Demonstration Support Vector Machines

W. Grossmann

```
#-------------------------------------
#Classification CRM data
#-------------------------------------
#Libraries
#-------------------------------------
library("e1071")        # SVM

## Warning: package 'e1071' was built under R version 3.2.5

library("pROC") # for ROC

## Warning: package 'pROC' was built under R version 3.2.5

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

##
## Attaching package: 'pROC'

## Die folgenden Objekte sind maskiert von 'package:stats':
##
##     cov, smooth, var

#-------------------------------------
#Data
#Data from customers of a company: a number of sociodemographic
#characteristics, Sales in the last year, Average sales per year  #during the
relationship to the company, Duration of the relationship,
#Usage of a specific service (USES9),
#Split into training and test sample.
#-------------------------------------

data<-read.csv("CRM_class.csv",header=T,sep=";")
dim(data)

## [1] 1311   10

variable.names(data)

##  [1] "ID"            "Age_gr"        "sex"           "student"
##  [5] "office"        "Sales"         "Average_Sales" "Duration_CR"
##  [9] "UseS9"         "Train"

summary(data[,6:8])

##      Sales         Average_Sales      Duration_CR
##  Min.   :   0.00   Min.   :  0.010   Min.   : 12.03
##  1st Qu.:  17.48   1st Qu.:  1.605   1st Qu.: 31.80
##  Median :  58.03   Median :  4.510   Median : 68.50
```

```
##  Mean   : 197.57   Mean   : 12.089   Mean   : 78.38
##  3rd Qu.: 178.01   3rd Qu.: 12.630   3rd Qu.:119.36
##  Max.   :5869.48   Max.   :348.350   Max.   :188.13
```

```
table(data$UseS9,data$Train)
```

```
##
##       0   1
##   0 190 373
##   1 216 532
```

```
#-----------------------------------------------
#Selection of training data

Data.svm1<-as.data.frame(subset(data, Train==1))
attach(Data.svm1)
#-----------------------------------------------
# Support Vector machine
#-----------------------------------------------

x<-subset(Data.svm1, select = Sales:Duration_CR) # predictors
y<-as.factor(UseS9) # if y factor classification, otherwise regression
Data.svm2<-as.data.frame(cbind(x,y))
mod.svm2<-svm(y~., probability = TRUE,
              data=Data.svm2, kernel="radial")
summary(mod.svm2)
