STOR 565 Spring 2018 Homework 6

Due on 01/31/2018 in Class

Coleman Breen

Remark. Credits for **Theoretical Part** and **Computational Part** are in total 100 pt. For **Computational Part**, please complete your answer in the **RMarkdown** file and summit your printed PDF homework created by it.

Comment

If dplyr and MASS are both loaded, you might need to specify `dplyr::select` to specify that you want the dplyr version of the `select` function.

Computational Part

About the data: Tree leaf images

We will attempt to identify trees based on image data of their leaves. This is a tough problem, though apps such as iNaturalist now do a pretty good job identifying plants from images taken on your phone.

The data set is from here: https://www.kaggle.com/c/leaf-classification/data

Images have been pre-processed, so the dataset inlcudes vectors for margin, shape and texture attributes for each of almost 1000 images. We will focus on the shape attributes, which describe the contours of the leaf in the image.

A helpful demonstration for SVM

http://uc-r.github.io/svm

Q1

(a) (3 points)

Load the `leaf_train` dataset.

(i) Subset the columns to include only `id`, `species` and the `shape` variables, which is most easily done using the dplyr `select` function and the sub-function `contains`. There should be 66 variables in all.

(ii) Then create a new variable `genus` by extracting the first part of the species name. You can use the following code, assuming your data objects are named in a compatible way. You will probably want to load the data with `stringsAsFactors` as false.

(iii) Lastly, convert the genus variable to a factor.

```
#--> Part i)
leaf <- read.csv("leaf_train.csv", stringsAsFactors = FALSE)
leaf <- select(leaf, contains("id"), contains("species"), contains("shape"))

#--> Part ii)
leaf$genus <- str_split(leaf$species, "_", simplify = TRUE)[, 1]

#--> Part iii)
leaf$genus <- as.factor(leaf$genus)
```

(iv) Display your resulting data frame and the result of `summary(leaf$genus)`, which should give the number of observations of each genus. **Display only the id, species and first two species variables in your output, and only five rows of the data, eg by using the head function.**

```
#--> Part iv)
summary(leaf$genus)
```

```
##         Acer        Alnus  Arundinaria       Betula   Callicarpa
##          100           50           10           20           10
##     Castanea       Celtis       Cercis       Cornus      Cotinus
##           10           10           10           30           10
##    Crataegus      Cytisus   Eucalyptus        Fagus       Ginkgo
##           10           10           30           10           10
##         Ilex  Liquidambar Liriodendron  Lithocarpus     Magnolia
##           20           10           10           20           20
##        Morus         Olea  Phildelphus      Populus       Prunus
##           10           10           10           30           20
##   Pterocarya      Quercus Rhododendron        Salix       Sorbus
##           10          380           10           20           10
##        Tilia        Ulmus     Viburnum      Zelkova
##           30           10           20           10
```

```
head(leaf[,1:4], 5)
```

```
##   id               species     shape1     shape2
## 1  1           Acer_Opalus 0.00064671 0.00060945
## 2  2 Pterocarya_Stenoptera 0.00074942 0.00069461
## 3  3  Quercus_Hartwissiana 0.00097311 0.00091025
## 4  5       Tilia_Tomentosa 0.00045312 0.00046534
## 5  6    Quercus_Variabilis 0.00068161 0.00059775
```

(v) Randomly split your data into test and training sets. About 35 percent of the data should be in the test set. Display a summary of genus labels in the training set.

Note: In the rare event that one class in the training data is not represented, you may reduce the test set percentage to 30 percent and resample.

```
#--> Part v)
set.seed(919) # Petey Pablo
n <- nrow(leaf)
s <- floor(n * 0.65)
rows <- 1:n

#--> Sampling
train_index <- sample(rows, s, replace = F)
test_index <- rows[-train_index]
length(unique(c(train_index, test_index))) # Gut check that all numbers are accounted for
```

```
## [1] 990
```

```
#--> Split the data and check that all classes are represented in train
train <- leaf[train_index, ]
test <- leaf[test_index, ]

#--> Check
length(unique(train$genus))
```

```
## [1] 34
```

```
length(unique(leaf$genus))
```

```
## [1] 34
```

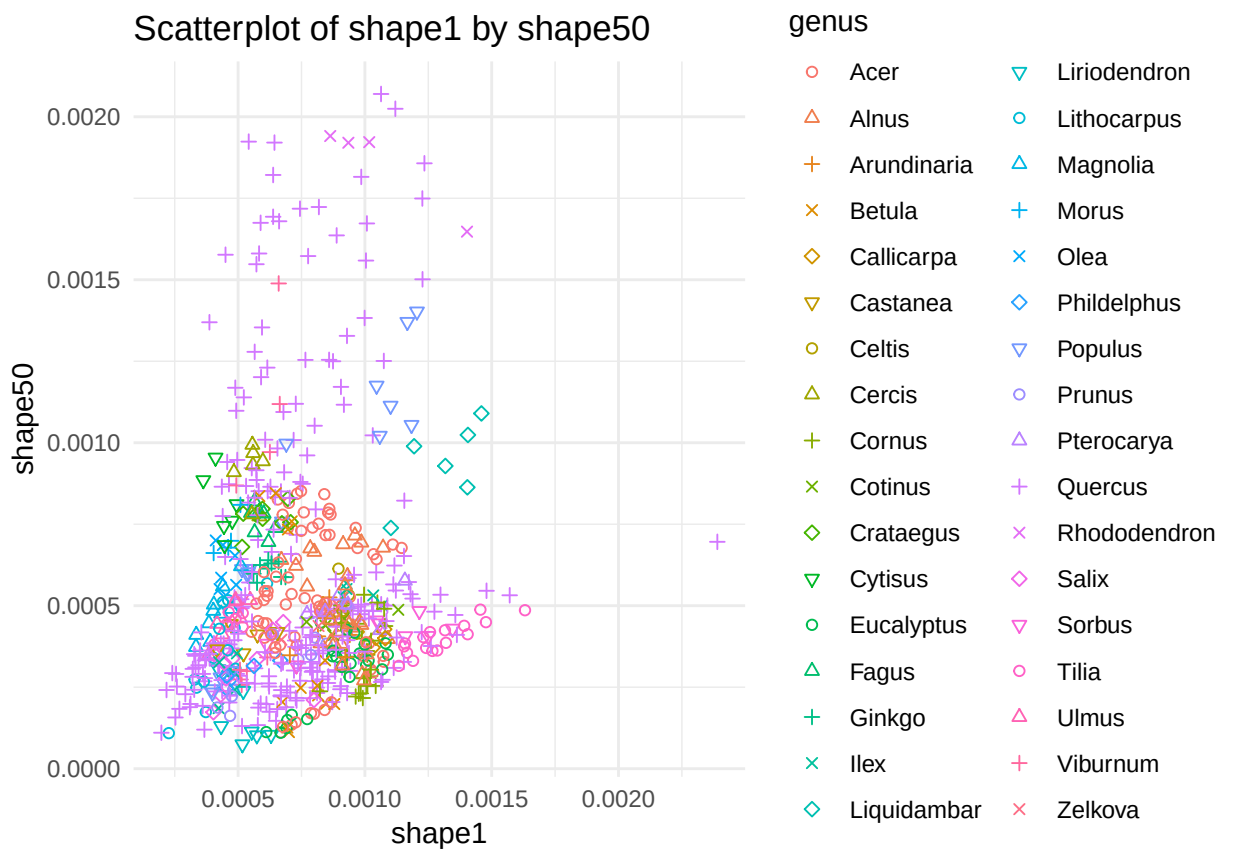
The 990 represents the number of unique integers contained in the union of the `train_index` and `test_index`. Because it is 990, each of our observations is in either the train or test set. The first 34 is the number of unique genera in the train set. The second 34 is the number of unique genera in the entire leaf dataset. Therefore, each of the genera has at least one observation in the train set.

