accuracy(dfB1))
```

```
## [1] 0.8066298
```

3. Classification Error Rate Function

Classification Error Rate (incorrect predictions over all predicitons:

```
ClassificationErrorRate <- function(df) {

li <- TableToDf(df)
Error <- (li$FP + li$FN)/(li$TP + li$TN + li$FP + li$FN)

return (Error)


}
print(ClassificationErrorRate(dfB1))
```

```
## [1] 0.1933702
```

Accuracy + error rate = 1

```
print(as.numeric(ClassificationErrorRate(dfB1)) + as.numeric(Accuracy(dfB1)))
```

```
## [1] 1
```

4. Precision Function (True positives/All who tested positive):

```
Precision <- function(df) {

li <- TableToDf(df)
Precision <- (li$TP)/(li$TP + li$FP)

return(Precision)

}
print(Precision(dfB1))
```

```
## [1] 0.84375
```

5. Sensitivity/Recall Function (True positives/All who are positive):

```
Sensitivity <- function(df) {

li <- TableToDf(df)


  Sensitivity <- (li$TP)/(li$TP + li$FN)



return(Sensitivity)

}
print(Sensitivity(dfB1))
```

```
## [1] 0.4736842
```

6. Specificity Function (True negatives/All who are negative):

```
Specificity <- function(df) {

li <- TableToDf(df)
Specificity <- (li$TN)/(li$TN + li$FP)

return(Specificity)

}
print(Specificity(dfB1))
```

```
## [1] 0.9596774
```

7. F1 score Function:

```
F1 <- function(df) {


p <- Precision(dfB1)
s <- Sensitivity(dfB1)

F1 <- (2*p*s)/(p+s)

return(F1)

}

print(F1(dfB1))
```

```
## [1] 0.6067416
```

8. Bounds on the F1 score

Precision (P) and Sensitivity (S) themselves are always <= 1 because their denominators are a simple sum which include the numerators. Therefore, $P*S < P$ and $P*S < S$. Therefore, $2PS < P + S$.

9. ROC curve Function

```
ROC <- function(df) {

list1 <- list()

dfNew <- data.frame(class = numeric(),
 scored.class = numeric())

dfFinal <- data.frame(Specificity = numeric(),
 Sensitivity = numeric(), Area = numeric(), Width=numeric())

spec_prev = 0
sens_prev = 0

for (i in 1:100) {
  for (j in 1:length(df$class))
  {

  dfNew[j, 1] = df$class[j]
  dfNew[j, 2]  <- ifelse(df$scored.probability[j] > i/100, 1, 0)
  }

  sens <- Sensitivity(dfNew)
```

```
  spec <- 1 - Specificity(dfNew)

  width = spec_prev - spec
  ave_sens = (sens + sens_prev)/sens

  dfFinal[i,1] <- spec
  dfFinal[i,2] <- sens
  dfFinal[i,3] <- sens*width
  dfFinal[i,4] <- width

  spec_prev = spec
  sens_prev = sens
}

dfFinal <- na.omit(dfFinal)

AUC <- sum(dfFinal$Area)/sum(dfFinal$Width)
print (paste("AUC:", AUC))

g <- ggplot(dfFinal, aes(x=Specificity, y=Sensitivity)) +
             geom_line()

x <- list(g, dfFinal$Area)

