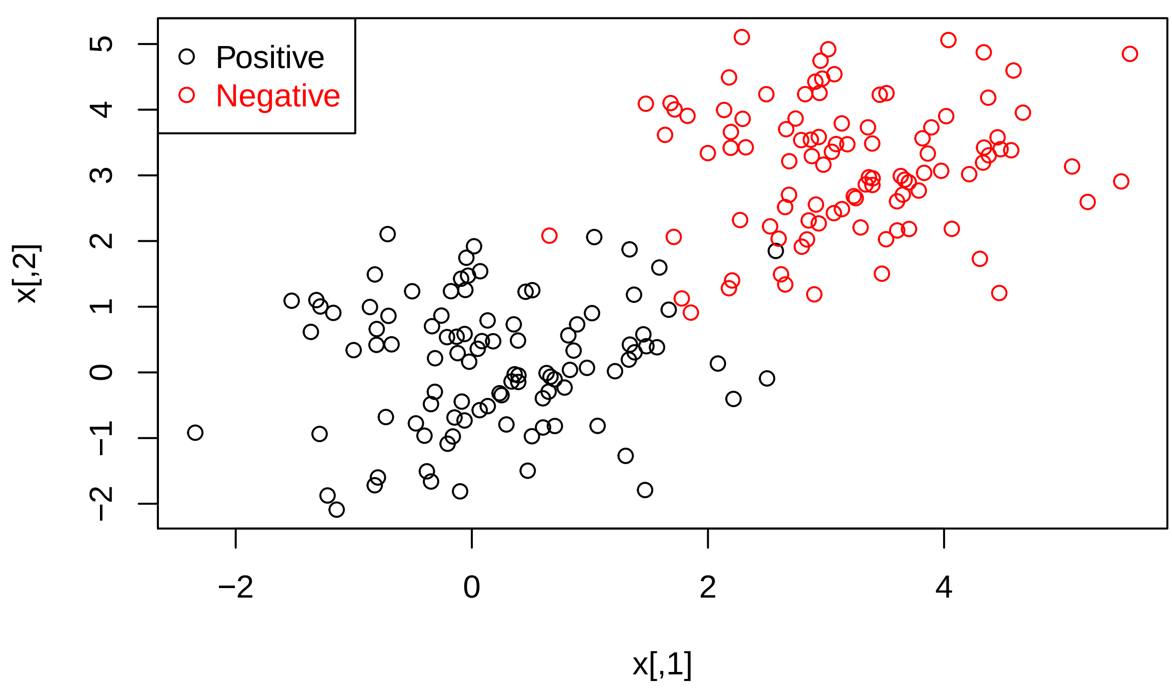
Lab 8 Solutions

Linear separable dataset.

```
#####Generate dataset###

set.seed(300)
x.pos <- matrix(rnorm(100*2,mean=0), nrow = 100, ncol = 2)
set.seed(300)
x.neg <- matrix(rnorm(100*2,mean=3), nrow = 100, ncol = 2)
y <- c(rep(1,100),rep(-1,100))
x <- rbind(x.pos,x.neg)
dat <- data.frame(x = x, y = as.factor(y))


plot(x, col = ifelse(y>0,1,2))
legend("topleft",c("Positive","Negative"),col=seq(2),pch=1,text.col=seq(2))
```




```
#####Training set and test set#####
set.seed(400)
train <-  sample(200, 200*0.7)
dat.train <-  dat[train,]
x.test <- x[-train,]
y.test <- y[-train]
```

1) Install and load the library `e1071`.

```
#install.packages("e1071")
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.4.2
```

2) Build a linear SVM model with `cost = 1`.

```
svmfit.train.C1 <- svm(y ~ .,
                       data = dat.train,
                       kernel = "linear",
                       cost = 1,
                       scale = FALSE)
```

3) Report how many support vectors are from each class respectively.

