Module 4 Technique Practice

Jeff Hackmeister

2025-05-04

Introduction and Data Exploration

To examine the use of Support Vector Machines for classification methods, we'll be using the Iris dataset from base R to build a predictor for Iris species based on sepal and petal width and length. The dataset contains 150 observations, 50 of each species, and 4 measurement variables.

Support Vector Machines (SVM) are supervised learning models finds a hyper plane to separate the data between classes (Classifying Data Using Support Vector Machines (SVM) in R, 2018).

For this practice, we'll use the four numeric variables in the dataset to create the classification model.

To begin, we will load the necessary libraries and the dataset.

```r
library(e1071)
library(tidyverse)
library(GGally)
data(iris)
attach(iris)
```

To confirm the data matches our expectations from the documentation, we will examine the structure of the dataset.

```r
str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

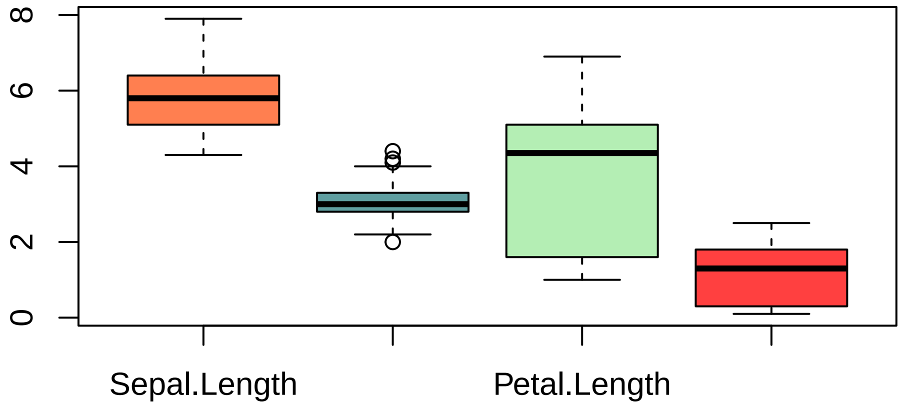
As expected, we have 150 observations of 5 variables. The first four are numeric and the third is a factor with 3 levels. For more detail, we will use the summary function to see the distribution of the numerical data.

```r
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
       Species  
 setosa    :50  
 versicolor:50  
 virginica :50  
```

And for a visual exploration of the data, we'll create a box plot.

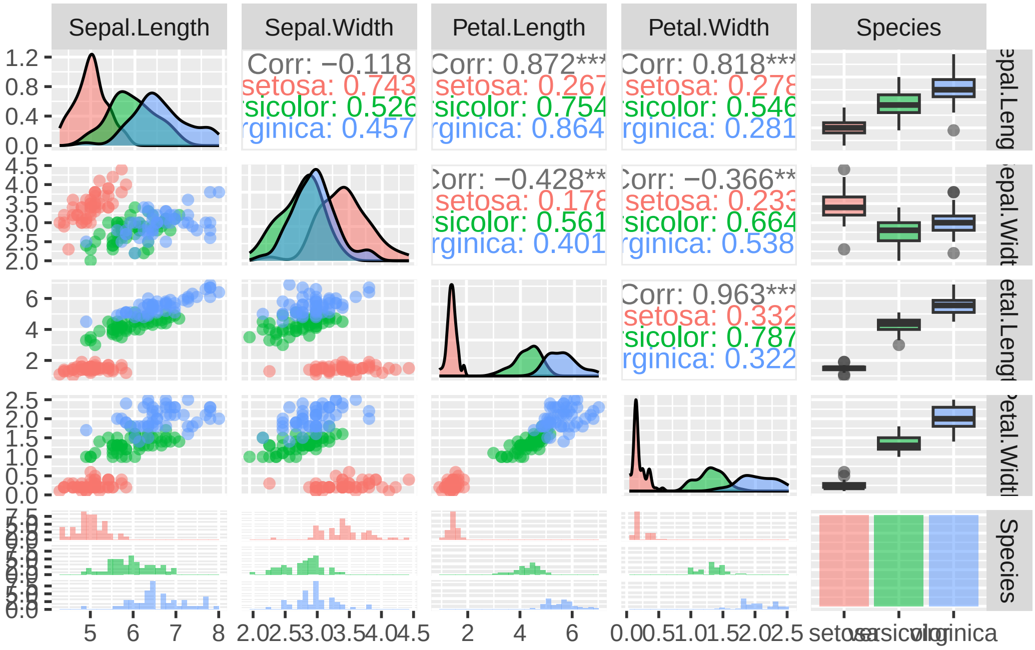
```
boxplot(iris[,-5], col = c("coral", "cadetblue", "darkseagreen2", "brown1"))
```



From this we can see that the data is pretty tightly distributed, and while there are a few outlier observations for sepal width, they are still relatively close to the rest of the data. We will keep this in mind as we go forward with the SVM.

To better understand the correlation between the variables, we will use the ggpairs function to create a pairs plot.

