Feature Selection

A J Dazzle

2022-11-15

```
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.1.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.3
```

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

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

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

```
data("Iris")
```

```
## Warning in data("Iris"): data set 'Iris' not found
```

```
View(iris)
```

calculate correlation matrix

```
correlationMatrix <- cor(iris[,1:3])
```

summarize the correlation matrix

```
print(correlationMatrix)
```

```
##              Sepal.Length Sepal.Width Petal.Length
## Sepal.Length    1.0000000  -0.1175698    0.8717538
## Sepal.Width    -0.1175698   1.0000000   -0.4284401
## Petal.Length    0.8717538  -0.4284401    1.0000000
```

find attributes that are highly corrected (ideally >0.75)

```
highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.5)
```

print indexes of highly correlated attributes

```
print(highlyCorrelated)
```

```
## [1] 3
```

ensure results are repeatable

```
set.seed(7)
```

prepare training scheme

```
?trainControl
```

```
## starting httpd help server ... done
```

```
control <- trainControl(method="repeatedcv", number=10, repeats=3)
```

train the model

```
?train
model <- train(Species~., data=iris, method="lvq", preProcess="scale",
               trControl=control)
```

estimate variable importance

```
?varImp
importance <- varImp(model, scale=FALSE)
```

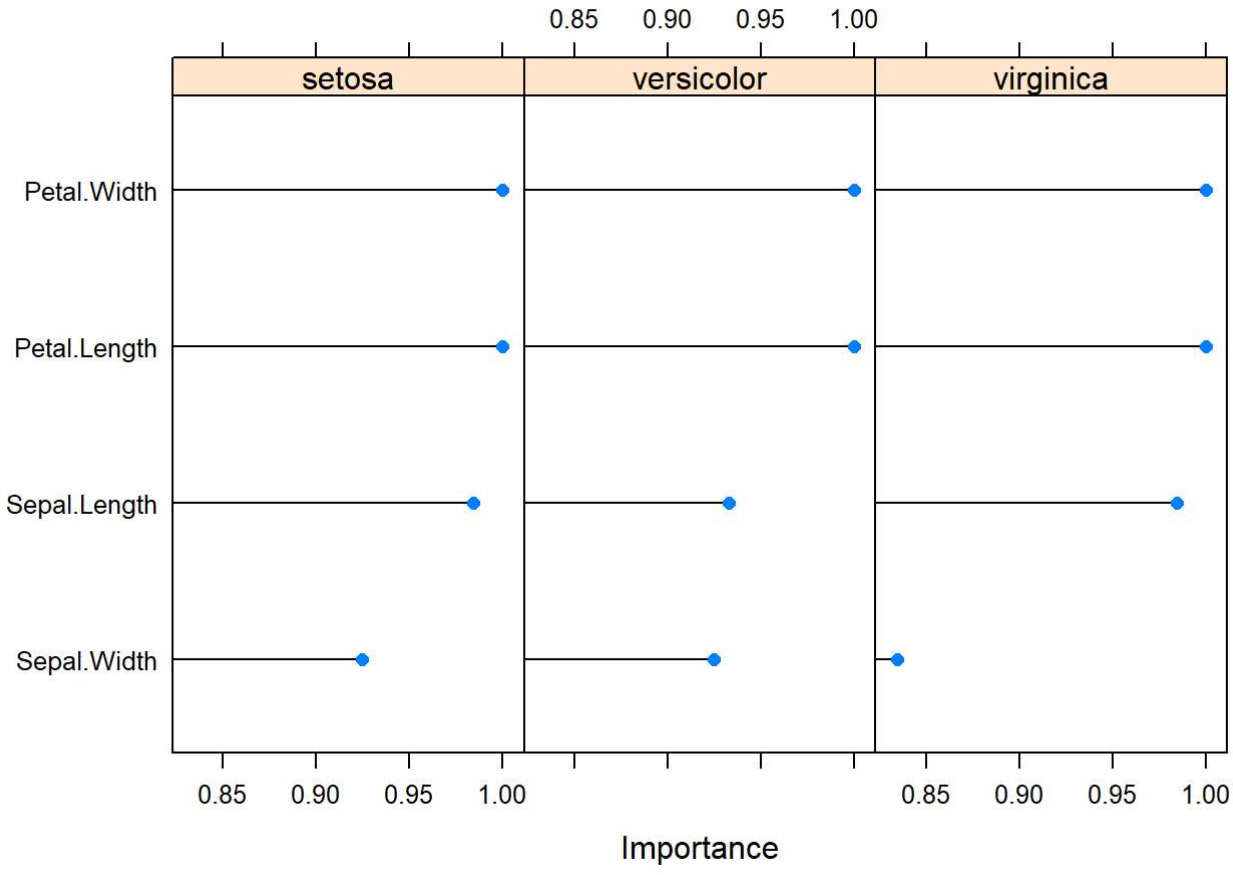
summarize importance

```
print(importance)
```

```
## ROC curve variable importance
##
##   variables are sorted by maximum importance across the classes
##              setosa versicolor virginica
## Petal.Width  1.0000     1.0000    1.0000
## Petal.Length 1.0000     1.0000    1.0000
## Sepal.Length 0.9846     0.9326    0.9846
## Sepal.Width  0.9248     0.9248    0.8344
```

plot importance

```
plot(importance)
```



#

define the control using a random forest selection function

```
control <- rfeControl(functions=rfFuncs, method="cv", number=10)
```

run the RFE algorithm

```
results <- rfe(iris[,1:3], iris[,3], sizes=c(1:3),
               rfeControl=control)
```

summarize the results

```
print(results)
```

```
##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold)
##
## Resampling performance over subset size:
##
##  Variables    RMSE Rsquared     MAE  RMSESD RsquaredSD   MAESD Selected
##          1 0.04556   0.9992 0.02161 0.02798   0.001164 0.01124        *
##          2 0.16585   0.9907 0.10884 0.09059   0.011897 0.05471
##          3 0.24869   0.9846 0.16565 0.09202   0.009355 0.05549
##
## The top 1 variables (out of 1):
##    Petal.Length
```

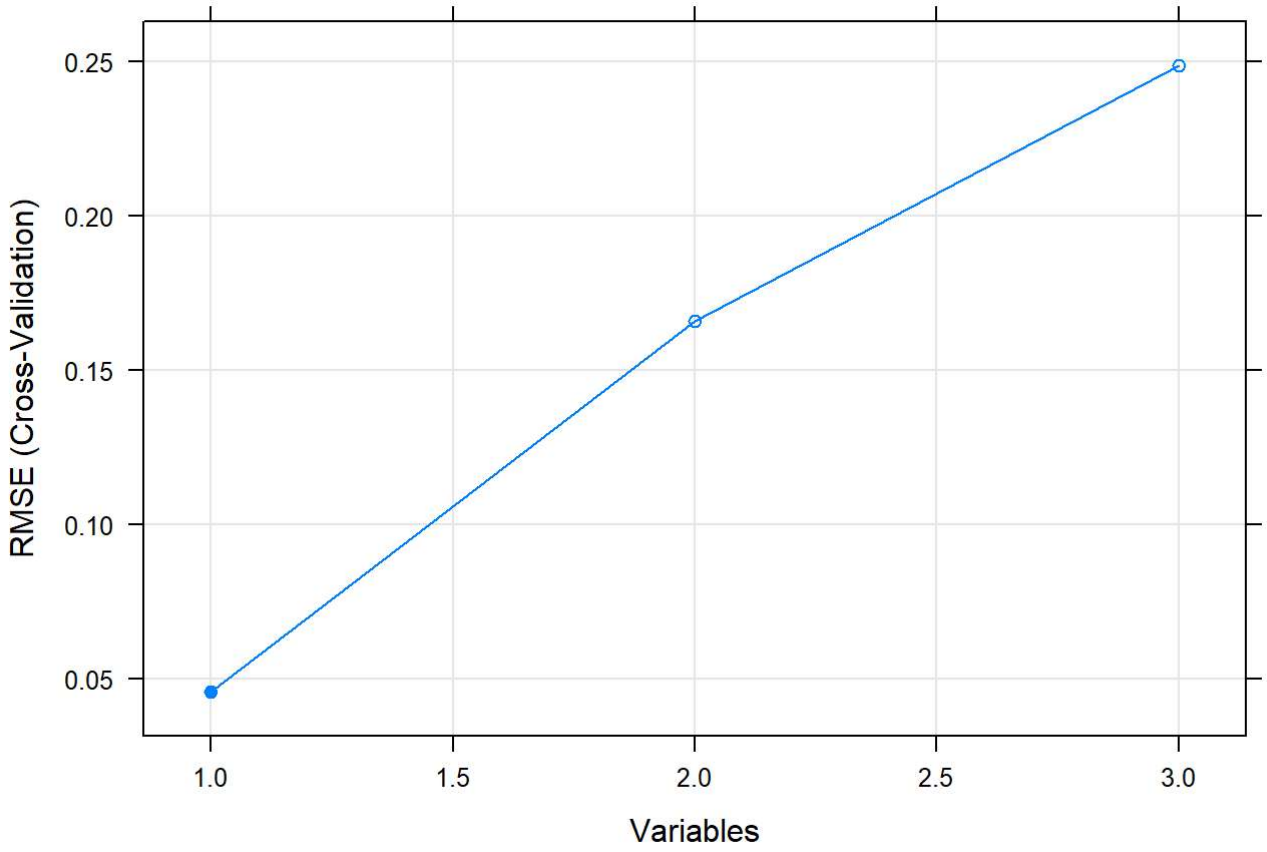
list the chosen features

```
predictors(results)
```

```
## [1] "Petal.Length"
```

plot the results

```
plot(results, type=c("g", "o"))
```


RMSE (Cross-Validation)
0.25
0.20
0.15
0.10
0.05
1.0
1.5
2.0
2.5
3.0
Variables