ASML: test intermedio 24/11/2023

Leonardo Egidi, Nicola Torelli

1. Quale delle seguenti definizioni descrive meglio il modo in cui viene utilizzata la regressione LASSO per la selezione delle covariate?

   (a) La regressione LASSO si adatta in modo iterativo a più modelli di regressione con più o meno covariate incluse e seleziona l'insieme finale di covariate in base al modello di regressione con le prestazioni migliori.
   (b) La regressione LASSO crea un elenco classificato di covariate che possono essere utilizzate per selezionare le prime n covariate da utilizzare nella creazione di modelli.
   (c) La regressione LASSO è identica alla regressione dei minimi quadrati (o regressione MLE), tranne per il fatto che, dopo aver adattato i coefficienti, passa attraverso un secondo passaggio per rimuovere eventuali caratteristiche con coefficienti troppo grandi. La grandezza dei coefficienti nel caso del lasso è determinata dal valore assoluto del coefficiente.
   (d) La regressione LASSO è una forma di regressione in cui la funzione obiettivo viene modificata per includere l'aggiunta di un termine proporzionale alla somma dei valori assoluti dei coefficienti. Solo le covariate sufficientemente predittive da giustificare la dimensione dei loro coefficienti saranno incluse nel modello finale.

2. Considerare il seguente codice R qui sotto (potete copiarlo e incollarlo nei vostri pc): quale delle seguenti affermazioni è sicuramente falsa?

   (a) Non si possono calcolare correttamente le stime del modello lineare.
   (b) Le stime del LASSO non sono calcolabili in forma chiusa, quindi non sono presenti nel codice.
   (c) Il codice serve a calcolare le stime OLS e ridge.
   (d) L'ultima riga serve a calcolare le stime LASSO con $\lambda = 2.5$.

```r
# dati simulati
p <- 5
n <- 100
X <- matrix(NA, n, p)
X[,1] <- rep(1, n)
X[,2] <- runif(n, -5,5)
X[,3] <- rnorm(n, 0,1)
X[,4] <- rgamma(n, 2,2)
X[,5] <- X[,3]-3*X[,2]
y <- rnorm(n, 1.3*X[,1]+ 0.2 * X[,2] + 2.5*X[,3] -1.5*X[,4] -2*X[,5] + rnorm(n, 0, 0.8), 1)

solve(t(X)%*%X)%*%t(X)%*%y
solve(t(X)%*%X + 2.5*diag(p) )%*%t(X)%*%y
```

3. Considerare la seguente matrice di predittori, con i-esima riga:

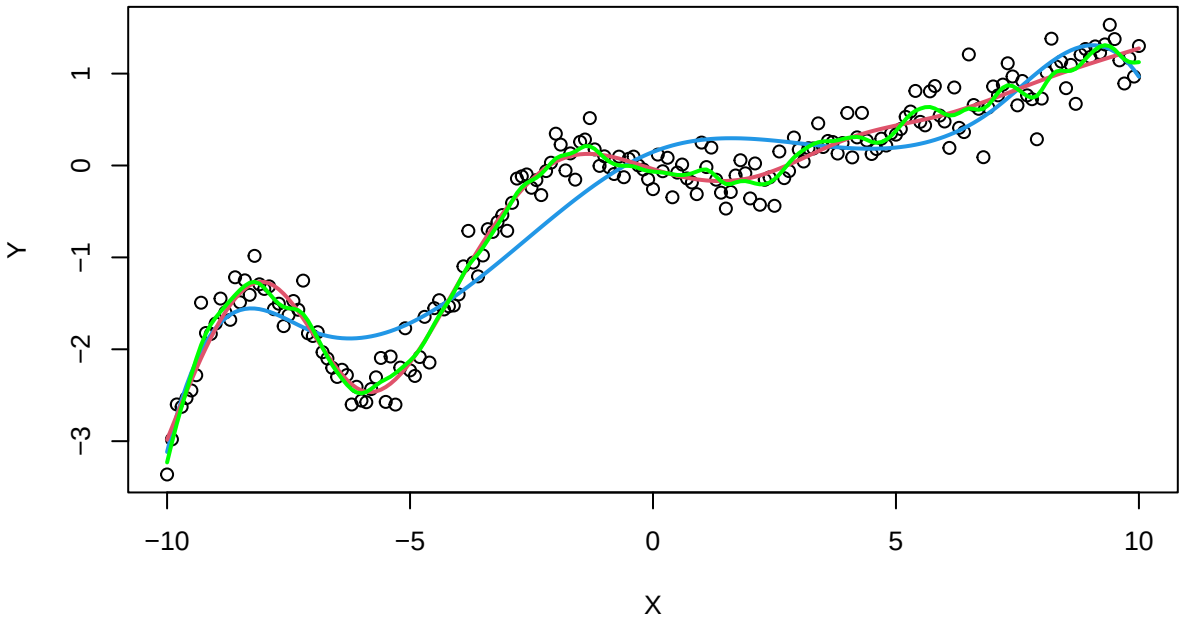
$$\begin{pmatrix} 1 & x_i & x_i^2 & (x_i-\xi_1)_+^2 & (x_i-\xi_2)_+^2 & \dots & (x_i-\xi_K)_+^2 \end{pmatrix},$$