(b) (2 points)

For the training data:

(i) Make a scatter plot of `shape1` by `shape50`, with some form of genus label. `ggplot2` is probably the best package for this, though you do not need to make the plot fancier than required to display the information above.

```
#--> Part i)
train %>%
  ggplot(aes(x=shape1, y=shape50, color=genus, shape=genus)) +
  geom_point() +
  scale_shape_manual(values = rep(1:6, 8)) +
  theme_minimal() +
  ggtitle("Scatterplot of shape1 by shape50")
```



(ii) Write two to three sentences discussing some possible implications of this plot for the SVM model.

Recall that we are trying to classify our observations into one of 34 genera (isn't that an obnoxious plural).

We can imagine a super simple SVM where we only consider `shape1` and `shape50`. Already, we can imagine a maximal margin classifier hyperplane that would split the Quercus (purple plus) samples from the Magnolias (blue triangle). However, with so many classes, and so many features, we can see it would be impossible to split the 34 genera by 33 hyperplanes in the `shape1` by `shape50` feature space. Therefore, we need to consider all of the shape features to create separating hyperplanes that will classify effectively.

(c) (15 points)

For the training data:

(i) Write a function, or use an available one, to choose the cost parameter for the SVM model on this training data with **linear kernel.** Use **shape variables as predictors only, genus as response**.

Use **5-fold cross validation.** Use the array of costs provided in the code below.

If you use a built-in function, you must state specifically how the best parameter value is chosen, for example by giving the error function minimized. Simply stating `classification error` is insufficient and will receive no points. You must state what that means. If using your own function, you may use any error function you like that is justified for classification problems.

See the demo linked above for help.

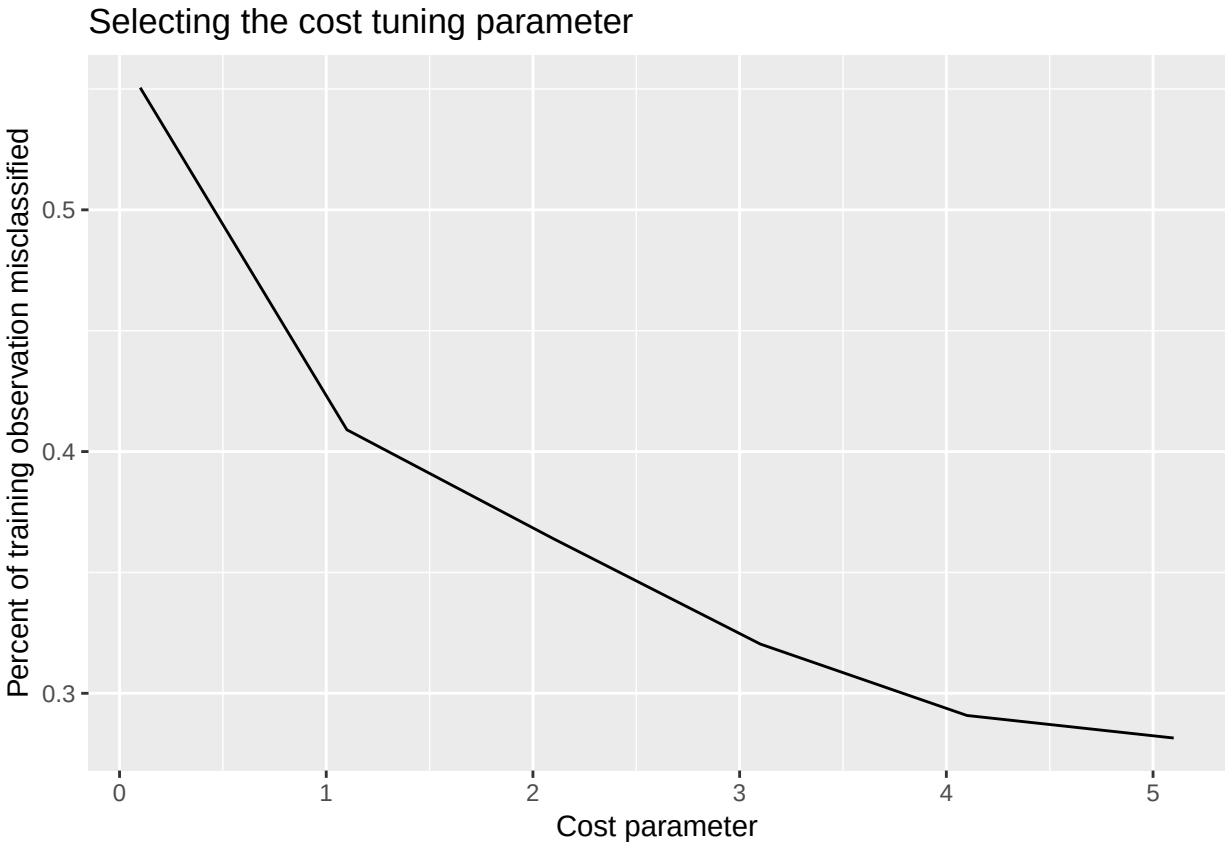
This might take some time to run. Do not knit your file at the last minute before the assignment is due.

```
#--> Given cost parameters to test
cost_out <- seq(from = 0.1, to =5.1, by = 1)
mis <- rep(NaN, length(cost_out))

#--> Loop thru costs
for (i in 1:length(cost_out)){
  #--> Fit model with 5-fold cv
  svm.model <- svm(formula=genus~., data=select(train, genus, contains("shape")),
                   kernel="linear", cost=cost_out[i], fold=5)
  #--> What percent were misclassified
  mis[i] <- (sum(svm.model$fitted != train$genus)) / (nrow(train))
}
```

Using 5-fold cross validation to fit the model for each of the six cost values, I used the misclassification rate as the error to minimize. Specifically, I used the percentage of the fitted values that were not classified as their actual genera. We want this error to be as low as possible (see below for value and plot).

(ii) Report the best value of cost chosen, and plot the errors by the cost values.

```
#--> Plot the cost vs. misclassification rate
temp <- as.data.frame(cbind(cost_out, mis))

temp %>%
  ggplot(aes(x=cost_out, y=mis)) +
  geom_line() +
  ylab("Percent of training observation misclassified") +
  xlab("Cost parameter") +
  ggtitle("Selecting the cost tuning parameter")
```



The best cost value is `5.1`. See above for the plot of the error metric over the cost parameters.

(iii) Write two or three sentences discussing some basic implications of your answer in (ii), using the concepts from class. Lecture 7 will be helpful.

The tuning parameter `c` is the "budget for training observations being on the wrong side." Based on our first plot, our observations are not easily separable by shape features. For example, Cornus and Tilia observations have near identical `shape1` and `shape50` values. Therefore, even in higher dimensions (dozens of shape features), our data are not cleanly separable. Having a higher `c` value allows us to embrace this messiness and create hyperplanes that will overall do a good job but perhaps miss a couple of anomolous points.