```
ggpairs(iris, ggplot2::aes(color = Species, alpha = 0.4))
```



From the scatter plots and the histograms along the bottom row, we can see that petal length and petal width appear to best divide the data by species. This will be helpful for possible refinement of our SVM as we continue.

Creating the SVM

To build the SVM, we will use the svm function from the e1071 package. By default, we will be using the classification method of an SVM, since we are working with categorical values, and a radial kernel.

```
model <- svm(Species ~ ., data = iris)

print(model)
```

```
Call:
svm(formula = Species ~ ., data = iris)
```

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  51
```

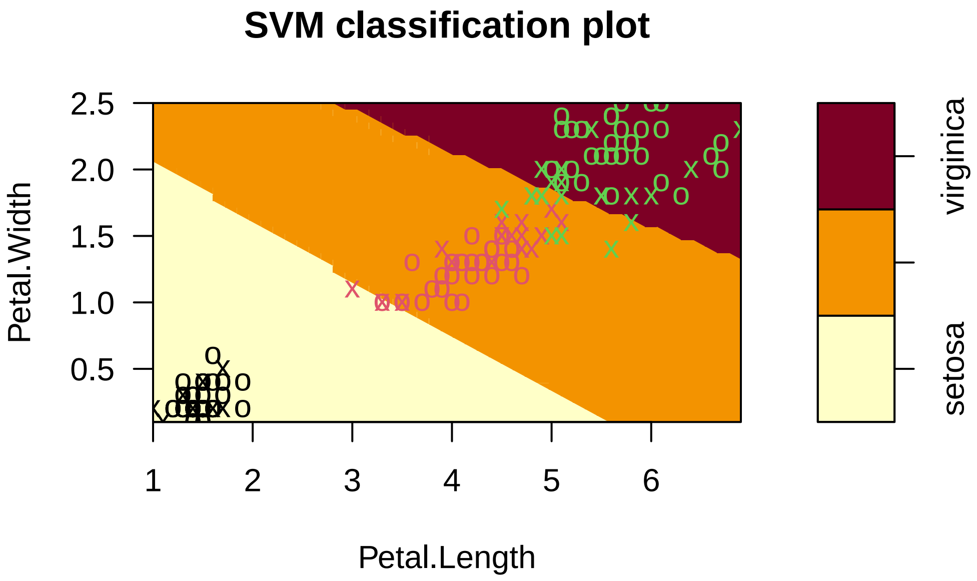
```r
summary(model)
```

```


Call:
svm(formula = Species ~ ., data = iris)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  51

 ( 8 22 21 )


Number of Classes:  3

Levels:
 setosa versicolor virginica
```

From the summary, we can see that the model produced 51 support vectors, these are the critical observations in the data that are closest to the hyper plane between our categories. Of these 51- 8 support setosa, 22 support versicolor, and 21 support virginica.

To better understand the division in the data, we can plot the model. We will use petal length and width as our axes.

```r
plot(model, data=iris,
     Petal.Width~Petal.Length,
     slice = list(Sepal.Width=3, Sepal.Length=4)
     )
```



In the plot, the 51 support vectors are Xs while the other observations are 0s. We can see a clear delineation for the setosas. There is some overlap between versicolor and virginica but these are promising results.

