DATA 609 HW 7

Daniel Moscoe

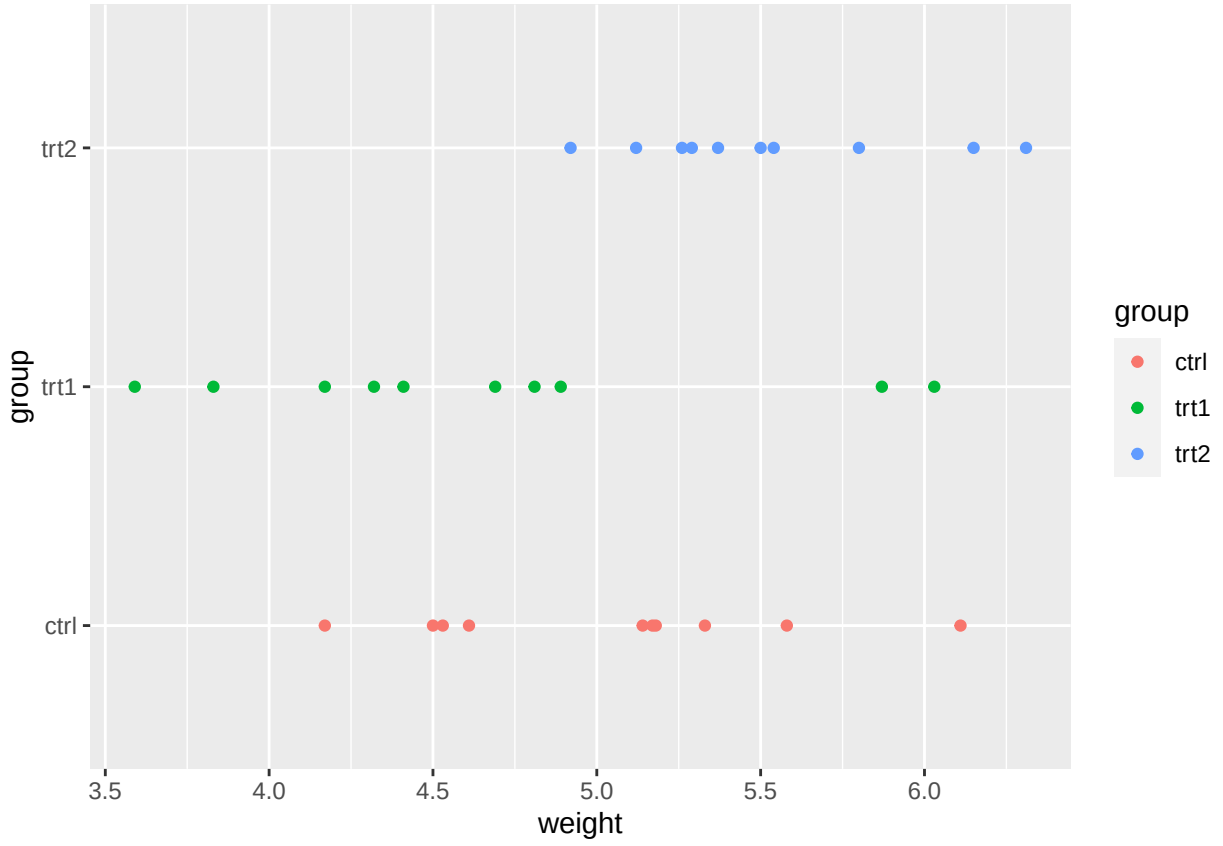
11/13/2021

Ex. 1. Use the `svm()` algorithm of the `e1071` package to carry out the support vector machine for the `PlantGrowth` data set. Then, discuss the number of support vectors/samples.

Response.

```
library(e1071)
library(tidyverse)
data(PlantGrowth)

ggplot(PlantGrowth) +
  geom_point(aes(x = weight, y = group, color = group))
```



Examining the data, we see that the groups are not separable based on weight due to overlap in each pair of groups. However, the plot also shows that mean weight for the trt1 group is less than mean weight for the other groups.

```
svm_model <- svm(group ~ weight,
                 data = PlantGrowth,
                 type = "C-classification",
                 kernel = "linear",
                 scale = FALSE)
svm_model
```

```
##
## Call:
## svm(formula = group ~ weight, data = PlantGrowth, type = "C-classification",
##     kernel = "linear", scale = FALSE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  28
```

The number of support vectors relative to the size of the dataset is very large: there are 28 support vectors for a dataset of 30 observations. The large number of support vectors shows that the svm's ability to distinguish groups with a linear kernel is poor. We can see this also by examining the model's accuracy:

```
pred_train <- predict(svm_model, PlantGrowth)
table(pred_train, PlantGrowth$group)
```

```
##
## pred_train ctrl trt1 trt2
##       ctrl    3    2    4
##       trt1    4    6    0
##       trt2    3    2    6
```

