Best Subset Selection

Data Mining
CLAMSES - University of Milano-Bicocca

Aldo Solari

Riferimenti bibliografici

- AS § 3.6.1
- HTF § 3.3, § 7.10.2

Si assuma che il modello generatore dei dati sia il modello lineare $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}$ dove $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{o}$ e $\mathbb{V}\mathrm{ar}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. In molti set di dati moderni, ci sono motivi per credere che siano presenti molte più variabili di quelle necessarie per spiegare la risposta. Sia S l'insieme degli indici delle variabili rivelanti, i.e.

$$S = \{k \in \{1, \dots, p\} : \beta_k^0 \neq 0\}$$

e sia $s = |S| \ll p$ la cardinalità dell'insieme S. L'errore quadratico medio di previsione (Fixed-X) per lo stimatore OLS è

$$\begin{aligned}
\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}}) &= \frac{1}{n}\mathbb{E}\|\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}\|_2^2 \\
&= \frac{1}{n}\mathbb{E}\{(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})\} \\
&= \frac{1}{n}\mathbb{E}[\mathrm{tr}\{(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\}] \\
&= \frac{1}{n}\mathrm{tr}[\mathbb{E}\{(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}^{\mathrm{OLS}})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\}] \\
&= \frac{1}{n}\mathrm{tr}[\mathbb{V}\mathrm{ar}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}})\mathbf{X}^\mathsf{T}\mathbf{X}] = \frac{p}{n}\sigma^2
\end{aligned}$$

Se potessimo identificare S e quindi adattare un modello lineare utilizzando solo queste variabili, otterremmo un errore di previsione di $s\sigma^2/n$ invece di $p\sigma^2/n$.

Inoltre, si può dimostrare che le stime dei parametri dal modello ridotto sono più accurate. Il modello più parsimonioso sarebbe anche più facile da interpretare.

Un approccio naturale per la ricerca di S è considerare tutti i 2^p possibili modelli di regressione, ognuno dei quali implica la regressione della variabile risposta su un diverso set di predittori $\mathbf{X}_M$, dove M è un sottoinsieme di $\{1, \ldots, p\}$

Algoritmo Best Subset Selection

Set B_0 as the null model (only intercept)

For $k = 1, \ldots, p$:

1. Fit all $\binom{p}{k}$ models that contain exactly k predictors
2. Pick the "best" among these $\binom{p}{k}$ models, and call it B_k, where "best" is defined having the smallest residual sum of squares RSS = $n\text{MSE}_{\text{Tr}}$

Select a single best model from among $B_0, B_1, \ldots, B_p$ using AIC, BIC, Cross-Validation, etc.

