Un semplice problema-tipo

Data Mining
CLAMSES - University of Milano-Bicocca

Aldo Solari

Riferimenti bibliografici

- AS §3.2

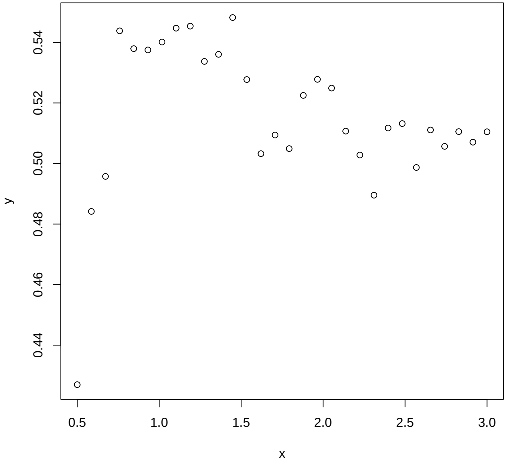
Descrizione del problema

- Si consideri il seguente problema illustrativo che ci servirà da propototipo per situazioni più complesse e realistiche.
- Ieri abbiamo raccolto $n = 30$ coppie di osservazioni, i dati di addestramento (training set)

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- Domani osserveremo nuove $n = 30$ coppie di osservazioni, i dati di verifica (test set)

$$(x_1, y_1^*), (x_2, y_2^*), \ldots, (x_n, y_n^*)$$


0.44
0.46
0.48
0.50
0.52
0.54
y
0.5
1.0
1.5
2.0
2.5
3.0
x

```
library(readr)
PATH <- "http://azzalini.stat.unipd.it/Book-DM/
yesterday.dat"
df <- read_table(PATH)
train <- data.frame(x=df$x, y=df$y.yesterday)
```

Dati simulati

I dati in realtà sono stati generati artificialmente da una legge del tipo

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

dove $\varepsilon_1, \ldots, \varepsilon_n$ sono variabili casuali (v.c.) indipendenti e identicamente distribute (i.i.d.) $N(0, \sigma^2)$ con $\sigma = 10^{-2}$, mentre f è una funzione che lasceremo non specificata, salvo per il fatto che si tratta di una funzione dall'andamento sostanzialmente regolare. Naturalmente per poter generare i dati è stata scelta una funzione specifica (e non è un polinomio).

Si noti che la v.c. viene indicata con Y_i, mentre la sua realizzazione (il valore osservato) con y_i. Inoltre si assume che $x_1, \ldots, x_n$ sono dei valori costanti (non casuali) fissati dallo sperimentatore.

Fixed-X setting

Per semplicità di ragionamento assumiamo che queste nuove y_i^* siano associate alle stesse ascisse x_i dei dati di ieri. Abbiamo quindi che domani osserveremo n coppie di dati (x_i, y_i^*) per $i = 1, \ldots, n$, i dati di verifica (test set) generati come

$$Y_i^* = f(x_i) + \varepsilon_i^*, \quad i = 1, \ldots, n$$

dove $\varepsilon_1^*, \ldots, \varepsilon_n^*$ sono i.i.d. $N(0, \sigma^2)$ con $\sigma = 10^{-2}$.

Le assunzioni fatte corrispondono al cosiddetto *Fixed-X setting*:

- i valori $x_1, \ldots, x_n$ del training set sono fissati (non casuali)
- i valori di x nel test set sono uguali ai valori di x nel training set

Regressione polinomiale

Si consideri un modello di regressione polinomiale di grado d:

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \ldots + \beta_{d+1} x^d$$

E' quindi possibile utilizzare i dati di addestramento (training set) per ottenere le stime $\hat{\beta}_1, \hat{\beta}_2, \ldots$ e quindi

$$\hat{f}(x) = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 x^2 + \ldots + \hat{\beta}_{d+1} x^d$$

per predire le nuove y_i^* che osserveremo domani utilizzando

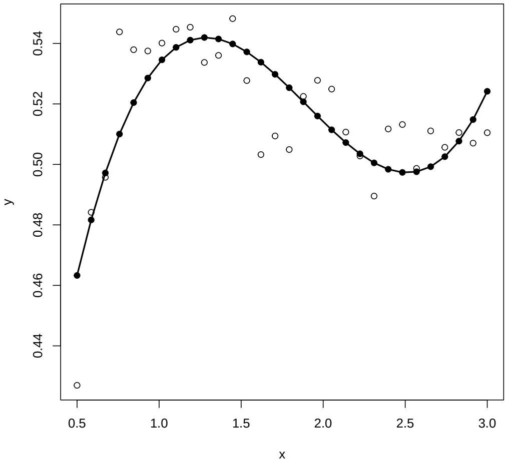
$$\hat{y}_i = \hat{f}(x_i), \quad i = 1, \ldots, n$$

Non avendo informazioni che ci guidino nella scelta del grado del polinomio, possiamo considerare tutti i gradi possibili con d tra 0 e $n-1$, quindi con un numero $p = d+1$ di parametri che varia da 1 a n, in aggiunta a σ.

