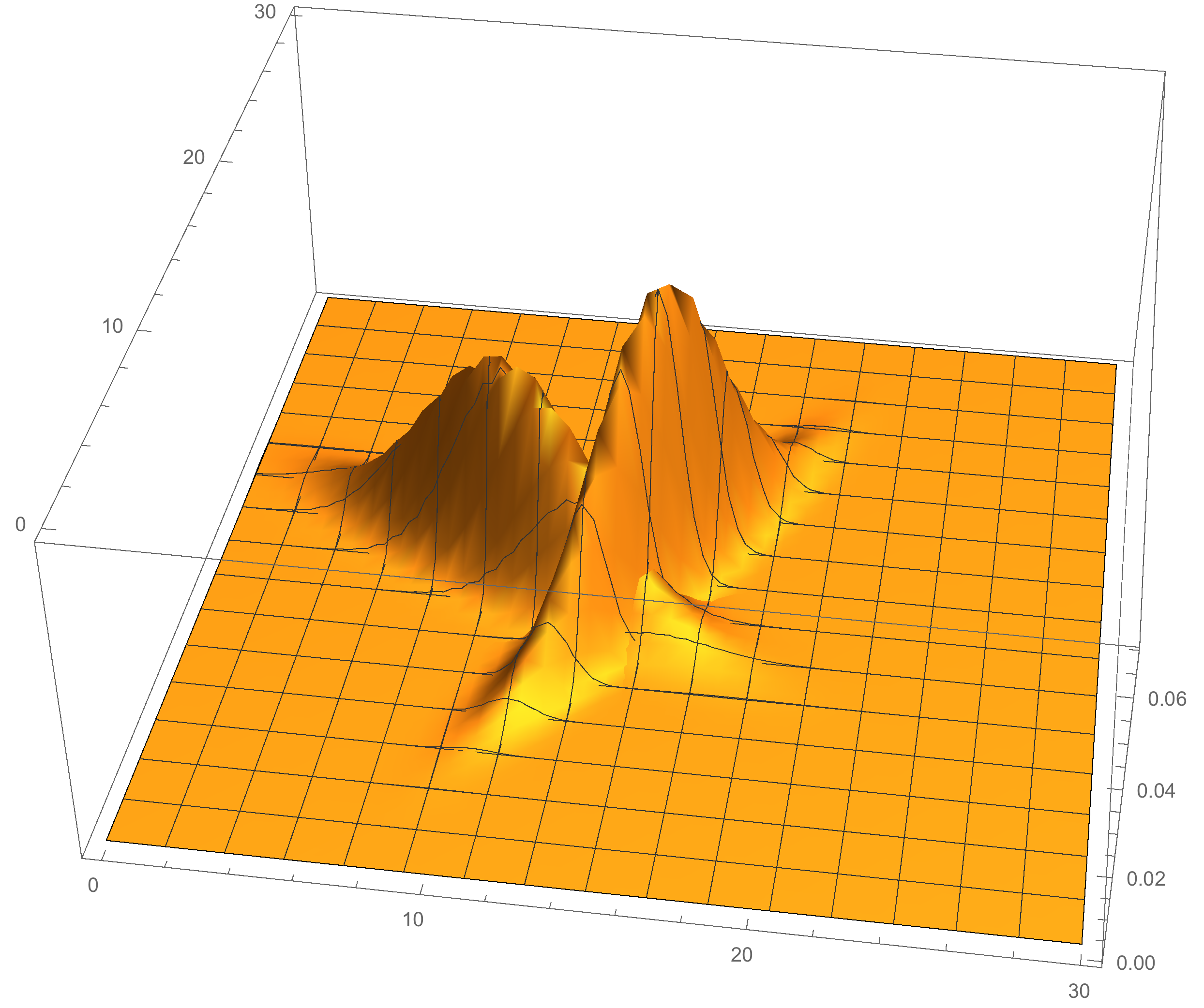
SVM Assignment 2

For this assignment we will examine what happens in a classification problem when the class attributes overlap. Suppose that we have two multinormal probability distributions that look as follows.



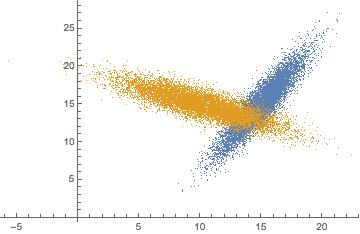
Each distribution is relatively narrow. They overlap in the area surrounding of x=14 and y=14. Let’s draw a random sample from each distribution, which will be a set of 2D points. For the distribution that slopes down to the right the first few random points are

