Ottimismo

Data Mining
CLAMSES - University of Milano-Bicocca

Aldo Solari

Riferimenti bibliografici

- AS §3.4, §3.5.1, §3.5.2, §3.5.3
- HTF §7.4, §7.5

Table of Contents

Visto che non conosciamo $f(x)$...

La conclusione della lezione precedente è stata che dobbiamo operare un compromesso tra le componenti di distorsione e di varianza. Operativamente però non possiamo utilizzare la conoscenza di $f(x)$, ovviamente ignota in pratica

Presenteremo due approcci per stimare l'errore di previsione e scegliere il modello:

- Il concetto di ottimismo e i criteri basati sull'informazione
- Il metodo della convalida incrociata

Stima dell'errore di previsione

Calcolare l'errore quadratico medio sugli stessi dati con i quali abbiamo stimato il modello non è sembrato molto utile

Si ricordi l'esempio della regressione polinomiale: $\mathrm{MSE}_{\mathrm{Tr}}$ decresce al crescere di d, fino ad arrivare a 0 per $d = n - 1$

Tuttavia se risultasse

$$\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}}) = \mathbb{E}(\mathrm{MSE}_{\mathrm{Tr}}) + \text{costante}$$

allora si potrebbe stimare l'errore (atteso) di previsione come

$$\begin{aligned}\widehat{\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}})} &= \widehat{\mathbb{E}(\mathrm{MSE}_{\mathrm{Tr}})} + \text{costante} \\ &= \mathrm{MSE}_{\mathrm{Tr}} + \text{costante}\end{aligned}$$

a patto di conoscere il valore della costante

Ottimismo

Si consideri il setting Fixed-X.

Chiameremo *ottimismo* la differenza

$$\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}}) - \mathbb{E}(\mathrm{MSE}_{\mathrm{Tr}}) = \frac{2}{n}\sum_{i=1}^{n}\mathbb{C}\mathrm{ov}(Y_i, \hat{f}(x_i)) = \mathrm{OptF}$$

Maggiore è la correlazione tra Y_i e $\hat{f}(x_i)$, maggiore è l'ottimismo

Abbiamo

$$\begin{aligned}\mathbb{E}[(Y_i - \hat{f}(x_i))^2] &= \mathbb{V}\mathrm{ar}(Y_i - \hat{f}(x_i)) + (\mathbb{E}[Y_i - \hat{f}(x_i)])^2 \\ &= \mathbb{V}\mathrm{ar}(Y_i) + \mathbb{V}\mathrm{ar}(\hat{f}(x_i)) - \\ &- 2\mathbb{C}\mathrm{ov}(Y_i, \hat{f}(x_i)) + (\mathbb{E}[Y_i] - \mathbb{E}[\hat{f}(x_i)])^2\end{aligned}$$

e

$$\begin{aligned}\mathbb{E}[(Y_i^* - \hat{f}(x_i))^2] &= \mathbb{V}\mathrm{ar}(Y_i^* - \hat{f}(x_i)) + (\mathbb{E}[Y_i^* - \hat{f}(x_i)])^2 \\ &= \mathbb{V}\mathrm{ar}(Y_i^*) + \mathbb{V}\mathrm{ar}(\hat{f}(x_i)) - \\ &- 2\mathbb{C}\mathrm{ov}(Y_i^*, \hat{f}(x_i)) + (\mathbb{E}[Y_i^*] - \mathbb{E}[\hat{f}(x_i)])^2\end{aligned}$$

Si noti che Y_i^* è indipendente da Y_i, ma hanno la stessa distribuzione, quindi $\mathbb{E}(Y_i^*) = \mathbb{E}(Y_i)$, $\mathbb{V}\mathrm{ar}(Y_i^*) = \mathbb{V}\mathrm{ar}(Y_i)$ e $\mathbb{C}\mathrm{ov}(Y_i^*, \hat{f}(x_i)) = 0$. Allora

$$\begin{aligned} \mathbb{E}[(Y_i^* - \hat{f}(x_i))^2] &= \mathbb{V}\mathrm{ar}(Y_i) + \mathbb{V}\mathrm{ar}(\hat{f}(x_i)) + (\mathbb{E}[Y_i] - \mathbb{E}[\hat{f}(x_i)])^2 \\ &= \mathbb{E}[(Y_i - \hat{f}(x_i))^2] + 2\mathbb{C}\mathrm{ov}(Y_i, \hat{f}(x_i)) \end{aligned}$$

