Metodi non parametrici per regressione e classificazione: K-Nearest Neighbors

Metodi parametrici: regressione lineare

La regressione lineare è un esempio di metodo *parametrico*, in quanto assume una forma funzionale lineare per $f(X)$.
Metodi parametrici:

- semplici da stimare
- semplici da interpretare
- fanno assunzioni forti sulla forma di $f(X)$

Metodi parametrici e non parametrici

Se la forma funzionale specificata è molto distante da quella vera, il metodo parametrico avrà una performance insoddisfacente in termini di accuratezza delle previsioni.
Al contrario, i metodi non parametrici non assumono esplicitamente una forma per $f(X)$, e offrono un approccio alternativo e più flessibile per la regressione.
Il più semplice e conosciuto tra i metodi non parametrici è la regressione basata sui k-Nearest Neighbors.

k-Nearest Neighbors

Il metodo dei k-Nearest Neighbors, KNN, può essere utilizzato sia per problemi di regressione che di classificazione.
Per prevedere o classificare una nuova osservazione, tale metodo si basa sull'idea di trovare osservazioni "simili" nell'insieme di stima (training set).

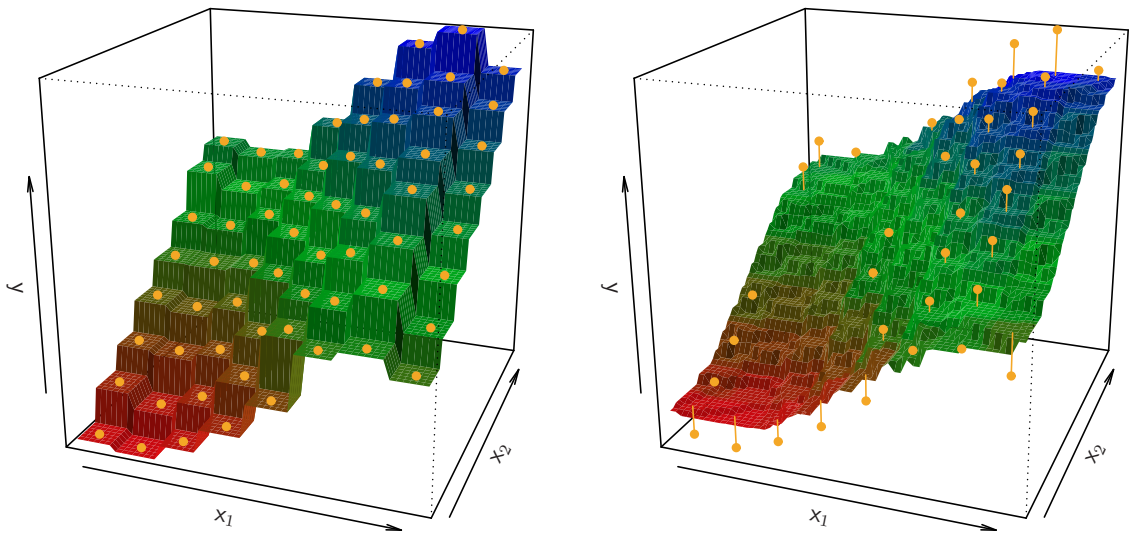
È un metodo data-driven altamente automatizzato, che identifica nell'insieme di stima le K osservazioni più vicine a quella che si vuole prevedere o classificare.

k-Nearest Neighbors: regressione

Dato un valore di k e un punto x_0, la regressione KNN identifica nel training set le k osservazioni più vicine, N_0

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in N_0} y_i$$

k-Nearest Neighbors: regressione



Previsioni con KNN per $p = 2$, $k = 1$ (sinistra) e $k = 9$ (destra). Con K più basso si ha previsione più flessibile con distorsione bassa e varianza alta, in quanto la previsione è effettuata su una sola osservazione.

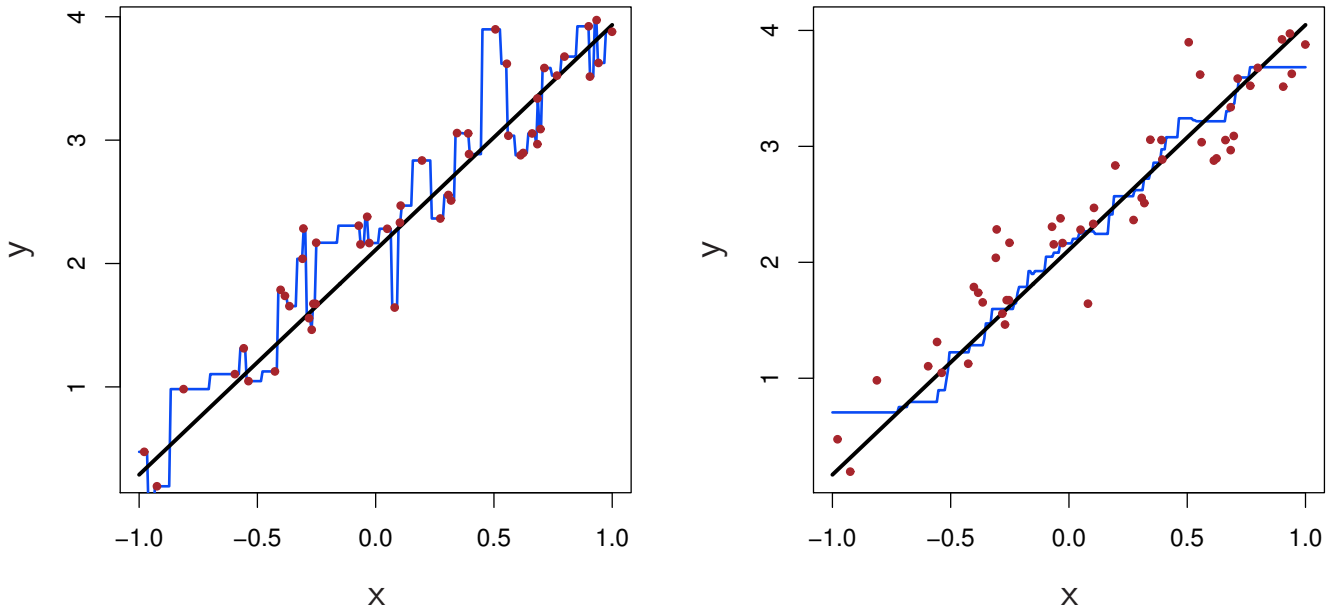
k-Nearest Neighbors: regressione

In generale il valore ottimale di k è collegato al trade-off tra varianza e distorsione.

- k piccolo $\rightarrow$ bassa distorsione e varianza elevata
- k elevato $\rightarrow$ bassa varianza (previsione più 'liscia') e distorsione elevata -può non essere catturata la struttura locale di $f(X)$-

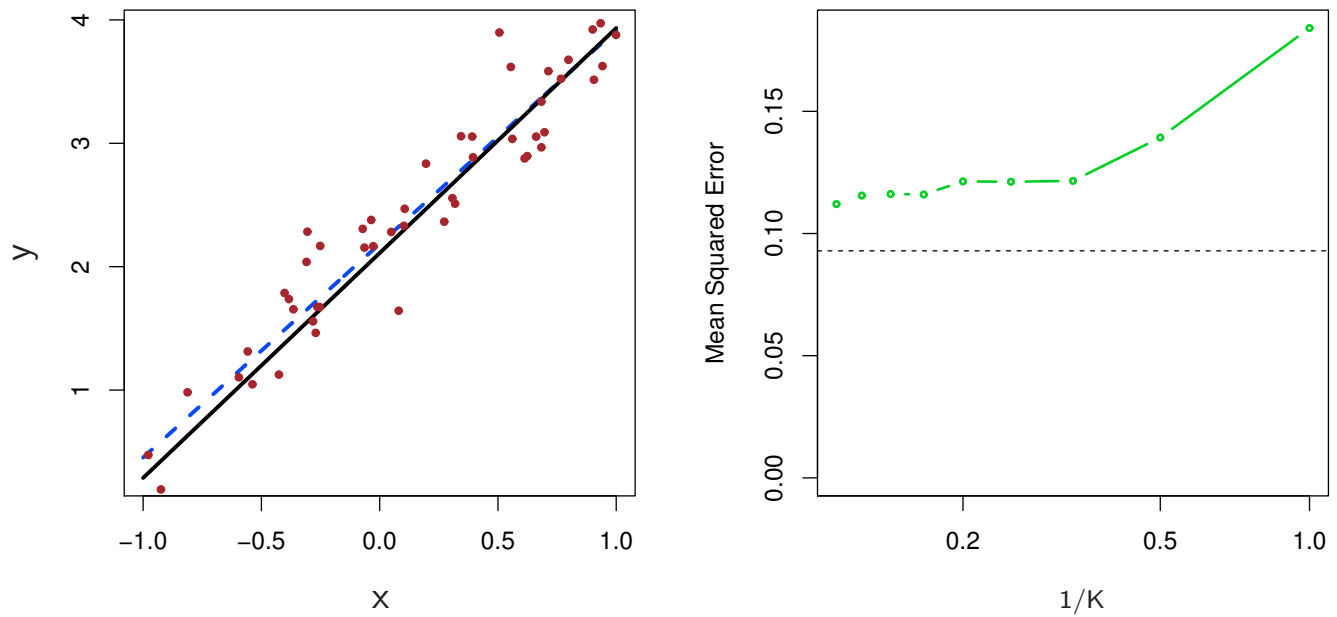
k-Nearest Neighbors: regressione

L'approccio parametrico è preferibile a quello non parametrico se la forma parametrica scelta è vicina alla vera forma di f.



