Statistical Learning

Prova d'esame

8 Maggio 2023

Tempo a disposizione: 180 minuti

Problema 1

Si risponda alle seguenti domande:

Sia $y = X\beta + \epsilon$, dove

$$X = \begin{bmatrix} -0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ 0.5 & 0.5 \end{bmatrix}, \quad \beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \epsilon \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix} \right)$$

Calcolare il valore di λ che minimizza $\text{MSE}(\hat{\beta}_\lambda)$ per lo stimatore *ridge regression*

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

Riportare il risultato arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
X = matrix(
  c(-0.5, -0.5,
    -0.5,  0.5,
     0.5, -0.5,
     0.5, 0.5),byrow=TRUE, ncol=2)
p = ncol(X)
beta = c(1,1)
sigma2 = 0.5
lambda = c(p*sigma2/crossprod(beta))
a = lambda
round(a,3)
```

```
## [1] 0.5
```

Sia $y = (-1.4, -0.2, 1.1, 1)^t$ una realizzazione del modello specificato al punto precedente. Calcolare la stima $\tilde{\beta}_\lambda$ dello stimatore *lasso*

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

con il valore di λ determinato al punto precedente (Attenzione: nel problema di minimo che definisce lo stimatore lasso non c'è il fattore $\frac{1}{2}$). Riportare il valore del primo elemento di $\tilde{\beta}_\lambda$, arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
soft_thresh <- function(a, b)
  {
    a[abs(a) <= b] <- 0
    a[a > 0] <- a[a > 0] - b
```

```
    a[a < 0] <- a[a < 0] + b
    a
}
Xty = crossprod(X,y)
beta_tilde = soft_thresh(a=Xty, b=lambda/2)
b = beta_tilde[1]
round(b,3)
```

```
## [1] 1.6
```

Calcolare la stima $\bar{\beta}_\lambda$ dello stimatore *elastic net*

$$\bar{\beta}_\lambda = \min_\beta \|y - X\beta\|^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1$$

con il valori di λ_1 e λ_2 entrambi pari al valore di λ utilizzato al punto precedente. Nel caso ortogonale, i.e. $X^tX = I_p$, lo stimatore *elastic net* è pari a $\bar{\beta}_\lambda = (\frac{1}{1+\lambda_2})S_{\lambda_1/2}(X^ty)$ dove $S_a(b)$ è l'operatore *soft-thresholding*.

Riportare il valore del primo elemento di $\bar{\beta}_\lambda$, arrotondando al **terzo decimale** (si ricordi l'uso della **virgola** per i decimali).

```
#c.
beta_bar = (1/(lambda+1))*soft_thresh(a=Xty, b=lambda/2)
c = beta_bar[1]
round(c,3)
```

```
## [1] 1.067
```

Problema 2

Si considerino i seguenti dati:

Player	n_i	s_i	pi_i
Baines	415	118	0.289
Barfield	476	117	0.256
Biggio	555	153	0.287
Bonds	519	156	0.297

di $p = 4$ giocatori di baseball, dove n_i e s_i indicano rispettivamente il numero di volte a battuta e il numero di battute valide, mentre π_i indica la vera media battuta (calcolata su tutta la carriera di ciascun giocatore).

Sia Z_i la variabile aleatoria Binomiale$(n_i, \pi_i)/n_i$, e si supponga che $Z_1, \ldots, Z_p$ siano indipendenti.

Si consideri valida la seguente approssimazione

$$X_i = \sqrt{n_i}\arcsin(2Z_i - 1) \approx N(\mu_i, 1)$$

dove $\mu_i = \sqrt{n_i}\arcsin(2\pi_i - 1)$.

