Seminar 4

Yufeng Deng Yuanqing Wang

**Task 1**

First, just have a brief look at the dataset:

labels <- c('Positive','Negative')

diagnosis <- c(sum(data$Patient\_Diagnosis[which(data$Patient\_Diagnosis==1)]),

nrow(data)-sum(data$Patient\_Diagnosis[which(data$Patient\_Diagnosis==1)]))

rater1 <- c(sum(data$Rater1[which(data$Rater1==1)]),

nrow(data)-sum(data$Rater1[which(data$Rater1==1)]))

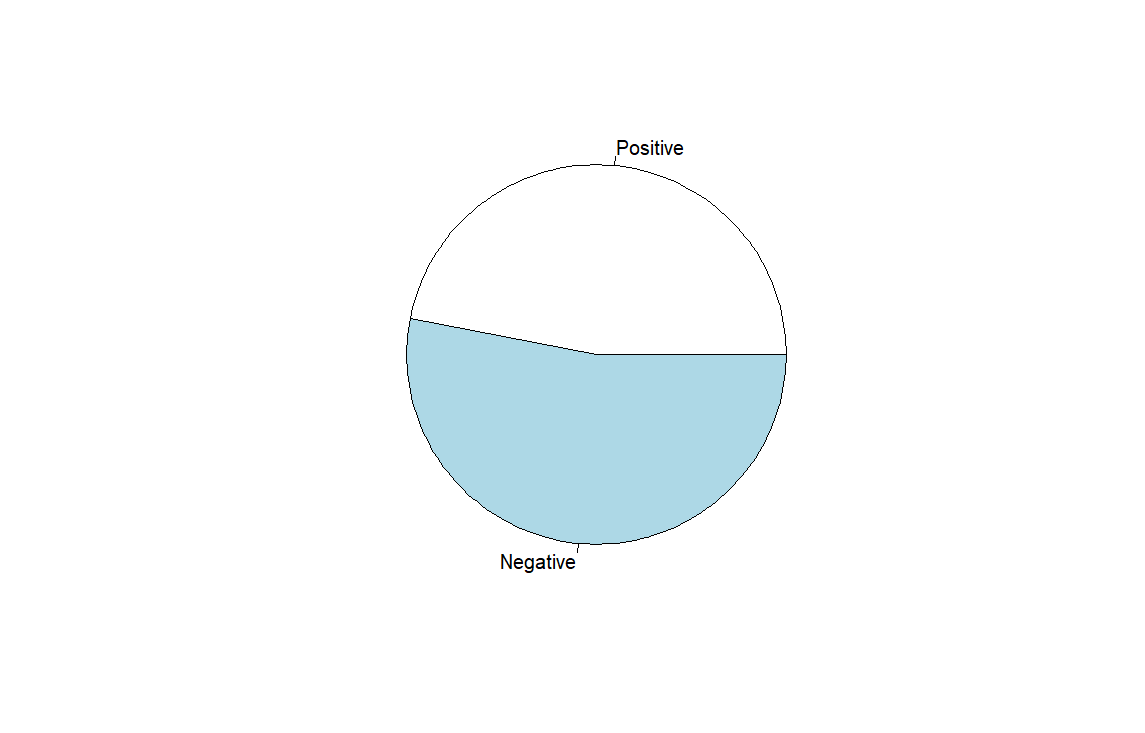
rater2 <- c(sum(data$Rater2[which(data$Rater2==1)]),

nrow(data)-sum(data$Rater1[which(data$Rater1==1)]))

pie(diagnosis,labels)

pie(rater1,labels)

pie(rater2,labels)