Se facciamo la media su tutte le n osservazioni

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - \hat{f}(x_i))^2\right] = \left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(x_i))^2\right] + \frac{2}{n}\sum_{i=1}^{n}\mathbb{C}\mathrm{ov}(Y_i, \hat{f}(x_i))$$

$$\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}}) = \mathbb{E}(\mathrm{MSE}_{\mathrm{Tr}}) + \frac{2}{n}\sum_{i=1}^{n}\mathbb{C}\mathrm{ov}(Y_i, \hat{f}(x_i))$$

Ottimismo per il modello lineare

Si noti che

$$\frac{2}{n}\sum_{i=1}^{n}\mathbb{C}\mathrm{ov}(Y_i,\hat{f}(x_i)) = \frac{2}{n}\mathrm{tr}(\mathbb{C}\mathrm{ov}(\mathbf{Y},\hat{\mathbf{f}}))$$

dove $\mathbf{Y} = (Y_1,\ldots,Y_n)^\mathsf{T}$ e $\hat{\mathbf{f}} = (\hat{f}(x_1),\ldots,\hat{f}(x_n))^\mathsf{T}$

I valori previsti dal modello lineare si possono esprimere come

$$\hat{\mathbf{f}} = \mathbf{HY}$$

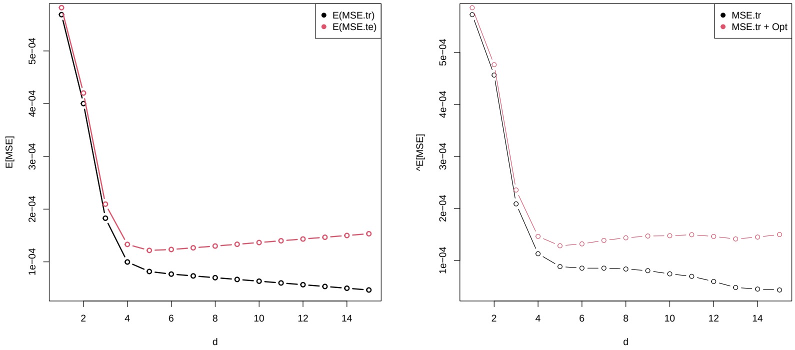
dove $\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$ è la matrice di proiezione (hat matrix), simmetrica e idempotente

Abbiamo

$$\mathrm{tr}\{\mathbb{C}\mathrm{ov}(\mathbf{Y},\mathbf{HY})\} = \mathrm{tr}\{\mathbb{C}\mathrm{ov}(\mathbf{Y},\mathbf{Y})\mathbf{H}^\mathsf{T}\} = \mathrm{tr}\{\sigma^2\mathbf{I}_n\mathbf{H}\} = \sigma^2\mathrm{tr}\{\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\}$$

e quindi

$$\mathrm{OptF} = \frac{2\sigma^2 p}{n}$$


E(MSE.tr)
E(MSE.te)
E[MSE]
5e−04
4e−04
3e−04
2e−04
1e−04
2
4
6
8
10
12
14
d
MSE.tr
MSE.tr + Opt
^E[MSE]
5e−04
4e−04
3e−04
2e−04
1e−04
2
4
6
8
10
12
14
d

Cp di Mallows

$\mathrm{OptF} = \frac{2\sigma^2 p}{n}$ richiede la conoscenza di σ^2, che però è un valore incognito

Possiamo però sostituirlo con la sua stima

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p} = \frac{\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{n-p}$$