```

```
##
## Call:
## svm(formula = y ~ ., data = Data.svm2, probability = TRUE, kernel =
"radial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.3333333
##
## Number of Support Vectors:  642
##
##  ( 313 329 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
ls(mod.svm2)
```

```
##  [1] "call"            "coef0"           "coefs"
##  [4] "compprob"        "cost"            "decision.values"
```

```
##  [7] "degree"          "epsilon"         "fitted"
## [10] "gamma"           "index"           "kernel"
## [13] "labels"          "levels"          "na.action"
## [16] "nclasses"        "nSV"             "nu"
## [19] "probA"           "probB"           "rho"
## [22] "scaled"          "sigma"           "sparse"
## [25] "SV"              "terms"           "tot.nSV"
## [28] "type"            "x.scale"         "y.scale"
```

```
pred<-fitted(mod.svm2)
summary(pred)
```

```
##   0   1
## 422 483
```

```
#-------------------------------------------
#Confusion Matrix
table(pred,Data.svm2[,4])
```

```
##
## pred   0   1
##    0 265 157
##    1 108 375
```

```
#accuracy
640/905
```

```
## [1] 0.7071823
```

```
#-------------------------------------------
#Prediction probabilities

pred2<-predict(mod.svm2,x, decision.values= TRUE, probability = TRUE)

summary(pred2)
```

```
##   0   1
## 374 531
```

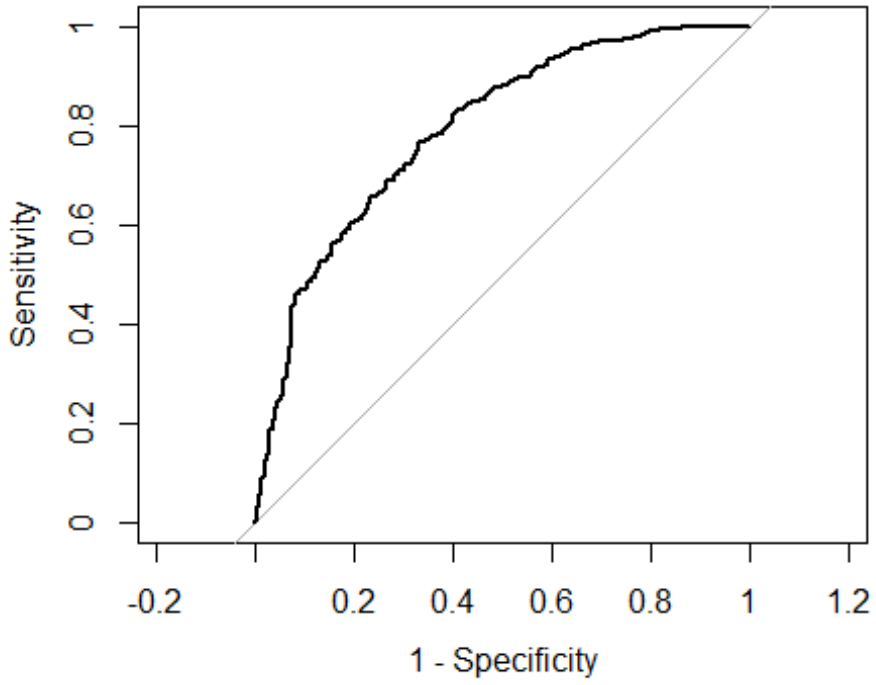
```
attr(pred2, "decision.values")[1:5,]
```

```
##         1         3         4         6        10
## 0.9173597 1.1070073 1.0462603 1.1479924 0.8430822
```

```
prob<-attr(pred2, "probabilities")
head(prob)
```

```
##            0         1
## 1  0.6949305 0.3050695
## 3  0.7385743 0.2614257
## 4  0.7250414 0.2749586
## 6  0.7474579 0.2525421
## 10 0.6767658 0.3232342
## 11 0.7186111 0.2813889
```

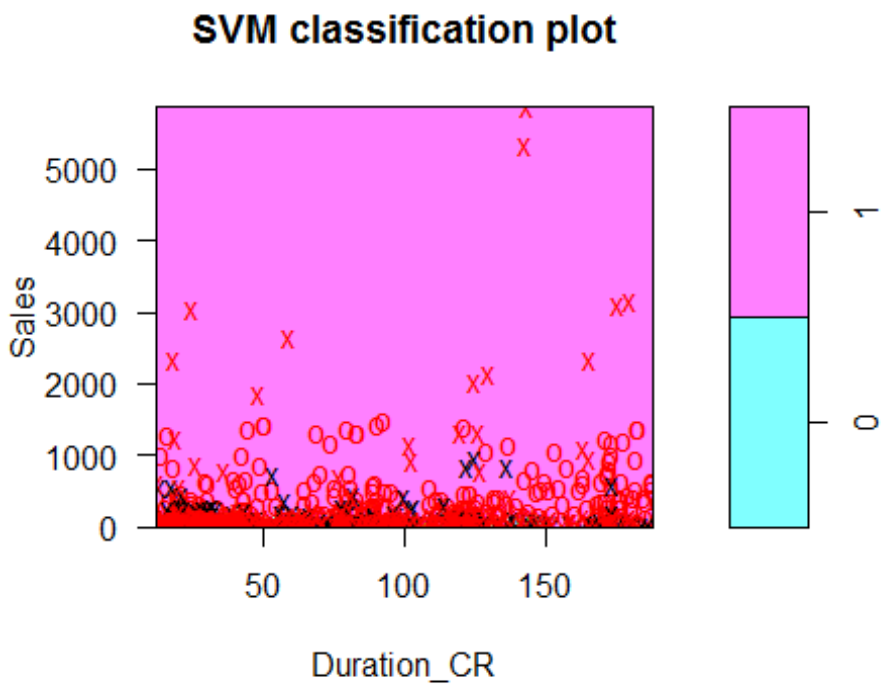
```
#-------------------------------------------------
# Plotting the ROC Curve
plot.roc(y,prob[,2],legacy.axes=TRUE)
```




```
auc(UseS9,prob[,2])
```

```
## Area under the curve: 0.7906
```

```
#-------------------------------------------------
#plotting the model(x is used for support vectors, o is used for data)
plot(mod.svm2, Data.svm2,Sales~Duration_CR,
     slice=list(Average_Sales=5))
```




```