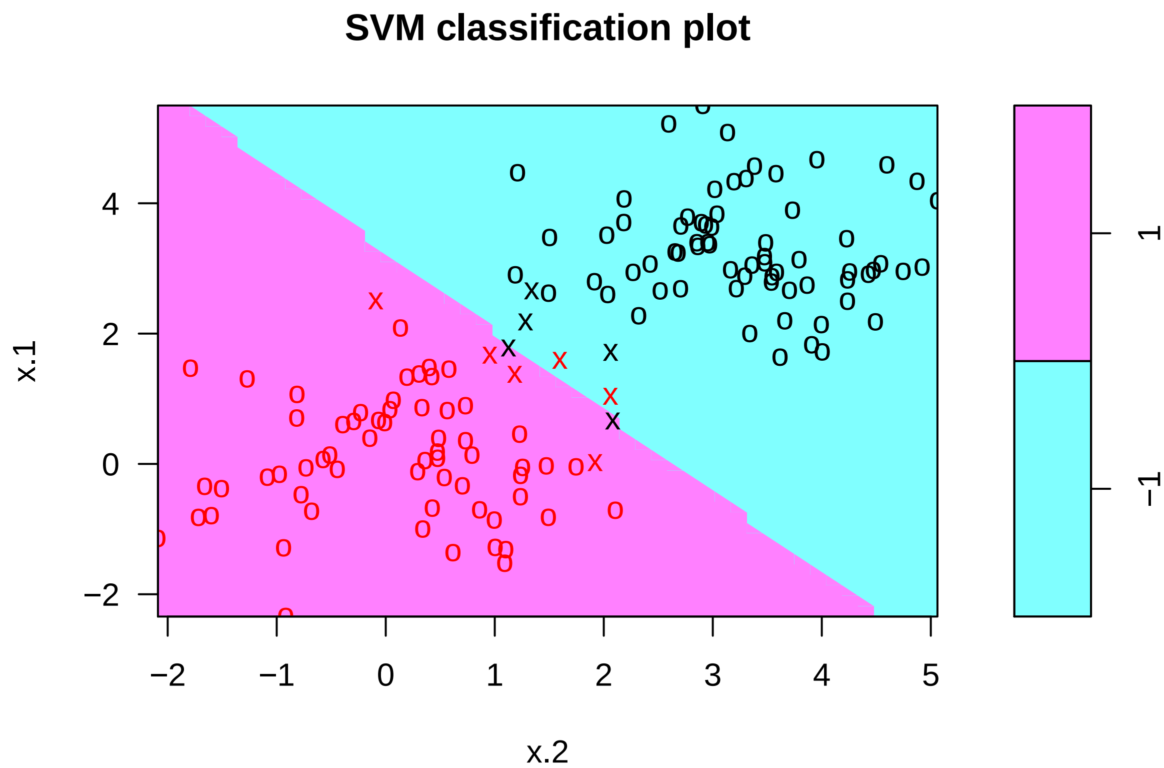
```
summary(svmfit.train.C1)
```

```
##
## Call:
## svm(formula = y ~ ., data = dat.train, kernel = "linear", cost = 1,
##     scale = FALSE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##       gamma:  0.5
##
## Number of Support Vectors:  11
##
##  ( 6 5 )
##
##
## Number of Classes:  2
##
## Levels:
##  -1 1
```

There are 11 support vectors, 6 from the class of `y = -1` and 5 from the class of`y = 1`.

4) Plot the svm model and check how many support vectors are on the wrong side of the boundary and how many data points are very close to the margin.

```
plot(svmfit.train.C1, dat.train)
```



There are 4 points on the wrong side of the boundary.

There are quite a few very close to the margin, at least 5 data points on the pink side and at least 2 on the cyan side.

5) Predict the class label of y on the test set and estimate the test error rate.

```
y.pred.C1 <- predict(svmfit.train.C1, newdata = x.test)
mean(y.pred.C1 != y.test)
```

```
## [1] 0.05
```