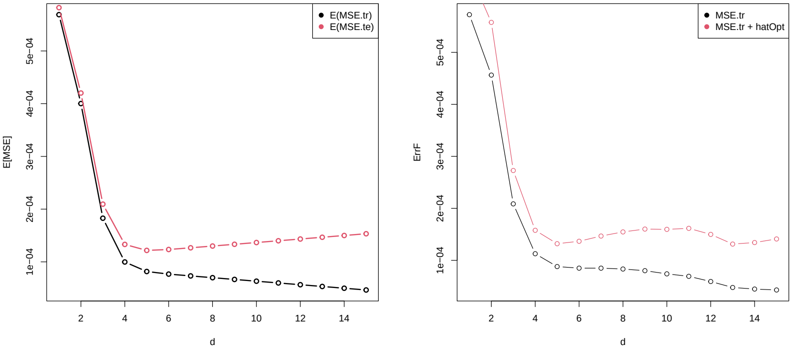
dove $\mathrm{RSS} = n\mathrm{MSE}_{\mathrm{Tr}}$ è la somma dei quadrati dei residui (Residual Sum of Squares)

Quindi una stima per l'errore di previsione è data da

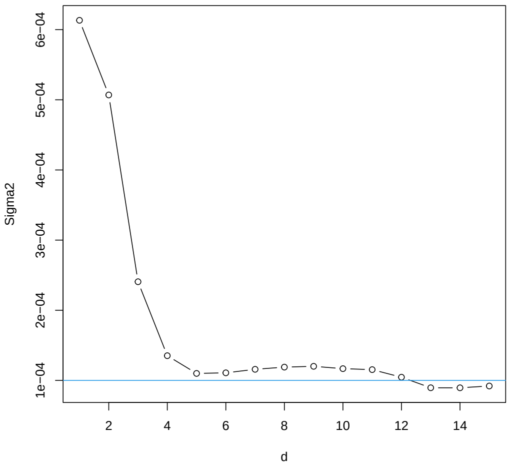
$$\mathbb{E}(\widehat{\mathrm{MSE}_{\mathrm{Te}}}) = \mathrm{MSE}_{\mathrm{Tr}} + \widehat{\mathrm{OptF}}$$

Questo stimatore è noto come Cp di Mallows:

$$\mathrm{Cp} = \mathrm{MSE}_{\mathrm{Tr}} + \frac{2\hat{\sigma}^2 p}{n}$$



E(MSE.tr)
E(MSE.te)
E[MSE]
5e−04
4e−04
3e−04
2e−04
1e−04
2
4
6
8
10
12
14
d
MSE.tr
MSE.tr + hatOpt
ErrF



Sigma2
6e−04
5e−04
4e−04
3e−04
2e−04
1e−04
2
4
6
8
10
12
14
d

Table of Contents

Criteri basati sull'informazione: AIC e BIC

AIC è definito come

$$\text{AIC} = -2 \cdot \text{loglikelihood}(\hat{\beta}, \hat{\sigma}^2) + 2p$$

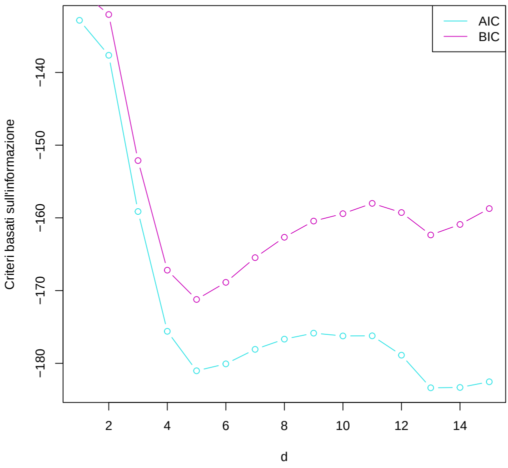
dove per il modello lineare $-2 \cdot \text{loglikelihood}(\hat{\beta}, \hat{\sigma}^2) = n \log(\text{MSE}_{\text{Tr}})$

Per i modelli lineari, Cp e AIC sono proporzionali, e quindi il valore più piccolo per Cp corrisponde al valore più piccolo per AIC

BIC è definito come

$$\text{BIC} = -2 \cdot \text{loglikelihood}(\hat{\beta}, \hat{\sigma}^2) + \log(n)p$$

Dal momento che $\log(n) > 2$ per $n > 7$, BIC generalmente seleziona modelli meno complessi rispetto a quelli selezionati da AIC



