**Exercise – SVM 2**

Begin by importing the attached files for week 6 into R, here is my process, keep in mind that rd1, rd2, srd1, and srd2 are just x, y data points: *code in blue, comments in green & outputs in black or as screen shots*.

> library(e1071)

> getwd() #confirming my location

[1] "/Users/homecomputer"

> srd1<-read.csv("srd1.csv", header=FALSE) #importing the srd1 file

> srd2<-read.csv("srd2.csv", header=FALSE) #importing the srd2 file

> y1<-rep(0,1000) # creates a vector of 1,000 zeros

> y2<-rep(1,1000) # creates a vector of 1,000 ones

> y <-factor(c(y1,y2)) # concatenates y1 & y2 and makes the results into factors

> x<-rbind(srd1,srd2) # create a 2,000 x 2 matrix using row bind

> model <-svm(x,y,type="C-classification") # trains

> p <-predict(model,x) # run this to generate predictions

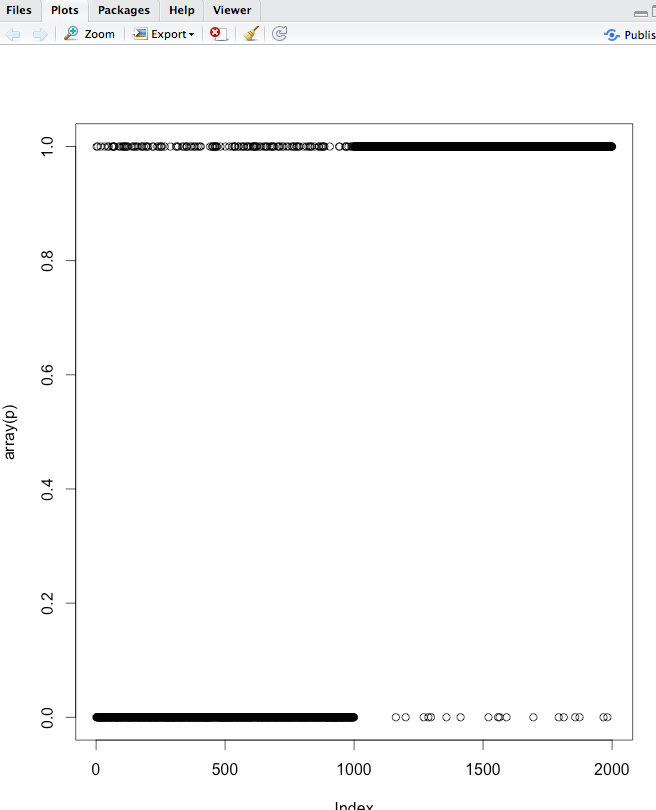
> summary(p) # the model predicted 883 0’s and 1,117 1’s

0 1

883 1117

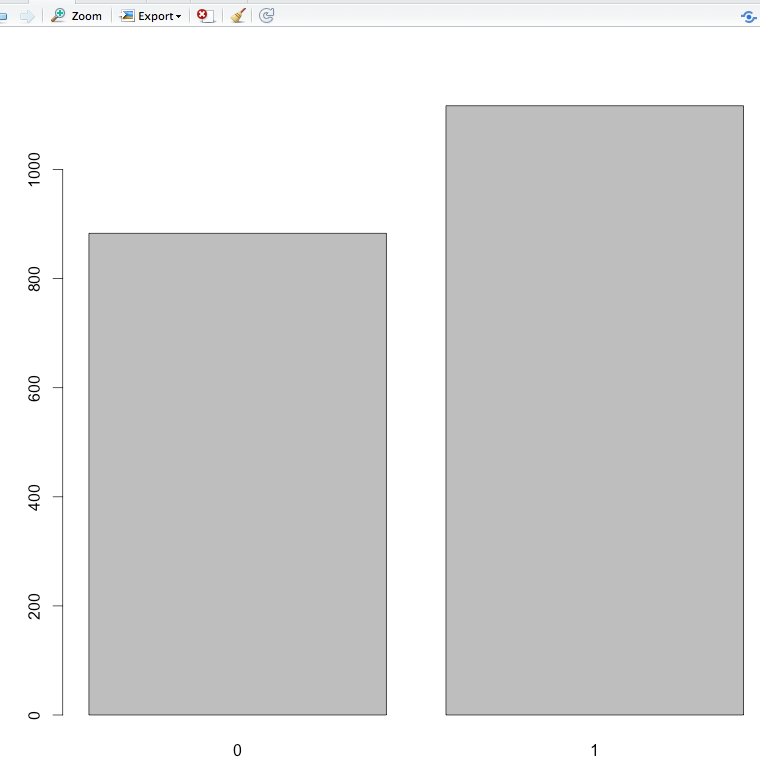
> plot(array(p)) # changes the predictions to an array & plots them

here is the output:



The above is not very perceptive as everything is so tightly packed, so we will see if there are other ways to display the predictions. You can get pretty fancy here, or you can make a simple bar chart:

The command > plot(p) will give you a simple bar chart:



Even more useful view:

> f = plot(srd1, col="red")

> points(srd2, col="blue")



The next section is a modified example used by a student in a prior class using the ScatterPlot3D Library. Here the ambiguous points become obvious.

> # install.packages(scatterplot3d) # initial install, do not need to run if you already have scatterplot3d installed

> library(scatterplot3d) # prompt scatterplot3d

> i <- as.matrix(as.numeric(predict(model, x[1 : 1000, ])) - 1) # Predicts the first thousand

> j <- as.matrix(as.numeric(predict(model, x[1001 : 2000, ]))) - 2 # Predicts the second thousand

> j <- abs(j) # Converts the -1s to 1s in the second thousand

> z <- c(i, j) # Concatenates i and j into a 2000 x 1 matrix

> scatterplot3d(x[,1], x[,2], z, pch=1, highlight.3d=TRUE, type="h", main="SVM\_2", xlab = "x", ylab = "y", zlab = "ambiguity", angle = 65) # Creates the scatterplot with vertical lines



#3. Train the svm model using the 1,000 length data and then predict the 10,000 length data…Discuss some ideas of why you think cross validation are important.

> rd1 <- read.csv("rd1.csv",header=FALSE)

> rd2 <- read.csv("rd2.csv", header=FALSE)

> x2<-rbind(rd1,rd2)

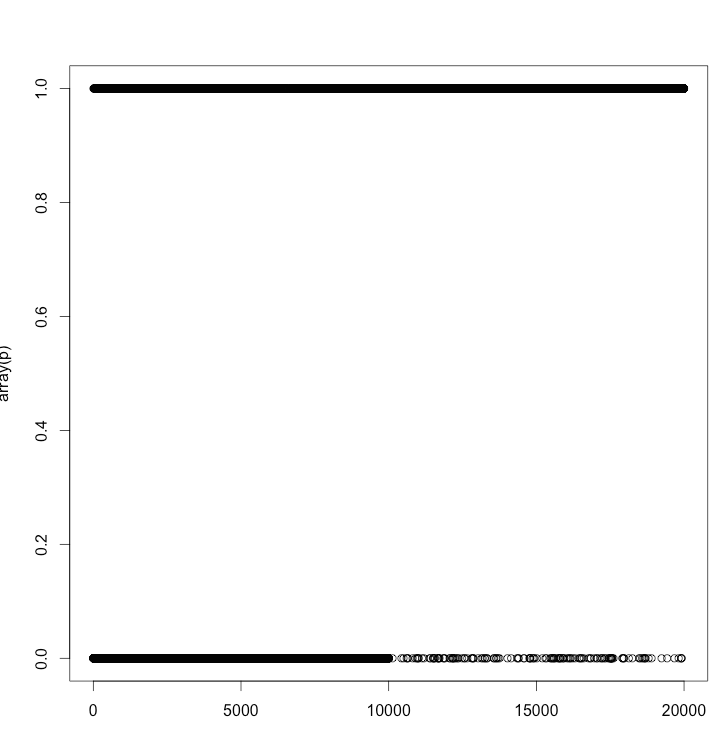
> p <-predict(model,x2)

> summary(p) # the model predicted 8,745 0’s and 11,255 1’s

0 1

8745 11255

> plot(array(p))



> z = plot(rd1, col="red")

> points(rd2, col="blue")



Part 3 is asking us to use the 1,000 length data to predict the 10,000 length data.

When we create a new model - in this case we are trying to predict the 10,000 length data - and we use the same data for training and testing, we need to limit problems like over-fitting. A goal of cross validation is to use a sample of the total dataset for training.

#4. Any way to remove the ambiguity? Share some thoughts.