(d) (15 points)

(i) Run the SVM model on the **training data** with **linear kernel** and the cost determined in part (c). If you are unable to do part (c), use a cost of 1, the default. Report a summary of the fitted class label counts.

```
#--> (i) Fit model
svm.model <- svm(formula=genus~., data=select(train, genus, contains("shape")),
                 kernel="linear", cost=cost_out[length(cost_out)], fold=5)
summary(svm.model$fitted)
```

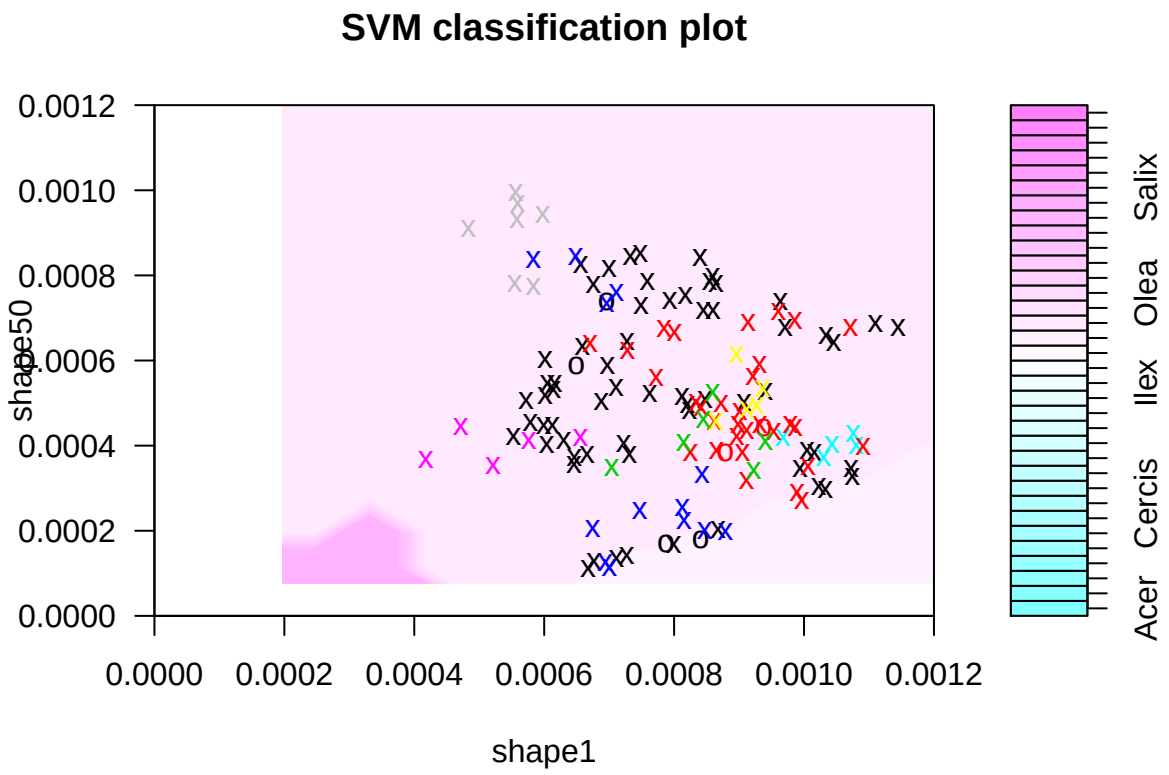
```
##         Acer        Alnus  Arundinaria       Betula   Callicarpa
##           42           24            2            4            1
##     Castanea       Celtis       Cercis       Cornus      Cotinus
##            5            0            0           22            0
##    Crataegus      Cytisus   Eucalyptus        Fagus       Ginkgo
##            4            6           12            0            6
```

```
##          Ilex  Liquidambar Liriodendron  Lithocarpus     Magnolia
##             5            6           7            9            7
##        Morus         Olea  Phildelphus      Populus       Prunus
##             3            0           5            6            1
##   Pterocarya      Quercus Rhododendron        Salix       Sorbus
##             2          419           4           10            5
##        Tilia        Ulmus     Viburnum      Zelkova
##            16            7           0            3
```

(ii) Create a classification plot from the model, plotting the variables `shape50` by `shape1`. See `?plot.svm`. In your plot statement, use the argument `xlim = c(0, 0.0012), ylim = c(0, 0.0012)`.

See the linked demo for an explanation of the plot. Write two sentences explaining what you see **using concepts and terminology from class.**

```
#--> (ii) Classification plot
plot(svm.model, select(train, genus, contains("shape")),
     shape50 ~ shape1, xlim = c(0, 0.0012), ylim = c(0, 0.0012))
```



This plot attempts to show separating boundaries projected onto the `shape1` by `shape50` space. We can see three shades of purple, each corresponding to different classes (genera). Because this is a projection, these are not necessarily "decision boundaries" but rather "decision gradients." For example, an observation falling into the bottom left corner would be more likely to be classified as Morus. There is no clear maximal margin classifier (hyperplane) that lies in this plane, so we don't get any straight lines.

(iii) Predict outcomes based on your model in (i) for the test data. Display a confusion matrix and compute sensitivity, specificity statistics. You may use the function demonstrated in class.

Warning: the confusion matrix will be awkward to display. Don't worry about it so much.

The sensitivity and specificity are good summaries.

```
#--> (iii) Confusion matrix
pred <- predict(svm.model, newdata=test)
# confusionMatrix(test$genus, pred) # confusion matrix not included because it is 34 by 34
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
confusion <- confusionMatrix(test$genus, pred)
confusion
```

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     Acer Alnus Arundinaria Betula Callicarpa Castanea Celtis
##   Acer           16     0           0      0           0        0      0
##   Alnus           2     5           0      0           0        0      0
##   Arundinaria     0     0           0      0           0        0      0
##   Betula          0     0           0      0           0        0      0
##   Callicarpa      0     0           0      0           0        0      0
##   Castanea        0     0           0      0           0        3      0
##   Celtis          1     0           0      0           0        0      0
##   Cercis          0     0           0      0           0        0      0
##   Cornus          0     0           0      0           0        0      0
##   Cotinus         0     0           0      0           0        0      0
##   Crataegus       0     0           0      0           0        0      0
##   Cytisus         0     0           0      0           0        0      0
##   Eucalyptus      0     0           0      0           0        0      0
##   Fagus           0     0           0      0           0        0      0
##   Ginkgo          0     0           0      0           0        0      0
##   Ilex            0     0           0      0           0        0      0
##   Liquidambar     0     0           0      0           0        0      0
##   Liriodendron    0     0           0      0           0        0      0
##   Lithocarpus     0     0           0      0           0        0      0
##   Magnolia        0     0           0      0           0        0      0
##   Morus           0     0           0      0           0        0      0
##   Olea            0     0           0      0           0        0      0
##   Phildelphus     0     0           0      0           0        0      0
##   Populus         0     0           0      0           0        0      0
##   Prunus          0     0           0      0           0        0      0
##   Pterocarya      0     0           0      0           0        0      0
##   Quercus         0     0           0      0           0        0      0
##   Rhododendron    0     0           0      0           0        0      0
##   Salix           0     0           0      0           0        0      0
##   Sorbus          0     0           0      0           0        0      0
##   Tilia           0     0           0      0           0        0      0
##   Ulmus           0     0           0      0           0        0      0
##   Viburnum        0     0           0      0           0        0      0
##   Zelkova         0     1           0      0           0        0      0
```