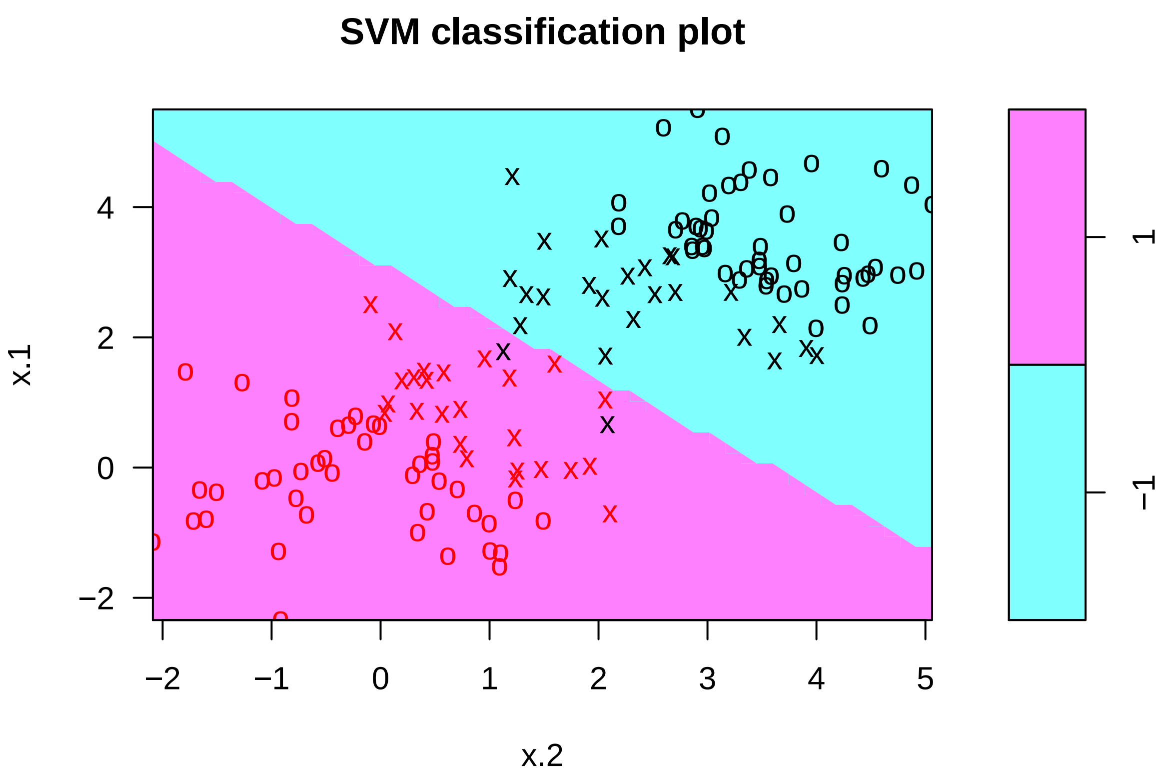
6) Now try a smaller `cost = 0.01` and a larger `cost = 1e5` and repeat step 2)-5).

```
#######cost = 0.01########
svmfit.train.C001 <- svm(y ~ .,
                         data = dat.train,
                         kernel = "linear",
                         cost = 0.01,
                         scale = FALSE)
summary(svmfit.train.C001)
```

```
##
## Call:
## svm(formula = y ~ ., data = dat.train, kernel = "linear", cost = 0.01,
##     scale = FALSE)
##
```

```
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.01
##       gamma:  0.5
##
## Number of Support Vectors:  50
##
##  ( 25 25 )
##
##
## Number of Classes:  2
##
## Levels:
##  -1 1
```

```
plot(svmfit.train.C001,dat.train)
```




```
y.pred.C001 <- predict(svmfit.train.C001, newdata = x.test)
mean(y.pred.C001 != y.test)
```