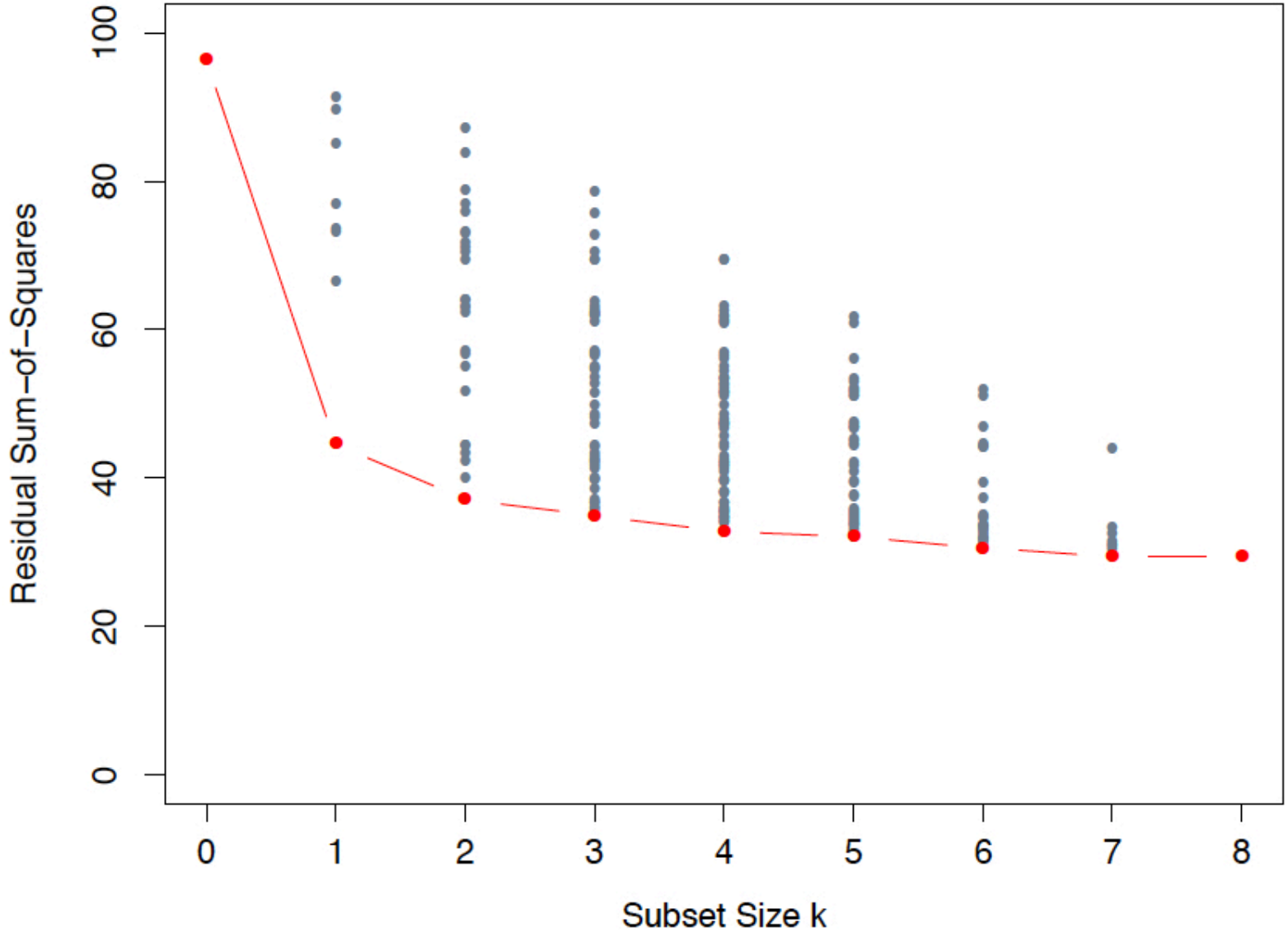


Figura 3.5 in Hastie, Tibshirani and Friedman.

Algoritmo Backward stepwise selection

La soluzione è sub-ottimale rispetto a Best Subset, ma è più efficiente dal punto di vista computazionale. È applicabile solo quando $n > p$.

Set S_p as the full model (all p predictors)

For $k = p, p-1, \ldots, 1$:

1. Consider all k models that contain all but one of the predictors in S_k, for a total of $k-1$ predictors
2. Choose the ”best” among these k models and call it S_{k-1}, where “best” is defined having the smallest RSS

Select a single best model from among $S_0, S_1, \ldots, S_p$ using AIC, BIC, cross-validation, etc.

Algoritmo Forward stepwise selection

Applicabile anche quando $n < p$, individua la sequenza $S_0, S_1, \ldots, S_{n-1}$

Set S_0 as the null model (only intercept)

For $k = 0, \ldots, \min(n-1, p-1)$:

1. Consider all $p - k$ models that augment the predictors in S_k with one additional predictor
2. Choose the "best" among these $p - k$ models and call it S_{k+1}, where "best" is defined having the smallest RSS

Select a single best model from among $S_0, S_1, S_2, \ldots$ using AIC, BIC, cross-validation, etc.

Algoritmo Forward with AIC-based stopping rule

Set S_0 as the null model and $k = 0$.

