3.2 交叉验证

徐平峰

长春工业大学

2021/3/27

Boston房价数据

```
setwd("E:/teaching_plan_notes/msa11091083/rmd")#设置工作目录
w=read.csv("MVAPureData/BostonHousing2.csv")#读取Boston房价数据
w=w[,-c(1:5)]#去掉前5列变量，只分析后14个变量
a=lm(cmedv~., data=w)#以cmedv为因变量，其他为自变量，拟合最小二乘回归
a$coef#最小二乘估计
```

```
#   (Intercept)          crim            zn         indus          chas
#  3.637189e+01 -1.062004e-01  4.772368e-02  2.325237e-02  2.691727e+00
#           nox            rm           age           dis           rad
# -1.774262e+01  3.789395e+00  5.749168e-04 -1.501794e+00  3.037606e-01
#           tax       ptratio             b         lstat
# -1.270462e-02 -9.239118e-01  9.228445e-03 -5.306619e-01
```

```
(newdata=as.data.frame(w[1,-1]))#新的数据
```

```
#      crim zn indus chas   nox    rm  age  dis rad tax ptratio     b lstat
# 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.09   1 296    15.3 396.9  4.98
```

```
predict(a, newdata=newdata)#预测新数据的
```

```
#        1
# 29.97626
```

- 10折交叉验证

```
set.seed(1010)#
(n=nrow(w))#w数据的函数为n
```

```
# [1] 506
```

```
Z=10#10折
n/Z#平均每组的样本数
```

```
# [1] 50.6
```

```
ceiling(n/Z)#取整数
```

```
# [1] 51
```

```
(1:Z)#从1到Z的整数
```

```
#  [1]  1  2  3  4  5  6  7  8  9 10
```

```
(m=rep(1:Z,ceiling(n/Z)))#将1:Z重复ceiling(n/Z)次
```

```
#   [1]  1  2  3  4  5  6  7  8  9 10  1  2  3  4  5  6  7  8  9 10  1  2  3
#  [26]  6  7  8  9 10  1  2  3  4  5  6  7  8  9 10  1  2  3  4  5  6  7  8
#  [51]  1  2  3  4  5  6  7  8  9 10  1  2  3  4  5  6  7  8  9 10  1  2  3
#  [76]  6  7  8  9 10  1  2  3  4  5  6  7  8  9 10  1  2  3  4  5  6  7  8
# [101]  1  2  3  4  5  6  7  8  9 10  1  2  3  4  5  6  7  8  9 10  1  2  3
# [126]  6  7  8  9 10  1  2  3  4  5  6  7  8  9 10  1  2  3  4  5  6  7  8
```

- 10折交叉验证

```
(m1=sample(m))#得到m中的数的一个随机的全排列
```

```
#   [1]  8  6  4 10  4 10  9  2  2  5  6  3  7  5  6  1  3  9  1  8  1  1  9
#  [26]  4  2  5  8  3  5  2  1  6  4  3  9  6  8  9  9 10  8  7  1 10  6  4
#  [51]  5 10  9  9  7  6 10  1  3 10  7  8  2  6  5  2  6  5  1  9 10  7 10
#  [76]  6  8  3  6  3  4  2  9  3  4  7  1  1  1  3  9  1  7  2  1  9  1  4
# [101]  7  4  2  3  4  3  4  5 10  7  2  6  3 10  4  6  9 10  5  7  2 10  8
# [126]  2  7  1  9  4  5  6  7  8  1  5  4  6  7  2  5  7  7  9  2  2  1  1
# [151]  6  4  4  8 10  5  8  9  2  5  3  4  2  8  2  1  3 10  9  8  4  3  4
# [176]  5 10  6  2  7  6  4  7  9  8  6  5  1  6 10  6  9  7  8  4 10  3  5
# [201] 10  7  6  3 10  8  6  2  5  7  3  3  3  9  1  4  3  1  8 10  6  2  4
# [226]  5  6 10  3  8  5  9  7  3  2  3  1  2  5  9  4  7  1  8  2  1  1  2
# [251]  3  5  3  4  1  2  5  3  8  7  9  2  6  4  7  8  4  7  9 10  1  2  8
# [276] 10  9  9  9  1  9  9 10  5  6  6  8  6  7 10  4  8  2 10  8  9  4 10
# [301]  3  4  5  9  5  7  4  6  4  5  9  6  4  1  5  5  8  4  1  7  8  4  3
# [326]  5  5  6  2  2  9  9 10  5  3  7  3  2  4  2  5  5  1  6  8  9  1  5
# [351]  8  6  1  8  7  2  5  4 10 10  2  4  5  5  8  4  3  9  9 10  4  1  9
# [376]  6  3  1  1  6  2  2  7  7  8  5  9  6  1  2  1  2 10  8  9  7  7  8
# [401]  7 10  8  4  8  9  5  4  3 10  8  6  2  8  8  2  1 10  8  7  3  2  1
# [426]  2  8  2  5  3  1  3  8 10  9  2 10  9  3  5  5  3  7  9  6  5  6 10
# [451] 10  9  5  7  1  2  7  7  1  9  4 10  3  2  3  3 10  9  4  8  3 10  2
# [476]  1 10  4 10  8  7  8  5  4  4  6  4  3  3 10  6  7  5  1  6  2  2  7
# [501]  8  6  1  8  7  3  5  1  6  8
```