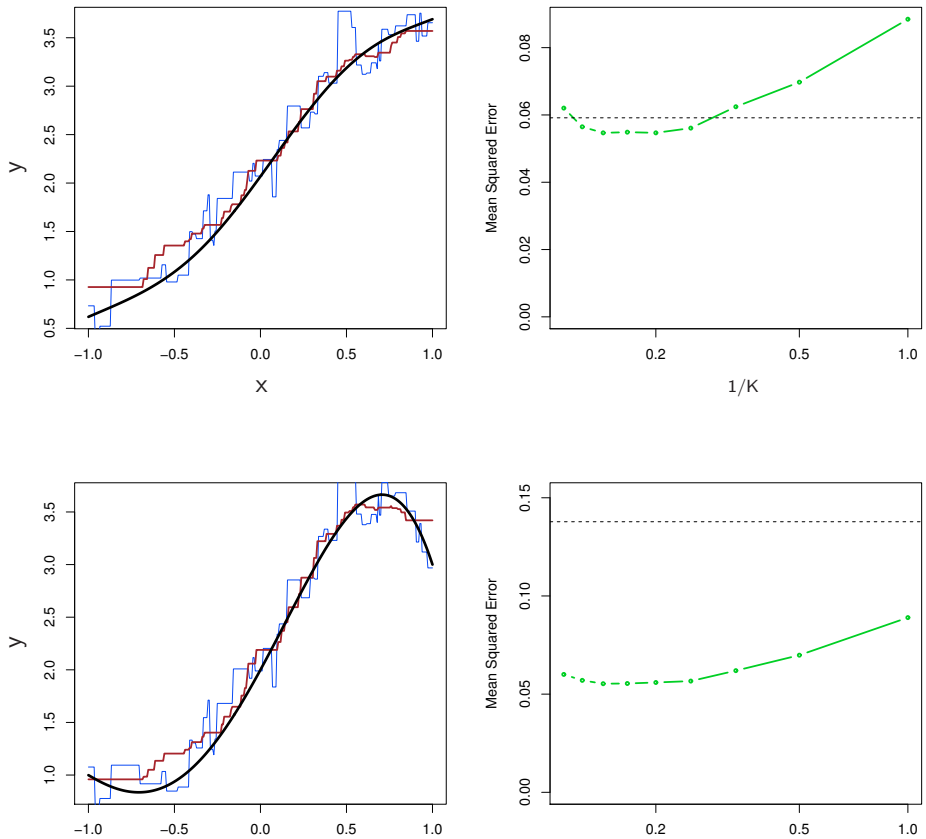
Confronto fra KNN con $k = 1$ (sinistra) e $k = 9$ (destra).
La vera relazione è lineare e quindi l'approccio non parametrico può difficilmente competere con quello parametrico.

k-Nearest Neighbors: regressione



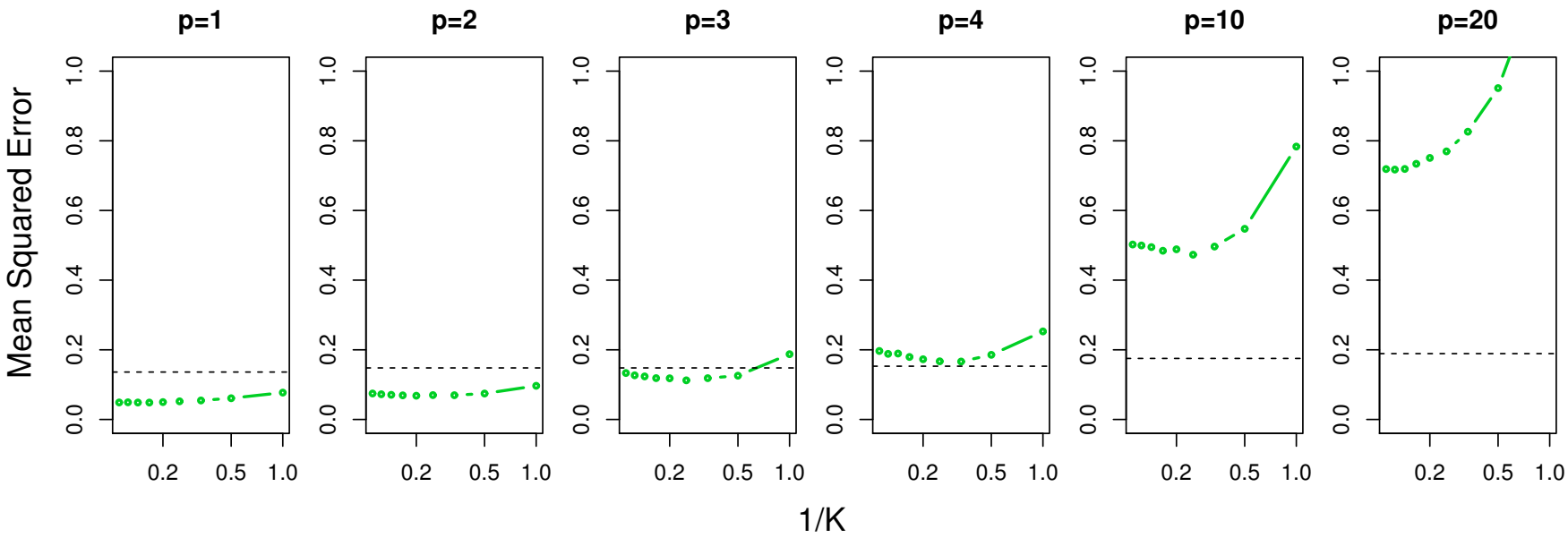
Retta di regressione (linea tratteggiata).
Test MSE per retta di regressione (linea tratteggiata) e KNN (linea verde) come funzione di $1/k$.
I migliori risultati per il KNN vengono raggiunti con un valore elevato di k.

k-Nearest Neighbors: regressione



Relazioni nonlineari e KNN con $k = 1$ (linea blu) e $k = 9$ (linea rossa). In funzione della non linearità di f cambia la performance di KNN rispetto al modello lineare. Più la relazione è nonlineare, migliore sarà la performance di KNN con un k elevato.

k-Nearest Neighbors: regressione

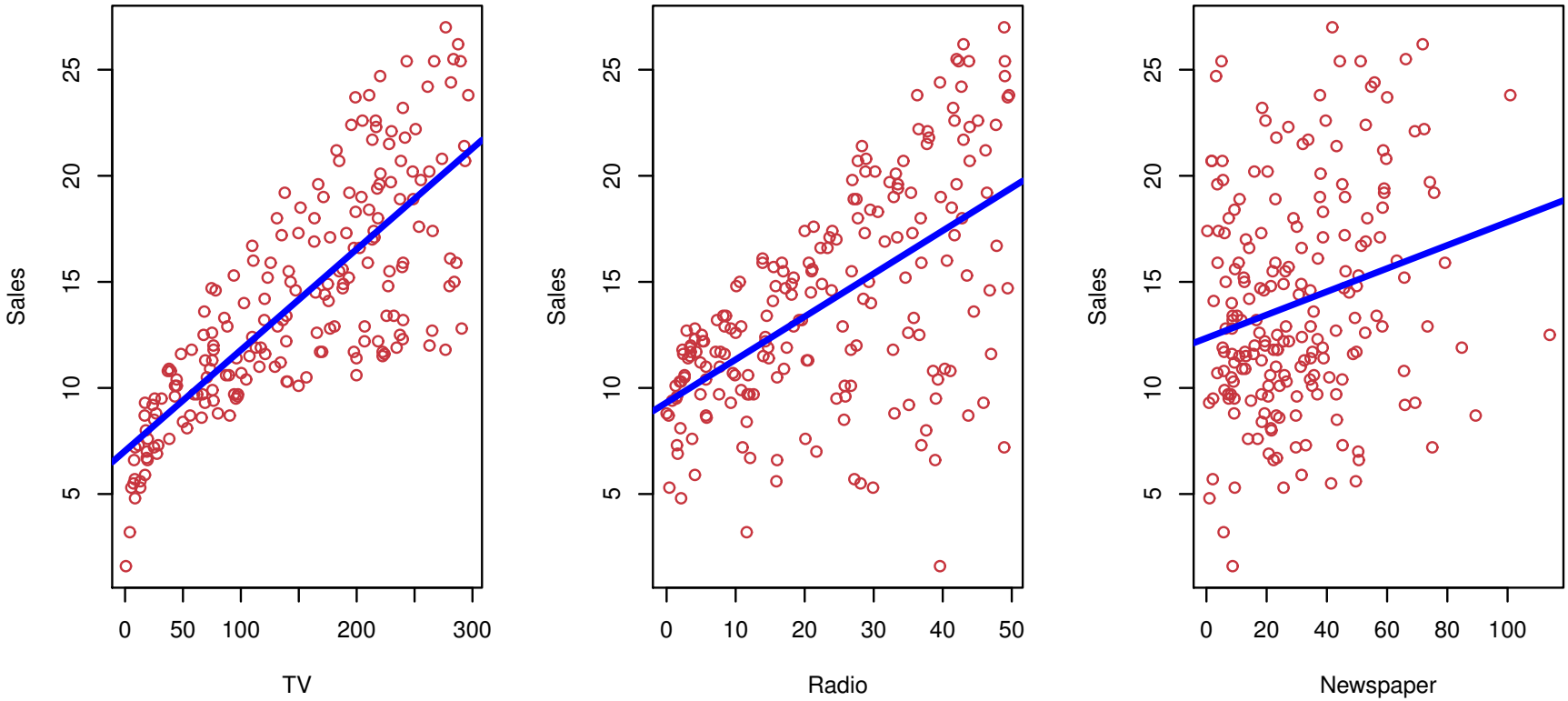


Al crescere del numero di variabili p, la performance di KNN diminuisce velocemente in termini di test MSE

Questo perchè al crescere di p è sempre più difficile trovare i "vicini" di una data osservazione. . . *maledizione delle dimensionalità*

k-Nearest Neighbors: esempio

Vendite in migliaia di unità, in funzione di budget in `tv`, `radio`, `giornali` per 200 mercati differenti.