   e supponiamo di voler stimare il seguente modello $y_i = \sum_j \beta_j b_j(x_i) + \epsilon_i$, dove $\xi_1, \xi_2, \dots, \xi_K$ sono i cosiddetti nodi. Si tratta di:

(a) Una spline cubica con $K + 2$ gdl.
(b) Una spline quadratica con $K + 3$ gdl.
(c) Una spline quadratica con $K + 4$ gdl.
(d) Una spline naturale perché c'è un vincolo sul coefficiente cubico.

4. Quale di queste affermazioni sul seguente grafico è sicuramente vera:

(a) La curva blu non può essere una spline cubica perché si distanzia troppo dai dati.
(b) La curva verde ha meno gradi di libertà delle altre.
(c) La curva rossa non è una spline cubica naturale.
(d) La curva verde è quella con maggior varianza.



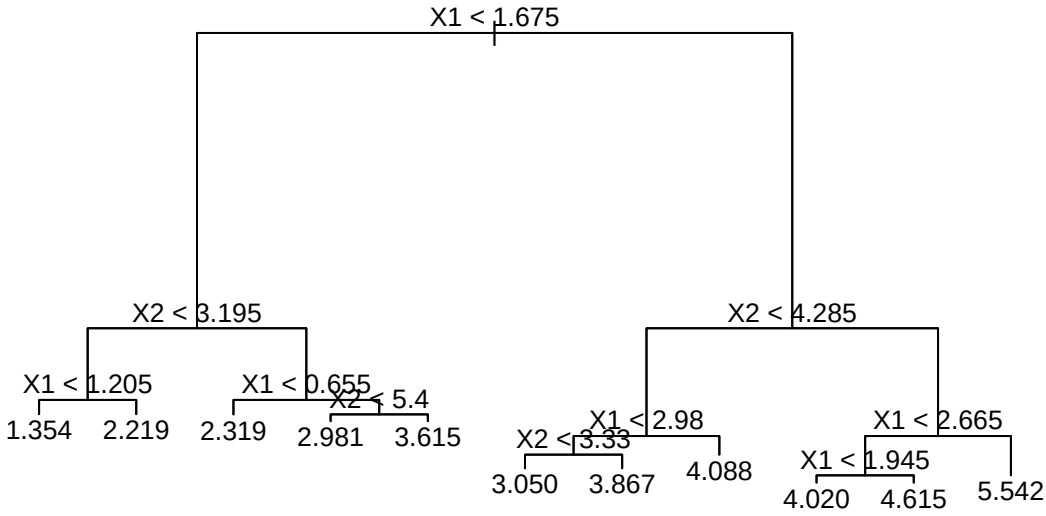
5. Quale delle seguenti affermazioni riguardanti l'algoritmo dell'albero di regressione è vera?