- 10折交叉验证

```
(mm=m1[1:n])#取m2的从1到第n个元素，记为mm
```

```
#    [1]  8  6  4 10  4 10  9  2  2  5  6  3  7  5  6  1  3  9  1  8  1  1  9
#   [26]  4  2  5  8  3  5  2  1  6  4  3  9  6  8  9  9 10  8  7  1 10  6  4
#   [51]  5 10  9  9  7  6 10  1  3 10  7  8  2  6  5  2  6  5  1  9 10  7 10
#   [76]  6  8  3  6  3  4  2  9  3  4  7  1  1  1  3  9  1  7  2  1  9  1  4
#  [101]  7  4  2  3  4  3  4  5 10  7  2  6  3 10  4  6  9 10  5  7  2 10  8
#  [126]  2  7  1  9  4  5  6  7  8  1  5  4  6  7  2  5  7  7  9  2  2  1  1
#  [151]  6  4  4  8 10  5  8  9  2  5  3  4  2  8  2  1  3 10  9  8  4  3  4
#  [176]  5 10  6  2  7  6  4  7  9  8  6  5  1  6 10  6  9  7  8  4 10  3  5
#  [201] 10  7  6  3 10  8  6  2  5  7  3  3  3  9  1  4  3  1  8 10  6  2  4
#  [226]  5  6 10  3  8  5  9  7  3  2  3  1  2  5  9  4  7  1  8  2  1  1  2
#  [251]  3  5  3  4  1  2  5  3  8  7  9  2  6  4  7  8  4  7  9 10  1  2  8
#  [276] 10  9  9  9  1  9  9 10  5  6  6  8  6  7 10  4  8  2 10  8  9  4 10
#  [301]  3  4  5  9  5  7  4  6  4  5  9  6  4  1  5  5  8  4  1  7  8  4  3
#  [326]  5  5  6  2  2  9  9 10  5  3  7  3  2  4  2  5  5  1  6  8  9  1  5
#  [351]  8  6  1  8  7  2  5  4 10 10  2  4  5  5  8  4  3  9  9 10  4  1  9
#  [376]  6  3  1  1  6  2  2  7  7  8  5  9  6  1  2  1  2 10  8  9  7  7  8
#  [401]  7 10  8  4  8  9  5  4  3 10  8  6  2  8  8  2  1 10  8  7  3  2  1
#  [426]  2  8  2  5  3  1  3  8 10  9  2 10  9  3  5  5  3  7  9  6  5  6 10
#  [451] 10  9  5  7  1  2  7  7  1  9  4 10  3  2  3  3 10  9  4  8  3 10  2
#  [476]  1 10  4 10  8  7  8  5  4  4  6  4  3  3 10  6  7  5  1  6  2  2  7
#  [501]  8  6  1  8  7  3
```

- 10折交叉验证

```
N=1:n#N为1，2，...，n的整数序列
(N[mm==1])#找出mm中数值等于1的位置
```

```
#  [1]  16  19  21  22  24  33  45  58  69  87  88  89  92  95  97 128 135 14
# [20] 166 188 215 218 237 243 246 247 255 271 280 314 319 343 347 353 372 37
# [39] 389 391 399 417 423 431 455 459 476 494 500 503
```

```
which(mm==1)#也可以用这个命令
```

```
#  [1]  16  19  21  22  24  33  45  58  69  87  88  89  92  95  97 128 135 14
# [20] 166 188 215 218 237 243 246 247 255 271 280 314 319 343 347 353 372 37
# [39] 389 391 399 417 423 431 455 459 476 494 500 503
```

```
which(mm==2)
```

```
#  [1]   8   9  27  32  63  66  82  94 100 103 111 121 126 140 145 146 159 16
# [20] 179 208 222 235 238 245 248 256 262 272 293 329 330 338 340 356 361 38
# [39] 390 392 413 416 422 426 428 436 456 464 473 496 497
```

```
K=list();#K为list型变量，其中可以包含其他任何类型的变量作为元素
for(i in 1:Z)K[[i]]=N[mm==i]
```