1. Consider all $p - k$ models that augment the predictors in S_k with one additional predictor.
2. Choose the ”best” among these $p - k$ models and call it S_{k+1}, where ”best” is defined having the smallest AIC.
3. If AIC(S_{k+1}) $<$ AIC(S_k), set $k = k + 1$ and go to 1., otherwise STOP

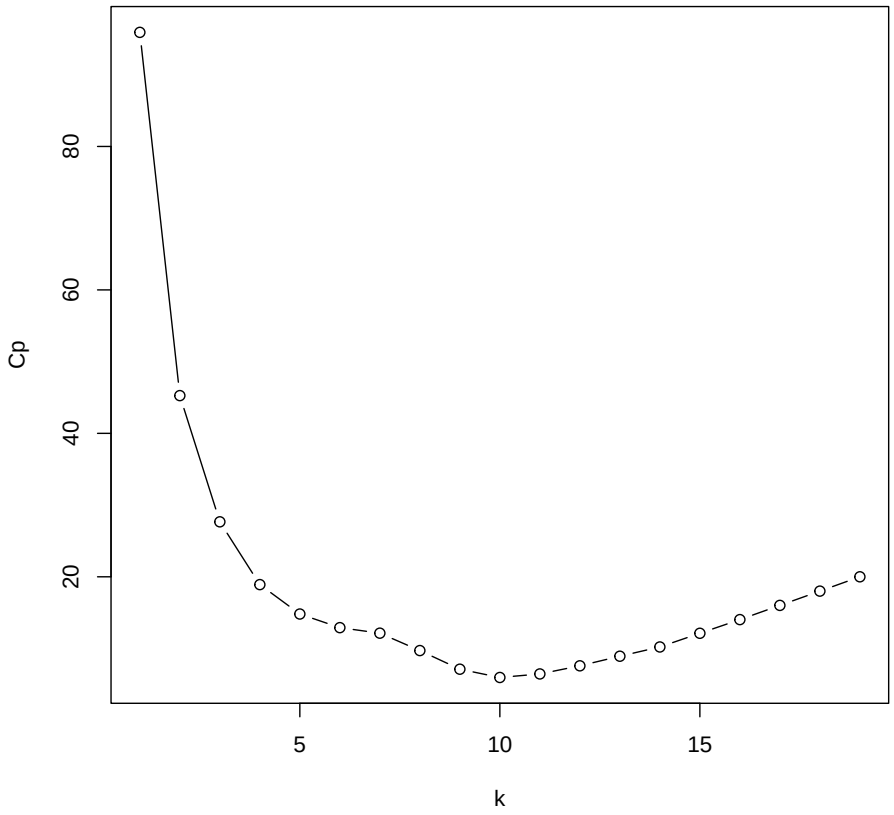
Si considerino i dati Hitters:

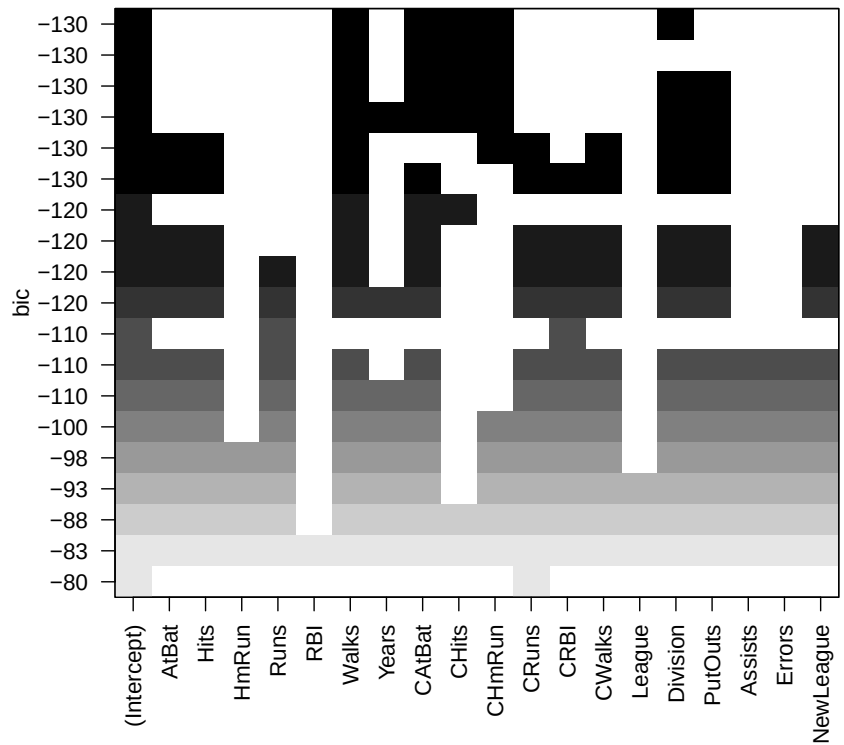
```
library(ISLR)
data(Hitters)
Hitters = Hitters[complete.cases(Hitters),]
Hitters[,"League"]=(Hitters[,"League"]=="A")*1
Hitters[,"Division"]=(Hitters[,"Division"]=="E")*1
Hitters[,"NewLeague"]=(Hitters[,"NewLeague"]=="A")*1
set.seed(123)
n = 163
train.id = sample(nrow(Hitters),n)
train = Hitters[train.id,]
names(train)[19] = "y"
p = ncol(X)
test = Hitters[-train.id,]
names(test)[19] = "y"
m = nrow(X.star)
```