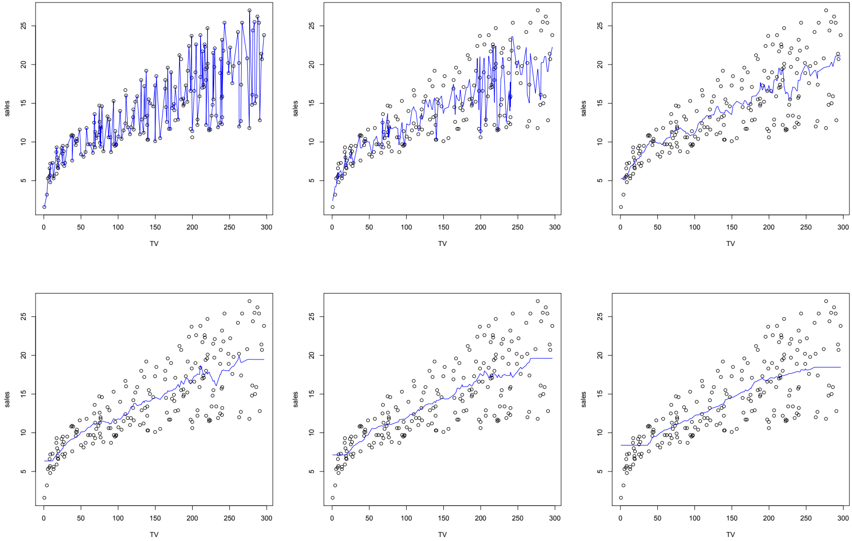
Retta di regressione stimata per tv, radio, giornali.

k-Nearest Neighbors: esempio

Vogliamo studiare la performance di un KNN con alcuni valori di k considerando inizialmente la sola variabile `tv`