{11.9015, 13.5505}, {8.49426, 15.6236}, {10.8917, 16.0151}, {7.97075, 16.8332}

For the distribution that slopes up to the right the first few points are

{11.1335, 10.6196}, {14.3089, 16.1996}, {14.971, 17.0967}, {20.575, 24.7863}

If we plot random points from created from each probability distribution we see

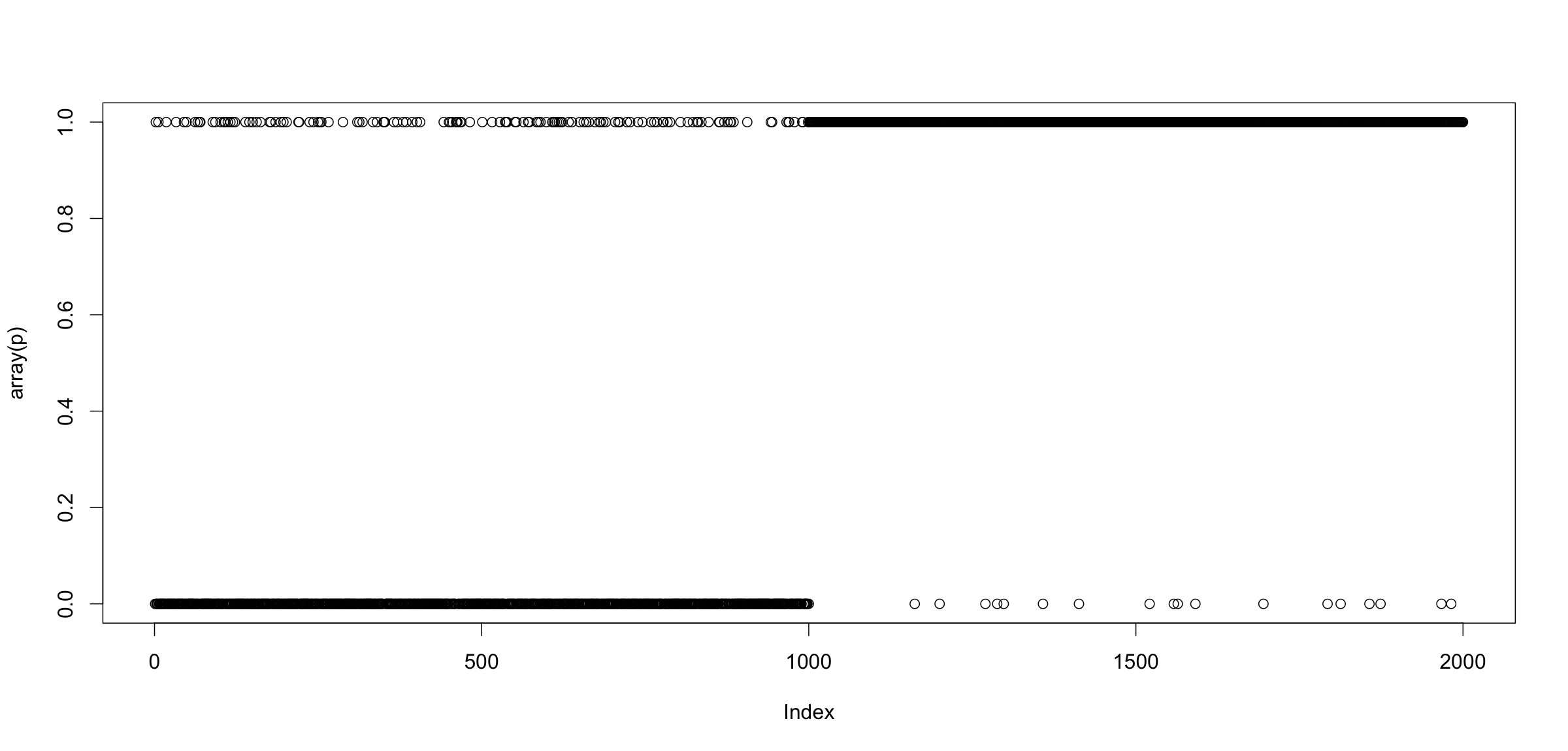


Here we see the overlap in the data sampled from the two distributions again around the x=14 and y=14 coordinates.

Included you will find two sets of data. The data srd1 and srd2 are samples of size 1000 from each distribution. The data rd1 and rd2 are samples of size 10,000 from each distribution. Let’s first look at the srd data. Examine the following R code.

1. library(e1071)
2. srd1 <- read.csv("srd1.csv",header=FALSE)
3. srd2 <- read.csv("srd2.csv",header=FALSE)
4. y1 <- rep(0,1000)
5. y2 <- rep(1,1000)
6. y <- factor(c(y1,y2))
7. x <- rbind(srd1,srd2)
8. model <- svm(x,y,type="C-classification")
9. p <- predict(model,x)
10. plot(array(p))

Lines 2 and 3 read the csv files for srd1 and srd2. Line 4 creates class labels 0 for srd1. Line 5 creates class labels 1 for srd2. Line 6 concatenates y1 and y2 and makes the results into factors for the svm model. Line 7 concatenates the two data frames, srd1 and srd2. Line 8 runs to svm model. Line 9 generates the predictions and line 10 canges the predictions to an array and plots them. This plot looks like.



Ideally, the model would predict 0 for the first 1000 values and 1 for the last 1000 values. However, you see a number of 1’s predicted in the first 1000 numbers and a fewer number of 0’s predicted in the second set of 1000 numbers.

1. Recreate the R part of this experiment using your computer.
2. See if you can find other ways to display the predictions.
3. Train the svm model using the 1000 length data and then predict the 10,000 length data. Compare the predictions with the data. See the ideas of cross validation.
4. Is there anyway to remove the ambiguity between the two classes give the existing data? What about if you could add additional measurements for each data point?