data science HW5

Amanda Howarth

5/9/2020

```
library(mlbench)
library(caret)
library(e1071)
library(ISLR)
```

Create atraining set containing a random sample of 800 observations, and a test set containing theremaining observations.

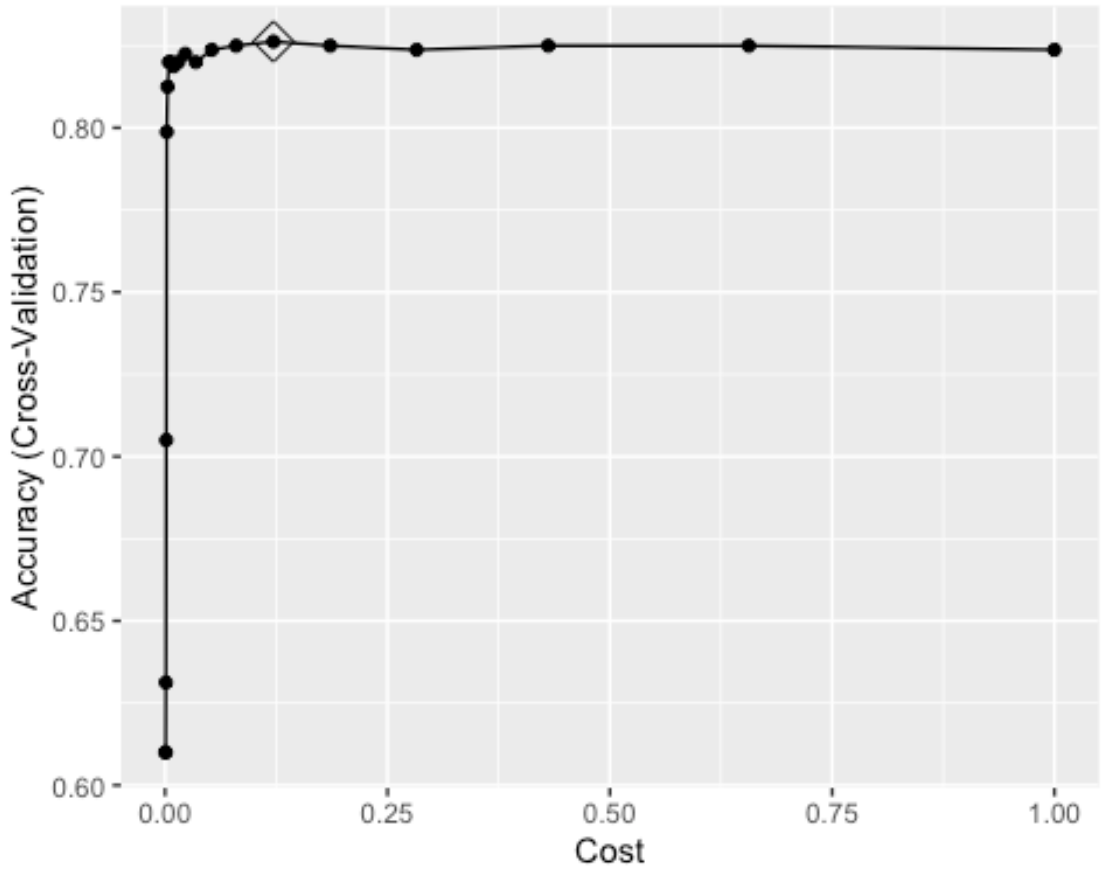
```
data(OJ)

set.seed(1)
rowTrain <- createDataPartition(y = OJ$Purchase,
                                p = 0.747,
                                list = FALSE)
```