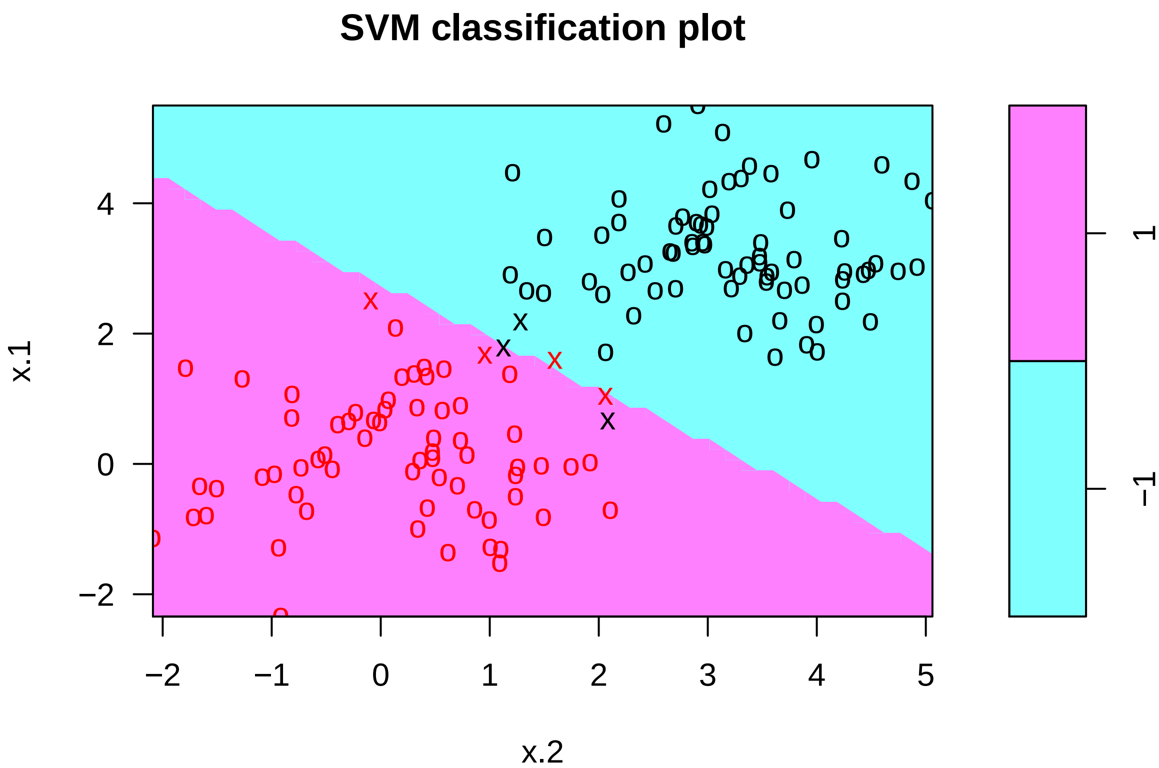
```
## [1] 0.03333333
```

```
#######cost = 1e5########
svmfit.train.C1e5 <- svm(y ~ .,
                         data = dat.train,
                         kernel = "linear",
```

```r
                         cost = 1e5,
                         scale = FALSE)
summary(svmfit.train.C1e5)
```

```
##
## Call:
## svm(formula = y ~ ., data = dat.train, kernel = "linear", cost = 1e+05,
##     scale = FALSE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1e+05
##       gamma:  0.5
##
## Number of Support Vectors:  7
##
##  ( 4 3 )
##
##
## Number of Classes:  2
##
## Levels:
##  -1 1
```

```r
plot(svmfit.train.C1e5,dat.train)
```




```
y.pred.C1e5 <- predict(svmfit.train.C1e5, newdata = x.test)
mean(y.pred.C1e5 != y.test)
```

```
## [1] 0.05
```

7) Use `tune()` function to select the best model (tune the parameter `C` or `cost`). Set seed to be 1.

```
set.seed(1)
tune.out <- tune(svm,
                 y ~ .,
                 data = dat.train,
                 kernel = "linear",
                 ranges = list(cost = c(0.001,0.01,0.1,1,5,10,100,1000,10000,1e5)))
summary(tune.out)
```

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##  0.01
##
## - best performance: 0.01428571
```

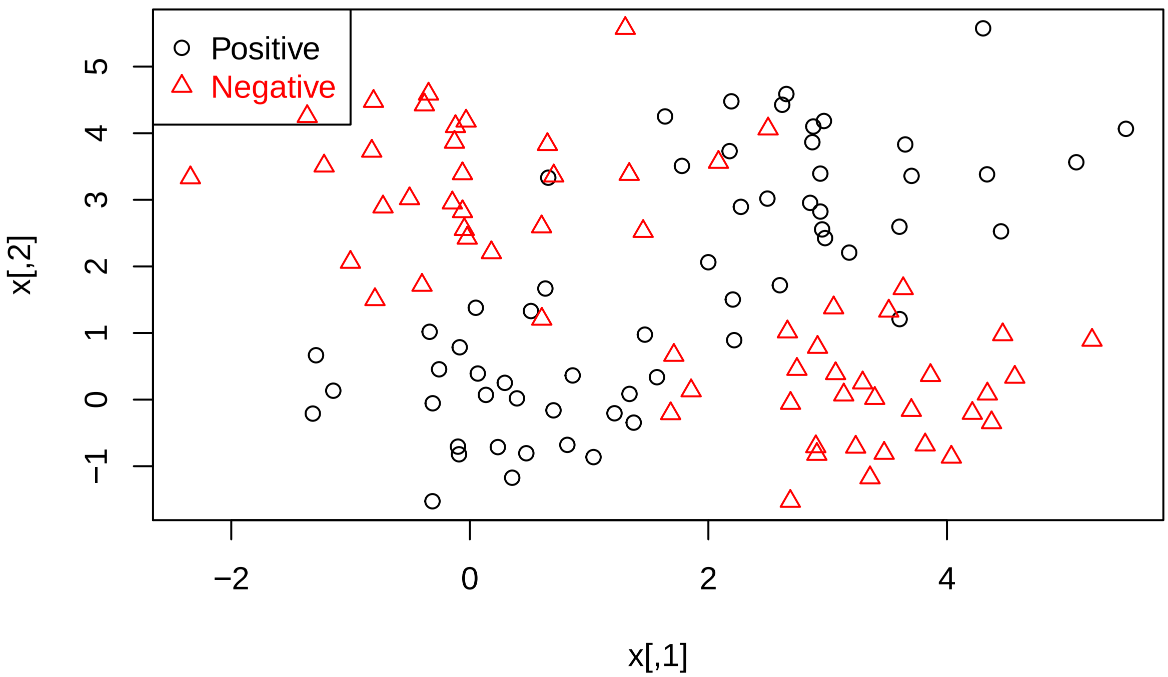
```
##
## - Detailed performance results:
##     cost      error dispersion
## 1  1e-03 0.47857143 0.08940468
## 2  1e-02 0.01428571 0.03011693
## 3  1e-01 0.01428571 0.03011693
## 4  1e+00 0.02857143 0.03688556
## 5  5e+00 0.02857143 0.03688556
## 6  1e+01 0.02857143 0.03688556
## 7  1e+02 0.02857143 0.03688556
## 8  1e+03 0.02857143 0.03688556
## 9  1e+04 0.02857143 0.03688556
## 10 1e+05 0.02857143 0.03688556
```

Linearly inseparable dataset.