Esercizio 1

- Stimare il modello di regressione polinomiale di grado $d = 3$
- Aggiungere al diagramma di dispersione di (x_i, y_i), $i = 1, \ldots, n$ i valori previsti dal modello.
- Calcolare l'errore quadratico medio sui dati di training (Training Mean Squared Error)

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

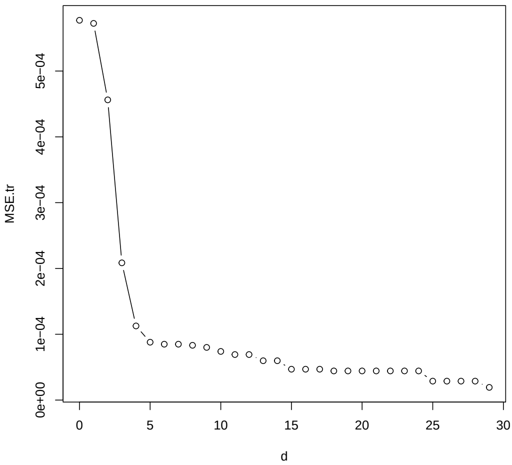


MSE_Tr = 0.0002085353

Esercizio 2

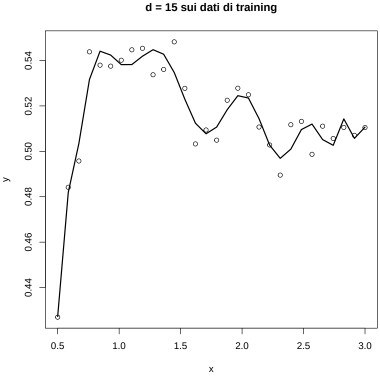
Sia $\mathrm{MSE}_{\mathrm{Tr}}(d)$ l'errore quadratico medio sui dati di training per la regressione polinomiale di grado d.

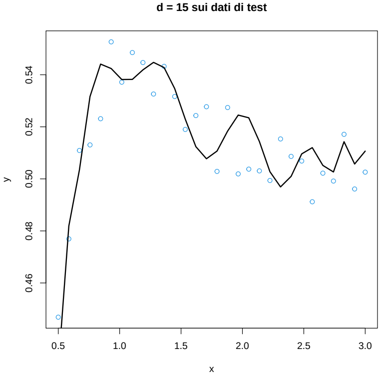
Costruire il grafico $(d, \mathrm{MSE}_{\mathrm{Tr}}(d))$ per $d = 0, 1, \ldots, 29$.



MSE.tr
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
0
5
10
15
20
25
30
d

Sovra-adattamento (overfitting)





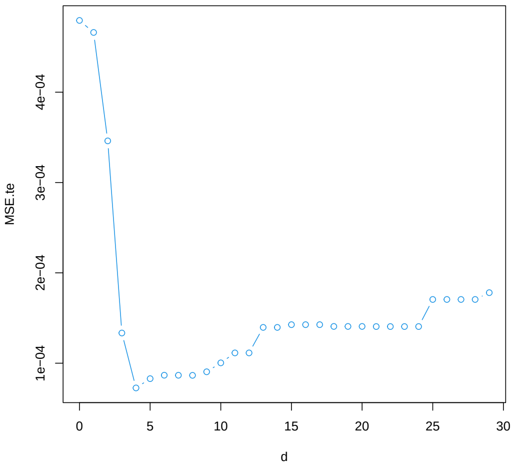
Esercizio 3

Si decida il grado d da utilizzare per prevedere i dati di domani, con l'obiettivo di minimizzare l'errore di previsione, ovvero l'errore quadratico medio sui dati di test (Test Mean Squared Error)

$$\mathrm{MSE}_{\mathrm{Te}} = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{f}(x_i))^2$$

Si noti che $\mathrm{MSE}_{\mathrm{Te}}$ sarà calcolabile solo domani (ovvero dopo aver fatto le previsioni), a differenza dell'errore quadratico medio sui dati di training $\mathrm{MSE}_{\mathrm{Tr}}$, che si può calcolare già oggi avendo a disposizione i dati di ieri.