```
##              Reference
## Prediction     Cercis Cornus Cotinus Crataegus Cytisus Eucalyptus Fagus
##   Acer              0      1       0         0       0          0     0
##   Alnus             0      0       0         0       0          0     0
##   Arundinaria       0      0       0         0       0          0     0
##   Betula            0      0       0         0       0          0     0
##   Callicarpa        0      0       0         0       0          0     0
##   Castanea          0      0       0         0       0          0     0
##   Celtis            0      0       0         0       0          0     0
##   Cercis            0      0       0         0       0          0     0
##   Cornus            0      2       0         0       0          0     0
##   Cotinus           0      0       0         0       0          0     0
##   Crataegus         0      0       0         0       0          0     0
##   Cytisus           0      0       0         0       3          0     0
##   Eucalyptus        0      0       0         0       0          2     0
##   Fagus             0      0       0         0       0          0     0
##   Ginkgo            0      0       0         0       0          0     0
##   Ilex              0      0       0         0       0          0     0
##   Liquidambar       0      0       0         0       0          0     0
##   Liriodendron      0      0       0         0       0          0     0
##   Lithocarpus       0      0       0         0       0          0     0
##   Magnolia          0      0       0         0       0          0     0
##   Morus             0      0       0         0       0          0     0
##   Olea              0      0       0         0       0          0     0
##   Phildelphus       0      0       0         0       0          0     0
##   Populus           0      0       0         0       0          0     0
##   Prunus            0      0       0         0       0          0     0
##   Pterocarya        0      0       0         0       0          0     0
##   Quercus           0      0       0         0       2          1     0
##   Rhododendron      0      0       0         0       0          0     0
##   Salix             0      0       0         0       0          0     0
##   Sorbus            0      0       0         0       0          0     0
##   Tilia             0      0       0         0       0          0     0
##   Ulmus             0      0       0         0       0          0     0
##   Viburnum          0      0       0         0       0          0     0
##   Zelkova           0      0       0         0       0          0     0
##              Reference
## Prediction     Ginkgo Ilex Liquidambar Liriodendron Lithocarpus Magnolia
##   Acer              0    0           0           0           0        0
##   Alnus             0    0           0           0           0        0
##   Arundinaria       0    0           0           0           0        0
##   Betula            0    0           0           0           0        0
##   Callicarpa        0    0           0           0           0        0
##   Castanea          0    0           0           0           0        0
##   Celtis            0    0           0           0           0        0
##   Cercis            0    0           0           0           0        0
##   Cornus            0    0           0           0           0        0
##   Cotinus           0    0           0           0           0        0
##   Crataegus         0    0           0           0           0        0
##   Cytisus           0    0           0           0           0        0
##   Eucalyptus        0    0           0           0           0        0
##   Fagus             0    0           0           0           0        0
##   Ginkgo            1    0           0           0           0        0
##   Ilex              0    3           0           0           0        0
```

```
##    Liquidambar       0    0           4           0           0        0
##    Liriodendron      0    0           0           3           0        0
##    Lithocarpus       0    0           0           0           4        1
##    Magnolia          0    0           0           0           1        3
##    Morus             0    0           0           0           0        0
##    Olea              0    0           0           0           0        0
##    Phildelphus       0    0           0           0           0        0
##    Populus           0    0           0           0           0        0
##    Prunus            0    0           0           0           0        0
##    Pterocarya        0    0           0           0           0        0
##    Quercus           0    0           0           0           2        1
##    Rhododendron      0    0           0           0           0        0
##    Salix             0    0           0           0           0        0
##    Sorbus            0    0           0           0           0        0
##    Tilia             0    0           0           0           0        0
##    Ulmus             0    0           0           0           0        0
##    Viburnum          0    0           0           0           0        0
##    Zelkova           0    0           0           0           0        0
##                Reference
## Prediction     Morus Olea Phildelphus Populus Prunus Pterocarya Quercus
##    Acer            0    0           0       0      0          0      16
##    Alnus           0    0           0       0      0          1       9
##    Arundinaria     0    0           0       0      0          0       4
##    Betula          0    0           0       0      0          0       7
##    Callicarpa      0    0           0       0      0          0       5
##    Castanea        0    0           0       0      0          0       0
##    Celtis          0    0           0       0      0          0       4
##    Cercis          0    0           0       0      0          0       3
##    Cornus          0    0           0       0      0          0       3
##    Cotinus         0    0           0       0      0          0       5
##    Crataegus       0    0           0       0      0          0       2
##    Cytisus         0    0           0       0      0          0       1
##    Eucalyptus      0    0           0       0      0          0       5
##    Fagus           0    0           0       0      0          0       3
##    Ginkgo          0    0           0       0      0          0       0
##    Ilex            0    0           0       0      0          0       6
##    Liquidambar     0    0           0       0      0          0       0
##    Liriodendron    0    0           0       0      0          0       0
##    Lithocarpus     0    0           0       0      0          0       2
##    Magnolia        0    0           0       0      0          0       6
##    Morus           3    0           0       0      0          0       3
##    Olea            0    0           0       0      0          0       3
##    Phildelphus     0    0           0       0      0          0       5
##    Populus         0    0           0       2      0          0      11
##    Prunus          0    0           0       0      0          0      10
##    Pterocarya      0    0           0       0      0          0       4
##    Quercus         0    0           0       0      0          0     110
##    Rhododendron    0    0           0       0      0          0       4
##    Salix           0    0           0       0      1          0       7
##    Sorbus          0    0           0       0      0          0       2
##    Tilia           0    0           0       0      0          0       2
##    Ulmus           0    0           0       0      0          0       1
##    Viburnum        0    0           0       0      0          0      11
##    Zelkova         0    0           0       0      0          0       1
```

```
##               Reference
## Prediction     Rhododendron Salix Sorbus Tilia Ulmus Viburnum Zelkova
##   Acer                    0     0      0     0     0        0       0
##   Alnus                   0     0      0     0     0        0       0
##   Arundinaria             0     0      0     0     0        0       0
##   Betula                  0     0      0     0     0        0       0
##   Callicarpa              0     0      0     0     0        0       0
##   Castanea                0     1      0     0     1        0       0
##   Celtis                  0     0      0     0     0        0       0
##   Cercis                  0     0      0     0     0        0       0
##   Cornus                  0     0      0     0     0        0       0
##   Cotinus                 0     0      0     0     0        0       0
##   Crataegus               0     0      0     0     0        0       0
##   Cytisus                 0     0      0     0     0        0       0
##   Eucalyptus              0     0      0     0     0        0       0
##   Fagus                   0     0      0     0     0        0       0
##   Ginkgo                  0     0      0     0     0        0       0
##   Ilex                    0     0      0     0     0        0       0
##   Liquidambar             0     0      0     0     0        0       0
##   Liriodendron            0     0      0     0     0        0       0
##   Lithocarpus             0     0      0     0     0        0       0
##   Magnolia                0     0      0     0     0        0       0
##   Morus                   0     0      0     0     0        0       0
##   Olea                    0     0      0     0     0        0       0
##   Phildelphus             0     0      0     0     0        0       0
##   Populus                 1     0      0     0     0        0       0
##   Prunus                  0     0      0     0     0        0       0
##   Pterocarya              0     0      0     0     0        0       0
##   Quercus                 1     3      0     3     0        0       0
##   Rhododendron            2     0      0     0     0        0       0
##   Salix                   0     1      0     0     0        0       0
##   Sorbus                  0     0      0     0     0        0       0
##   Tilia                   0     0      0     4     0        0       0
##   Ulmus                   0     0      0     0     1        0       0
##   Viburnum                0     0      0     0     0        0       0
##   Zelkova                 0     0      0     0     0        0       5
##
## Overall Statistics
##
##                Accuracy : 0.5101
##                  95% CI : (0.4562, 0.5638)
##     No Information Rate : 0.7349
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3288
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Acer Class: Alnus Class: Arundinaria
## Sensitivity              0.84211      0.83333                 NA
## Specificity              0.94817      0.96481            0.98847
## Pos Pred Value           0.48485      0.29412                 NA
## Neg Pred Value           0.99045      0.99697                 NA
```