 图表

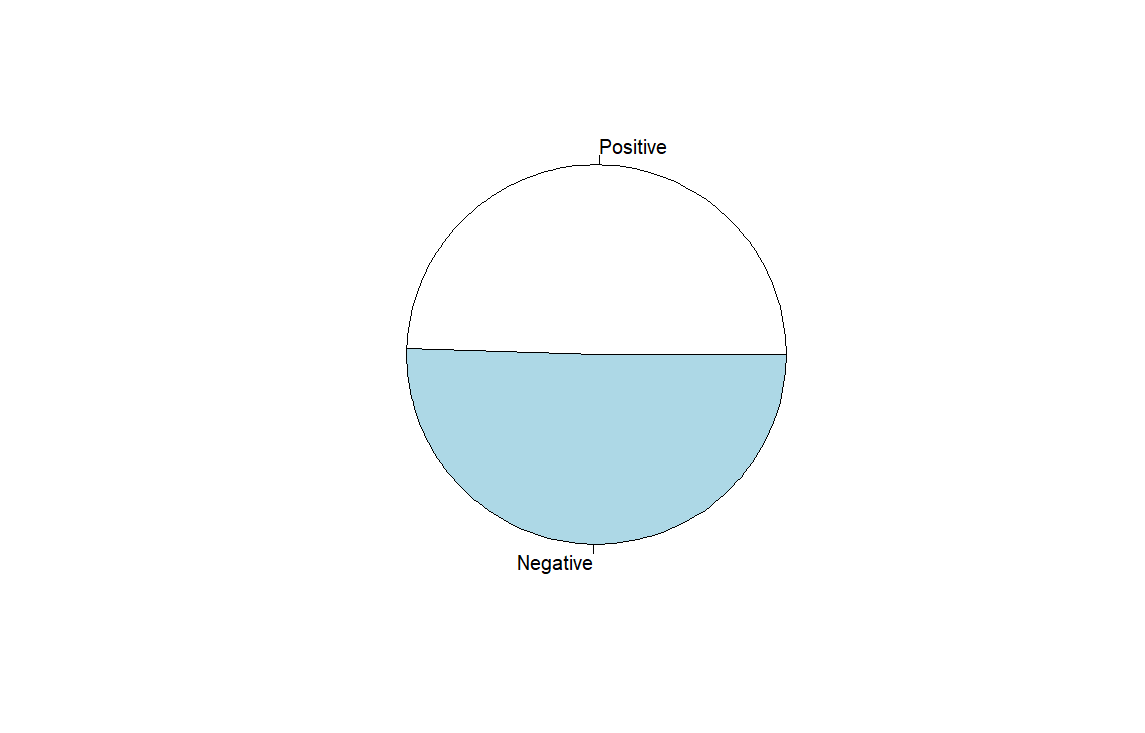
低可信度描述已自动生成 

Figure 1.1 Distribution by the label. Diagnosis - Rater1 - Rater2

For Agreement analysis, confusion matrixes are needed to be provided.

# between diagnosis and raters

test1 <- get\_Metrics(data$Patient\_Diagnosis,data$Rater1)

print(test1$confusion\_matrix)

test2 <- get\_Metrics(data$Patient\_Diagnosis,data$Rater2)

print(test2$confusion\_matrix)

# between two raters

test\_agree <- get\_Metrics(data$Rater1,data$Rater2)

agree\_matrix <- test\_agree$confusion\_matrix

Table 1.1 Confusion Matrix between Diagnosis and Rater 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Pos R1 | Neg R1 | Total |
| Diseased | 43 | 4 | 47 |
| Healthy | 4 | 49 | 53 |
| Total | 47 | 53 | 100 |

Table 1.2 Confusion Matrix between Diagnosis and Rater 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Pos R2 | Neg R2 | Total |
| Diseased | 43 | 4 | 47 |
| Healthy | 4 | 49 | 53 |
| Total | 47 | 53 | 100 |

Table 1.3 Confusion Matrix between Rater 1 and Rater 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Pos R2 | Neg R2 | Total |
| Pos R1 | 39 | 8 | 47 |
| Neg R1 | 13 | 40 | 53 |
| Total | 52 | 48 | 100 |

To test the agreement of Rater 1 and Rater 2, we calculated the kappa value:

# kappa value

data\_2col <- data.frame(Rater1 = data$Rater1,Rater2 = data$Rater2)

result <- kappa2(data\_2col)

print(result)

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 100

Raters = 2

Kappa = 0.581

z = 5.84

p-value = 5.25e-09

The kappa value is 0.581, which means that the two Raters have moderate agreement. It’s too low in medical field, which always requires kappa value higher than 0.7.