Predictions

Next, we will use the model to make predictions using the predict function to evaluate the effectiveness of the model.

```
table(Predicted = predict(model, newdata = iris),
      Actual = iris$Species)
```

```
            Actual
Predicted     setosa versicolor virginica
  setosa          50          0         0
  versicolor       0         48         2
  virginica        0          2        48
```

These are strong results for the model. The original dataset was a even distribution between the three species (50 observations each) and the SVM model correctly identified all 50 setosas, and 48 of each the versicolor and virginica categories. These results match the results from the plot above. Next, we'll take a look at the decision values from the model. These are the mathematical outputs from the prediction and represent the distance from the decision boundary.

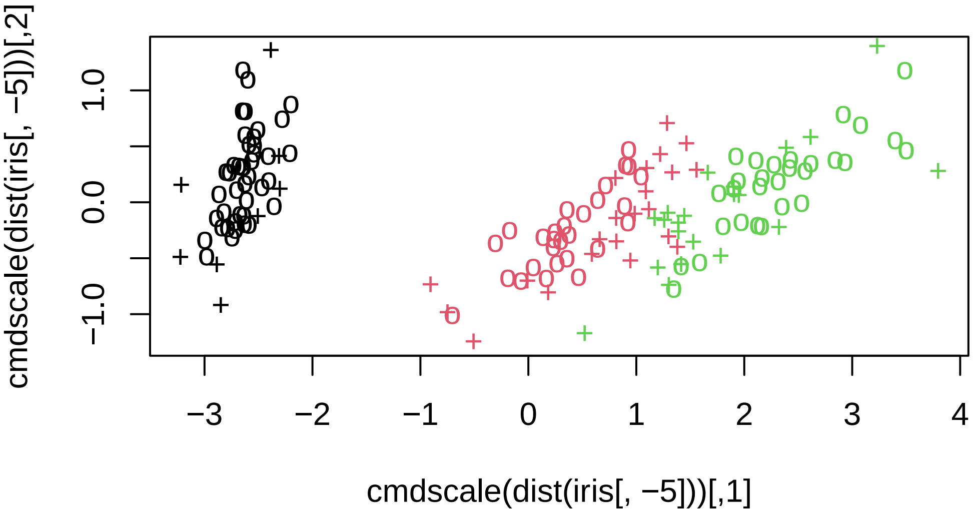
```
pred <- predict(model, iris[, 1:4], decision.values = TRUE)
attr(pred, "decision.values")[1:4,]
```

```
  setosa/versicolor setosa/virginica versicolor/virginica
1          1.196152         1.091757            0.6708810
2          1.064621         1.056185            0.8483518
3          1.180842         1.074542            0.6439798
4          1.110699         1.053012            0.6782041
```

All of the values in the table above are positive, meaning that when choosing between the first and second category, the model always chose the first. The larger values for setosa/versicolor and setosa/virginica indicate higher confidence in those decisions than when choosing between versicolor/virginica. This further validates the findings that there is a clear boundary for setosa, while there is more crossover between virginica and veriscolor.

For another look at the division, we'll use the cdmscale function to flatten the 4-dimensional model (created by 4 different variables) into a 2-dimensional plot.

```
plot(cmdscale(dist(iris[,-5])),
     col = as.integer(iris[,5]),
     pch = c("o","+")[1:150 %in% model$index + 1])
```



In this plot, the support vectors are marked by a +, while the other observations are shown as o. This plot clearly shows the clear classification of the setosa category and a less clear definition of the others.

Results and Recommendations

The results of our SVM for classification are quite promising for a relatively small dataset. The overlap between versicolor and virginica does leave considerable room for improvement. The initial dataset only provided 4 variables, it is entirely possible that the introduction of additional variables such as petal color or growing region could produce a even strong classification model. Additionally, including more observations - ideally from a different geographical region - could improve the strength and accurarcy of the classification.

References

[1] Classifying data using Support Vector Machines(SVMs) in R. (2018, August 28). GeeksforGeeks.

https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/

[2] e1071 package - RDocumentation. (2024). Rdocumentation.org.

https://www.rdocumentation.org/packages/e1071/versions/1.7-16

[3] iris function - RDocumentation. (n.d.). Www.rdocumentation.org.

https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/iris

[4] svm function - RDocumentation. (2024). Rdocumentation.org.

https://www.rdocumentation.org/packages/e1071/versions/1.7-16/topics/svm