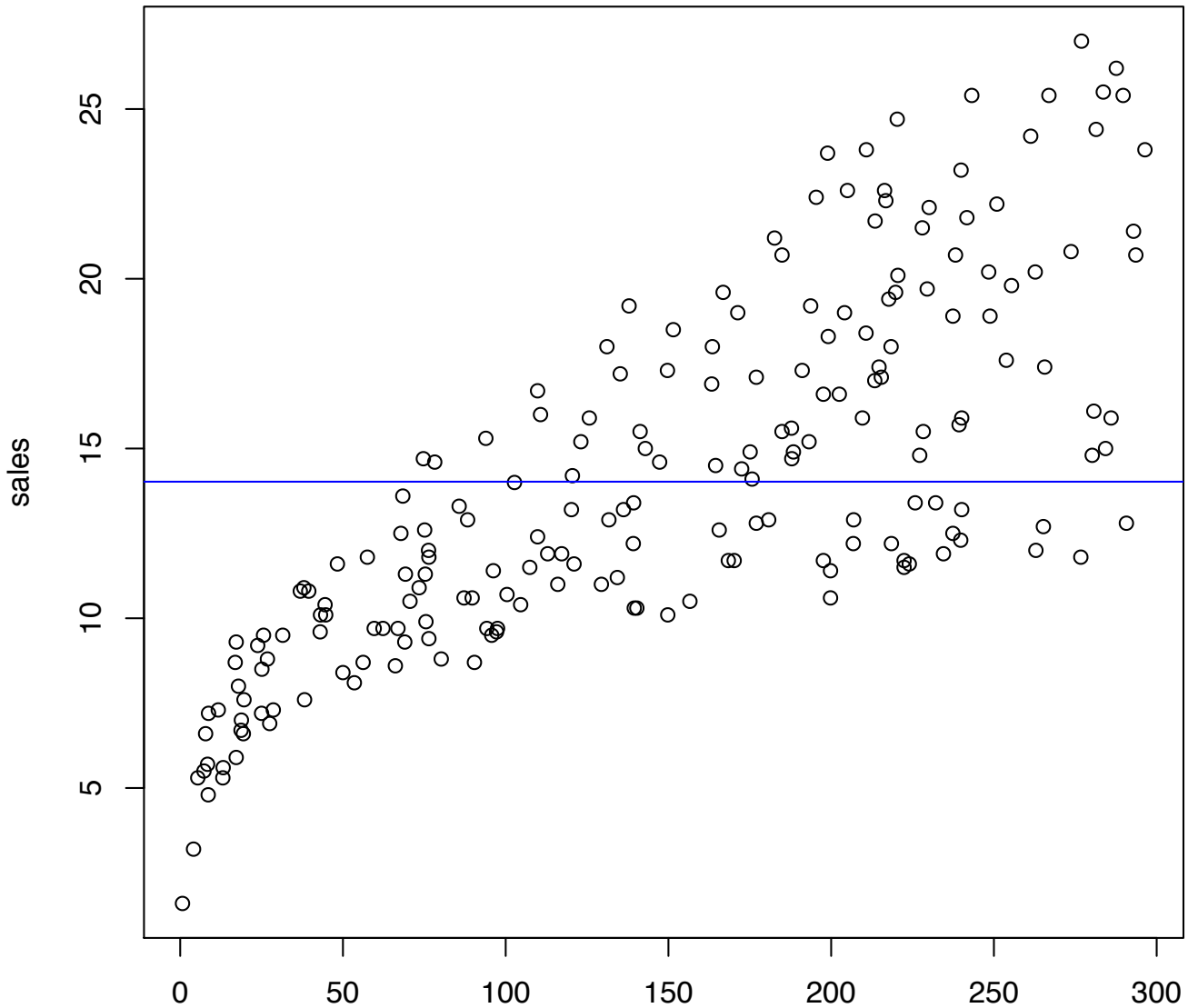
k-Nearest Neighbors: esempio

tutti i dati, k=1, 2, 10, 20, 30, 50

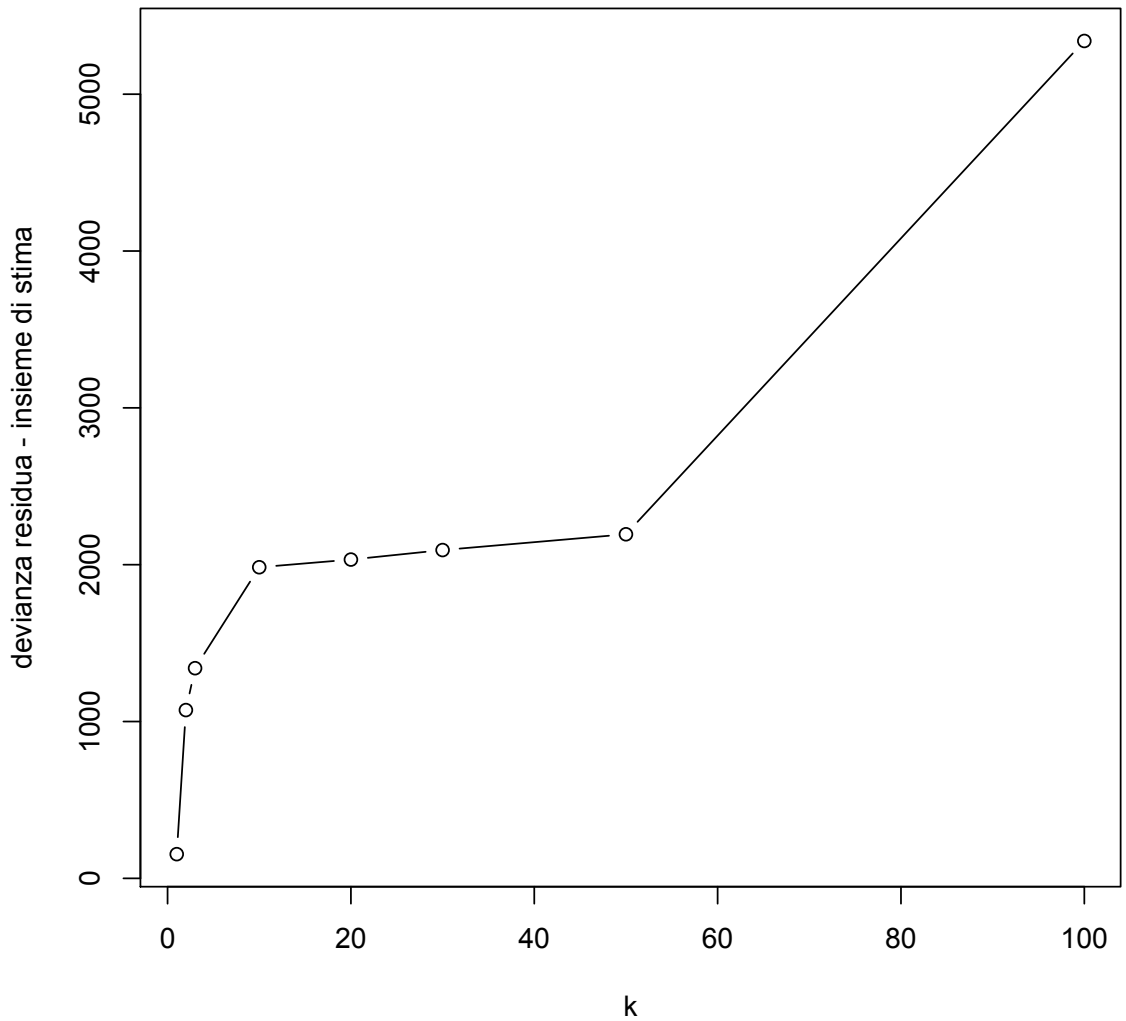


k-Nearest Neighbors: esempio

k=200



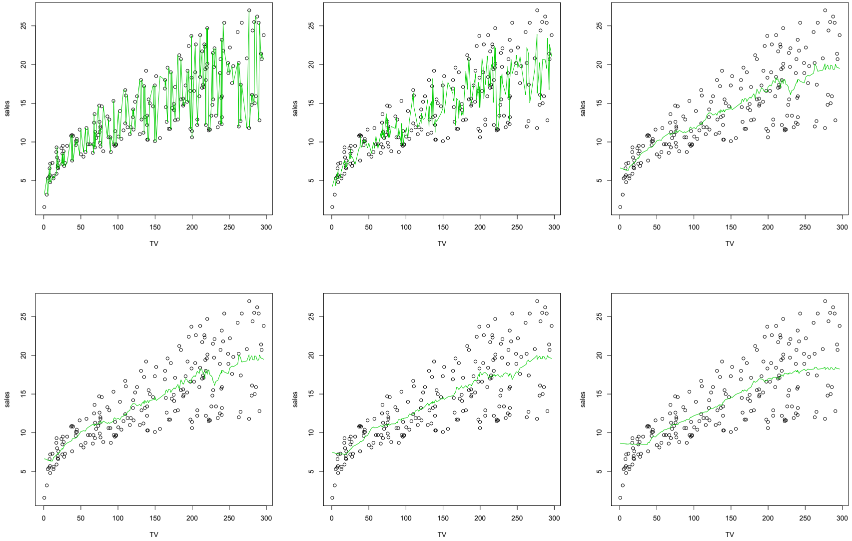
k-Nearest Neighbors: esempio



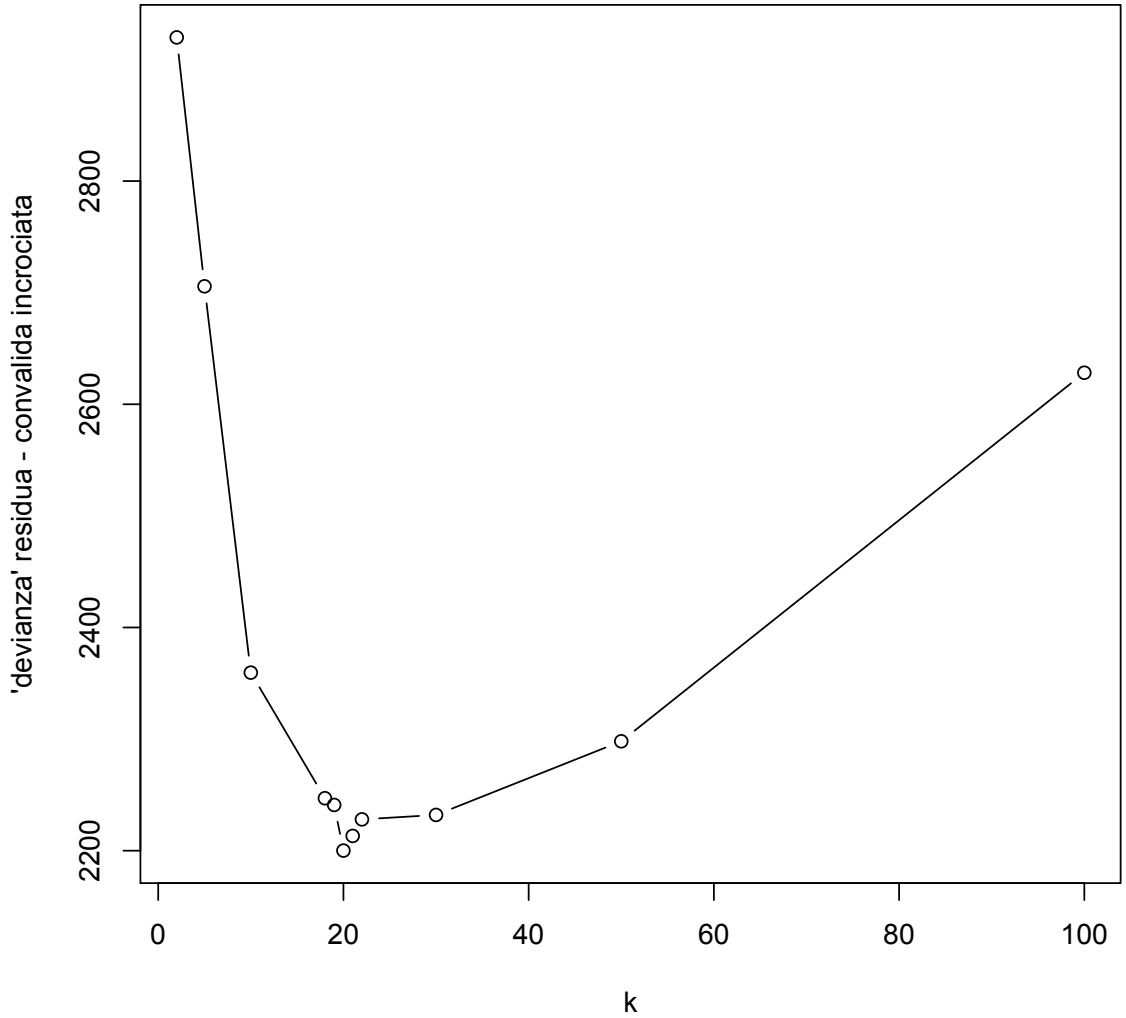
La performance del KNN peggiora al crescere di k

k-Nearest Neighbors: esempio

Convalida incrociata leave-one-out, k=1, 2, 10, 20, 30, 50



k-Nearest Neighbors: esempio

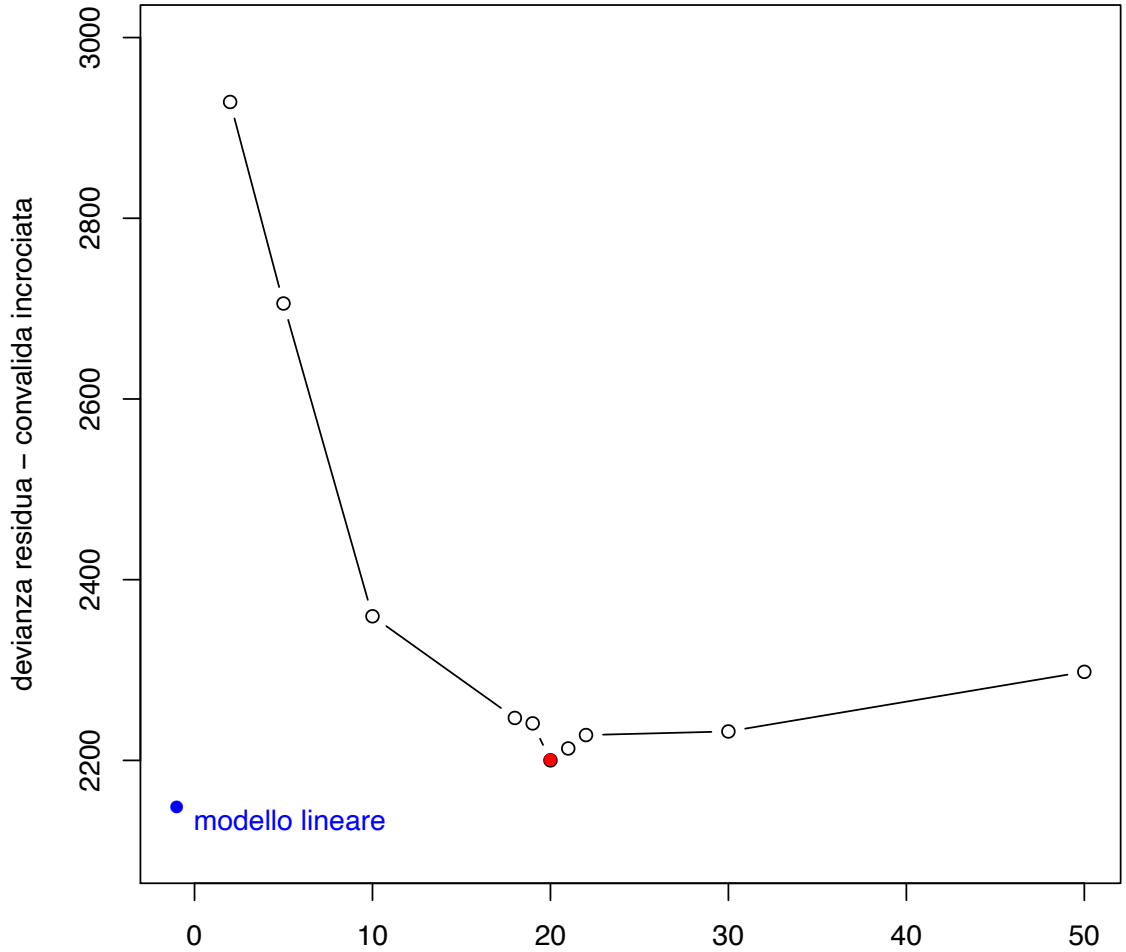


Esiste un minimo...