To test if there is a systematic difference between Rater 1 and Rater 2, we can apply the McNemar’s test:

# McNemar's test

agree\_matrix\_2dim <- get\_2dim\_matrix(agree\_matrix)

result <- mcnemar.test(agree\_matrix\_2dim)

print(result)

McNemar's Chi-squared test with continuity correction

data: agree\_matrix\_2dim

McNemar's chi-squared = 0.7619, df = 1, p-value = 0.3827

Since the p-value is higher than 0.05, it can be considered as there is no systematic difference between Rater 1 and Rater 2.

To test differences in sensitivity & specificity between the two Raters, we still need to apply the McNemar’s test:

# Test difference in sensitivity

subpop\_diseased <- data.frame(data[which(data$Patient\_Diagnosis == 0),])

test\_subdis <- get\_Metrics(subpop\_diseased$Rater1,subpop\_diseased$Rater2)

subdis\_matrix\_2dim <- get\_2dim\_matrix(test\_subdis$confusion\_matrix)

result <- mcnemar.test(subdis\_matrix\_2dim)

print(result)

McNemar's Chi-squared test with continuity correction

data: subdis\_matrix\_2dim

McNemar's chi-squared = 1.2308, df = 1, p-value = 0.2673

Table 1.4 subpopulation of diseased subjects

|  |  |  |  |
| --- | --- | --- | --- |
|  | Pos R2 | Neg R2 | Total |
| Pos R1 | 0 | 4 | 4 |
| Neg R1 | 9 | 40 | 49 |
| Total | 9 | 44 | 53 |

# For specificity, look at the subpopulation of healthy subjects.

subpop\_healthy <- data.frame(data[which(data$Patient\_Diagnosis == 1),])

test\_subheal <- get\_Metrics(subpop\_healthy$Rater1,subpop\_healthy$Rater2)

subheal\_matrix\_2dim <- get\_2dim\_matrix(test\_subheal$confusion\_matrix)

result <- mcnemar.test(subheal\_matrix\_2dim)

print(result)

McNemar's Chi-squared test

data: subheal\_matrix\_2dim

McNemar's chi-squared = 0, df = 1, p-value = 1

Table 1.5 subpopulation of healthy subjects

|  |  |  |  |
| --- | --- | --- | --- |
|  | Pos R2 | Neg R2 | Total |
| Pos R1 | 39 | 4 | 43 |
| Neg R1 | 4 | 0 | 4 |
| Total | 43 | 4 | 47 |

According to the results, we can conclude that there are no differences in Specificity and Sensitivity between Rater 1 and Rater 2.

Two functions are used:

get\_Metrics <- function(diagnosis,rater){# generate the confusion matrix and metrics

TP <- sum(diagnosis == 1 & rater == 1)

FP <- sum(diagnosis == 0 & rater == 1)

TN <- sum(diagnosis == 0 & rater == 0)

FN <- sum(diagnosis == 1 & rater == 0)

sumr1 <- sum(TP,FN)

sumr2 <- sum(FP,TN)

sumc1 <- sum(TP,FP)

sumc2 <- sum(FN,TN)

sum <- sum(TP,FP,FN,TN)

table <- matrix(c(TP,FN,sumr1,FP,TN,sumr2,sumc1,sumc2,sum),

nrow=3,

byrow=TRUE)

colnames(table) <- (c('PosTest','NegTest','Total'))

rownames(table) <- (c('Diseased','Healthy','Total'))

# df <- data.frame(table)

output <- list(confusion\_matrix=table,

Sensitivity=TP/sumr1,

Specificity=TN/sumr2,

PPV=TP/sumc1,

NPV=TN/sumc2,

Accuracy=(TP+TN)/sum)

return (output)

}

get\_2dim\_matrix <- function(confmatrix){# turn a 3 dim matrix to a 2 dim one

table <- matrix(c(confmatrix[1,1],confmatrix[1,2],confmatrix[2,1],confmatrix[2,2]),

nrow = 2,

byrow = TRUE,

dimnames = list("Rater1" = c("Positive", "Negative"),

"Rater2" = c("Positive", "Negative")))

return(table)

}

**Task4**

To analyze the performance of the model, we can calculate the AUC of its ROC curve.

roc\_obj <- roc(data$labels\_obs,data$prob\_pred)

AUC <- roc\_obj$auc

print(AUC)

Area under the curve: 0.8533

The AUC of ROC curve is 0.8533, which can be considered clinically useful1.