```r
#####Generate dataset###

set.seed(300)
x.pos1 <- matrix(rnorm(30*2,mean=0), nrow=30, ncol=2)
x.pos2 <- matrix(rnorm(30*2,mean=3), nrow=30, ncol=2)
set.seed(300)
x.neg1 <- matrix(c(rnorm(30,mean=0)+3,rnorm(30,mean=0)),nrow=30,ncol=2)
x.neg2<- matrix(c(rnorm(30,mean=3)-3,rnorm(30,mean=3)),nrow=30,ncol=2)
y <- c(rep(1,60),rep(-1,60))
x <- rbind(x.pos1, x.pos2, x.neg1, x.neg2)
dat2 <- data.frame(x = x, y = as.factor(y))


plot(x,
     col = ifelse(y > 0, 1, 2),
     pch = ifelse(y > 0, 1, 2))
legend("topleft",
       c("Positive","Negative"),
       col = seq(2),
       pch = 1:2,
       text.col = seq(2))
```




```
#####Training set and test set#####
set.seed(400)
train <- sample(120, 120*0.7)
#There are 120 data points in total and we take 70% to be training set
dat2.train <- dat2[train,]
x.test <- x[-train,]
y.test <- y[-train]
```

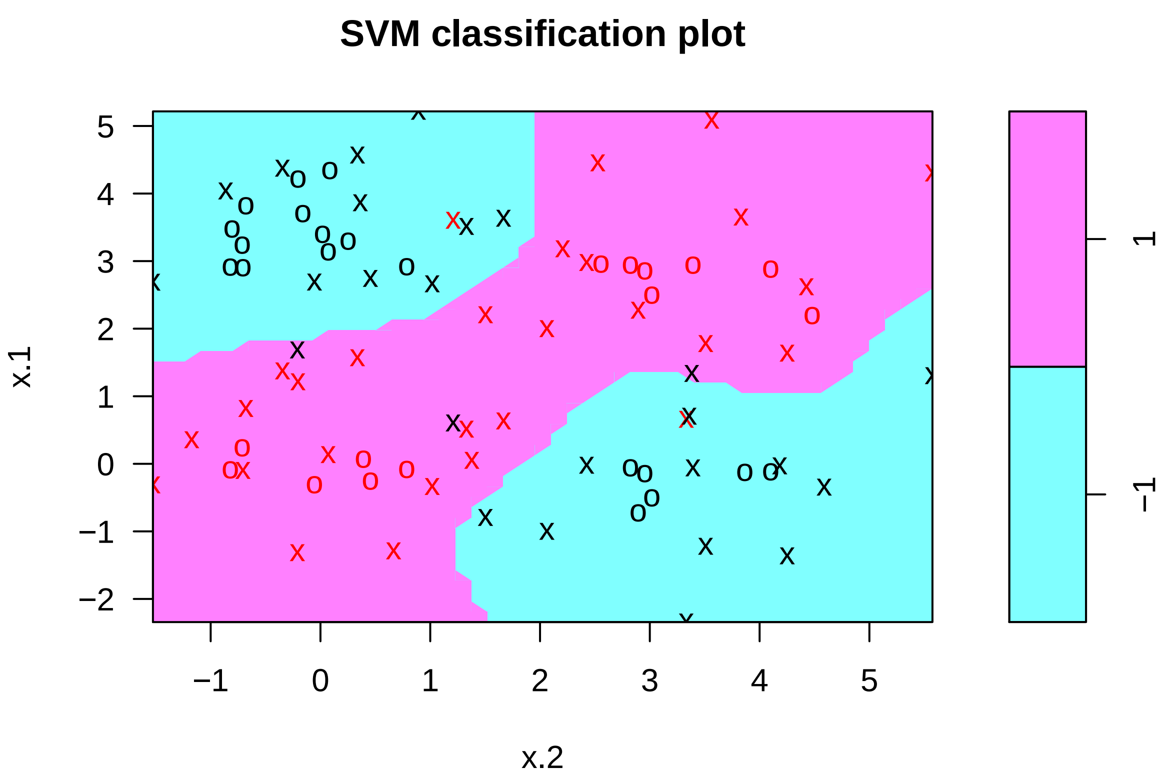
8) Build an SVM model with a radial kernel, `gamma = 1` and `cost = 1`. In this model, a) see the summary; 2) plot the svm and 3) estimate the test error rate.