(a) L'algoritmo dell'albero di regressione non produce una varianza elevata poiché le suddivisioni nei dati vengono determinate in sequenza.
(b) L'algoritmo dell'albero di regressione suddivide ad ogni passaggio i dati in sottoinsiemi al fine di massimizzare la riduzione della RSS risultante da ciascuna suddivisione. Pertanto, l'algoritmo dell'albero di regressione produce un modello che fornirà previsioni ottimali e minimizzerà la perdita attesa.
(c) L'algoritmo dell'albero di regressione può dar luogo a modelli ad alta distorsione perché non è vincolato da ipotesi parametriche sui dati sottostanti.
(d) L'algoritmo dell'albero di regressione si adatta in modo più preciso se il set di addestramento è piccolo.

6. Analizzare il fit di quest'albero per la variabile risposta Y. Cosa si può concludere? (Una sola risposta è sicuramente giusta)

(a) Quando $X_2 < 3.33$, allora la previsione corretta è sempre 3.05.

(b) Il valore previsto 1.354 derivante dallo split $X_1 < 1.205$ è un nodo interno dell'albero.
(c) Il predittore X_3 non è determinante per prevedere Y.
(d) Probabilmente in un modello lineare X_2 e X_1 sono collineari.



7. Cosa ottengo con il seguente codice R? (Potete copiarlo e incollarlo nei vostri PC)

(a) La distribuzione della media nel campione bootstrap.
(b) Un intervallo di confidenza per la media con metodo bootstrap.
(c) Un intervallo di confidenza per la standard deviation con metodo bootstrap non-parametrico.
(d) Un intervallo di confidenza per la standard deviation con metodo bootstrap parametrico.

```r
y <- rnorm(50, 0,4) # genero dati casuali
n <- length(y)
set.seed(1989)
B <- 10^4
b.sample <- matrix(NA, nrow = B, ncol = n)
for(i in 1:B) b.sample[i,] <- sample(y, n, replace = TRUE)
stat <- apply(b.sample, 1, sd )
quantile(stat, c(0.025, 0.975))
```

8. Quale delle seguenti frasi descrive la *cross-validation*?

(a) Dividere i dati in modo casuale in due gruppi e utilizzare un gruppo per addestrare l'algoritmo e il resto per la sua valutazione.
(b) Dividere i dati in modo casuale in K gruppi e utilizzare solo i $K - 1$ gruppi per addestrare l'algoritmo e i restanti per la convalida.

(c) Dividere i dati in modo casuale in K gruppi e utilizzare i $K-1$ gruppi per addestrare l'algoritmo e i restanti per la valutazione. Ripetere questo passaggio K volte e ogni volta tralasciare uno dei gruppi. Combinare la valutazione ottenuta nelle K ripetizioni.
(d) Gli stessi dati utilizzati per addestrare l'algoritmo vengono utilizzati per la sua valutazione.

9. Quale delle seguenti affermazioni è sicuramente falsa?

(a) Un *random forest* è un esempio di *bagging* che utilizza alberi di regressione come modelli di base.
(b) Il bagging è un algoritmo flessibile in grado di abbassare la varianza aumentando il numero di iterazioni.
(c) La riduzione del numero di iterazioni potrebbe controllare l'adattamento eccessivo, o *overfitting*, negli algoritmi di *boosting*.
(d) Il modo in cui il *bagging* valuta i predittori ad ogni split produce singoli alberi di regressione piuttosto correlati.

10. Per controllare il *trade-off* bias-varianza in un modello predittivo occorre:

(a) aumentare la complessità del modello.
(b) mantenere il modello il più semplice possibile.
(c) utilizzare tecniche di regolarizzazione (come LASSO nei modelli di regressione).
(d) provare a raccogliere più dati.