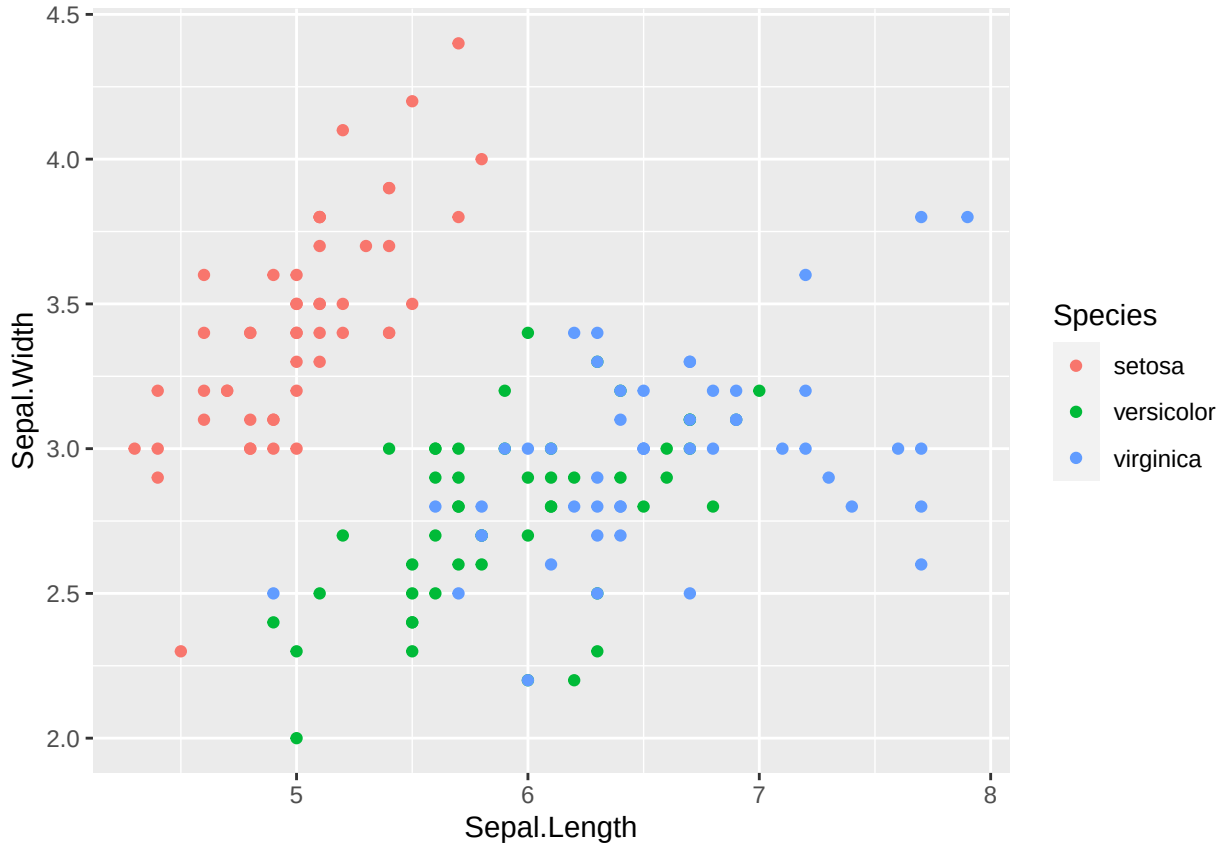
The overall accuracy is only 50%.

Ex. 2. Do a similar SVM analysis as that in the previous question using the `iris` data set. Discuss the number of support vectors/samples.

Response.

```
data(iris)

ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point()
```




```
ggplot(iris, aes(x = Petal.Length, y = Petal.Width, color = Species)) +
  geom_point()
```

Scatterplots of the Petal columns and the Sepal columns show that the Species are clustered with little overlap. This is especially evident in the plot of the Petal columns. This clustering suggests that an svm model will perform well.

```
svm_model <- svm(Species ~ .,
                 data = iris,
                 type = "C-classification",
                 kernel = "linear",
                 scale = FALSE)
svm_model
```

```
##
## Call:
## svm(formula = Species ~ ., data = iris, type = "C-classification",
##     kernel = "linear", scale = FALSE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  27
```

The number of support vectors is 27, or 18% of the data set. The number of support vectors relative to the total number of observations is much lower for this data than for `PlantGrowth`. Examining accuracy:

```
pred_train <- predict(svm_model, iris)
table(pred_train, iris$Species)
```

```
##
## pred_train   setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         49         0
##   virginica       0          1        50
```

The overall accuracy is over 99%.

Ex. 3. Use the `iris` data set (or any other data set) to select 80% of the samples for training `svm()`, then use the remaining 20% for validation. Discuss your results.

Response. How do results for the `iris` data set change when we set aside 20% of observations for testing?

```
set.seed(857)
training_rows <- sample(nrow(iris), 0.80 * nrow(iris), replace = FALSE)
iris_train <- iris[training_rows,]
iris_test <- iris[-training_rows,]

svm_model <- svm(Species ~ .,
                 data = iris_train,
                 type = "C-classification",
                 kernel = "linear",
                 scale = FALSE)

pred_train <- predict(svm_model, iris_train)
train_cm <- table(pred_train, iris_train$Species)
train_cm
```

```
##
## pred_train   setosa versicolor virginica
##   setosa         40          0         0
##   versicolor      0         35         0
##   virginica       0          2        43
```

Accuracy for the training set is 100%.

```
pred_test <- predict(svm_model, iris_test)
test_cm <- table(pred_test, iris_test$Species)
test_cm
```

```
##
## pred_test    setosa versicolor virginica
##   setosa         10          0         0
##   versicolor      0         13         0
##   virginica       0          0         7
```

Accuracy for the test set is also 100%.

Withholding 20% of the data from the training set did not impair model performance. The number of support vectors for the model declined slightly, to 15% of the training set.