Using `caret`

Question A

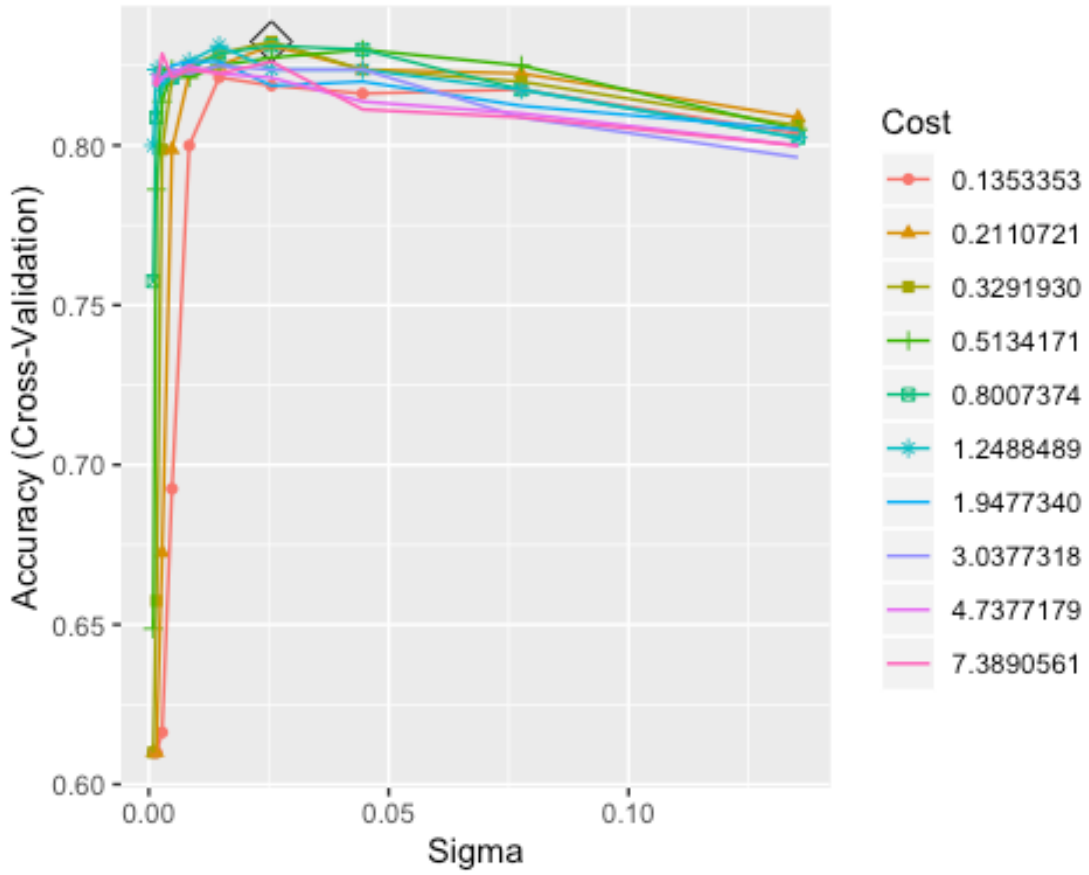
Fit a support vector classifier (linear kernel) to the training data with Purchase as the response and the other variables as predictors. What are the training and test error rates?

```
ctrl <- trainControl(method = "cv")

set.seed(1)
svml.fit <- train(Purchase~.,
                  data = OJ[rowTrain,],
                  method = "svmLinear2",
                  preProcess = c("center", "scale"),
                  tuneGrid = data.frame(cost = exp(seq(-8,0,len=20))),
                  trControl = ctrl)

ggplot(svml.fit, highlight = TRUE)
```