return (x)
}
```

10. Provide all metrics

```
print(paste("Accuracy: ", Accuracy(dfB1)))
```

```
## [1] "Accuracy:  0.806629834254144"
```

```
print(paste("Classification Error: ", ClassificationErrorRate(dfB1)))
```

```
## [1] "Classification Error:  0.193370165745856"
```

```
print(paste("Precision: ", Precision(dfB1)))
```

```
## [1] "Precision:  0.84375"
```

```
print(paste("Sensitivity: ", Sensitivity(dfB1)))
```

```
## [1] "Sensitivity:  0.473684210526316"
```

```
print(paste("Specificity:", Specificity(dfB1)))
```

```
## [1] "Specificity: 0.959677419354839"
```

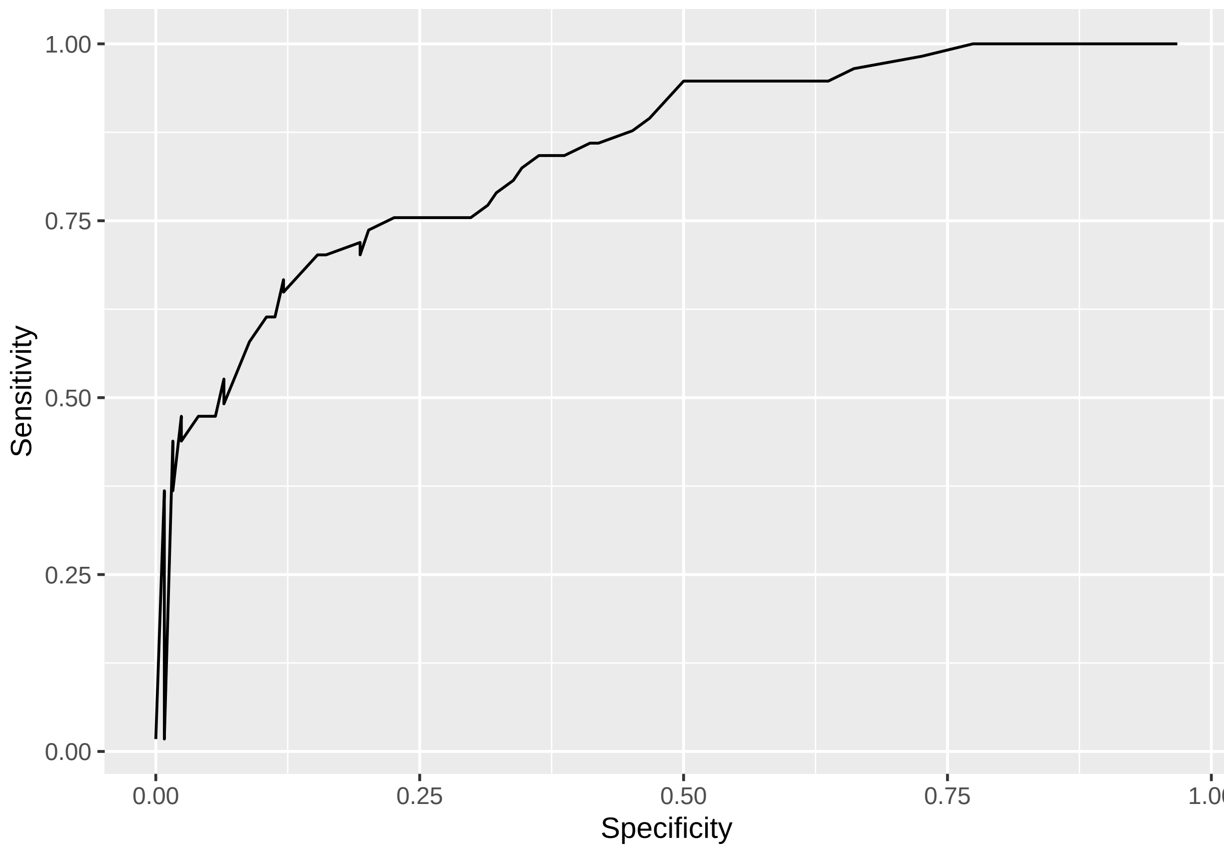
```
print(paste("F1:", F1(dfB1)))
```

```
## [1] "F1: 0.606741573033708"
```

```
x1 <- ROC(dfB)
```

```
## [1] "AUC: 0.839930404523706"
```

```
print(x1[[1]])
```



11. The caret package provides statistics that match ours.

