Validation_CrossValidation.R

vijaykalmath

2022-01-04

```
# Validation and Cross Validation  Approach

# We must use only the training examples to perform all aspects of the model-fitting including the vari

# If the full data set is used to perform the best subset selection step , the validation set errors an


# set seed for reproducibility
set.seed(1)

library(ISLR)
library(leaps)

Hitters = na.omit(Hitters)


train = sample(c(TRUE,FALSE),nrow(Hitters),rep=TRUE)

test = (!train)

regfit.best = regsubsets(Salary~.,data=Hitters[train,],nvmax = 19 )

summary(regfit.best)
```

```
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., data = Hitters[train, ], nvmax = 19)
## 19 Variables  (and intercept)
##            Forced in Forced out
## AtBat          FALSE      FALSE
## Hits           FALSE      FALSE
## HmRun          FALSE      FALSE
## Runs           FALSE      FALSE
## RBI            FALSE      FALSE
## Walks          FALSE      FALSE
## Years          FALSE      FALSE
## CAtBat         FALSE      FALSE
## CHits          FALSE      FALSE
## CHmRun         FALSE      FALSE
## CRuns          FALSE      FALSE
## CRBI           FALSE      FALSE
## CWalks         FALSE      FALSE
## LeagueN        FALSE      FALSE
```

```
## DivisionW      FALSE      FALSE
## PutOuts        FALSE      FALSE
## Assists        FALSE      FALSE
## Errors         FALSE      FALSE
## NewLeagueN     FALSE      FALSE
## 1 subsets of each size up to 19
## Selection Algorithm: exhaustive
##           AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI
## 1  ( 1 )  " "   " "  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "
## 2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "
## 3  ( 1 )  " "   " "  " "   " "  " " " "*"   " "   "*"    "*"   " "    " "   " "
## 4  ( 1 )  " "   " "  " "   " "  " " " "*"   " "   "*"    "*"   " "    " "   " "
## 5  ( 1 )  " "   " "  " "   " "  " " " "*"   " "   "*"    "*"   " "    " "   " "
## 6  ( 1 )  " "   " "  " "   " "  " " " "*"   " "   "*"    "*"   " "    "*"   " "
## 7  ( 1 )  "*"   "*"  " "   " "  " " " "*"   " "   " "    " "   " "    "*"   " "
## 8  ( 1 )  "*"   "*"  "*"   " "  " " " "*"   " "   " "    " "   " "    "*"   " "
## 9  ( 1 )  "*"   "*"  "*"   " "  " " " "*"   " "   " "    " "   "*"    "*"   " "
## 10  ( 1 ) "*"   "*"  "*"   " "  " " " "*"   " "   "*"    " "   " "    "*"   "*"
## 11  ( 1 ) "*"   "*"  "*"   " "  " " " "*"   " "   "*"    " "   " "    "*"   "*"
## 12  ( 1 ) "*"   "*"  "*"   " "  " " " "*"   " "   "*"    " "   " "    "*"   "*"
## 13  ( 1 ) "*"   "*"  "*"   " "  " " " "*"   " "   "*"    " "   "*"    "*"   "*"
## 14  ( 1 ) "*"   "*"  "*"   " "  " " " "*"   "*"   "*"    " "   "*"    "*"   "*"
## 15  ( 1 ) "*"   "*"  "*"   " "  " " " "*"   "*"   "*"    " "   "*"    "*"   "*"
## 16  ( 1 ) "*"   "*"  "*"   "*"  " " " "*"   "*"   "*"    "*"   "*"    "*"   "*"
## 17  ( 1 ) "*"   "*"  "*"   "*"  " " " "*"   "*"   "*"    "*"   "*"    "*"   "*"
## 18  ( 1 ) "*"   "*"  "*"   "*"  " " " "*"   "*"   "*"    "*"   "*"    "*"   "*"
## 19  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"    "*"   "*"    "*"   "*"
##           CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN
## 1  ( 1 )  " "    " "     " "       " "     " "     " "    " "
## 2  ( 1 )  " "    " "     " "       " "     " "     " "    " "
## 3  ( 1 )  " "    " "     " "       " "     " "     " "    " "
## 4  ( 1 )  " "    " "     "*"       " "     " "     " "    " "
## 5  ( 1 )  " "    " "     "*"       "*"     " "     " "    " "
## 6  ( 1 )  " "    " "     "*"       "*"     " "     " "    " "
## 7  ( 1 )  "*"    " "     "*"       "*"     " "     " "    " "
## 8  ( 1 )  "*"    " "     "*"       "*"     " "     " "    " "
## 9  ( 1 )  "*"    " "     "*"       "*"     " "     " "    " "
## 10  ( 1 ) "*"    " "     "*"       "*"     " "     " "    " "
## 11  ( 1 ) "*"    "*"     "*"       "*"     " "     " "    " "
## 12  ( 1 ) "*"    "*"     "*"       "*"     "*"     " "    " "
## 13  ( 1 ) "*"    "*"     "*"       "*"     "*"     " "    " "
## 14  ( 1 ) "*"    " "     "*"       "*"     "*"     " "    "*"
## 15  ( 1 ) "*"    "*"     "*"       "*"     "*"     " "    "*"
## 16  ( 1 ) "*"    " "     "*"       "*"     "*"     " "    "*"
## 17  ( 1 ) "*"    "*"     "*"       "*"     "*"     " "    "*"
## 18  ( 1 ) "*"    "*"     "*"       "*"     "*"     "*"    "*"
## 19  ( 1 ) "*"    "*"     "*"       "*"     "*"     "*"    "*"
```

```
test.mat = model.matrix(Salary~.,data=Hitters[test,])

val.error  = rep(NA,19)

for (i in 1:19){
  coefi  = coef(regfit.best,id=i)
```

```
  pred <- test.mat[, names(coefi)] %*% coefi

  val.error[i]= mean((Hitters$Salary[test]-pred)^2)
}

val.error
```

```
##  [1] 164377.3 144405.5 152175.7 145198.4 137902.1 139175.7 126849.0 136191.4
##  [9] 132889.6 135434.9 136963.3 140694.9 140690.9 141951.2 141508.2 142164.4
## [17] 141767.4 142339.6 142238.2
```