AIC
BIC
Criteri basati sull'informazione
−140
−150
−160
−170
−180
2
4
6
8
10
12
14
d

Table of Contents

Metodo della convalida incrociata

Come notato in precedenza, stimare un modello e valutarne la performance sugli stessi dati produce un risultato troppo ottimistico

Il metodo della convalida incrociata (Cross-Validation, abbreviato CV) valuta le previsioni del modello su dati ”nuovi” al fine di fornire una stima $\widehat{\mathrm{Err}}$ dell’errore di previsione $\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}})$

L’idea alla base del metodo è di dividere le osservazioni (data-split): una parte dei dati (training set) è utilizzata per addestrare il modello, e i dati rimanenti (test set) sono utilizzati per misurare la performance del modello

Una delle principali caratteristiche della CV è la sua universalità. CV è un metodo non parametrico che può essere applicato a qualsiasi algoritmo/modello. Questa universalità non è condivisa ad es. da Cp, che è specifico della regressione lineare

Validation set

Una semplice soluzione consiste nel dividere casualmente le n osservazioni in due parti: un insieme di addestramento e un insieme di verifica

Si stima il modello $\hat{f}^{-V}$ sull'insieme delle osservazioni di addestramento $T \subset \{1, \ldots, n\}$, e lo si utilizza per prevedere le osservazioni sull'insieme di verifica $V = \{1, \ldots, n\} \setminus T$

Questo approccio fornisce una stima dell'errore di previsione (atteso)

$$\widehat{\mathrm{Err}} = \frac{1}{\#V} \sum_{i \in V} (y_i - \hat{f}^{-V}(x_i))^2$$

Tuttavia questo procedimento riduce la numerosità delle osservazioni (ma questo non è un problema se n è veramente elevato)

Se n non è molto grande, tuttavia, questa stima può essere molto variabile

Convalida incrociata

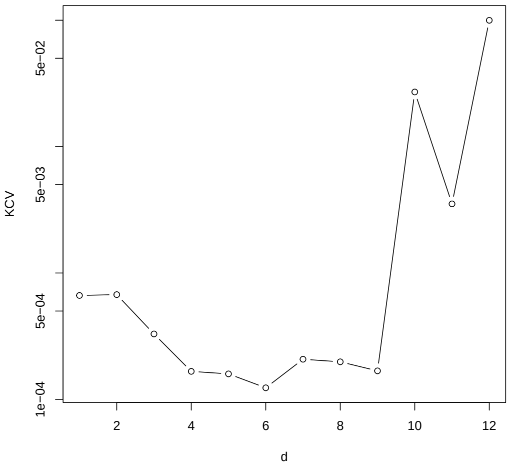
Un modo per superare parzialmente questa arbitrarietà è dividere i dati in parti uguali $V_1, \ldots, V_K$.

Nel metodo della convalida incrociata con K porzioni (K-fold cross-validation) utilizziamo le osservazioni $i \notin V_k$ per addestrare il modello e le osservazioni $i \in V_k$ per valutarlo:

$$\frac{1}{\# V_k} \sum_{i \in V_k} (y_i - \hat{f}^{-V_k}(x_i))^2$$

e alla fine calcoliamo la media delle stime per stimare l'errore di previsione atteso:

$$\widehat{\mathrm{Err}} = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{\# V_k} \sum_{i \in V_k} (y_i - \hat{f}^{-V_k}(x_i))^2 \right]$$



KCV
5e−02
5e−03
5e−04
1e−04
2
4
6
8
10
12
d

Leave-one-out cross validation

Nella leave-one-out cross validation (LOOCV), ciascuna osservazione viene esclusa a turno dall'insieme delle osservazioni per essere utilizzata per la verifica della previsione

Per $i = 1, \ldots, n$:

- Escludere l' i-sima osservazione (x_i, y_i) - Utilizzare le rimandenti $n - 1$ osservazioni per stimare il modello $\hat{f}^{-i}$ e valutarlo sull'osservazione esclusa (x_i, y_i), calcolando $(y_i - \hat{f}^{-i}(x_i))^2$ - Infine, calcolare la media

$$\widehat{\mathrm{Err}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{-i}(x_i))^2$$

Si noti che LOOCV è un caso particolare di K-fold CV che corrisponde a $K = n$.

