Linear Model Selection and Regularization - 24/10

Linear Model Selection and Regularization

Il modello lineare standard

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

è comunemente usato per descrivere la relazione tra la risposta Y e l'insieme dei predittori $X_1, \ldots, X_p$.

Questo modello è tipicamente fittato usando least squares (come del resto abbiamo fatto noi), che potrebbe però non funzionare bene con un p grande o in presenza di multicollinearity.

Discutiamo alcune procedure di fitting alternative che possono migliorare le performance del modello lineare.

Ci sono due ragioni per le quali si potrebbe preferire di **non** usare solo la stima OLS (ordinary least squares):

- accuratezza della previsione;
- interpretabilità del modello.

Prediction accuracy

A seconda della differenza di dimensioni tra il numero di osservazioni n e il numero di predittori p cambia l'accuratezza del modello.

- $n >> p$, se il numero di osservazioni è significativamente maggiore del numero di variabili allora le stime least squares tendono ad avere bassa varianza, il che migliora le performance sulle test-observations.
- $n \cong p$, in questo caso il fit a minimi quadrati può avere alta varianza e potrebbe risultare in overfitting e stime di bassa qualità sulle osservazioni di test.
- $n < p$, in questo caso non è più vero che c'è un'unica stima dei coefficienti least squares e quindi la varianza è infinita e il metodo non può essere utilizzato.

Model interpretability

Quando abbiamo un grande numero di predittori X nel modello ce ne saranno molti che hanno effetto piccolo, se non nullo, su Y.

Lasciare queste variabili nel modello rende più difficile vedere le vere relazioni tra i predittori e la variabile dipendente (vedere the big picture, la visione di insieme) ed è difficile apprezzare l'effetto delle "variabili rilevanti" che descrivono Y.

Il modello sarebbe più facile da interpretare rimuovendo le variabili non importanti, cioè settando i loro coefficienti a 0.

Metodi da accostare a least squares

Ci sono molte alternative, classiche e moderne, al solo fit least squares. Noi approfondiamo 3 classi di metodi.

- **Subset Selection**. Questo approccio consiste nell'identificare un sottoinsieme dei predittori p che noi crediamo essere relazionati alla risposta. A questo punto effettuiamo il fit del modello usando least squares sul set ridotto di variabili.

- **Shrinkage**. Questo approccio consiste nel fare il fit del modello usando tutti i p preditorri, però i coefficienti stimati sono "ridotti" (shrunken) verso 0 rispetto alle stime least squares.
  Questa riduzione (anche nota come
  regolarizzazione) ha l'effetto di ridurre la varianza. Sulla base del tipo di shrinkage effettuato alcuni coefficienti potrebbero essere stimati pari esattamente a 0, di conseguenza i metodi di shrinkage possono anche effettuare variable selection.
- **Dimension Reduction**. Questo approccio consiste nel proiettare i predittori p in un sottospazio M-dimensionale dove M<p. Questo è ottenuto calcolando M diverse combinazioni lineari, o proiezioni, delle variabili. Poi queste M proiezioni sono usate come predittori per fare il fit di un modello di regressione lineare tramite least squares.

Subset Selection

Consideriamo ora dei metodi per selezionare sottoinsiemi di predittori.

Best Subset Selection

In questo approccio eseguiamo una regressione lineare per ogni possibile combinazione degli X predittori.

Al fine di scegliere il modello "migliore", un approccio semplice è selezionare il sottoinsieme con l'RSS minimo o, equivalentemente, con il più grande R^2.

Sfortunatamente si può dimostrare che il modello che include tutte le variabili avrà sempre il più grande valore R^2 e il più piccolo RSS, visto che le quantità sono related al training error.

Introdurremo più avanti C_p, AIC, BIC e adjusted R^2.