```
# min value of val.error

min(val.error)
```

```
## [1] 126849
```

```
# Associated number of variables count

which.min(val.error)
```

```
## [1] 7
```

```
# Get Coefficients of Best Validation error Model
coef(regfit.best,id=which.min(val.error))
```

```
##  (Intercept)        AtBat         Hits        Walks        CRuns       CWalks
##   67.1085369   -2.1462987    7.0149547    8.0716640    1.2425113   -0.8337844
##    DivisionW      PutOuts
## -118.4364998    0.2526925
```

```
# Cross Validation Approach

###
predict.regsubsets <- function(object, newdata, id, ...) {
  form <- as.formula(object$call[[2]])
  mat <- model.matrix(form, newdata)
  coefi <- coef(object, id = id)
  xvars <- names(coefi)
  mat[, xvars] %*% coefi
}



# Define number of K-Fold Cross Validation
k <- 10

n <- nrow(Hitters)

folds <- sample(rep(1:k, length = n))
```

```
cv.errors <- matrix(NA, k, 19,
                    dimnames = list(NULL, paste(1:19)))
for (j in 1:k) {
  best.fit <- regsubsets(Salary ~ .,
                         data = Hitters[folds != j, ],
                         nvmax = 19)
  for (i in 1:19) {
    pred <- predict(best.fit, Hitters[folds == j, ], id = i)
    cv.errors[j, i] <-
      mean((Hitters$Salary[folds == j] - pred)^2)
  }
}

mean.cv.errors <- apply(cv.errors, 2, mean)

mean.cv.errors
```

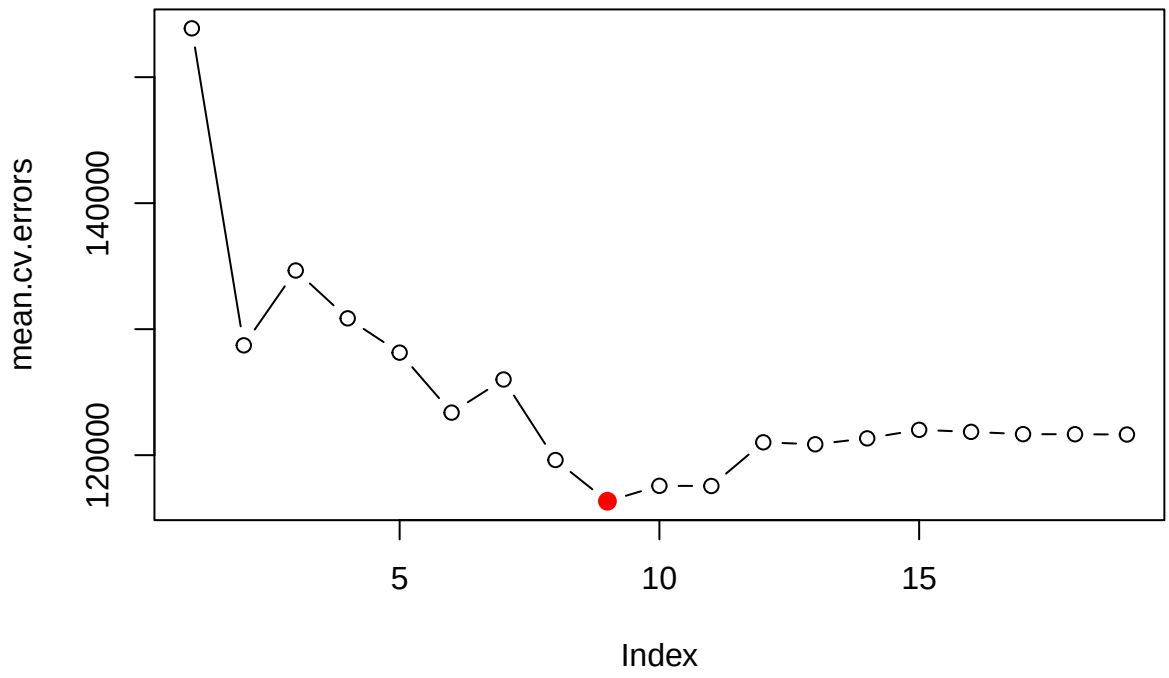
```
##        1        2        3        4        5        6        7        8
## 153874.7 128718.1 134656.1 130860.0 128137.0 123377.3 126006.9 119616.2
##        9       10       11       12       13       14       15       16
## 116343.9 117567.5 117554.5 121021.0 120864.0 121334.4 122011.5 121850.3
##       17       18       19
## 121668.9 121660.7 121640.2
```

```
which.min(mean.cv.errors)
```

```
## 9
## 9
```

```
par(mfrow = c(1, 1))

plot(mean.cv.errors, type = "b") + points(which.min(mean.cv.errors),min(mean.cv.errors),col="red",pch=2
```




```
## integer(0)
```

```
reg.best <- regsubsets(Salary ~ ., data = Hitters,
                       nvmax = 19)

coef(reg.best, which.min(mean.cv.errors))
```

```
##   (Intercept)         AtBat          Hits         Walks        CAtBat
##  146.24960033   -1.93676754    6.65672102    5.55204413   -0.09953904
##         CRuns          CRBI        CWalks     DivisionW       PutOuts
##    1.25067124    0.66176849   -0.77798498 -115.34950146    0.27773062
```