Compromesso tra varianza e distorsione

k-Nearest Neighbors: esempio

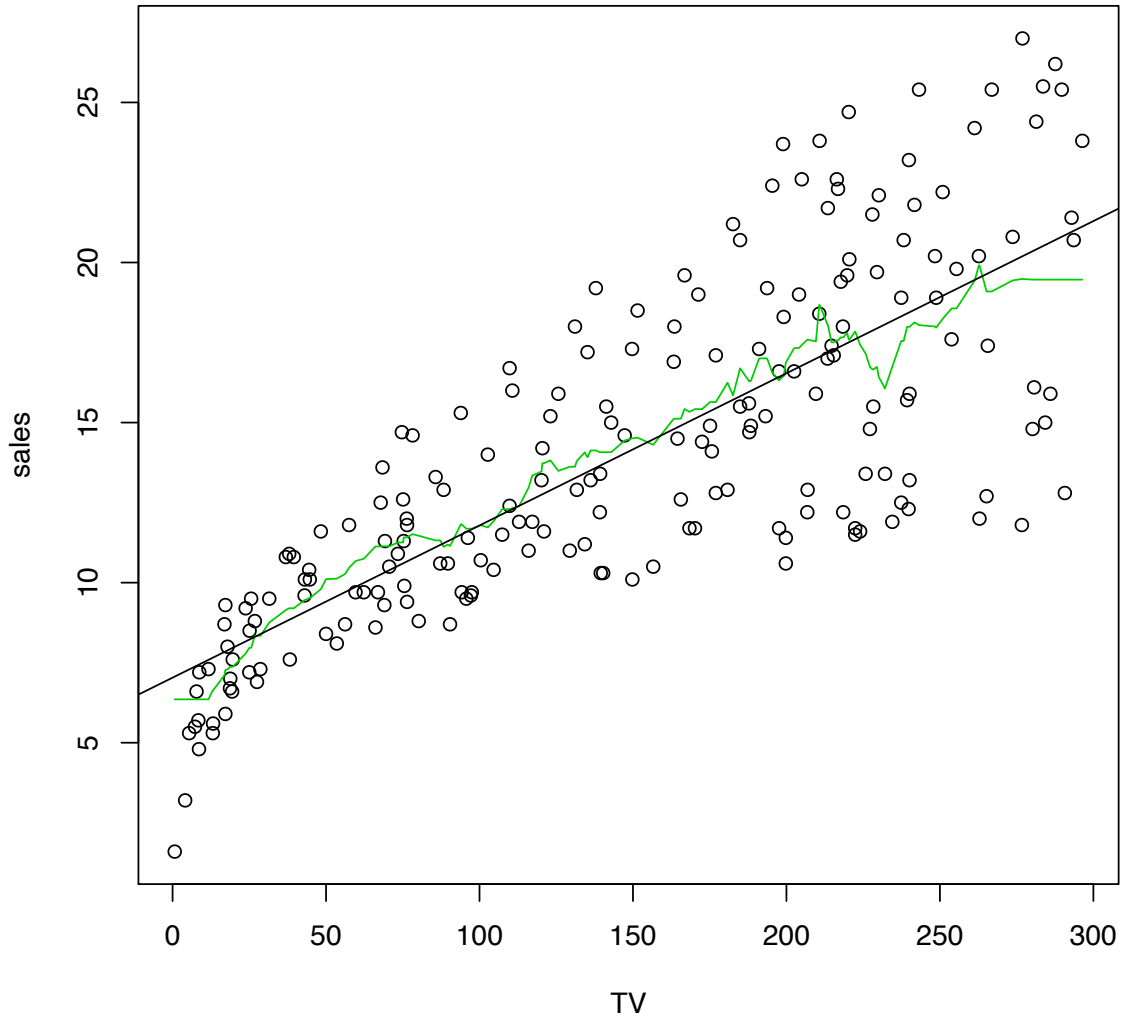
Performance di modello lineare e KNN con la sola variabile `tv`



Nel caso di `tv` il modello lineare ha una performance migliore del KNN per qualsiasi valore di k

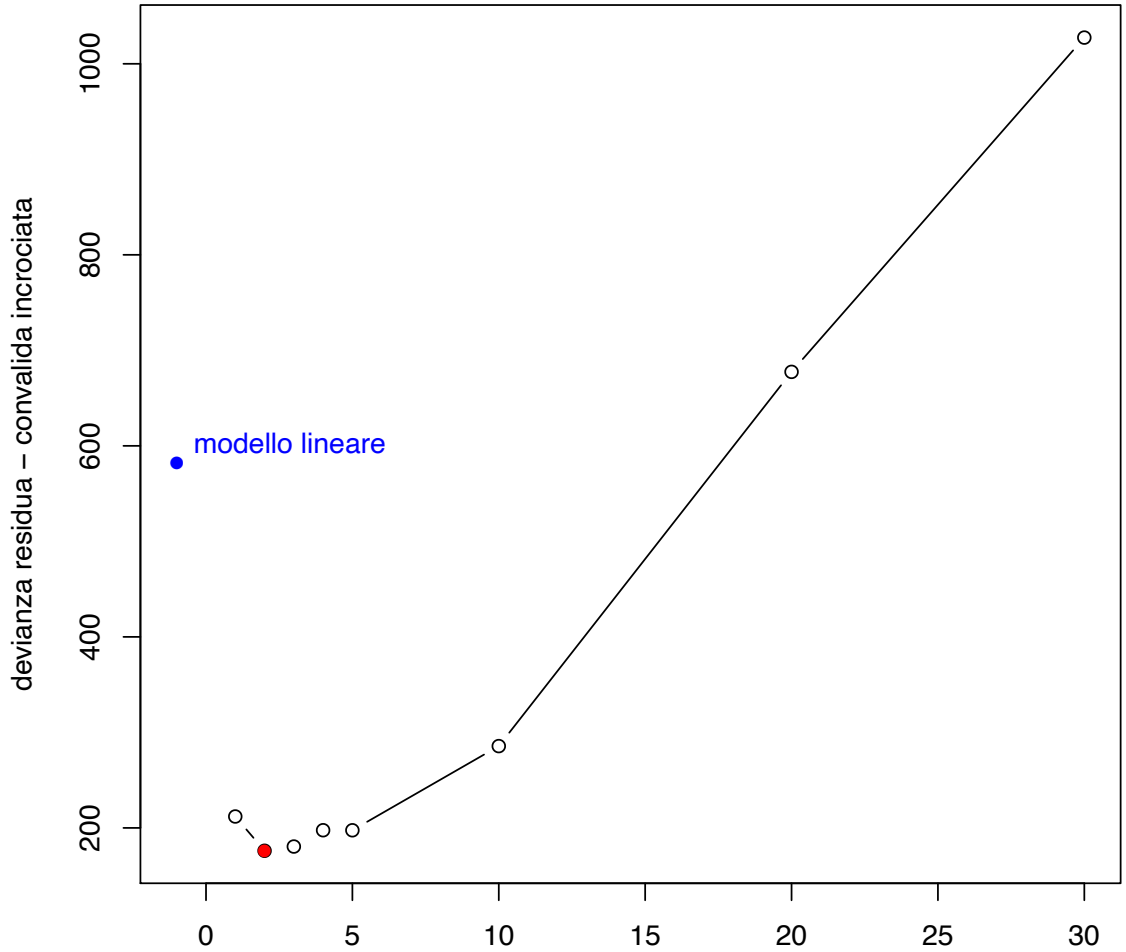
k-Nearest Neighbors: esempio

Modello lineare e KNN-20 con la sola variabile `tv`



k-Nearest Neighbors: esempio

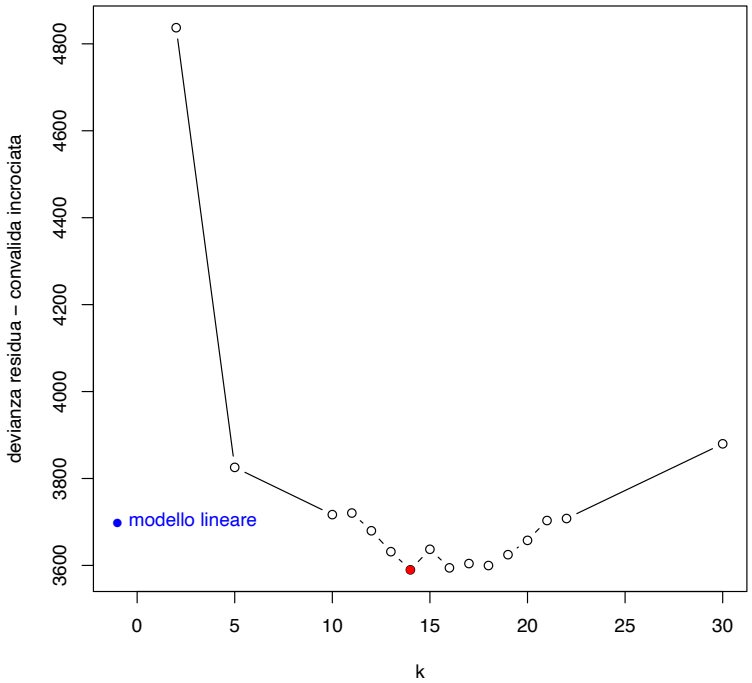
Performance di modello lineare e KNN con `tv` e `radio`