Algoritmo Best subset selection

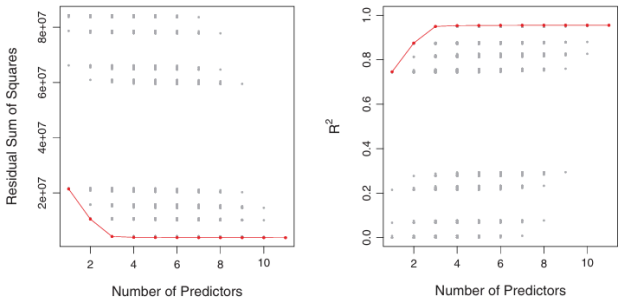
1. Sia M_0 tale che denota il null model, che contiene 0 predittori. Questo modello semplicemente predice la media campionaria per ogni osservazione.
2. Per $k = 1, 2, \dots p$:
   a. si esegue il fit di tutti i $\binom{p}{k}$ modelli che contengono esattamente k predittori.
   b. si sceglie il migliore tra i $\binom{p}{k}$ modelli e lo chiamiamo M_k. In questo contesto best è definito come avente il minor RSS, o equivalentemente il più grande R^2 (a ogni passo di questo ciclo tutti i modelli analizzati hanno lo stesso numero k di predittori quindi valutare l'RSS o l'R^2 funziona).
3. Si selezione il singolo miglior modello tra $M_0, \dots, M_P$ usando cross-validated prediction error, C_p (AIC), BIC, o adjusted R^2.

GENERALIZZAZIONE DI BEST SUBSET SELECTION

La stessa idea di best subset selection si applica ad altri tipi di modelli non basati sulla regressione least squares, come la logistic regression.

La deviance, uguale a -2log(maximazed likelihood), gioca il ruolo di RSS per una classe di modelli più ampia: più piccola è la deviance e migliore è il fit.

Best subset selection: Credit data example



Per ogni possibile modello contenente un sottoinsieme dei 10 predittori nel data set Credit sono mostrati l'RSS e l'R^2.

La frontiera rossa traccia il miglior modello per un dato numero di predittori secondo l'RSS e l'R^2.

Nonostante il data set contenga solo 10 predittori l'asse delle x va da 1 a 11, visto che una delle variabili è categorical e richiede 3 valori, portando alla creazione di 2 dummy variables.

Altre misure di comparazione

Abbiamo visto che RSS e R^2 **non** sono adatti per la selezione del modello migliore in una collection di modelli con diversi numeri di predittori.

Al fine di selezionare il miglior modello rispetto al test error dobbiamo stimare il test error stesso.

Ci sono due approcci comuni:

- Misure di comparazione alternative a RSS e $\mathbb{R}^2$, possiamo indirettamente stimare il test error facendo correzioni al training error per tenere in conto il bias dovuto all'overfitting. Ad esempio C_P, AIC, BIC e adjusted R^2. Questi metodi aggiungono una penalità all'RSS per il numero di variabili (cioè la complessità) del modello;
- Validation e Cross-Validation, possiamo direttamente stimare il test error usando un approccio validation-set o un approccio cross-validation.

Misure di comparazione alternative a RSS e $\mathbb{R}^2$

Mallow's C_p (unbiased estimator of MSE)

Per un modello fitted via least squares contenente d predittori, la stima C_p del test MSE è

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

dove $\hat{\sigma}^2$ è una stima di $\sigma^2 = Var(\epsilon)$.

La statistica C_p aggiunge una penalità di $2d\hat{\sigma}^2$ al training RSS al fine di correggere (adjust) per il fatto che il training error tende a sottostimare il test error.

> 💡 Possiamo dimostrare che se $\hat{\sigma}^2$ è lo stimatore unbiased di σ^2 allora C_p è lo stimatore unbiased del test MSE.

Per i dati di Credit: C_p seleziona 6 variabili: income, limit, rating, cards, age e student.

Akaike Information Criterion (AIC)

Il criterio AIC è definito per una ampia classe di modelli il cui fit è fatto per mezzo di **maximum likelihood**:

$$AIC = -2logL + 2d$$

dove L è il valore massimizzato della likelihood.

Nel caso del modello lineare con errori Gaussiani maximum likelihood e least squares sono la stessa cosa.

In questo caso AIC è dato da

$$\text{AIC} \propto \text{RSS} + 2d\hat{\sigma}^2$$

dove, per semplicità, si è omessa una costante additiva.

> 💡 Per i modelli least squares, C_p e AIC sono equivalenti.

Quindi per credit data: anche AIC seleziona le 6 variabili income, limit, rating, cards, age e student.

> 💡 Un modello con un AIC più piccolo è migliore, NON più piccolo in valore assoluto. Quindi valori più negativi sono preferibili (-180 è meglio di -160).
>
> https://www.reddit.com/r/AskStatistics/comments/5ydt2c/if_my_aic_and_bic_are_negative_does_that_mean/

Bayesian information Criterion (BIC)

BIC è derivato da un punto di vista Bayesiano ma alla fine è simile a C_p (ed AIC).