```
## Prevalence               0.05476      0.01729            0.00000
## Detection Rate           0.04611      0.01441            0.00000
## Detection Prevalence     0.09510      0.04899            0.01153
## Balanced Accuracy        0.89514      0.89907                 NA
##                      Class: Betula Class: Callicarpa Class: Castanea
## Sensitivity                     NA                NA        1.000000
## Specificity                0.97983           0.98559        0.994186
## Pos Pred Value                  NA                NA        0.600000
## Neg Pred Value                  NA                NA        1.000000
## Prevalence                 0.00000           0.00000        0.008646
## Detection Rate             0.00000           0.00000        0.008646
## Detection Prevalence       0.02017           0.01441        0.014409
## Balanced Accuracy               NA                NA        0.997093
##                      Class: Celtis Class: Cercis Class: Cornus
## Sensitivity                     NA            NA      0.666667
## Specificity                0.98559      0.991354      0.991279
## Pos Pred Value                  NA            NA      0.400000
## Neg Pred Value                  NA            NA      0.997076
## Prevalence                 0.00000      0.000000      0.008646
## Detection Rate             0.00000      0.000000      0.005764
## Detection Prevalence       0.01441      0.008646      0.014409
## Balanced Accuracy               NA            NA      0.828973
##                      Class: Cotinus Class: Crataegus Class: Cytisus
## Sensitivity                      NA               NA       0.600000
## Specificity                 0.98559         0.994236       0.997076
## Pos Pred Value                   NA               NA       0.750000
## Neg Pred Value                   NA               NA       0.994169
## Prevalence                 0.00000         0.000000       0.014409
## Detection Rate             0.00000         0.000000       0.008646
## Detection Prevalence       0.01441         0.005764       0.011527
## Balanced Accuracy                NA               NA       0.798538
##                      Class: Eucalyptus Class: Fagus Class: Ginkgo
## Sensitivity                   0.666667           NA      1.000000
## Specificity                   0.985465     0.991354      1.000000
## Pos Pred Value                0.285714           NA      1.000000
## Neg Pred Value                0.997059           NA      1.000000
## Prevalence                    0.008646     0.000000      0.002882
## Detection Rate                0.005764     0.000000      0.002882
## Detection Prevalence          0.020173     0.008646      0.002882
## Balanced Accuracy             0.826066           NA      1.000000
##                      Class: Ilex Class: Liquidambar Class: Liriodendron
## Sensitivity             1.000000            1.00000             1.000000
## Specificity             0.982558            1.00000             1.000000
## Pos Pred Value          0.333333            1.00000             1.000000
## Neg Pred Value          1.000000            1.00000             1.000000
## Prevalence              0.008646            0.01153             0.008646
## Detection Rate          0.008646            0.01153             0.008646
## Detection Prevalence    0.025937            0.01153             0.008646
## Balanced Accuracy       0.991279            1.00000             1.000000
##                      Class: Lithocarpus Class: Magnolia Class: Morus
## Sensitivity                     0.57143        0.600000     1.000000
## Specificity                     0.99118        0.979532     0.991279
## Pos Pred Value                  0.57143        0.300000     0.500000
## Neg Pred Value                  0.99118        0.994065     1.000000
```

```
## Prevalence                     0.02017        0.014409     0.008646
## Detection Rate                 0.01153        0.008646     0.008646
## Detection Prevalence           0.02017        0.028818     0.017291
## Balanced Accuracy              0.78130        0.789766     0.995640
##                      Class: Olea Class: Phildelphus Class: Populus
## Sensitivity                   NA                 NA       1.000000
## Specificity             0.991354            0.98559       0.965217
## Pos Pred Value                NA                 NA       0.142857
## Neg Pred Value                NA                 NA       1.000000
## Prevalence              0.000000            0.00000       0.005764
## Detection Rate          0.000000            0.00000       0.005764
## Detection Prevalence    0.008646            0.01441       0.040346
## Balanced Accuracy             NA                 NA       0.982609
##                      Class: Prunus Class: Pterocarya Class: Quercus
## Sensitivity               0.000000          0.000000         0.4314
## Specificity               0.971098          0.988439         0.8587
## Pos Pred Value            0.000000          0.000000         0.8943
## Neg Pred Value            0.997033          0.997085         0.3527
## Prevalence                0.002882          0.002882         0.7349
## Detection Rate            0.000000          0.000000         0.3170
## Detection Prevalence      0.028818          0.011527         0.3545
## Balanced Accuracy         0.485549          0.494220         0.6450
##                      Class: Rhododendron Class: Salix Class: Sorbus
## Sensitivity                     0.500000     0.200000            NA
## Specificity                     0.988338     0.976608      0.994236
## Pos Pred Value                  0.333333     0.111111            NA
## Neg Pred Value                  0.994135     0.988166            NA
## Prevalence                      0.011527     0.014409      0.000000
## Detection Rate                  0.005764     0.002882      0.000000
## Detection Prevalence            0.017291     0.025937      0.005764
## Balanced Accuracy               0.744169     0.588304            NA
##                      Class: Tilia Class: Ulmus Class: Viburnum
## Sensitivity               0.57143     0.500000              NA
## Specificity               0.99412     0.997101          0.9683
## Pos Pred Value            0.66667     0.500000              NA
## Neg Pred Value            0.99120     0.997101              NA
## Prevalence                0.02017     0.005764          0.0000
## Detection Rate            0.01153     0.002882          0.0000
## Detection Prevalence      0.01729     0.005764          0.0317
## Balanced Accuracy         0.78277     0.748551              NA
##                      Class: Zelkova
## Sensitivity                 1.00000
## Specificity                 0.99415
## Pos Pred Value              0.71429
## Neg Pred Value              1.00000
## Prevalence                  0.01441
## Detection Rate              0.01441
## Detection Prevalence        0.02017
## Balanced Accuracy           0.99708
```

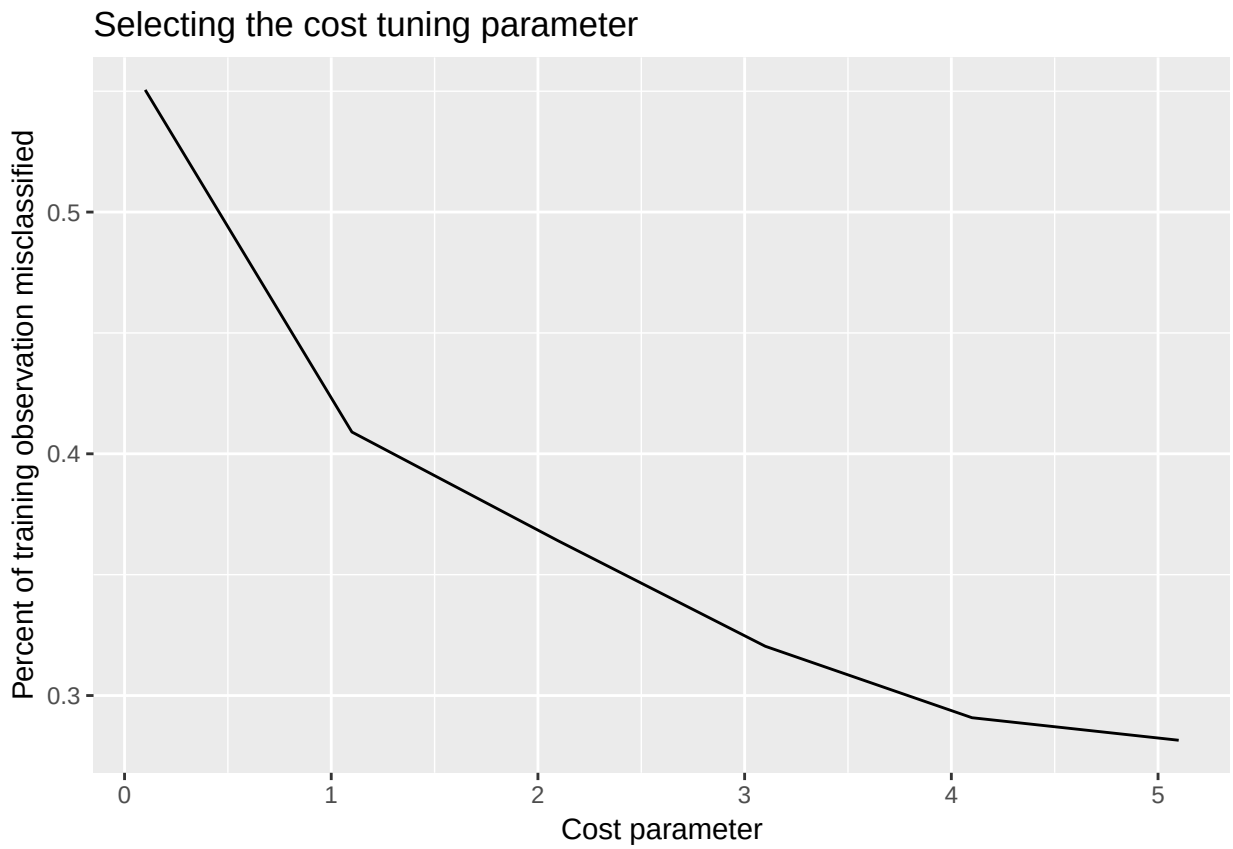
The above confusion matrix gives sensitivity and specificity.

(e) (15 points)

This question will use a non-linear kernel for the SVM and compare results.

(i) Modify your function in part (c) to find the optimal cost value for the SVM on the **training data** with **radial kernel** with gamma parameter 0.55. Use the same cost range. Report the optimal cost.