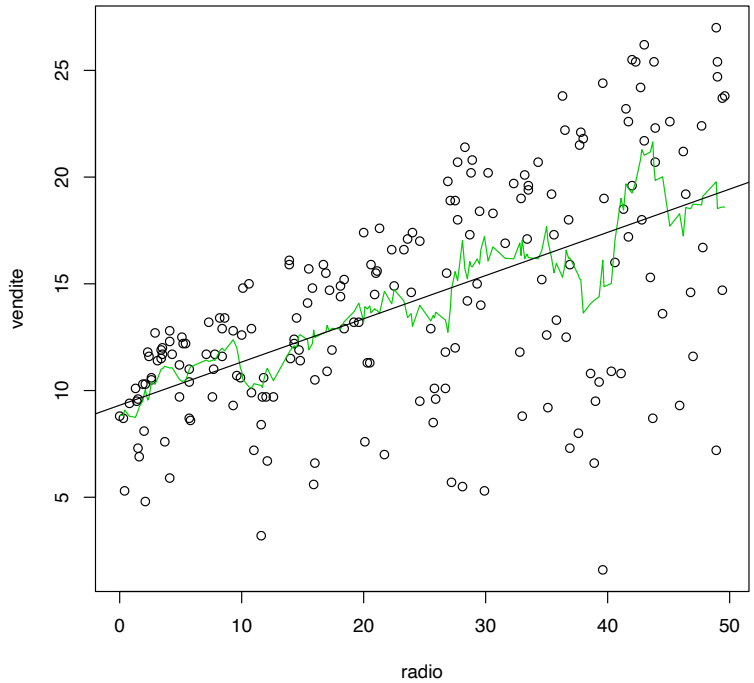


L'aggiunta della variabile `radio` fa migliorare considerevolmente la performance del KNN. Il minimo viene raggiunto per $k = 2$

k-Nearest Neighbors: esempio

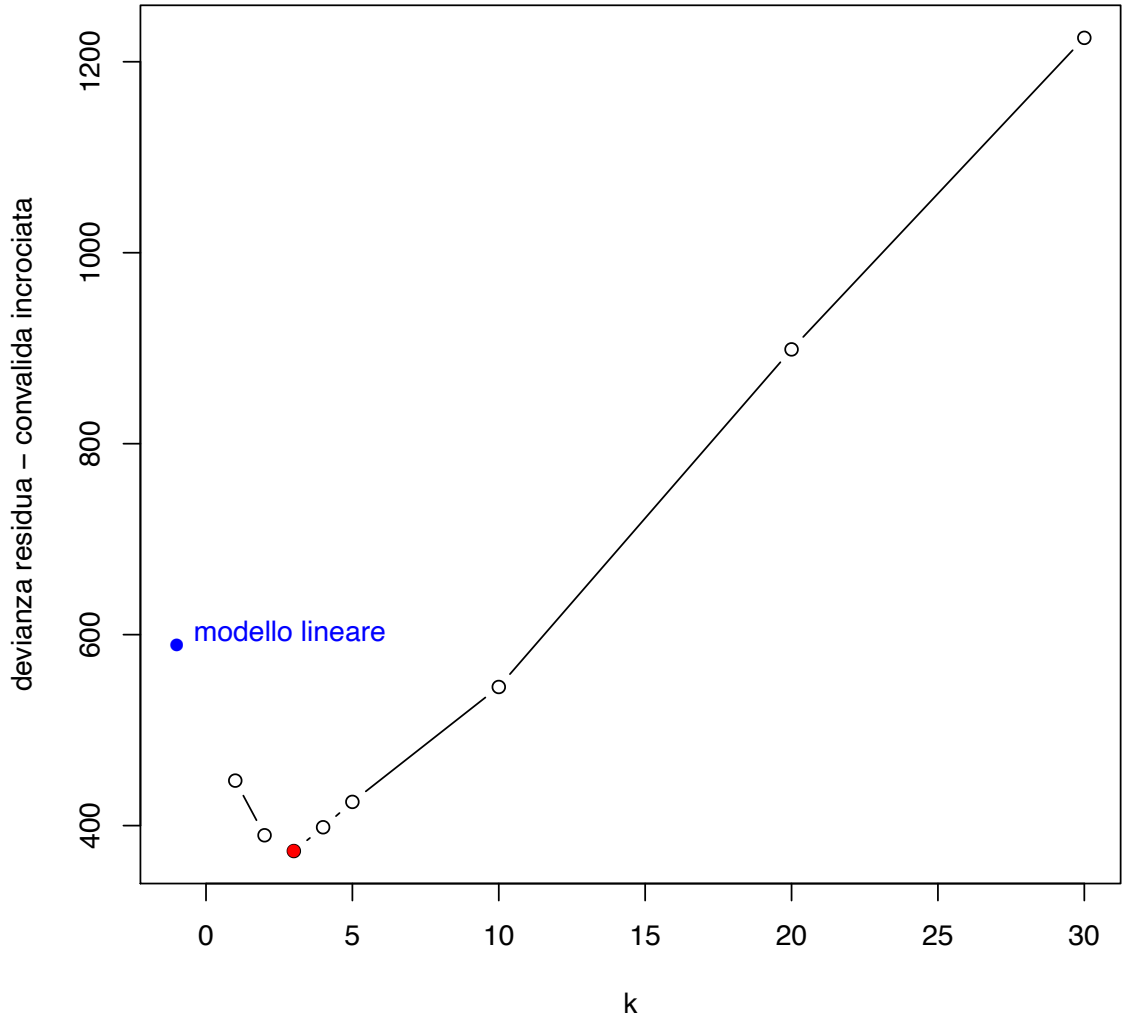
Performance di modello lineare e KNN con la sola variabile `radio`





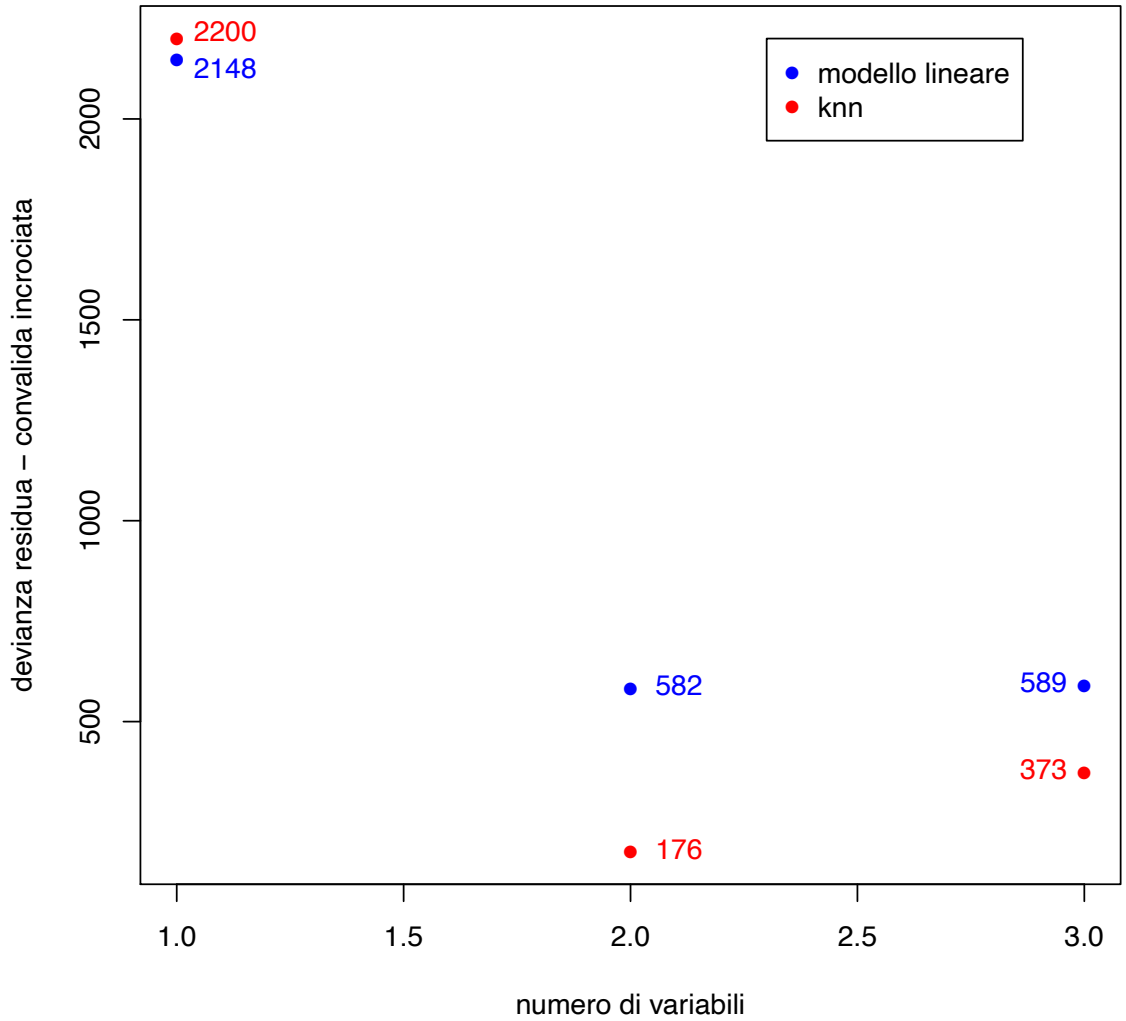
k-Nearest Neighbors: esempio

Aggiungiamo la variabile `giornali`



Il KNN risulta ad ogni modo migliore del modello lineare.

k-Nearest Neighbors: esempio



Classificazione

Nel caso della classificazione, la variabile da prevedere non è più numerica. Qui l'accuratezza del modello può essere valutata attraverso un tasso di errore del tipo

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

dove $\hat{y}_i$ è la classe prevista con il metodo di classificazione prescelto. $I(y_i \neq \hat{y}_i)$ è una variabile indicatrice che assume valore 1 quando $y_i \neq \hat{y}_i$ e 0 quando $y_i = \hat{y}_i$. In questo modo si ha una misura delle classificazioni errate calcolate sull'insieme di stima.