- 10折交叉验证

```
K[[2]]#第2份的样本
```

```
#  [1]   8   9  27  32  63  66  82  94 100 103 111 121 126 140 145 146 159 16
# [20] 179 208 222 235 238 245 248 256 262 272 293 329 330 338 340 356 361 38
# [39] 390 392 413 416 422 426 428 436 456 464 473 496 497
```

```
K[[6]]#第6份的样本
```

```
#  [1]   2  11  15  25  34  38  47  56  64  67  76  79  99 112 116 132 138 15
# [20] 178 181 186 189 191 203 207 221 227 249 263 275 285 286 288 308 312 32
# [39] 352 376 380 388 412 425 445 447 486 491 495 502
```

- 10折交叉验证

```
K#显示所有份的样本
```

```
# [[1]]
#  [1]  16  19  21  22  24  33  45  58  69  87  88  89  92  95  97 128 135 14
# [20] 166 188 215 218 237 243 246 247 255 271 280 314 319 343 347 353 372 37
# [39] 389 391 399 417 423 431 455 459 476 494 500 503
#
# [[2]]
#  [1]   8   9  27  32  63  66  82  94 100 103 111 121 126 140 145 146 159 16
# [20] 179 208 222 235 238 245 248 256 262 272 293 329 330 338 340 356 361 38
# [39] 390 392 413 416 422 426 428 436 456 464 473 496 497
#
# [[3]]
#  [1]  12  17  30  36  49  59  78  80  84  90 104 106 113 161 167 172 197 20
# [20] 212 213 217 229 234 236 251 253 258 274 301 323 335 337 350 367 374 37
# [39] 409 421 430 432 439 442 463 465 466 471 488 489 506
#
# [[4]]
#  [1]   3   5  26  35  48  81  85  98 102 105 107 115 130 137 152 153 162 17
# [20] 182 195 216 223 241 250 254 264 267 291 297 302 307 309 313 318 322 32
# [39] 358 362 366 371 404 408 449 461 469 478 484 485 487
#
# [[5]]
#  [1]  10  14  28  31  51  65  68  75 108 119 125 131 136 141 156 160 176 18
# [20] 200 209 226 231 239 252 257 284 303 305 310 315 316 326 327 334 341 34
# [39] 357 363 364 386 407 429 440 441 446 453 483 493
#
# [[6]]
#  [1]   2  11  15  25  34  38  47  56  64  67  76  79  99 112 116 132 138 15
# [20] 178 181 186 189 191 203 207 221 227 249 263 275 285 286 288 308 312 32
# [39] 352 376 380 388 412 425 445 447 486 491 495 502
```

- 10折交叉验证

```
CV=function(data=w,Z=10,seed=1010){#总结为以下代码，并将过程定义为函数MV
  n=nrow(data);N=1:n;Z=10
  set.seed(seed)
  mm=sample(rep(1:Z,ceiling(n/Z))[N])
  K=list();for(i in 1:Z)K[[i]]=N[mm==i]
  return(K)
}
```

- 10折交叉验证

```
Z=10;#10折
mm=CV(w,seed=1010)#将样本随机分成10份，其中seed为随机数的种子
y_test=NULL#令y_test为空变量#也可以用y_test<-NULL
y_test->y_pred#将y_test赋值给变量y_pred
# 以上两行可以合并为下面的语句
y_test=NULL->y_pred#令y_test为空变量，并将其赋值给变量y_pred
for(i in 1:Z){
  #以去除第i份样本的所有其他样本为训练数据，拟合最小二乘回归
  at=lm(cmedv~.,w[-mm[[i]],])
  #以第i份样本为测试数据，并将其观测值合并到y_test中
  y_test=c(y_test,w$cmedv[mm[[i]]])
  #用拟合的模型预测第i份样本，得到预测值，并将预测值合并到y_pred中
  y_pred=c(y_pred,predict(at,w[mm[[i]],]))
}#每次循环，y_test和y_pred逐渐增多
```

- 10折交叉验证的结果

```
(sse=sum((y_test-y_pred)^2))#计算MSE_cv
```

```
# [1] 11820.88
```

```
(sst=sum((y_test-mean(y_test))^2))#计算SST_cv
```

```
# [1] 42577.74
```

```
(nmse=sse/sst);#计算标准化的均方误差
```

```
# [1] 0.2776305
```

```
(r2=1-nmse);#计算R^2_cv
```

```
# [1] 0.7223695
```

```
print(c(nmse,r2))
```

```
# [1] 0.2776305 0.7223695
```

蟹蟹

本幻灯片由 R 包 **xaringan** 生成；

查克拉来自于 remark.js、**knitr**、以及 R Markdown。