caret

Yue Jin

11/8/2017

The caret package is a set of tools for building machine learning models in R. The name "caret" stands for Classification And REgression Training. As the name implies, the caret package gives you a toolkit for building classification models and regression models. Moreover, caret provides you with essential tools for:

There are many different modeling functions in R. Some have different syntax for model training and/or prediction. The package started off as a way to provide a uniform interface the functions themselves, as well as a way to standardize common tasks (such parameter tuning and variable importance).

One of the primary tools in the package is the train function which can be used to - evaluate, using resampling, the effect of model tuning parameters on performance - choose the "optimal" model across these parameters - estimate model performance from a training set

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
ind<-sample(1:nrow(mtcars),0.75*nrow(mtcars))
ind<-createDataPartition(
  y=mtcars$wt,
  ## the outcome data are needed
  p = .75,
  ## The percentage of data in the training set
  list = FALSE
  ## The format of the results
  )
mtcars_train<-mtcars[ind,]
mtcars_test<-mtcars[-ind,]
# create training index
```

```
# train(formula, dataset, method)
mtcarsFit <- train(mpg ~ wt,
                         mtcars_train,
                         method = "lm",
                         ## Center and scale the predictors for the training
                         preProc = c("center", "scale")
                        )
c <- coef(mtcarsFit$finalModel)[1]
b <- coef(mtcarsFit$finalModel)[2]
```

Other customization: However, we would probably like to customize it in a few ways: - tuning grid/parameter search method - the type of resampling used. The simple bootstrap is used by default. We can use cross validation - the methods for measuring performance. If unspecified, overall accuracy and the Kappa statistic are computed. For regression models, root mean squared error and R2 are computed.

Prediction

```
pred <- predict(mtcarsFit, newdata = mtcars_test)
pred
```

```
##       Mazda RX4 Wag            Merc 280         Merc 450SLC
##           22.204031           19.168536           17.341866
## Lincoln Continental           Fiat X1-9        Lotus Europa
##            8.509382           27.254235           29.521454
##      Ford Pantera L       Maserati Bora
##           20.619126           18.470103
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.