Si giustifichi la scelta effettuata.




```
test <- data.frame(x=df$x, y=df$y.tomorrow)
```

La regressione polinomiale è un caso particolare del modello lineare (in notazione matriciale)

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

dove $\underset{n\times 1}{\mathbf{y}} = (y_1, \ldots, y_n)^\mathsf{T}$ è il vettore risposta di dimensione $n \times 1$, $\underset{p\times 1}{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ è il vettore dei coefficienti di dimensione $p \times 1$ e $\underset{n\times p}{\mathbf{X}}$ è la matrice del disegno di dimensione $n \times p$, ovvero

$$\underset{n\times 1}{\mathbf{y}} = \left[\begin{array}{c} y_1 \\ \cdots \\ y_i \\ \cdots \\ y_n \end{array}\right] \qquad \underset{n\times p}{\mathbf{X}} = \left[\begin{array}{c} x_1^\mathsf{T} \\ x_2^\mathsf{T} \\ \cdots \\ x_i^\mathsf{T} \\ \cdots \\ x_n^\mathsf{T} \end{array}\right] = \left[\begin{array}{cccccc} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{array}\right]$$

e infine $\underset{n\times 1}{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T}$ ha distribuzione Normale n-variata $N_n(\mathbf{o}, \sigma^2\mathbf{I}_n)$ dove $\mathbf{I}_n$ indica la matrice identità con n righe.

La matrice del disegno

	Intercept	x	x^2	x^3
1	1	0.50	0.25	0.12
2	1	0.59	0.34	0.20
3	1	0.67	0.45	0.30
4	1	0.76	0.58	0.44
5	1	0.84	0.71	0.60
6	1	0.93	0.87	0.81
...				

La matrice del disegno del modello di regressione polinomiale di terzo grado ha dimensione $p = 4$ perchè include l'intercetta 1 e i termini x, x^2, x^3

Polinomi ortogonali

```
fit <- lm( y ~ poly(x, degree=3, raw=FALSE), train)
X = model.matrix(fit)
colnames(X) = c("Intercept","x1","x2","x3")
round( t(X) %*% X, 8)
```

```
          Intercept x1 x2 x3
Intercept        30  0  0  0
x1                0  1  0  0
x2                0  0  1  0
x3                0  0  0  1
```

Per una spiegazione di come la funzione `poly` costruisce i polinomi ortogonali si veda
`https://stackoverflow.com/questions/39031172/`
`how-poly-generates-orthogonal-polynomials-how-to-understan`
`39051154#39051154`

Di conseguenza, se $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}_p$, si ottiene

$$\hat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\mathbf{y}$$

Esercizio 4

- Le previsioni $\hat{y}_i$ ottenute con la regressione polinomiale utilizzando polinomi tradizionali (`raw = TRUE`) o ortogonali (`raw = FALSE`) sono le stesse? Perchè?
- Sul mio computer, se stimo la regressione polinomiale utilizzando i polinomi tradizionali (`raw = TRUE`) ottengo alcuni `NA` nelle stime dei coefficienti per $d \geq 12$. Perchè?
- Se invece stimo la regressione polinomiale utilizzando i polinomi ortogonali (`raw = FALSE`) ottengo un messaggio di errore per $d \geq 24$. Perchè?
- Cosa vi aspettate di ottenere (in termini di valori previsti $\hat{y}_i$) se utilizzate il polinomio di grado $n - 1$?

Yesterday's data and tomorrow's data

I dati sono disponibili all'indirizzo web `http://azzalini.stat.unipd.it/Book-DM/`. In particolare:

- `http://azzalini.stat.unipd.it/Book-DM/yesterday.dat` dove (`x`, `y.yesterday`) sono i dati di training (x_i, y_i) e (`x`, `y.tomorrow`) sono i dati di test (x_i, y_i^*)
- i valori della vera funzione f (`f.true`) in corrispondenza ai punti specificati (`x.30`) e il valore vero di σ (`sqm.true <- 0.01`): `http://azzalini.stat.unipd.it/Book-DM/f_true.R`

Esercizio 5

Si supponga di conoscere la vera f e di aver osservato i dati di ieri. Si decida il grado d da utilizzare per prevedere i dati di domani, dopo-domani, dopo-dopo-domani etc. ovvero il grado d che minimizza

$$\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}}|\mathrm{Training}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Y_i^* - \hat{f}(x_i))^2 | Y_1 = y_1, \ldots, Y_n = y_n\right]$$

dove il valore atteso è rispetto alle v.c. $Y_1^*, \ldots, Y_n^*$.

Si commenti il risultato con riferimento alla risposta fornite all'Esercizio 3.

Per rispondere a questa domanda, potete utilizzare i dati `x` e `y.yesterday`, `f.true` e `sqm.true`.

Esercizio 6

Si supponga di conoscere la vera f. Si decida il grado d da utilizzare per prevedere generici dati di domani con generici dati di ieri , con l'obiettivo di minimizzare il valore atteso dell'errore di previsione, ovvero

$$\mathbb{E}[\mathrm{MSE}_{\mathrm{Te}}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(y_i^* - \hat{f}(x_i))^2]$$

dove il valore atteso è rispetto alle v.c. $Y_1, \ldots, Y_n$ e $Y_1^*, \ldots, Y_n^*$.

Si commenti il risultato con riferimento alla risposta fornite agli Esercizi 3 e 5.

Per rispondere a questa domanda, potete utilizzare i dati `x` , `f.true` e `sqm.true`.