1. Sia $\hat{\pi}^{\text{MLE}}$ la stima di massima verosimiglianza per $\pi = (\pi_1, \ldots, \pi_p)$. Riportare il valore della stima per Barfield.

```
# a.
z_i = s_i / n_i
a = round(z_i[row_i], 3 )
a
```

```
## [1] 0.246
```

2. Sia $\hat{\pi}^{\text{JS}}$ la stima secondo James-Stein per π (per $X_i \sim N(\mu_i, 1)$, $i = 1, \ldots, p$, utilizzare lo stimatore $\hat{\mu}_{JS}^{\bar{X}}$). Riportare il valore della stima per Barfield.

```
# b.
x_i = sqrt(n_i) * asin(2*z_i-1)
x_bar <- mean(x_i)
S <- sum((x_i-x_bar)^2)
mu_i_js = x_bar + (1 - ((p-3)/S)) * (x_i - x_bar)
pi_i_js = 0.5 * ( 1+sin(mu_i_js/sqrt(n_i)))
b = round(pi_i_js[row_i], 3)
b
```

```
## [1] 0.252
```

3. Sia $\hat{\pi}^{*}$ la stima secondo l'oracolo per π (per $X_i \sim N(\mu_i, 1)$, $i = 1, \ldots, p$, l'oracolo conosce il vero valore di $\|\mu\|^2$). Riportare il valore della stima per Barfield.

```
# c.
mu_i = sqrt(n_i) * asin(2*pi_i-1)
mu_i_star = (sum(mu_i^2)/(p+sum(mu_i^2))) * x_i
pi_i_star = 0.5 * ( 1+sin(mu_i_star/sqrt(n_i)))
c = round(pi_i_star[row_i], 3 )
c
```

```
## [1] 0.248
```

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 `longley` presente nella libreria `datasets`. La variabile risposta è `Employed`, i predittori sono le variabili rimanenti.

Utilizzare i dati di training $(x_1, y_1), \ldots, (x_n, y_n)$ (le prime 15 righe corrispondenti agli anni 1947-1961) e il test point x_{n+1} (la 16ma riga, corrispondente all'anno 1962) per verificare se il valore candidato $y_{n+1} = 70.551$ (si tratta proprio del valore della risposta nell'anno 1962) è incluso oppure no nell'intervallo di previsione di livello $1 - \alpha = 0.75$ costruito con l'algoritmo *Full Conformal*. Utilizzare come modello la regressione *Best Subset Selection* con criterio *BIC* implementata dalla funzione `regsubsets` presente nella libreria `leaps`.

Riportare :

a) Il valore critico R_α

b) Gli anni corrispondenti alle osservazioni tali che $R_i > R_\alpha$, $i = 1, \ldots, n + 1$.

```
rm(list=ls())
library(leaps)
library(MASS)

X = longley[-16,-7]
y = longley[-16,7]
xnew = longley[16,-7]
ynew = longley[16,7]

XX = rbind(X,xnew)
yy = c(y,ynew)
```

```r
alpha = 0.25

fit = regsubsets(x=XX, y=yy)
summary_fit = summary(fit)
best = which.min(summary_fit$bic)
names_best = names(coef(fit,id=best))[-1]
fit_best = lm(yy ~ as.matrix(XX[,names_best]))
res =  abs(residuals(fit_best))
o = order(res)
c = ceiling((1-alpha)*length(yy))
r = res[o][c]
# a.
a = r
a
```

```
##        16
## 0.2790681
```

```r
# b.
b = row.names(XX)[which(res > r)]
b
```

```
## [1] "1950" "1951" "1956" "1961"
```

Problema 4

a) La seguente figura (disponibile solo sulla piattaforma esamionline) rappresenta l'esito della procedura *knockoff*. Quanto vale la stima $\widehat{\text{FDP}}(\hat{S}_\tau)$ per $\tau = 2$? Quanto vale il vero valore di $\text{FDP}(\hat{S}_\tau)$?

b) E' possibile garantire che la *screening property* venga soddisfatta con probabilità 1?

c) Sia $\pi_j^n = \text{pr}(j \in \hat{S}_n)$ la probabilità di inclusione della j-sima variabile nell'*active set* $\hat{S}_n$ di un *lasso*, i.e. $\hat{S}_n = \{j \in \{1, \ldots, p\} : \hat{\beta}_j(\lambda) \neq 0\}$ dove $\hat{\beta}_j(\lambda)$ è lo stimatore *lasso* per β_j basato su n osservazioni e un certo valore $\lambda \geq 0$. Proporre uno stimatore non distorto per π_j^n.