To determine the threshold, we tried to use the features of ROC curve:

# Using ROC to determine the threshold

roc\_data <- data.frame(

Spec = roc\_obj$specificities, # Specificity

Sens = roc\_obj$sensitivities, # Sensitivities

Thresholds = roc\_obj$thresholds

)

# draw ROC

ggplot(roc\_data, aes(x = Spec, y = Sens)) +

geom\_line(color = "blue", size = 1) +

geom\_abline(slope = 1, intercept = 1, linetype = "dashed", color = "red") + # 参考线

labs(

title = "ROC Curve",

x = "Specificity",

y = "Sensitivity"

) +

theme\_minimal() +

scale\_x\_reverse()

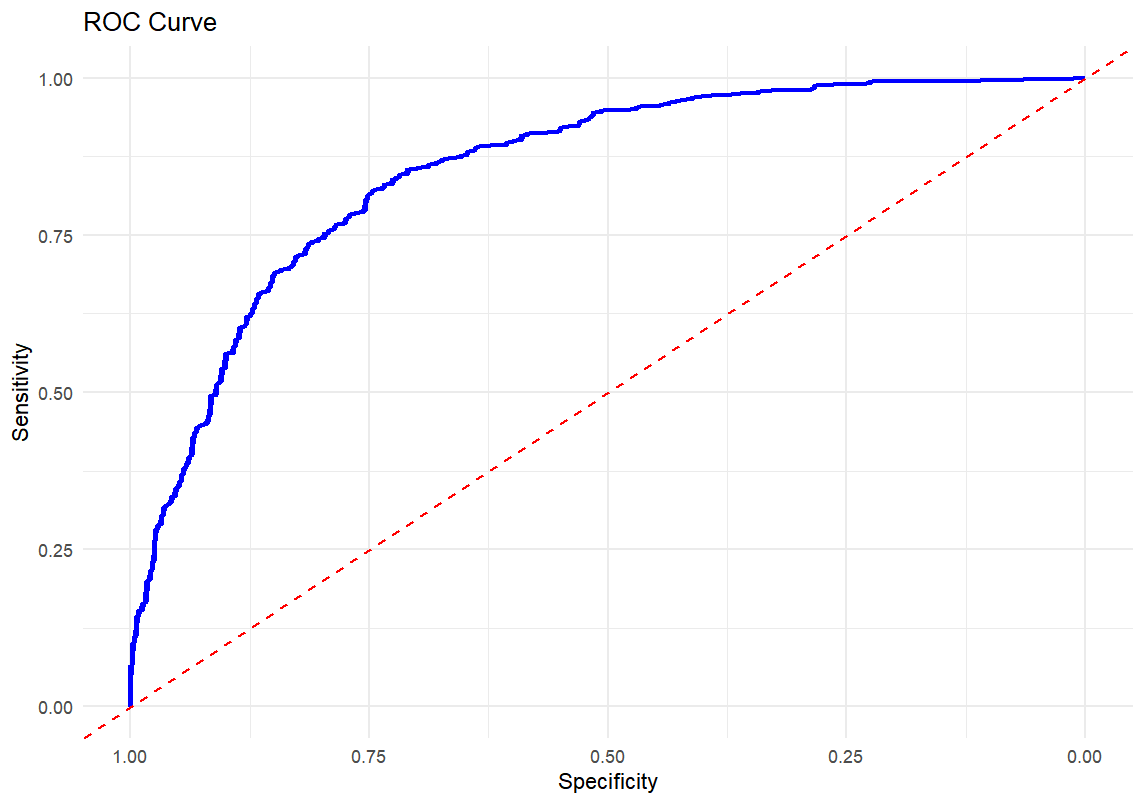


Figure 4.1 ROC curve

To find the optimal threshold, we assume that the threshold which can obtain the maximum of (Specificity + Sensitivity) is the optimal threshold.