```
#--> Given cost parameters to test
cost_out <- seq(from = 0.1, to =5.1, by = 1)
mis.radial <- rep(NaN, length(cost_out))

#--> Loop thru costs
for (i in 1:length(cost_out)){
  #--> Fit model with 5-fold cv
  svm.model.radial <- svm(formula=genus~., data=select(train, genus, contains("shape")),
                  kernel="radial", cost=cost_out[i], fold=5, gamma=0.55)
  #--> What percent were misclassified
  mis.radial[i] <- (sum(svm.model.radial$fitted != train$genus)) / (nrow(train))
}

#--> Plot the cost vs. misclassification rate
temp <- as.data.frame(cbind(cost_out, mis.radial))

temp %>%
  ggplot(aes(x=cost_out, y=mis)) +
  geom_line() +
  ylab("Percent of training observation misclassified") +
  xlab("Cost parameter") +
  ggtitle("Selecting the cost tuning parameter")
```



```
print(cbind(cost_out, mis.radial))
```

```
##      cost_out mis.radial
## [1,]      0.1 0.60031104
## [2,]      1.1 0.07620529
## [3,]      2.1 0.00777605
## [4,]      3.1 0.00155521
## [5,]      4.1 0.00155521
## [6,]      5.1 0.00155521
```

The optimum cost parameter is a three-way tie between 3.1, 4.1, and 5.1. We'll use `c` = 5.1.

(ii) Run the radial SVM model with these optimal parameters on the training data.

```
#--> (i) Fit model
svm.model.radial <- svm(formula=genus~., data=select(train, genus, contains("shape")),
                kernel="radial", cost=cost_out[length(cost_out)], fold=5)
summary(svm.model.radial$fitted)
```

```
##         Acer        Alnus  Arundinaria       Betula   Callicarpa
##           62           37            0            0            0
##     Castanea       Celtis       Cercis       Cornus      Cotinus
##            5            0            0           34            0
##    Crataegus      Cytisus   Eucalyptus        Fagus       Ginkgo
##            2            5            5            1            8
##         Ilex  Liquidambar Liriodendron  Lithocarpus     Magnolia
##            2            6            6            1            3
##        Morus         Olea  Phildelphus      Populus       Prunus
##            0            9            0            6            0
##   Pterocarya      Quercus Rhododendron        Salix       Sorbus
##            0          421            4            1            1
##        Tilia        Ulmus     Viburnum      Zelkova
##           19            5            0            0
```

(iii) Repeat part (d)(iii) but for the radial SVM model instead of the linear one.

(Predict outcomes based on your model in (i) for the test data. Display a confusion matrix and compute sensitivity, specificity statistics. You may use the function demonstrated in class.)