```
svmfit.radial.G1C1 <- svm(y ~ ., data = dat2.train,
                          kernel = "radial",
                          gamma = 1, cost = 1,
                          scale = F)
summary(svmfit.radial.G1C1)
```

```
##
## Call:
## svm(formula = y ~ ., data = dat2.train, kernel = "radial", gamma = 1,
##     cost = 1, scale = F)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
```

```
##        cost:  1
##       gamma:  1
##
## Number of Support Vectors:  53
##
##  ( 28 25 )
##
##
## Number of Classes:  2
##
## Levels:
##  -1 1
```

```
plot(svmfit.radial.G1C1, data = dat2.train)
```




```
y.pred.radial.G1C1 <- predict(svmfit.radial.G1C1, newdata = x.test)
mean(y.pred.radial.G1C1 != y.test)
```

```
## [1] 0.1388889
```

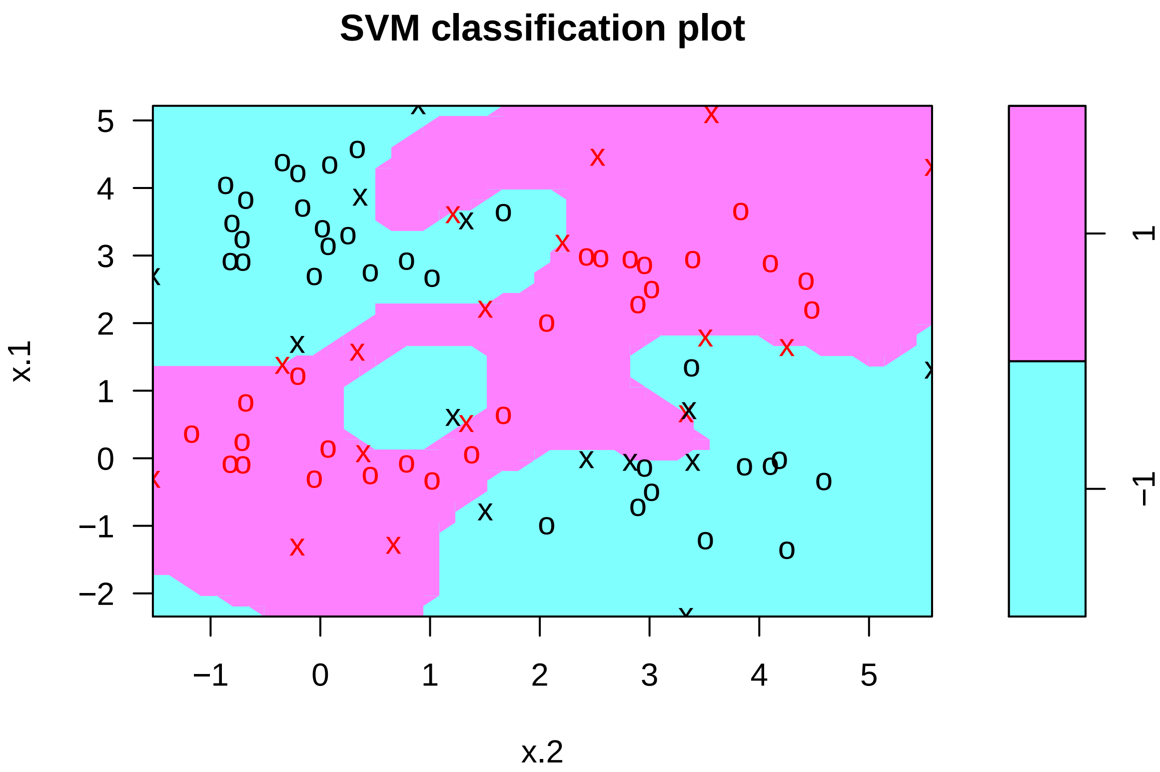
9) Build an SVM model with a radial kernel, gamma = 1 and cost = 1e5. In this model, a) see the summary; 2) plot the svm and 3) estimate the test error rate.