Classificazione

Se l'interesse è invece misurare l'errore su un set di osservazioni del tipo (x_0, y_0) possiamo considerare, un *test error*

$$Ave(I(y_0 \neq \hat{y}_0))$$

dove $\hat{y}_0$ è la previsione ottenuta applicando il metodo di classificazione.
Un buon metodo di classificazione è quello per cui tale misura risulta la più piccola possibile.

Classificazione

- Si ricorda che strumenti tipici per la valutazione di un metodo di classificazione sono la matrice di confusione e le misure di accuratezza derivanti (sensibilità e specificità di un modello).
- Inoltre, la probabilità di appartenere a una certa classe è legata alla definizione di una soglia (tipicamente 0.5)
- Tale valore può comunque essere aggiustato in modo da accettare più errate classificazioni laddove risulta meno costoso.

Classificazione

È possibile dimostrare che l'errore $Ave(I(y_0 \neq \hat{y}_0))$ viene minimizzato, in media, da un classificatore che assegna *ogni osservazione alla classe più probabile, dati certi valori dei predittori*. Assegneremo un'osservazione con predittore x_0 alla classe j per la quale

$$Pr(Y = j|X = x_0)$$

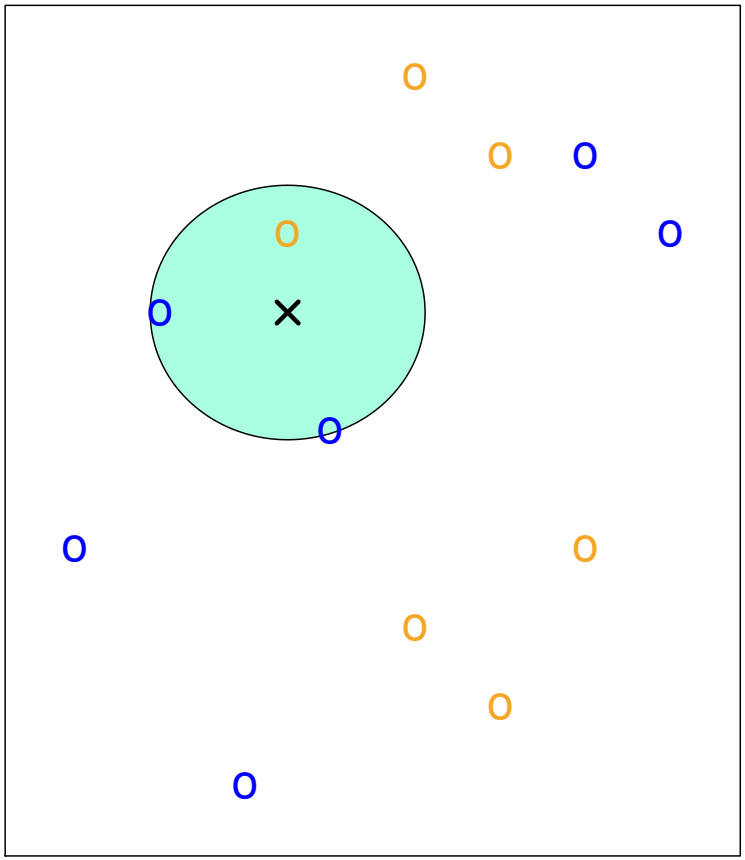
è la più alta. Da notare che tale probabilità è condizionata al valore di x_0. Questo semplice metodo viene denominato classificatore di Bayes.

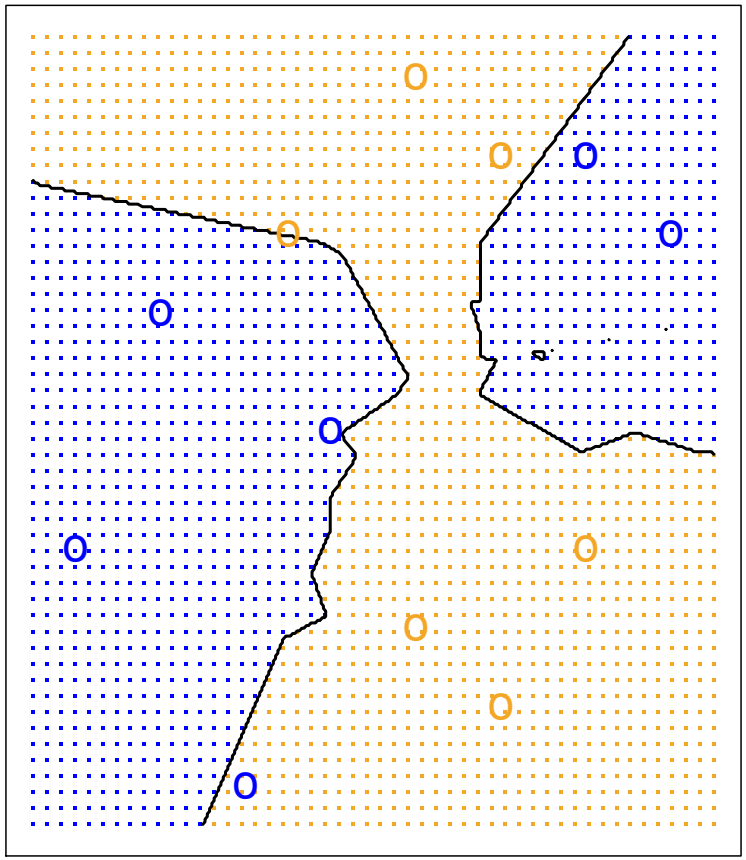
k-Nearest Neighbors: classificazione

- In teoria vorremmo sempre poter utilizzare il classificatore di Bayes, ma con dati reali non conosciamo la distribuzione condizionata di Y dato X e quindi la sua applicazione diviene impossibile.
- Tra i metodi alternativi troviamo il metodo k-Nearest Neighbors.
- Dato un k positivo e un'osservazione x_0, il classificatore KNN identifica i k punti più vicini a x_0 nel training set, rappresentati da N_0.
- Stima poi la probabilità condizionata per la classe j come la frazione di punti in N_0 la cui risposta è uguale a j