#-----------------------------------------------
# the indices of firts 10 support vectors
mod.svm2$index[1:10]
```

```
##  [1]  1  5  7  8  9 11 12 16 21 22
```

```
#-------------------------------------------------------------------
#SVM with modified cost parameter for the violation of the contraints

mod.svmC2<-svm(y~., probability = TRUE,
               data=Data.svm2, cost = 2, kernel="radial")
summary(mod.svmC2)
```

```
##
## Call:
## svm(formula = y ~ ., data = Data.svm2, probability = TRUE, cost = 2,
##     kernel = "radial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  2
##       gamma:  0.3333333
##
## Number of Support Vectors:  623
##
##  ( 302 321 )
##
```

```
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
predC2<-fitted(mod.svmC2)
length(predC2)
```

```
## [1] 905
```

```
table(predC2,Data.svm2[,4])
```

```
##
## predC2   0   1
##      0 263 153
##      1 110 379
```

```
#--------------------------------------
#Application to test data

Data.Test<-as.data.frame(subset(data, Train==0))
newdata<-subset(Data.Test, select = Sales:Duration_CR)

pred.new<-predict(mod.svm2,newdata)
table(pred.new, Data.Test[,9])
```

```
##
## pred.new   0   1
##        0 125  68
##        1  65 148
```

```
#Accuracy
273/406
```

```
## [1] 0.6724138
```

```
#-----------------------------------------------
#SVM for all data
Data.all<- as.data.frame(data)
attach(Data.all)
```

```
## Die folgenden Objekte sind maskiert von Data.svm1:
##
##     Age_gr, Average_Sales, Duration_CR, ID, office, Sales, sex,
##     student, Train, UseS9
```

```
x1<-subset(Data.all, select = Sales:Duration_CR)
y1<-as.factor(UseS9)
mod.svm3<-svm(x1,y1, probability = TRUE,
              data=Data.all, kernel="radial")
summary(mod.svm3)
```

```
##
## Call:
```

```
## svm.default(x = x1, y = y1, kernel = "radial", probability = TRUE,
##     data = Data.all)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.3333333
##
## Number of Support Vectors:  928
##
##  ( 450 478 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```
#-----------------------------------------------
#Confusion Matrix
pred3<-fitted(mod.svm3)
length(pred3)
```

```
## [1] 1311
```

```
table(pred3,Data.all[,9])
```

```
##
## pred3   0   1
##     0 391 225
##     1 172 523
```