```
dfB1$class <- as.factor(dfB1$class)
dfB1$scored.class <- as.factor(dfB1$scored.class)

confusionMatrix(data = dfB1$scored.class, reference = dfB1$class)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
```

```
##          0 119  30
##          1   5  27
##
##                Accuracy : 0.8066
##                  95% CI : (0.7415, 0.8615)
##     No Information Rate : 0.6851
##     P-Value [Acc > NIR] : 0.0001712
##
##                   Kappa : 0.4916
##
##  Mcnemar's Test P-Value : 4.976e-05
##
##             Sensitivity : 0.9597
##             Specificity : 0.4737
##          Pos Pred Value : 0.7987
##          Neg Pred Value : 0.8438
##              Prevalence : 0.6851
##          Detection Rate : 0.6575
##    Detection Prevalence : 0.8232
##       Balanced Accuracy : 0.7167
##
##        'Positive' Class : 0
##
```

12. The pROC can be used to create an ROC curve. This curve matches our own.

```
roc(class ~ scored.probability, dfB)
```

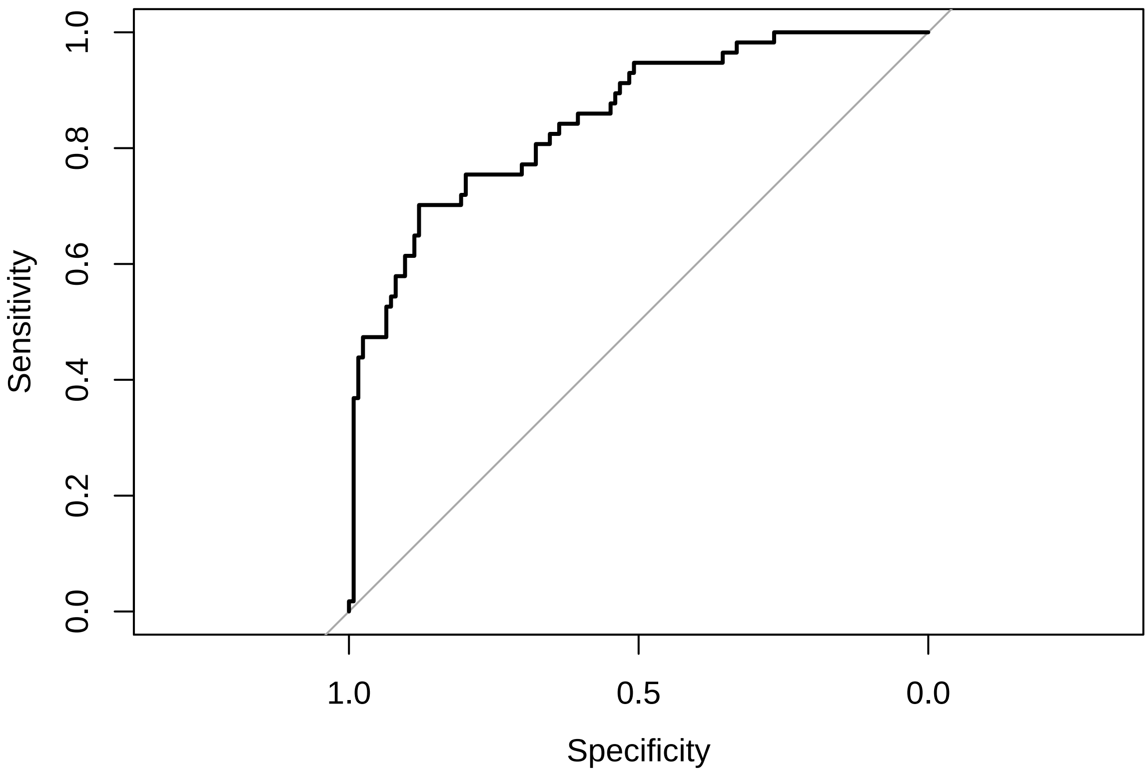
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
##
## Call:
## roc.formula(formula = class ~ scored.probability, data = dfB)
##
## Data: scored.probability in 124 controls (class 0) < 57 cases (class 1).
## Area under the curve: 0.8503
```

```
roc1 <- roc(dfB$class,
            dfB$scored.probability, plot=TRUE)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity