Feature Selection Approach

ashish dutt

February 9, 2018

Feature Selection Approach

Finding the most important predictor variables (of features) that explains major part of variance of the response variable is key to identify and build high performing models.

Boruta package

```
# Install the package if your using this library for the first time
# install.packages("Boruta", dependencies = TRUE)
# load the package
library(Boruta)
```

```
## Loading required package: ranger
```

```
# load the BostonHousing dataset
data("BostonHousing", package = "mlbench")
str(BostonHousing)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : num  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ b      : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Decide if a variable is important or not using Boruta

```
boruta_output <- Boruta(medv ~ ., data=na.omit(BostonHousing), doTrace=2)
```

```
##  1. run of importance source...
##  2. run of importance source...
##  3. run of importance source...
##  4. run of importance source...
##  5. run of importance source...
##  6. run of importance source...
##  7. run of importance source...
```

```
##  8. run of importance source...
##  9. run of importance source...
##  10. run of importance source...
##  11. run of importance source...
## After 11 iterations, +10 secs:
##  confirmed 13 attributes: age, b, chas, crim, dis and 8 more;
##  no more attributes left.
```

```
boruta_signif <- names(boruta_output$finalDecision[boruta_output$finalDecision %in% c("Confirmed", "Ten
print(boruta_signif)  # significant variables
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "b"       "lstat"
```

Plot the variable importance

```
plot(boruta_output, cex.axis=.7, las=2, xlab="", main="Variable Importance")  # plot variable importanc
```