Lab13 - Cross Validation e Bootstrap na prática

Machine Learning usando o R - Análise Macro

Thalles Quinaglia Liduares

2022-09-01

Irei utilizar o dataset `wage1` do pacote `wooldridge` para modelar o retorno salarial em função dos anos de educação dos individuos.

Upload pacotes

```
library(wooldridge)
library(boot)
```

Upload base de dados

```
data<-wooldridge::wage1

attach(data)
```

Partição da amostra em treino e teste

```
treino<-sample(526, 421)
```

Modelo simples

```
mod1<-lm(lwage~educ, data, subset = treino)

summary(mod1)
```

```
##
## Call:
## lm(formula = lwage ~ educ, data = data, subset = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.21345 -0.36961 -0.07449  0.29323  1.52152
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.563206   0.115026   4.896  1.4e-06 ***
## educ        0.084614   0.008932   9.473  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4925 on 419 degrees of freedom
## Multiple R-squared:  0.1764, Adjusted R-squared:  0.1744
## F-statistic: 89.73 on 1 and 419 DF,  p-value: < 2.2e-16
```

```
mean((lwage-predict(mod1, data))[-treino]^2)
```

```
## [1] 0.1825805
```

Modelo quadratico

```
mod2<-lm(lwage~poly(educ,2), data, subset=treino)

summary(mod2)
```

```
##
## Call:
## lm(formula = lwage ~ poly(educ, 2), data = data, subset = treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1546 -0.3389 -0.1093  0.3129  1.5804
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.62932    0.02353  69.248  < 2e-16 ***
## poly(educ, 2)1  5.02582    0.56097   8.959  < 2e-16 ***
## poly(educ, 2)2  2.57180    0.59841   4.298 2.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4825 on 418 degrees of freedom
## Multiple R-squared:  0.2112, Adjusted R-squared:  0.2075
## F-statistic: 55.97 on 2 and 418 DF,  p-value: < 2.2e-16
```

EQM modelo 2

```
mean((lwage-predict(mod2,data))[-treino]^2)
```

```
## [1] 0.1839318
```

Modelo cubico

```
mod3<-lm(lwage~poly(educ,3), data, subset=treino)

summary(mod3)
```

```
##
## Call:
## lm(formula = lwage ~ poly(educ, 3), data = data, subset = treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1484 -0.3349 -0.1025  0.3142  1.5866
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.62818    0.02354  69.159  < 2e-16 ***
## poly(educ, 3)1  5.08457    0.56316   9.029  < 2e-16 ***
## poly(educ, 3)2  2.43615    0.61003   3.993 7.69e-05 ***
## poly(educ, 3)3  0.67594    0.59573   1.135    0.257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4823 on 417 degrees of freedom
## Multiple R-squared:  0.2137, Adjusted R-squared:  0.208
## F-statistic: 37.77 on 3 and 417 DF,  p-value: < 2.2e-16
```

EQM modelo cubico

```
mean((lwage-predict(mod3, data))[-treino]^2)
```

```
## [1] 0.1856669
```

O modelo que performa o menor EQM é o quadratico.

k-fold cross validation

```
set.seed(3108)

v_erro_10<-rep(0,10)

for (i in 1:10){
  glm.fit=glm(lwage~poly(educ,i),data=data)
  v_erro_10[i]=cv.glm(data,glm.fit, K=10)$delta[1]
}

v_erro_10
```

```
##  [1] 0.2323132 0.2240438 0.2234156 0.2239069 0.2253518 0.2251812 0.2277157
##  [8] 0.2246789 0.2333677 0.2249320
```

O modelo quadrático apresenta um EQM relativamente baixo em relaçao aos demais e menos complexo em termos de interpretabilidade dos coeficientes, portanto é o modelo escolhido.

Bootstrap

```
boot.fn=function(data,index)
return(coef(lm(lwage~educ,data=data,subset=index)))
boot.fn(data,1:526)
```

```
## (Intercept)        educ
##  0.58377267  0.08274437
```

```
set.seed(0109)

boot.fn(data,sample(526,526,replace=T))
```

```
## (Intercept)        educ
##  0.49204389  0.09181553
```

```
boot(data,boot.fn,1000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = data, statistic = boot.fn, R = 1000)
##
##
## Bootstrap Statistics :
##       original        bias    std. error
## t1* 0.58377267 -0.0044412131 0.099120744
## t2* 0.08274437  0.0003556734 0.007850796
```