Statistical Learning

Prova d'esame

11 Aprile 2022

Tempo a disposizione: 150 minuti

Problema 1

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 $\hat{\pi}^{\mathrm{MLE}}$ 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 $\hat{\pi}^{\mathrm{JS}}$ la stima secondo James-Stein per π. 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] 0.248
```

3. Sia $\hat{\pi}^*$ la stima secondo l'oracolo per π (si supponga che l'oracolo conosca 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] 0.247
```

Problema 2

Si risponda alle seguenti domande.

1. Si consideri il modello lineare semplice con intercetta e una covariata x_i. I dati sono

$$\{(y_i, x_i)_{i=1}^4\} = \{(1.4, 0), (1.4, -2), (0.8, 0), (0.4, 2)\}$$

Riportare il valore di λ che corrisponde alla stima ridge $\hat{\beta}_\lambda = (1, -1/24)^t$ senza penalizzare l'intercetta.

```
# a.
x = c(0,-2,0,2)
y = c(1.4,1.4,0.8,0.4)
beta = -1/m
xtx = crossprod(x)
xty = crossprod(x,y)
l = c(xty/beta - xtx)
a = round(l,2)
a
```

```
## [1] 40
```

2. Calcolare la stima ridge $\hat{\beta}(\lambda) = (\hat{\beta}_1(\lambda), \hat{\beta}_2(\lambda))^t$ con $\lambda = 0$ per i seguenti dati con $n = 1$ e $p = 2$:

$$X = [0.3, 0.7], y = [0.2], X^tX = \begin{bmatrix} 0.09 & -0.21 \\ -0.21 & 0.49 \end{bmatrix}, X^ty = \begin{bmatrix} 0.06 \\ -0.14 \end{bmatrix}$$

Riportare il valore $\hat{\beta}_2(\lambda)$.

```
# b.
x = matrix(c(0.3,-0.7),ncol=2)
y = 0.2
svd_x = svd(x)
v = svd_x$v
d = svd_x$d
u = svd_x$u
beta_hat = v %*% ( crossprod(u,y) / d )
b = round(beta_hat[j],2)
b
```

```
## [1] -0.24
```

3. Sia

$$X = \begin{bmatrix} -1 & 2 \\ 0 & 1 \\ 2 & -1 \\ 1 & 0 \end{bmatrix}$$

Calcolare $\mathbb{V}\mathrm{ar}(\hat{\beta}_2(\lambda))$ con λ pari al valore trovato al primo punto dell'esercizio, ipotizzando $\sigma^2 = 40$.

```r
# c.
X = matrix(c(-1,2,0,1,2,-1,1,0), byrow=T, ncol=2)
svd_x = svd(X)
v = svd_x$v
d = svd_x$d
u = svd_x$u
var_beta_hat = diag(sigma2 * v %*% diag(d^2/(d^2 + 1)^2) %*% t(v))
c = round(var_beta_hat[j],2)
c
```

```
## [1] 0.1
```

Problema 3

Si consegni il file .R che produce le risposte alle domande richieste. Si risponda inoltre alle domande aperte direttamente in tale file, avendo cura di commentare con un cancelletto (#) quanto scritto.

1. Scrivere il codice R della funzione `my_stability` che calcola le frequenze relative di selezione $\hat{\pi}_j$, $j = 1, \ldots, p$ dell'algoritmo *Complementary Pairs Stability Selection*. Utilizzare come metodo per calcolare $\hat{S}_{n/2}$ la regressione *forward*, impostata in modo da selezionare q variabili, utilizzando la funzione `step` presente nella libreria `stats`.

```r
# Compute complementary pairs stability selection.
# Args:
  # X: A numeric data matrix.
  # y: Response vector.
  # B: Number of resamples.
  # q: Number of variables selected by forward selection.
# Returns:
# Selection probabilities vector of length ncol(X).
my_stability = function(X, y, B, q){
...
}
```

2. Applicare la funzione al dataset `Boston` presente nella libreria `MASS`, utilizzando come variabile risposta `medv` e le rimanenti variabili come predittori, specificando $B = 50$ e $q = 6$.

3. Calcolare l'insieme di predittori stabili $\hat{S}_{\text{stab}}$ utilizzando la soglia $\tau = 0.9$. Calcolare il limite superiore del numero atteso di errori di I tipo e commentare il risultato.

```r
rm(list=ls())
set.seed(123)
# Compute complementary pairs stability selection.
##
# Args:
  # X: A numeric data matrix.
  # y: Response vector.
  # B: Number of resamples.
  # q: Number of variables selected by forward selection.
# Returns:
# Selection probabilities vector of length ncol(X).
my_stability = function(X, y, B, q){
  p = ncol(X)
  n = nrow(X)
  if (is.null(names(X))) colnames(X) <- paste0("x",1:p)
  yX = data.frame(X, y=y)
```

```r
  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_null <- lm(y ~ 1, data = yX, subset=I)
  fit_full <- lm(y ~., data = yX, subset=I)
  fit = step(fit_null,
        scope=list(upper=fit_full),
        direction = "forward", k=0, trace = 0,
        steps = q)
  S_mat[(2*b-1),] = (colnames(X) %in% names(coef(fit))[-1])

  fit_null <- lm(y ~ 1, data = yX, subset=!I)
  fit_full <- lm(y ~., data = yX, subset=!I)
  fit = step(fit_null,
             scope=list(upper=fit_full),
             direction = "forward", k=0, trace = 0,
             steps = q)
  S_mat[2*b,] = (colnames(X) %in% names(coef(fit))[-1])

  }

  pi_hat = colMeans(S_mat)
  names(pi_hat) = colnames(X)
  return(pi_hat)
}

library(MASS)
y = Boston$medv
X = Boston[,-14]
p = ncol(X)
q = 6
pi_hat = my_stability(X, y, B=50, q=q)
pi_hat
```

```
##    crim      zn   indus    chas     nox      rm     age     dis     rad     tax
##    0.09    0.11    0.01    0.42    0.76    1.00    0.01    0.97    0.03    0.07
## ptratio   black   lstat
##    1.00    0.53    1.00
```

```r
tau = 0.9
S_hat = which(pi_hat  > tau)
S_hat
```

```
##      rm     dis ptratio   lstat
##       6       8      11      13
```

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

```
## [1] 3.461538
```

Problema 4

Si risponda alle seguenti domande :

1. Per quale motivo è preferibile utilizzare una *B-spline basis* rispetto alla *truncated power basis*?
2. Si consideri l'algoritmo *single split* per la selezione delle variabili. Si spieghi perché il Teorema che garantisce il controllo del FWER richiede che la procedura di selezione delle variabili soddisfi la *screening property*.
3. Si supponga di avere a disposizione un training set con n osservazioni e p predittori, con $n < p$, e si vuole costruire un intervallo di previsione per una nuova osservazione garantendo una probabilità di copertura marginale del 90%. Si consideri la seguente procedura a due stadi: a) Sulla base dei dati di training, selezionare $q < n$ predittori con un algoritmo di selezione delle variabili; b) Sul dataset di training ridotto contenente i q predittori selezionati al passo a), utilizzare l'algoritmo *split conformal prediction*. Questa procedura garantisce probabilità di copertura marginale del 90%? Motivare la risposta.