Per il modello a minimi quadrati con d predittori, il BIC è, a scanso di costanti irrilevanti, dato da

$$\text{BIC} = -2\log L + \log(n)\, d \implies \text{BIC} \propto \text{RSS} + \log(n) d\hat{\sigma}^2$$

Visto che log(n)>2 per ogni n>7, la BIC statistic generalmente piazza una penalità più alta su modelli con molte variabili, e quindi risulta nella selezione di modelli più piccoli di C_p.

Per credit data: BIC sceglie un modello che contiene solo i 4 predittori income, limit, cards e student.

Adjusted R^2

Ricordiamo il solito R^2 definito come

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}, \text{ where } \text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Per un modello least squares con d variabili, la adjusted R^2 statistic è calcolata come

$$\text{adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

💡 Un valore alto della adjusted R^2 indica un modello con un piccolo test error.

Per credit data adjusted R^2 seleziona un modello contenente le 7 variabili income, limit, rating, cards, age, student e gender.

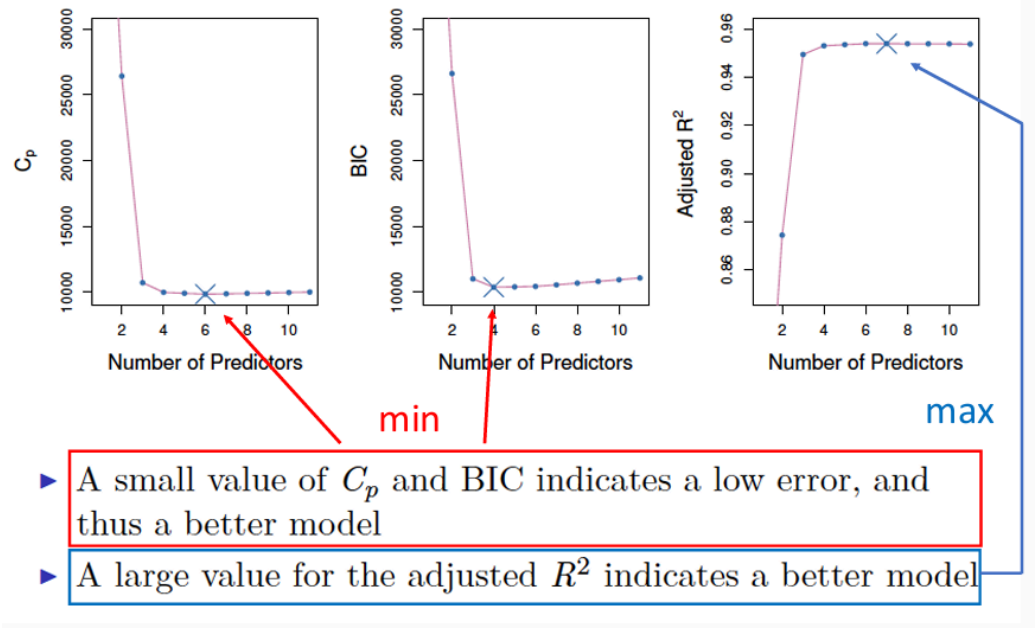
Differentemente da C_p, AIC e BIC, per i quali un piccolo valore indica un modello con un basso test error, un grande valore di adjusted R^2 indica un modello con un piccolo test error.

Massimizzare l'adjusted R^2 è equivalente a minimizzare $\frac{RSS}{n-d-1}$.

Mentre RSS decresce sempre all'aumentare del numero di variabili nel modello, $\frac{RSS}{n-d-1}$ potrebbe crescere o decrescere a causa della presenza di d al denominatore.

Differentemente da R^2 statistic, l'adjusted R^2 statistic "paga un prezzo" per l'inclusione di variabili non necessarie nel modello.

C_p, BIC, adjusted R^2 per Credit Data



C_p, AIC, BIC, adjusted R^2

C_p, AIC e BIC hanno tutti rigorose giustificazioni teoriche.

Adjusted R^2 è abbastanza intuitivo ma non motivato in teoria della statistica tanto quanto C_p, AIC e BIC.

Tutte queste misure sono semplici da calcolare e possono fornire stime del test error ma nessuna di loro è perfetta.