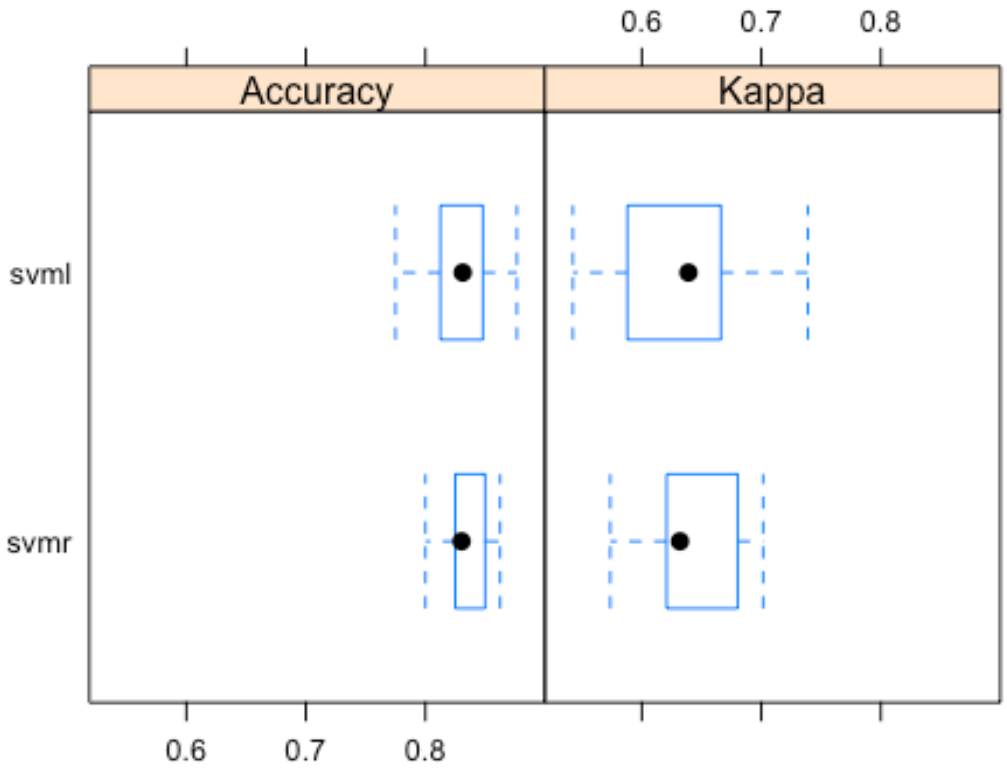


using kernels

```
svmr.grid <- expand.grid(C = exp(seq(-2,2,len=10)),
                         sigma = exp(seq(-7,-2,len=10)))
set.seed(1)
svmr.fit <- train(Purchase~., OJ,
                  subset = rowTrain,
                  method = "svmRadial",
                  preProcess = c("center", "scale"),
                  tuneGrid = svmr.grid,
                  trControl = ctrl)

ggplot(svmr.fit, highlight = TRUE)
```




```
resamp <- resamples(list(svmr = svmr.fit, svml = svml.fit))
bwplot(resamp)
```




```
summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: svmr, svml
## Number of resamples: 10
##
## Accuracy
##       Min. 1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## svmr 0.800  0.8250 0.8302215 0.8324795 0.8468750 0.8625000    0
## svml 0.775  0.8125 0.8312500 0.8262144 0.8454509 0.8765432    0
##
## Kappa
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## svmr 0.5736176 0.6211467 0.6319311 0.6418182 0.6730285 0.7014925    0
## svml 0.5422759 0.5915963 0.6390081 0.6288604 0.6636691 0.7388781    0
```

Suppport vector classifier with linear boundary (SVML): TRAINING ERROR: 0.8262144

Support vector machine with radial kernel (SVMR): TRAINING ERROR: 0.8324795

The support vector machine with radial kernel model has better performance based on cross validation. However, both perform incredibly similarly. Ultimately, the best model appears to be SVMR.

We finally look at the test data performance.

```
pred.svml <- predict(svml.fit, newdata = OJ[-rowTrain,])
pred.svmr <- predict(svmr.fit, newdata = OJ[-rowTrain,])

confusionMatrix(data = pred.svml,
                reference = OJ$Purchase[-rowTrain])

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  CH  MM
##         CH 144  19
##         MM  21  86
##
##                Accuracy : 0.8519
##                  95% CI : (0.8038, 0.892)
##     No Information Rate : 0.6111
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6894
##
##  Mcnemar's Test P-Value : 0.8744
##
##             Sensitivity : 0.8727
##             Specificity : 0.8190
##          Pos Pred Value : 0.8834
##          Neg Pred Value : 0.8037
##              Prevalence : 0.6111
##          Detection Rate : 0.5333
##    Detection Prevalence : 0.6037
##       Balanced Accuracy : 0.8459
##
##        'Positive' Class : CH
##

confusionMatrix(data = pred.svmr,
                reference = OJ$Purchase[-rowTrain])

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  CH  MM
##         CH 148  24
##         MM  17  81
##
##                Accuracy : 0.8481
```

```
##                  95% CI : (0.7997, 0.8888)
##     No Information Rate : 0.6111
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6766
##
##  Mcnemar's Test P-Value : 0.3487
##
##             Sensitivity : 0.8970
##             Specificity : 0.7714
##          Pos Pred Value : 0.8605
##          Neg Pred Value : 0.8265
##              Prevalence : 0.6111
##          Detection Rate : 0.5481
##    Detection Prevalence : 0.6370
##       Balanced Accuracy : 0.8342
##
##        'Positive' Class : CH
##
```

Suppport vector classifier with linear boundary: TEST ERROR: 0.8519

Support vector machine with radial kernel: TEST ERROR: 0.8481