Statistical Learning

Prova d'esame

14 Settembre 2022

Tempo a disposizione: 150 minuti

Problema 1

Si consideri il modello $y = X\beta + \epsilon$, dove

$$y = \begin{bmatrix} -1.3 \\ 0.2 \\ -0.5 \\ -0.6 \end{bmatrix}, \quad X = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ -1 & -1 \\ -1 & 1 \end{bmatrix}$$

β è un vettore di p parametri incogniti e $\epsilon \sim N(0, \sigma^2 I_p)$.

```
rm(list=ls())
y = c(-1.3,.2,-.5,-.6)
X = matrix(c(.5,-.5,.5,.5,-.5,-.5,-.5,.5),
byrow=TRUE, ncol=2)
p = ncol(X)
```

a. Calcolare la stima OLS $\hat{\beta}$. Riportare il valore del primo elemento di $\hat{\beta}$.

```
beta_OLS = t(X) %*% y
round(beta_OLS[1],3)
```

```
## [1] 0
```

b. Si consideri lo stimatore *ridge regression*

$$\hat{\beta}_\lambda = \min_\beta \|y - X\beta\|^2 + \lambda\|\beta\|_2^2$$

Riportare il valore di λ che corrisponde la stima $\hat{\beta}_\lambda = 0.5\hat{\beta}$, ovvero il valore che rende la stima *rigde* pari alla stima OLS dimezzata.

```
lambda = 1
lambda
```

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

c. Si consideri lo stimatore *lasso*

$$\tilde{\beta}_\lambda = \min_\beta \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1$$

Calcolare la stima lasso per il valore di λ determinato al punto precedente. Riportare il valore del primo elemento di $\tilde{\beta}_\lambda$.

```
beta_tilde = (beta_OLS+sign(lambda-beta_OLS)*lambda)*(abs(beta_OLS) > lambda)*1
beta_tilde[1]
```

```
## [1] 0
```

Problema 2

Si consideri il dataset `mcycle`, presente nella libreria `MASS`, dove `accel` è la variabile risposta e `times` il predittore.

Costruire una base *B-splines* `B` di grado 2 con 50 intervalli equidistanti (il *range* da dividere è da `min(times)` a `max(times)`). Si consideri la regressione *P-splines* che utilizza la base `B`, un ordine delle differenze pari a 3 e un valore di λ pari a 10.

a. Calcolare il *mean squared error* $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, dove $\hat{y}_i$ è la stima per y_i ottenuta con la regressione *P-splines*.

b. Calcolare l'errore di convalida incrociata *Leave-One-Out*, ovvero $\text{LOO} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2$, dove $\hat{y}_i^{(-i)}$ è la stima per y_i ottenuta con la regressione *P-splines* rimuovendo l'i-sima osservazione.

```r
rm(list=ls())

lambda = 10
bdeg = 2
pord = 3

tpower <- function(x, t, deg){
  (x - t) ^ deg * (x > t)
}

bbase <- function(x, xl, xr, ndx, deg){
  dx <- (xr - xl) / ndx
  knots <- seq(xl - deg * dx, xr + deg * dx, by = dx)
  P <- outer(x, knots, tpower, deg)
  n <- dim(P)[2]
  Delta <- diff(diag(n), diff = deg + 1) / (gamma(deg + 1) * dx ^ deg)
  B <- (-1) ^ (deg + 1) * P %*% t(Delta)
  B
}

library(MASS)
data(mcycle)
x = mcycle$times
y = mcycle$accel

xl = min(x)
xr = max(x)
ndx = 50
B <- bbase(x, xl, xr, ndx, bdeg)
n = length(y)

D <- diag(ncol(B))
for (k in 1:pord) D <- diff(D)
P <- lambda * crossprod(D)
S <- B %*% solve(crossprod(B) + P) %*% t(B)
y_hat <- S %*% y
S_ii <- diag(S)
# a.
```

```
res <- (y - y_hat)
round(mean(res^2),3)
```

```
## [1] 457.448
```

```
# b.
wres <- res / (1 - S_ii)
round(mean(wres^2),3)
```

```
## [1] 545.642
```

Problema 3

Si consegni il file .R che produce le risposte alle domande richieste. Il codice deve essere **riproducibile** e, se eseguito, deve stampare in output **solo** il risultati richiesti dalle domande a) e b).

Si consideri il dataset `Boston` presente nella libreria `MASS`, utilizzando come variabile risposta `medv` e le rimanenti variabili come predittori.

Calcolare le frequenze relative di selezione $\hat{\pi}_j$, $j = 1, \ldots, p$ dell'algoritmo *Complementary Pairs Stability Selection*, utilizzando $B = 50$ ricampionamenti.

Utilizzare come metodo per calcolare $\hat{S}_{n/2}$ la regressione *Best Subset Selection*, impostata in modo da selezionare $q = 6$ variabili, utilizzando la funzione `regsubsets`` presente nella libreria`leaps'.

a. Riportare l'insieme di predittori stabili $\hat{S}_{\text{stab}}$ utilizzando la soglia $\tau = 0.9$.
b. Per i predittori stabili ottenuti al punto precedente, calcolare il limite superiore del numero atteso di errori di I tipo

```
rm(list=ls())

library(MASS)
library(leaps)

set.seed(123)

vnames = colnames(Boston)[-14]
p = length(vnames)
n = nrow(Boston)
q = 6
B = 50
S_mat = matrix(NA,ncol=p,nrow=2*B)

for (b in 1:B){

  I = as.logical(sample(rep(0:1, each=n/2)))

  fit_I <- regsubsets(medv~., Boston, subset=I, nvmax=q)
  S_mat[(2*b-1),] = (vnames %in% names(coef(fit_I, i=q))[-1])

  fit_notI <- regsubsets(medv~., Boston, subset=!I, nvmax=q)
  S_mat[2*b,] = (vnames %in% names(coef(fit_notI, i=q))[-1])

}

pi_hat = colMeans(S_mat)
```

```r
names(pi_hat) = vnames
tau = 0.95

pi_hat
```

```
##     crim       zn    indus     chas      nox       rm      age      dis      rad      tax
##     0.05     0.12     0.01     0.43     0.93     1.00     0.01     1.00     0.03     0.01
## ptratio    black    lstat
##     1.00     0.41     1.00
```

```r
S_hat = vnames[which(pi_hat  > tau)]
# a.
S_hat
```

```
## [1] "rm"      "dis"     "ptratio" "lstat"
```

```r
bound  = (1 / (2*tau -1)) * (q^2 / p)
# b.
bound
```

```
## [1] 3.076923
```

Problema 4

a. Per quale motivo è preferibile utilizzare una *B-spline basis* rispetto alla *truncated power basis*?
b. Cosa afferma il Teorema di Theobald (1974)?
c. Siano X_1, X_2, X_3 variabili aleatorie indipendenti con $X_i \sim N(\mu_i, 1)$ per $i = 1, 2, 3$. Lo stimatore $\hat{\mu} = (3, 2, 1)$ per $\mu = (\mu_1, \mu_2, \mu_3)$ è ammissibile? Si motivi la risposta.