💡 Le formule per AIC, BIC e C_p sono presentate nel caso di un model fit lineare usando least squares; tuttavia queste quantità possono essere anche definite per tipi più generali di modelli.

Validation e Cross-Validation

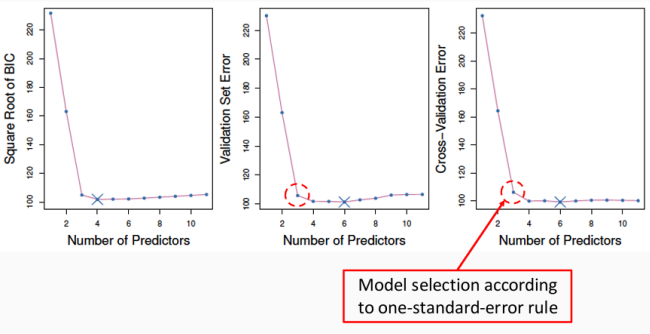
Ognuna delle procedure restituisce una sequenza di modelli M_k indicizzata per model size $k = 0, 1, 2, \ldots$. Il nostro lavoro è selezionare $\hat{k}$. Una volta selezionato restituiamo il modello $M_{\hat{k}}$.

Calcoliamo il validation set error o il cross-validation error per ogni modello M_k tenuto in considerazione e poi selezioniamo la k per la quale il test error stimato risultante è il più piccolo.

💡 Questa procedura ha un vantaggio relativamente a AIC, BIC, C_p e adjusted R^2 nel fatto che fornisce una stima diretta del test error e non richiede una stima della varianza di errore σ^2.

Può anche essere usato in un'ampia gamma di compiti di model selection, anche in casi dove è difficile capire quanti gradi di libertà ha il modello (ad esempio il numero di predittori del modello) o stimare la error variance σ^2.

Esempio credit card



Gli errori di validation sono stati calcolati selezionando casualmente 3/4 delle osservazioni come training set e lasciando il resto come validation set.

Gli errori di cross-validation sono stati calcolati usando k=10 folds.

In questo caso sia il metodo validation set che il metodo cross-validation hanno portato a modelli a 6 variabili.

Comunque tutti e tre gli approcci suggeriscono che i modelli a 4, 5 e 6 variabili sono più o meno equivalenti in termini dei loro test errors.

One-standard-error rule

In questo contesto possiamo selezione un modello usando la regola one-standard-error rule:

- per prima cosa calcoliamo lo standard error dei test MSE stimati per ogni model size;
- poi selezioniamo il modello più piccolo per il quale la stima del test error è entro 1 standard error dal punto più basso della curva.

> 💡 La logica è che se un insieme di modelli sembra essere più o meno buono allo stesso modo allora è sensato scegliere il modello più semplice, cioè quello con meno predittori.

In questo caso, applicando la regola one-standard-error all'approccio validation set o a quello cross-validation porta alla selezione del modello a tre variabili.

Stepwise selection

Best subset selection soffre di limitazioni computazionali, spesso non è possibile esaminare tutti i possibili modelli, visto che ce ne sono 2^p; ad esempio quando p = 40 ci sono più di un miliardo di modelli. Un algoritmo efficiente chiamato procedura leaps and bounds lo rende fattibile per p grande anche fino a 30 o 40.

Best subset selection potrebbe anche soffrire di problemi statistici quando p è grande: più grande è lo spazio di ricerca maggiore è la probabilità di trovare modelli che sembrano buoni sui dati di training anche se potrebbero non avere nessun potere predittivo sui futuri dati. Ne consegue che uno spazio di ricerca enorme può portare a overfitting e alta varianza delle stime dei coefficienti.

Per queste due ragioni si rende necessario un approccio automatizzato che cerca attraverso un sottoinsieme dei modelli. I metodi stepwise, che esplorano un insieme di modelli più ristretto, sono alternative attraenti a best subset selection.

Alcuni di questi concetti sono già stati affrontati in https://www.notion.so/2-Linear-Regression-B-112b965416158085a19ac287fabd1576?pvs=4#112b965416158051947cef4fc3e2f348 .

Forward stepwise selection

La Forward stepwise selection inizia con un modello senza predittori e poi aggiunge predittori al modello uno per volta fino a che tutti i predittori sono nel modello.

Ad ogni passo viene aggiunta al modello la variabile che da il più grande additional improvement.