```
#Accuracy
914/1311
```

```
## [1] 0.6971777
```

```
#-----------------------------------------------
#Probabilities

pred4<-predict(mod.svm3,x1, decision.values= TRUE, probability = TRUE)
summary(pred4)
```

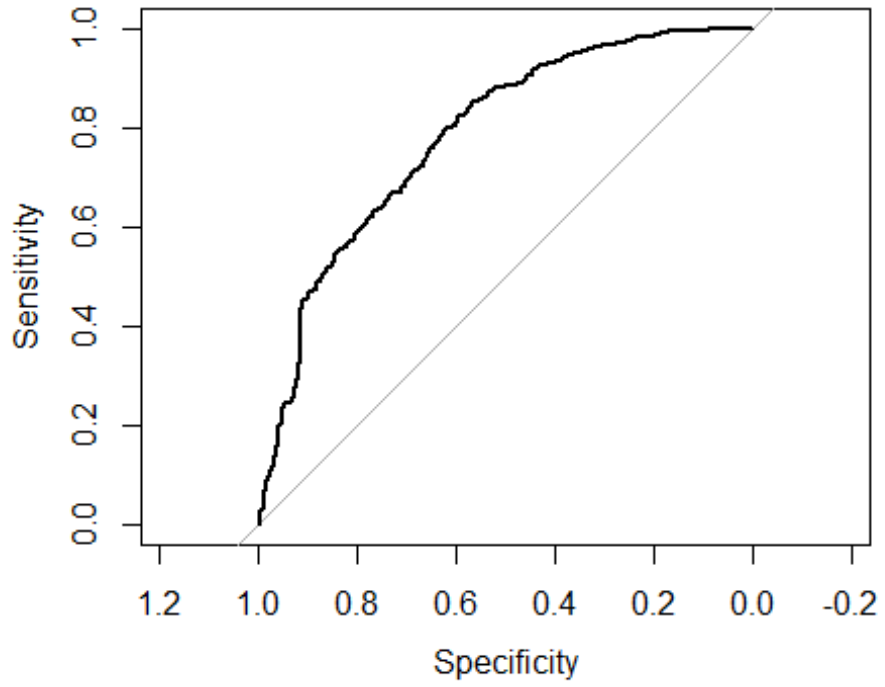
```
##   0   1
## 574 737
```

```
attr(pred4, "decision.values")[1:10,]
```

```
##         1         2         3         4         5         6         7
## 0.9734648 1.1664976 1.1736227 1.1185120 1.1483456 1.1710821 0.9984603
##         8         9        10
## 1.1705356 1.1620486 0.8914728
```

```r
prob1<-attr(pred4, "probabilities")

plot.roc(UseS9, prob1[,2])
```




```r
auc(UseS9,prob1[,2])
```

```
## Area under the curve: 0.7813
```

```r
#-----------------------------------------------------
#Cross Validation for all data

mod.svm4<-svm(x1,y1, probability = TRUE, cross = 10,
              data=Data.all, kernel="radial")
summary(mod.svm4)
```

```
##
## Call:
## svm.default(x = x1, y = y1, kernel = "radial", cross = 10, probability =
TRUE,
##     data = Data.all)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.3333333
##
## Number of Support Vectors:  928
```

```
##
##  ( 450 478 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
##
## 10-fold cross-validation on training data:
##
## Total Accuracy: 68.57361
## Single Accuracies:
##  65.64885 69.46565 67.93893 69.46565 71.75573 70.22901 58.77863 71.75573
74.0458 66.66667
```

```
#Cross validation for training sample

mod.svm5<-svm(x,y, probability = TRUE, cross = 10,
              data=Data.svm1, kernel="radial")

summary(mod.svm5)
```

```
##
## Call:
## svm.default(x = x, y = y, kernel = "radial", cross = 10, probability =
TRUE,
##     data = Data.svm1)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.3333333
##
## Number of Support Vectors:  642
##
##  ( 313 329 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
##
## 10-fold cross-validation on training data:
##
## Total Accuracy: 69.50276
## Single Accuracies:
##  68.88889 71.42857 65.55556 65.93407 65.55556 74.72527 68.88889 67.03297
78.88889 68.13187
```

```r
pred5<-fitted(mod.svm5)
length(pred5)
```

```
## [1] 905
```

```r
table(pred5, Data.svm1$UseS9)
```

```
##
## pred5   0   1
##     0 265 157
##     1 108 375
```