optimal\_idx <- which.max(roc\_obj$sensitivities + roc\_obj$specificities)

optimal\_threshold <- roc\_obj$thresholds[optimal\_idx]

print(paste("Optimal Threshold:", optimal\_threshold))

[1] "Optimal Threshold: 0.602483458602781"

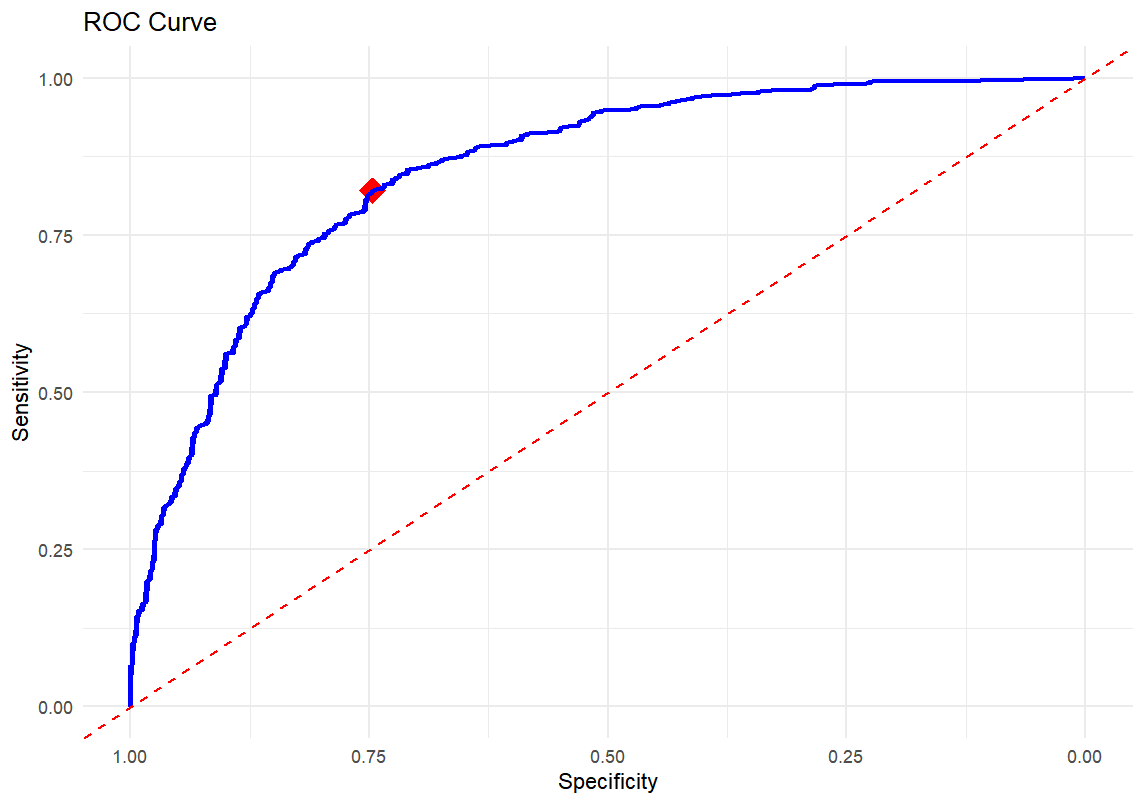


Figure 4.2 ROC curve with the optimal point

We can evaluate this way of choosing threshold by calculating several metrics:

binary\_pred = numeric(nrow(data))

for (i in data$X){

if (data$prob\_pred[i] >= optimal\_threshold){

binary\_pred[i] = 1

}

else{

binary\_pred[i] = 0

}

}

CM <- confusionMatrix(factor(binary\_pred),factor(data$labels\_obs))

print(CM$overall)

print(CM$byclass)

Table 4.1 Metrics of the optimal threshold

|  |  |
| --- | --- |
| Metrics | Value |
| Accuracy | 0.784 |
| Kappa | 0.568 |
| Sensitivity | 0.821 |
| Specificity | 0.746 |

Since we know that lowering FPR (1-Specificity) is the most important thing in medical diagnosis, we need to try some methods to increase Specificity. Here we force the Specificity to be 0.9 and 0.95.