```
#--> (iii) Confusion matrix
pred.radial <- predict(svm.model.radial, newdata=test)
# confusionMatrix(test$genus, pred) # confusion matrix not included because it is 34 by 34
library(caret)
confusion <- confusionMatrix(test$genus, pred.radial)
confusion
```

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     Acer Alnus Arundinaria Betula Callicarpa Castanea Celtis
##   Acer           20     1           0      0          0        0      0
##   Alnus           2    10           0      0          0        0      0
##   Arundinaria     0     0           0      0          0        0      0
##   Betula          0     0           0      0          0        0      0
##   Callicarpa      0     0           0      0          0        0      0
##   Castanea        0     0           0      0          0        2      0
##   Celtis          1     1           0      0          0        0      0
```

```
##   Cercis          0     0           0      0          0        0      0
##   Cornus          0     0           0      0          0        0      0
##   Cotinus         2     0           0      0          0        0      0
##   Crataegus       1     0           0      0          0        0      0
##   Cytisus         0     0           0      0          0        0      0
##   Eucalyptus      0     0           0      0          0        0      0
##   Fagus           0     0           0      0          0        0      0
##   Ginkgo          1     0           0      0          0        0      0
##   Ilex            0     0           0      0          0        0      0
##   Liquidambar     0     0           0      0          0        0      0
##   Liriodendron    0     0           0      0          0        0      0
##   Lithocarpus     0     0           0      0          0        0      0
##   Magnolia        0     0           0      0          0        0      0
##   Morus           0     0           0      0          0        0      0
##   Olea            0     0           0      0          0        0      0
##   Phildelphus     0     0           0      0          0        0      0
##   Populus         1     0           0      0          0        0      0
##   Prunus          0     0           0      0          0        0      0
##   Pterocarya      0     1           0      0          0        0      0
##   Quercus         4     2           0      0          0        0      0
##   Rhododendron    0     0           0      0          0        0      0
##   Salix           0     0           0      0          0        0      0
##   Sorbus          0     0           0      0          0        0      0
##   Tilia           0     0           0      0          0        0      0
##   Ulmus           0     0           0      0          0        0      0
##   Viburnum        0     0           0      0          0        0      0
##   Zelkova         0     3           0      0          0        0      0
##               Reference
## Prediction     Cercis Cornus Cotinus Crataegus Cytisus Eucalyptus Fagus
##   Acer              0      2       0         0       0          0     0
##   Alnus             0      1       0         0       0          0     0
##   Arundinaria       0      0       0         0       0          0     0
##   Betula            0      0       0         0       0          0     0
##   Callicarpa        0      1       0         0       0          0     0
##   Castanea          0      0       0         0       0          0     0
##   Celtis            0      0       0         0       0          0     0
##   Cercis            0      0       0         0       0          0     0
##   Cornus            0      4       0         0       0          0     0
##   Cotinus           0      0       0         0       0          0     0
##   Crataegus         0      0       0         0       0          0     0
##   Cytisus           0      0       0         0       3          0     0
##   Eucalyptus        0      1       0         0       0          1     0
##   Fagus             0      0       0         0       0          0     0
##   Ginkgo            0      0       0         0       0          0     0
##   Ilex              0      0       0         0       0          0     0
##   Liquidambar       0      0       0         0       0          0     0
##   Liriodendron      0      0       0         0       0          0     0
##   Lithocarpus       0      0       0         0       0          0     0
##   Magnolia          0      0       0         0       0          0     0
##   Morus             0      0       0         0       0          0     0
##   Olea              0      0       0         0       0          0     0
##   Phildelphus       0      0       0         0       0          0     0
##   Populus           0      0       0         1       0          0     0
##   Prunus            0      0       0         0       0          0     0
```

```
##    Pterocarya        0      0       0         0       0          0     0
##    Quercus           0      0       0         0       4          0     0
##    Rhododendron      0      0       0         0       0          0     0
##    Salix             0      0       0         0       0          0     0
##    Sorbus            0      0       0         0       0          0     0
##    Tilia             0      1       0         0       0          0     0
##    Ulmus             0      0       0         0       0          0     0
##    Viburnum          0      0       0         0       1          0     0
##    Zelkova           0      0       0         0       0          0     0
##                Reference
## Prediction      Ginkgo Ilex Liquidambar Liriodendron Lithocarpus Magnolia
##    Acer              0    0           0            0           0        0
##    Alnus             0    0           0            0           0        0
##    Arundinaria       0    0           0            0           0        0
##    Betula            0    0           0            0           0        0
##    Callicarpa        0    0           0            0           0        0
##    Castanea          0    0           0            0           0        0
##    Celtis            0    0           0            0           0        0
##    Cercis            0    0           0            0           0        0
##    Cornus            0    0           0            0           0        0
##    Cotinus           0    0           0            0           0        0
##    Crataegus         0    0           0            0           0        0
##    Cytisus           0    0           0            0           0        0
##    Eucalyptus        0    0           0            0           0        0
##    Fagus             0    0           0            0           0        0
##    Ginkgo            0    0           0            0           0        0
##    Ilex              0    2           0            0           0        0
##    Liquidambar       0    0           4            0           0        0
##    Liriodendron      0    0           0            3           0        0
##    Lithocarpus       0    0           0            0           0        3
##    Magnolia          0    0           0            0           0        2
##    Morus             0    0           0            0           0        0
##    Olea              0    0           0            0           0        0
##    Phildelphus       0    0           0            0           0        0
##    Populus           0    0           0            0           0        0
##    Prunus            0    0           0            0           0        0
##    Pterocarya        0    0           0            0           0        0
##    Quercus           0    0           0            0           0        0
##    Rhododendron      0    0           0            0           0        0
##    Salix             0    0           0            0           0        0
##    Sorbus            0    0           0            0           0        0
##    Tilia             0    0           0            0           0        0
##    Ulmus             0    0           0            0           0        0
##    Viburnum          0    0           0            0           0        0
##    Zelkova           0    0           0            0           0        0
##                Reference
## Prediction      Morus Olea Phildelphus Populus Prunus Pterocarya Quercus
##    Acer             0    0           0       0      0          0      10
##    Alnus            0    0           0       0      0          0       4
##    Arundinaria      0    0           0       0      0          0       4
##    Betula           0    0           0       0      0          0       7
##    Callicarpa       0    0           0       0      0          0       4
##    Castanea         0    0           0       0      0          0       3
##    Celtis           0    0           0       0      0          0       3
```

```
##    Cercis           0    0           0       0      0           0       3
##    Cornus           0    0           0       0      0           0       1
##    Cotinus          0    0           0       0      0           0       3
##    Crataegus        0    0           0       0      0           0       1
##    Cytisus          0    0           0       0      0           0       1
##    Eucalyptus       0    0           0       0      0           0       5
##    Fagus            0    0           0       0      0           0       3
##    Ginkgo           0    0           0       0      0           0       0
##    Ilex             0    0           0       0      0           0       7
##    Liquidambar      0    0           0       0      0           0       0
##    Liriodendron     0    0           0       0      0           0       0
##    Lithocarpus      0    0           0       0      0           0       4
##    Magnolia         0    0           0       0      0           0       8
##    Morus            0    3           0       0      0           0       3
##    Olea             0    1           0       0      0           0       2
##    Phildelphus      0    0           0       0      0           0       5
##    Populus          0    0           0       3      0           0       9
##    Prunus           0    0           0       0      0           0      10
##    Pterocarya       0    0           0       0      0           0       3
##    Quercus          0    0           0       1      0           0     109
##    Rhododendron     0    0           0       0      0           0       3
##    Salix            0    0           0       0      0           0       9
##    Sorbus           0    0           0       0      0           0       2
##    Tilia            0    0           0       0      0           0       3
##    Ulmus            0    0           0       0      0           0       2
##    Viburnum         0    0           0       