LOOCV per il modello lineare

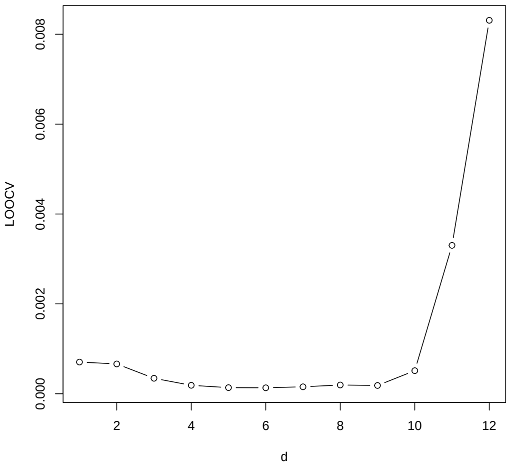
Per il modello lineare, c'è una scorciatoia per calcolare LOOCV:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}^{-i}(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(x_i)}{1-h_{ii}}\right)^2$$

$\underset{n\times p}{\mathbf{X}}$ è la matrice del disegno

h_{ii} è l' i-simo elemento diagonale della matrice di proiezione

$$\underset{n\times n}{\mathbf{H}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$$



LOOCV
0.000
0.002
0.004
0.006
0.008
2
4
6
8
10
12
d

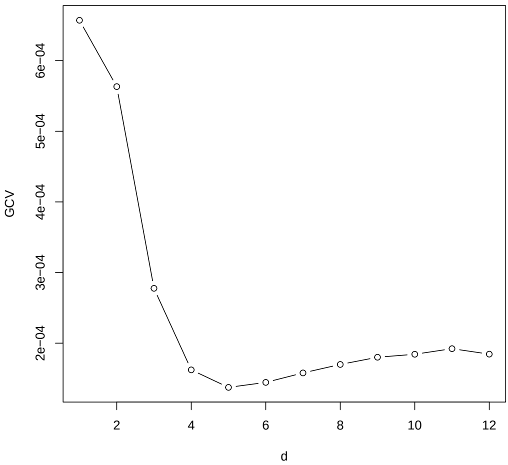
Metodo della convalida incrociata generalizzata

Nella convalida incrociata generalizzata calcoliamo

$$\widehat{\mathrm{Err}} = \frac{\mathrm{MSE}_{\mathrm{Tr}}}{\left(1 - \frac{p}{n}\right)^2}$$

dove stiamo approssimando h_{ii} con la sua media

$$\frac{1}{n} \sum_{i=1}^{n} h_{ii} = \frac{p}{n}$$


GCV
6e−04
5e−04
4e−04
3e−04
2e−04
2
4
6
8
10
12
d

La scelta di K

Una scelta comune per K oltre a $K = n$ è di scegliere $K = 5$ o $K = 10$

Tuttavia, trarre una conclusione generale sul CV è un compito quasi impossibile a causa della varietà delle situazioni che si possono incontrare

Compromesso distorsione-varianza per CV

Distorsione

K-fold CV con $K = 5$ o 10 fornisce una stima distorta (verso l'alto) dell'errore di previsione $\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}})$ perchè utilizza meno osservazioni nella stima del modello (4/5 or $9/10$ delle osservazioni)

LOOCV ha distorsione molto bassa (utilizza $n - 1$ osservazioni)

Varianza

Solitamente LOOCV è fortemente variabile perchè è la media di n quantità estremamente correlate (perchè le stime $\hat{f}^{-i}$ e $\hat{f}^{-l}$ si basano su $n - 2$ osservazioni comuni), e K-fold CV con $K = 5$ o 10 a meno variabilità perchè è la media di quantità meno correlate

Si ricordi che la varianza della somma di quantità fortemente correlate è maggiore di quella con quantità mediamente correlate:

$$\mathbb{V}\mathrm{ar}(A + B) = \mathbb{V}\mathrm{ar}(A) + \mathbb{V}\mathrm{ar}(B) + 2\mathbb{C}\mathrm{ov}(A, B)$$