Data 609 - Module7

Amit Kapoor

11/23/2021

Contents


```
# Libraries
library(e1071)
library(caret)
```

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

```
## Loading required package: ggplot2
```

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. [Install the e1071 package in R if needed.]

Solution

```
# PlantGrowth datset
data(PlantGrowth)
head(PlantGrowth)
```

```
##   weight group
## 1   4.17  ctrl
## 2   5.58  ctrl
## 3   5.18  ctrl
## 4   6.11  ctrl
## 5   4.50  ctrl
## 6   4.61  ctrl
```

```
svm_pg <- svm(group ~ ., data = PlantGrowth)
summary(svm_pg)
```

```
##
## Call:
## svm(formula = group ~ ., data = PlantGrowth)
##
##
## Parameters:
##    SVM-Type:  C-classification
```

```
##  SVM-Kernel:  radial
##        cost:  1
##
## Number of Support Vectors:  29
##
##  ( 10 9 10 )
##
##
## Number of Classes:  3
##
## Levels:
##  ctrl trt1 trt2
```

It comes out that there are 29 support vectors out of 30 samples that are closer to the hyperplane and influence the orientation and position of the hyperplane. There are 3 classes.

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.

Solution

```
# iris datset
data("iris")
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

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

```
##
## 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
```

There are 51 support vectors and 3 classes in this case out of 150 samples in the dataset. In this case there is very less data for support vectors as compared to `PlantGrowth` dataset.

Ex.3

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

Solution

```
set.seed(609)

# partitioning for training and validation
partition <- createDataPartition(iris$Species, p=0.80, list = FALSE)
training <- iris[partition,]
validation <- iris[-partition,]


svm_train_iris <- svm(Species ~ ., data = training)
summary(svm_train_iris)
```

```
##
## Call:
## svm(formula = Species ~ ., data = training)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##
## Number of Support Vectors:  45
##
##  ( 8 18 19 )
##
##
## Number of Classes:  3
##
## Levels:
##  setosa versicolor virginica
```

```
pred_valid <- predict(svm_train_iris, validation)
confusionMatrix(validation$Species, pred_valid)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         10          0         0
##   versicolor      0          9         1
##   virginica       0          0        10
##
## Overall Statistics
##
```

```
##                Accuracy : 0.9667
##                  95% CI : (0.8278, 0.9992)
##     No Information Rate : 0.3667
##     P-Value [Acc > NIR] : 4.476e-12
##
##                   Kappa : 0.95
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            1.0000           0.9091
## Specificity                 1.0000            0.9524           1.0000
## Pos Pred Value              1.0000            0.9000           1.0000
## Neg Pred Value              1.0000            1.0000           0.9500
## Prevalence                  0.3333            0.3000           0.3667
## Detection Rate              0.3333            0.3000           0.3333
## Detection Prevalence        0.3333            0.3333           0.3333
## Balanced Accuracy           1.0000            0.9762           0.9545
```

The accuracy of the model has comeout as 96.67%.Here the training model shows 45 support vectors.