Esercizio 1

- Calcolare RMSE_{Te}, l'errore di previsione sui dati di test RMSE_{Te} per il modello con tutte le variabili
- Calcolare RMSE_{Te} per il modello basato sul Best Subset identificato con il criterio AIC e BIC, utilizzando la funzione `regsubsets` presente nella libreria `leaps`
- Calcolare RMSE_{Te} per il modello identificato con l'algoritmo forward con AIC-stopping rule, utilizzando la funzione `step`

	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	CI
1 (1)											*	
2 (1)				*								*
3 (1)						*		*	*			
4 (1)						*		*	*	*		
5 (1)						*		*	*	*		
6 (1)						*		*	*	*		
7 (1)						*	*	*	*	*		
8 (1)	*	*				*				*	*	
9 (1)	*	*				*		*			*	*
10 (1)	*	*				*		*			*	*


Cp
20
40
60
80
5
10
15
k


bic
−130
−130
−130
−130
−130
−130
−120
−120
−120
−120
−110
−110
−110
−100
−98
−93
−88
−83
−80
(Intercept)
AtBat
Hits
HmRun
Runs
RBI
Walks
Years
CAtBat
CHits
CHmRun
CRuns
CRBI
CWalks
League
Division
PutOuts
Assists
Errors
NewLeague

Best Subset Selection con convalida incrociata

Poiché la selezione delle variabili fa parte del processo di costruzione del modello, la convalida incrociata dovrebbe tenere conto della variabilità della selezione durante il calcolo delle stime dell'errore del test.

Se l'intero set di dati viene utilizzato per trovare il Best Subset, la stima dell'errore sui dati di test tramite convalida incrociata sarà distorta verso il basso

Bisogna quindi eseguire la *best subset selection* all'interno di ogni iterazione di convalida incrociata. Si noti che a ogni iterazione vengono identificati sottoinsiemi potenzialmente diversi dei "migliori" k predittori. Si veda anche ISLR, § 6.5.3

Esercizio 2

- Calcolare RMSE_{Te} per il modello basato sul Best Subset identificato con un *validation set*
- Calcolare RMSE_{Te} per il modello basato sul Best Subset identificato con il metodo della convalida incrociata con $K = 10$

§ 7.10.2 HTF

Consider a scenario with $n = 50$ samples in two equal-sized classes, and $p = 5000$ quantitative predictors (standard Gaussian) that are independent of the class labels. The true (test) error rate of any classifier is 50%. Try the following approach:

1. choose the 100 predictors having highest correlation with the class labels
2. use a 1-nearest neighbors classifier, based on just these 100 selected predictors.
3. Use K-fold CV to estimate the test error of the final model

Leaving observations out after the variables have been selected does not correctly mimic the application of the classifier to a completely independent test set, since these predictors have already seen the left out observations. Here is the correct way to carry out cross-validation in this example:

1. Divide the observations into K cross-validation folds at random
2. For each fold $k = 1, \ldots, K$
    a. Find the best 100 predictors that have the largest (in absolute value) correlation with the class labels, using all of the observations except those in fold k
    b. Using just this subset of predictors, fit a 1-nearest neighbors classifier, using all of the observations except those in fold k
    c. Use the classifier to predict the class labels for the observations in fold k

Esercizio 3

Permutare i valori del vettore risposta del training set e del test set, in modo da rendere la variabile risposta indipendente da tutti i predittori. Utilizzare l'algoritmo *Best Subset Selection*. Quale *best subset* viene selezionato con i criteri AIC, BIC e cross-validation? Commentare il risultato.