```
svmfit.radial.G1C1e5 <- svm(y ~ .,data = dat2.train,
                            kernel = "radial",
                            gamma = 1,cost = 1e5,
                            scale = F)
```

```r
summary(svmfit.radial.G1C1e5)
```

```
##
## Call:
## svm(formula = y ~ ., data = dat2.train, kernel = "radial", gamma = 1,
##     cost = 1e+05, scale = F)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1e+05
##       gamma:  1
##
## Number of Support Vectors:  29
##
##  ( 16 13 )
##
##
## Number of Classes:  2
##
## Levels:
##  -1 1
```

```r
plot(svmfit.radial.G1C1e5,data = dat2.train)
```



```r
y.pred.radial.G1C1e5 <- predict(svmfit.radial.G1C1e5, newdata = x.test)
mean(y.pred.radial.G1C1e5 != y.test)
```

```
## [1] 0.2777778
```

10) Choose the best choice of gamma and cost for an SVM with a radial kernel.

```r
set.seed(1)
tune.out.radial <- tune(svm, y ~ .,
                        data = dat2.train, kernel = "radial",
                        ranges = list(cost = c(0.1,1,10,100,1000),
                                      gamma = c(0.5,1,2,3,4)))

summary(tune.out.radial)
```

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost gamma
##    10   0.5
##
## - best performance: 0.08333333
##
## - Detailed performance results:
##     cost gamma      error dispersion
## 1  1e-01   0.5 0.14305556 0.09329594
## 2  1e+00   0.5 0.09583333 0.09765117
## 3  1e+01   0.5 0.08333333 0.09777182
## 4  1e+02   0.5 0.11944444 0.09868824
## 5  1e+03   0.5 0.10694444 0.08760894
## 6  1e-01   1.0 0.10694444 0.10207257
## 7  1e+00   1.0 0.09583333 0.11081171
## 8  1e+01   1.0 0.10694444 0.10558195
## 9  1e+02   1.0 0.16805556 0.12116016
## 10 1e+03   1.0 0.21666667 0.10083669
## 11 1e-01   2.0 0.20833333 0.15686180
## 12 1e+00   2.0 0.09583333 0.11081171
## 13 1e+01   2.0 0.14305556 0.11360058
## 14 1e+02   2.0 0.22916667 0.11250095
## 15 1e+03   2.0 0.23888889 0.08013406
## 16 1e-01   3.0 0.31527778 0.19532977
## 17 1e+00   3.0 0.10694444 0.11786022
## 18 1e+01   3.0 0.13194444 0.12218188
## 19 1e+02   3.0 0.26388889 0.11527685
## 20 1e+03   3.0 0.24166667 0.12966269
## 21 1e-01   4.0 0.38750000 0.20822528
## 22 1e+00   4.0 0.10694444 0.11786022
## 23 1e+01   4.0 0.15694444 0.11767823
## 24 1e+02   4.0 0.25138889 0.07449245
## 25 1e+03   4.0 0.24166667 0.11861252
```

11) Use the best model to estimate the test error rate.

```r
y.pred.radial.best <- predict(tune.out.radial$best.model, newdata = x.test)
mean(y.pred.radial.best != y.test)
```

```
## [1] 0.1388889
```