# determine target

target\_specificity <- 0.9

# filter

filtered <- roc\_data[roc\_data$Spec > target\_specificity, ]

# find the max Sensitivity in this situation

threshold\_90 <- filtered$Threshold[which.max(filtered$Sens)]

print(paste("Threshold 0.9 Specificity:", threshold\_90))

[1] "Threshold 0.9 Specificity: 0.709243576435695"

[1] "Threshold 0.95 Specificity: 0.781006462207699"

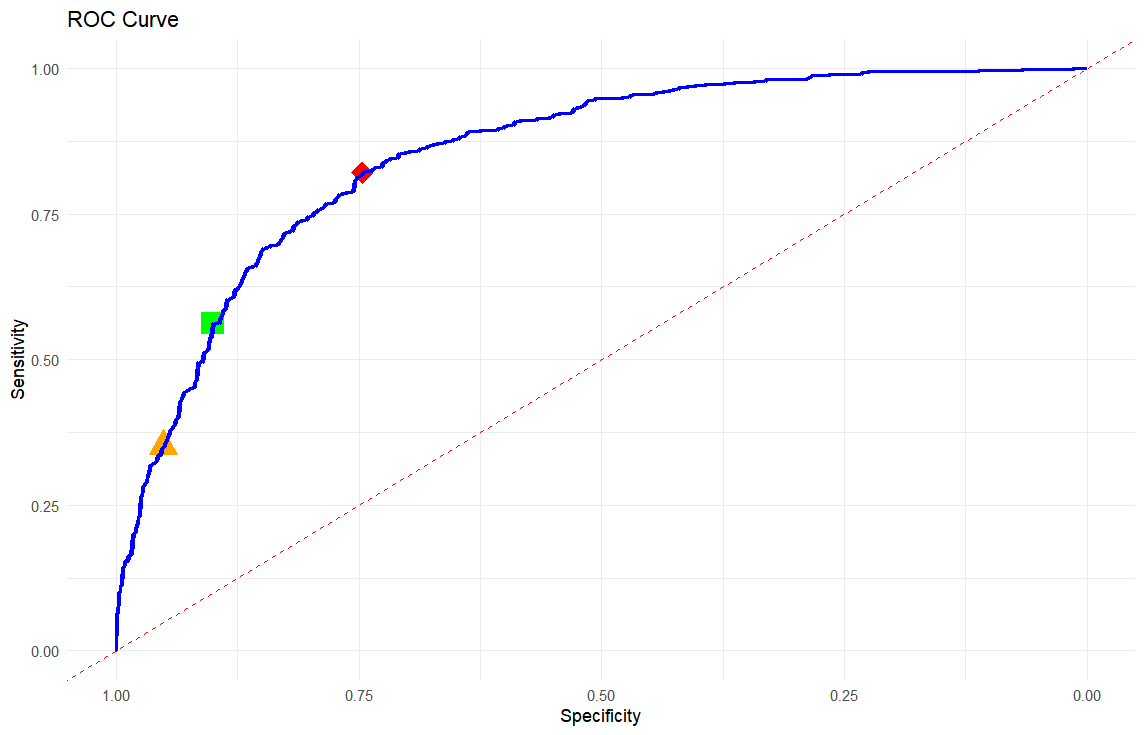


Figure 4.3 Three different thresholds (Specificity of orange, green: 0.95, 0.90)

We can evaluate them by metrics:

Table 4.2 Metrics of the three different thresholds

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | Optimal | 0.90 Specificity | 0.95 Specificity |
| Accuracy | 0.784 | 0.729 | 0.647 |
| Kappa | 0.568 | 0.461 | 0.300 |
| Sensitivity | 0.821 | 0.562 | 0.351 |
| Specificity | 0.746 | 0.901 | 0.951 |

It is obvious that when Specificity increases, the accuracy will decrease. It is hard to say which threshold is the best one.

**References**

1. Çorbacıoğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turk J Emerg Med*. 2023;23(4):195-198. doi:10.4103/tjem.tjem\_182\_23