Algoritmo 6.2 Forward stepwise selection

1. Sia M_0 tale che denota il null model, che non contiene predittori.
2. Per k=0, ..., p-1 (o fino al soddisfacimento di una stopping rule):
   a. si considerano tutti i p-k modelli che aumentano i predittori in M_k con un predittore aggiuntivo;
   b. si sceglie il migliore tra i p-k modelli e lo si chiama M_{k+1}. Per "migliore" qui si intende quello che ha il più piccolo RSS o il più grande R^2.
3. Si seleziona il modello migliore tra M_0, ..., M_p, usando il cross-validated prediction error, C_p (AIC), BIC o adjusted R^2.

Consiste nel fare il fit di un null model e di p-k modelli nella k-esima iterazione, per k=0, ..., p-1. Questo porta ad un numero totale di modelli:

$$1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$$

Quando p=20, la best subset selection richiede di fare il fit di 1,048,576 modelli, mentre la forward stepwise selection richede di fare il fit di soli 211 modelli.

💡 Però, la forward stepwise selection potrebbe non trovare il modello migliore possibile tra tutti 2^p modelli possibili.

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income` `student`, `limit`	`rating`, `income`, `student`, `limit`

Backward stepwise selection

Backward stepwise selection inizia con il modello least squares contenente tutti i p predittori e poi rimuove iterativamente il meno utile, uno per volta.

Algoritmo 6.3 Backward stepwise selection

1. Sia M_p il full model, tale che contiene tutti i p predittori.
2. Per k = p, p-1, ..., 1:
   a. si considerano tutti i k modelli che contengono tutti i predittori in M_k tranne 1, per un totale di k-1 predittori;
   b. si sceglie il best tra questi k modelli e lo si chiama M_{k-1}. Per "migliore" si intende quello avente il più piccolo RSS o il più grande R^2.
3. Si seleziona un singolo modello best tra M_0, ..., M_p usando il cross-validated prediction error, C_p (AIC), BIC o adjusted R^2.

Alternativamente si può anche usare come criterio di valutazione il p-value, si inizia con tutte le variabili nel modello, poi si rimuove la variabile con il più grande p-value, cioè la variabile che è meno significativa da un punto di vista statistico. Con questo abbiamo adattato (effettuato il fit) il nuovo modello a (p-1)-variabili.
Questa procedura continua fino al soddisfacimento di una
stopping rule, ad esempio ci si può fermare quando tutte le variabili rimanenti hanno un p-value al di sotto di un certo threshold.

Forward vs. backward stepwise selection

Entrambi i metodi cercano solo attraverso $1 + \frac{p(p+1)}{2}$ modelli e quindi possono essere applicati in contesti nei quali p è troppo grande per applicare best subset selection.

Entrambi i metodi non garantiscono di restituire il modello migliore contenente un sottoinsieme dei p predittori.

La Forward stepwise selection può essere applicata anche in casi ad alta dimensionalità dove $n \leq p$; in questo caso, possiamo costruire i sottomodelli solo fino a M_{n-1}.

La Backward selection richiede $n > p$ (in modo che l'intero modello possa essere fit).

Approcci Ibridi

Negli approcci ibridi le variabili sono aggiunte al modello sequenzialmente, similarmente a quanto avviene in forward selection.

Dopo l'aggiunta di ogni variabile il metodo rimuove ogni variabile che non apporta più un improvement al model fit.

Mixed Selection

Questa è una combinazione di forward e backward selection. Si inizia con nessuna variabile nel modello e, come nella forward selection, aggiungiamo la variabile che fornisce il miglior fit.

Proseguiamo aggiungendo variabili una alla volta.
Come notato nell'esempio Advertising, i p-values per le variabili possono crescere quando nuovi predittori vengono aggiunti al modello.
Di conseguenza,
se in un qualsiasi momento il p-value per una delle variabili nel modello sale al di sopra di un valore soglia quella variabile viene rimossa dal modello.
Questa procedura continua fino a che tutte le variabili nel modello hanno un
p-value sufficientemente basso e al contempo tutte le variabili al di fuori avrebbe un p-value grande se aggiunte al modello.

Forward selection è un approccio "greedy" e potrebbe includere nelle fasi iniziali variabili che più avanti divengono ridondanti. Si può rimediare a questo problema con la Mixed selection.