0      0           0      10
##    Zelkova          0    0           0       0      0           0       4
##                Reference
## Prediction      Rhododendron Salix Sorbus Tilia Ulmus Viburnum Zelkova
##    Acer                    0     0      0     0     0        0       0
##    Alnus                   0     0      0     0     0        0       0
##    Arundinaria             0     0      0     0     0        0       0
##    Betula                  0     0      0     0     0        0       0
##    Callicarpa              0     0      0     0     0        0       0
##    Castanea                0     0      0     0     0        0       0
##    Celtis                  0     0      0     0     0        0       0
##    Cercis                  0     0      0     0     0        0       0
##    Cornus                  0     0      0     0     0        0       0
##    Cotinus                 0     0      0     0     0        0       0
##    Crataegus               0     0      0     0     0        0       0
##    Cytisus                 0     0      0     0     0        0       0
##    Eucalyptus              0     0      0     0     0        0       0
##    Fagus                   0     0      0     0     0        0       0
##    Ginkgo                  0     0      0     0     0        0       0
##    Ilex                    0     0      0     0     0        0       0
##    Liquidambar             0     0      0     0     0        0       0
##    Liriodendron            0     0      0     0     0        0       0
##    Lithocarpus             0     0      0     0     0        0       0
##    Magnolia                0     0      0     0     0        0       0
##    Morus                   0     0      0     0     0        0       0
##    Olea                    0     0      0     0     0        0       0
##    Phildelphus             0     0      0     0     0        0       0
##    Populus                 0     0      0     0     0        0       0
##    Prunus                  0     0      0     0     0        0       0
```

```
##   Pterocarya              0     0      0     0     0        0       0
##   Quercus                 0     0      0     3     0        0       0
##   Rhododendron            3     0      0     0     0        0       0
##   Salix                   0     0      0     0     0        0       0
##   Sorbus                  0     0      0     0     0        0       0
##   Tilia                   0     0      0     2     0        0       0
##   Ulmus                   0     0      0     0     0        0       0
##   Viburnum                0     0      0     0     0        0       0
##   Zelkova                 0     0      0     0     0        0       0
##
## Overall Statistics
##
##                Accuracy : 0.487
##                  95% CI : (0.4333, 0.541)
##     No Information Rate : 0.7061
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3029
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Acer Class: Alnus Class: Arundinaria
## Sensitivity              0.62500      0.55556                 NA
## Specificity              0.95873      0.97872            0.98847
## Pos Pred Value           0.60606      0.58824                 NA
## Neg Pred Value           0.96178      0.97576                 NA
## Prevalence               0.09222      0.05187            0.00000
## Detection Rate           0.05764      0.02882            0.00000
## Detection Prevalence     0.09510      0.04899            0.01153
## Balanced Accuracy        0.79187      0.76714                 NA
##                      Class: Betula Class: Callicarpa Class: Castanea
## Sensitivity                     NA                NA        1.000000
## Specificity                0.97983           0.98559        0.991304
## Pos Pred Value                  NA                NA        0.400000
## Neg Pred Value                  NA                NA        1.000000
## Prevalence                 0.00000           0.00000        0.005764
## Detection Rate             0.00000           0.00000        0.005764
## Detection Prevalence       0.02017           0.01441        0.014409
## Balanced Accuracy               NA                NA        0.995652
##                      Class: Celtis Class: Cercis Class: Cornus
## Sensitivity                     NA            NA       0.40000
## Specificity                0.98559      0.991354       0.99703
## Pos Pred Value                  NA            NA       0.80000
## Neg Pred Value                  NA            NA       0.98246
## Prevalence                 0.00000      0.000000       0.02882
## Detection Rate             0.00000      0.000000       0.01153
## Detection Prevalence       0.01441      0.008646       0.01441
## Balanced Accuracy               NA            NA       0.69852
##                      Class: Cotinus Class: Crataegus Class: Cytisus
## Sensitivity                      NA         0.000000       0.375000
## Specificity                 0.98559         0.994220       0.997050
## Pos Pred Value                   NA         0.000000       0.750000
## Neg Pred Value                   NA         0.997101       0.985423
```

```
## Prevalence                  0.00000         0.002882       0.023055
## Detection Rate              0.00000         0.000000       0.008646
## Detection Prevalence        0.01441         0.005764       0.011527
## Balanced Accuracy               NA         0.497110       0.686025
##                      Class: Eucalyptus Class: Fagus Class: Ginkgo
## Sensitivity                   1.000000           NA            NA
## Specificity                   0.982659     0.991354      0.997118
## Pos Pred Value                0.142857           NA            NA
## Neg Pred Value                1.000000           NA            NA
## Prevalence                    0.002882     0.000000      0.000000
## Detection Rate                0.002882     0.000000      0.000000
## Detection Prevalence          0.020173     0.008646      0.002882
## Balanced Accuracy             0.991329           NA            NA
##                      Class: Ilex Class: Liquidambar Class: Liriodendron
## Sensitivity             1.000000            1.00000            1.000000
## Specificity             0.979710            1.00000            1.000000
## Pos Pred Value          0.222222            1.00000            1.000000
## Neg Pred Value          1.000000            1.00000            1.000000
## Prevalence              0.005764            0.01153            0.008646
## Detection Rate          0.005764            0.01153            0.008646
## Detection Prevalence    0.025937            0.01153            0.008646
## Balanced Accuracy       0.989855            1.00000            1.000000
##                      Class: Lithocarpus Class: Magnolia Class: Morus
## Sensitivity                      NA        0.400000           NA
## Specificity                 0.97983        0.976608      0.98271
## Pos Pred Value                   NA        0.200000           NA
## Neg Pred Value                   NA        0.991098           NA
## Prevalence                  0.00000        0.014409      0.00000
## Detection Rate              0.00000        0.005764      0.00000
## Detection Prevalence        0.02017        0.028818      0.01729
## Balanced Accuracy                NA        0.688304           NA
##                      Class: Olea Class: Phildelphus Class: Populus
## Sensitivity             0.250000                NA       0.750000
## Specificity             0.994169           0.98559       0.967930
## Pos Pred Value          0.333333                NA       0.214286
## Neg Pred Value          0.991279                NA       0.996997
## Prevalence              0.011527           0.00000       0.011527
## Detection Rate          0.002882           0.00000       0.008646
## Detection Prevalence    0.008646           0.01441       0.040346
## Balanced Accuracy       0.622085                NA       0.858965
##                      Class: Prunus Class: Pterocarya Class: Quercus
## Sensitivity                     NA                NA         0.4449
## Specificity                0.97118           0.98847         0.8627
## Pos Pred Value                  NA                NA         0.8862
## Neg Pred Value                  NA                NA         0.3929
## Prevalence                 0.00000           0.00000         0.7061
## Detection Rate             0.00000           0.00000         0.3141
## Detection Prevalence       0.02882           0.01153         0.3545
## Balanced Accuracy               NA                NA         0.6538
##                      Class: Rhododendron Class: Salix Class: Sorbus
## Sensitivity                     1.000000           NA           NA
## Specificity                     0.991279      0.97406     0.994236
## Pos Pred Value                  0.500000           NA           NA
## Neg Pred Value                  1.000000           NA           NA
```

```
## Prevalence                     0.008646      0.00000      0.000000
## Detection Rate                 0.008646      0.00000      0.000000
## Detection Prevalence           0.017291      0.02594      0.005764
## Balanced Accuracy              0.995640           NA            NA
##                      Class: Tilia Class: Ulmus Class: Viburnum
## Sensitivity              0.400000           NA              NA
## Specificity              0.988304     0.994236          0.9683
## Pos Pred Value           0.333333           NA              NA
## Neg Pred Value           0.991202           NA              NA
## Prevalence               0.014409     0.000000          0.0000
## Detection Rate           0.005764     0.000000          0.0000
## Detection Prevalence     0.017291     0.005764          0.0317
## Balanced Accuracy        0.694152           NA              NA
##                      Class: Zelkova
## Sensitivity                      NA
## Specificity                 0.97983
## Pos Pred Value                   NA
## Neg Pred Value                   NA
## Prevalence                 0.00000
## Detection Rate             0.00000
## Detection Prevalence       0.02017
## Balanced Accuracy                NA
```

(iv) Discuss briefly your results in (e)(iii) as compared to (d)(iii) **using concepts discussed in class**.

Our overall accuracy was slightly better for the linear kernel (51% accuracy) over the radial kernel (48.7%) when predicting for the test data. Because of the increasing complexitiy, radial kernel SVMs tend to overfit the training data, which we see here because we have a lower accuracy for the radial kernel. I would be curious to see what would have happened if we increased the `c` values for both radial and kernel SVMs because as discussed earlier, these observations have a lot of overlap. This makes it hard (if not impossible) to create a separating hyperplane if none of the training observations can be misclassified.