$$Pr(Y = j|X = x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j)$$

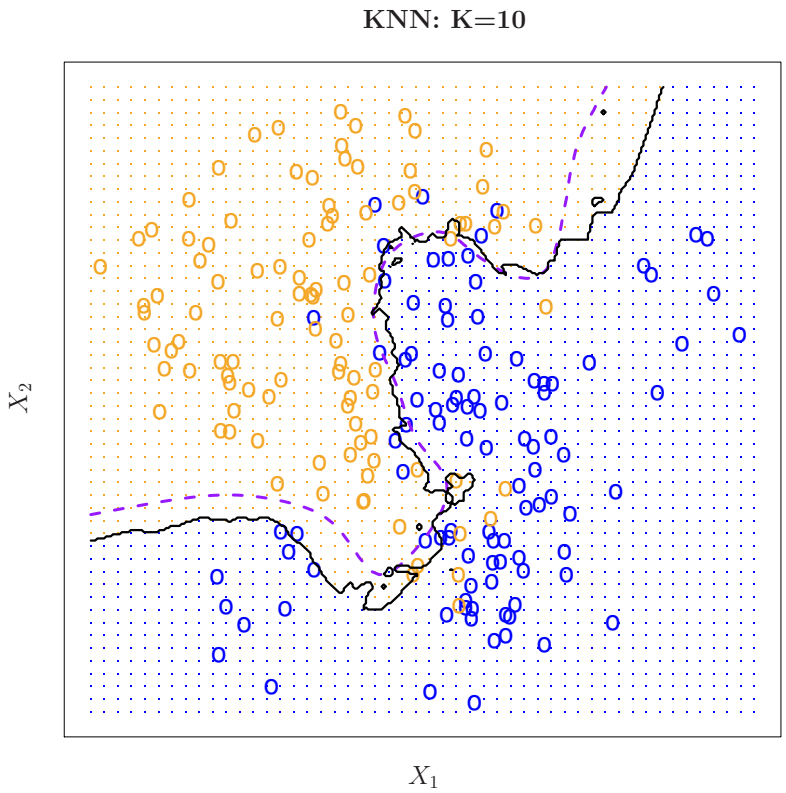
KNN: classificazione





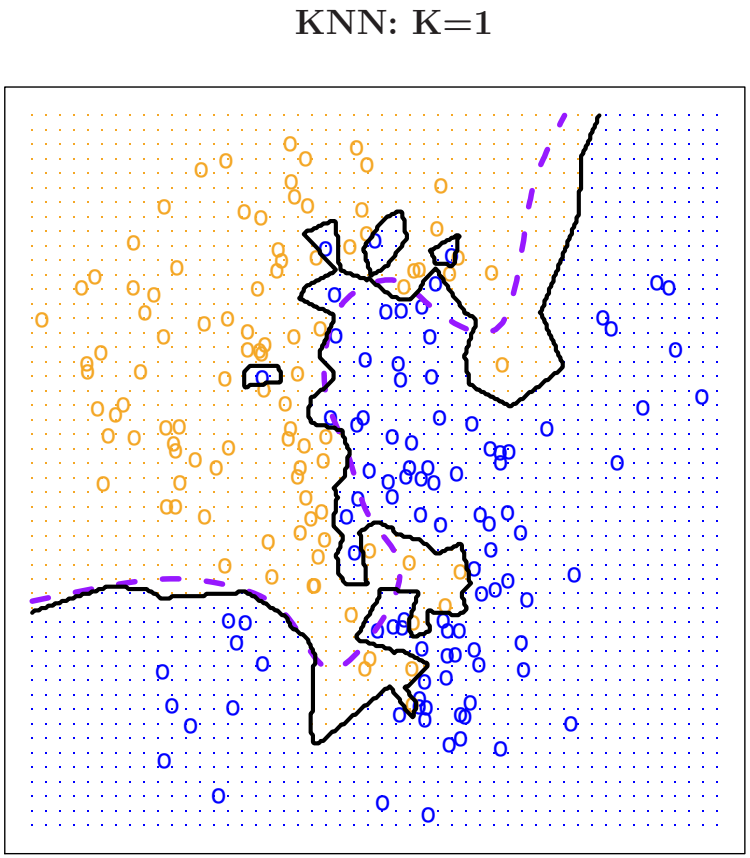
Esempio di KNN con $k = 3$ e frontiera di decisione.

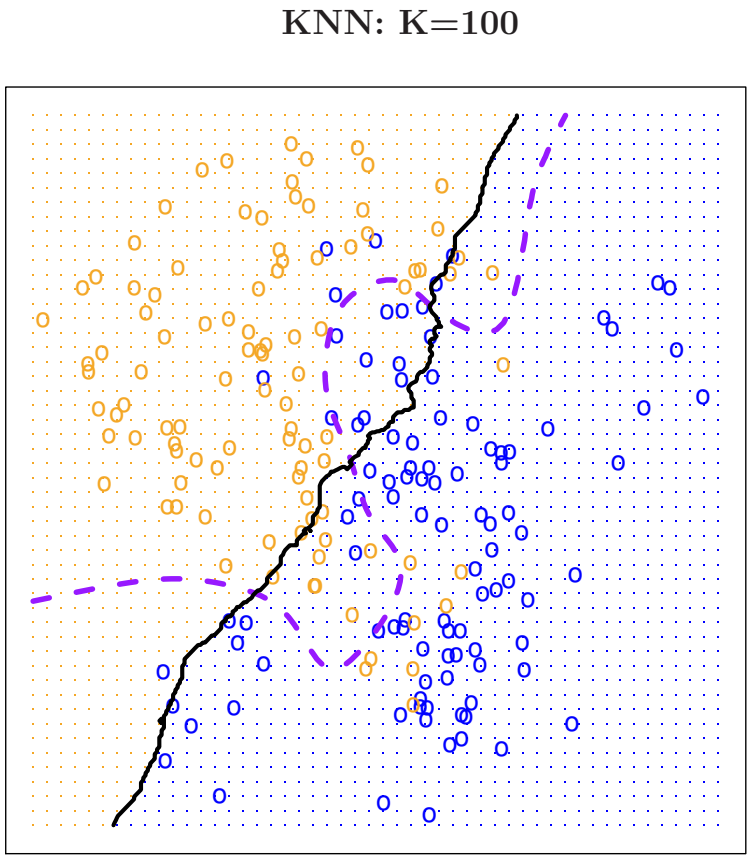
KNN: classificazione



Frontiera di decisione per $k = 10$. In questo caso la frontiera è molto simile a quella del classificatore 'di Bayes' (linea tratteggiata)

KNN: classificazione

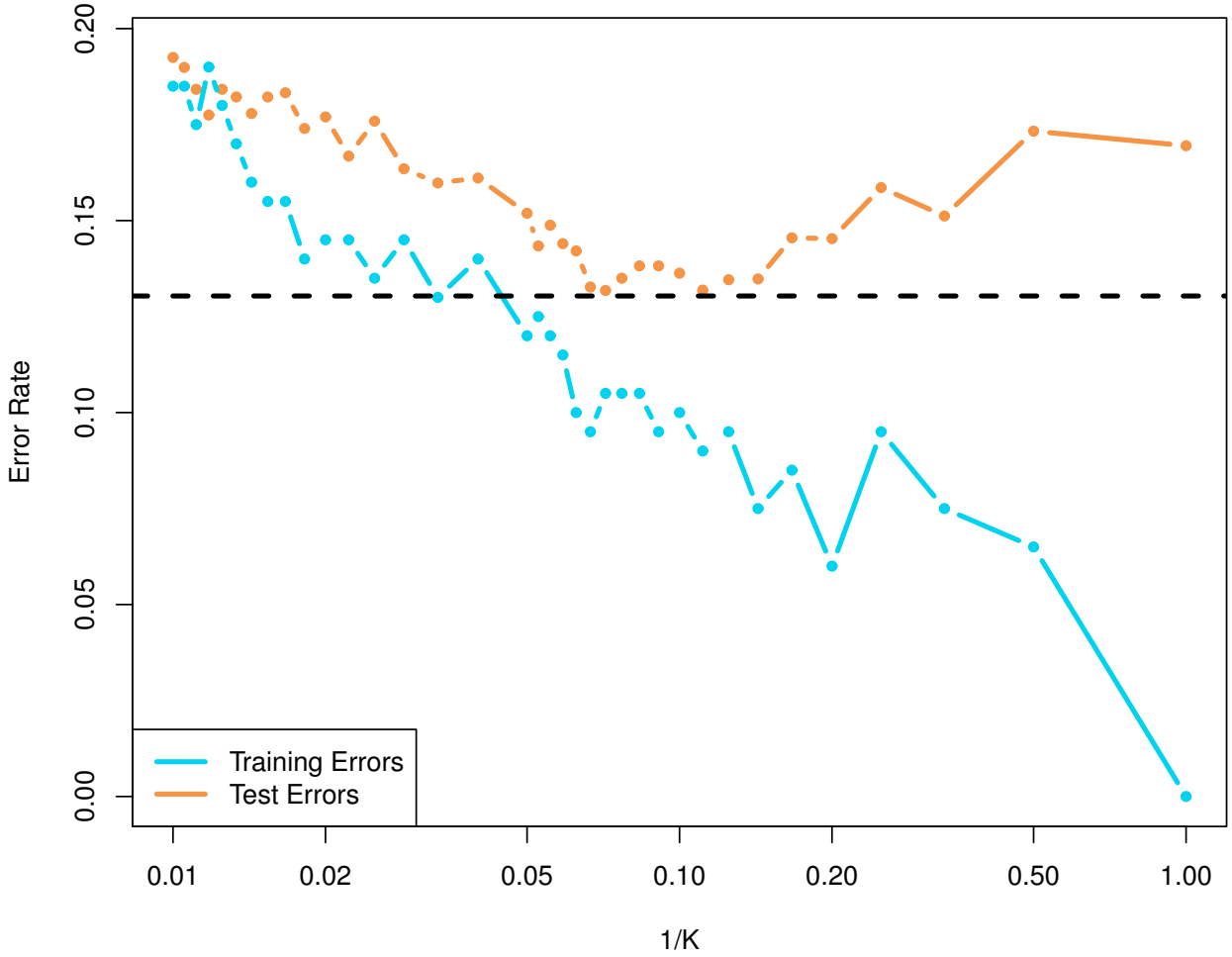




Frontiera di decisione per $k = 1$ e $k = 100$.

Nel primo caso è troppo flessibile, nel secondo è troppo rigida.

KNN: classificazione



Training e test error al crescere della flessibilità del modello $(1/k)$

k-Nearest Neighbors: classificazione

Il metodo KNN non fa assunzioni sulla relazione tra l'appartenenza a una classe Y e i predittori $X_1, X_2, ..., X_p$.
Piuttosto, apprende informazioni dalle similarità tra i valori dei predittori nel dataset.

Un aspetto centrale riguarda quindi la *misurazione della distanza* tra osservazioni basata sui predittori.
Una scelta frequente e comoda è quella di utilizzare la distanza euclidea.

k-Nearest Neighbors: classificazione

Dopo aver calcolato le distanze tra l'osservazione da classificare e le osservazioni presenti nel training set, è necessario stabilire una regola di classificazione, basata sulle classi di appartenenza delle osservazioni simili (neighbors).
Il caso più semplice è $k = 1$, nel quale si utilizza l'unica osservazione più vicina, la cui classe di appartenenza viene assegnata all'osservazione da classificare.
La stesso principio può essere ovviamente esteso al caso $k > 1$ in questo modo:

- trovare i KNN (vicini più vicini) per l'osservazione da classificare
- usare una regola di decisione per cui l'osservazione è assegnata alla classe prevalente fra i k più vicini.

KNN: esempio

Income	Lot_Size	Ownership
60.0	18.4	owner
85.5	16.8	owner
64.8	21.6	owner
61.5	20.8	owner
87.0	23.6	owner
110.1	19.2	owner
108.0	17.6	owner
82.8	22.4	owner
69.0	20.0	owner
93.0	20.8	owner
51.0	22.0	owner
81.0	20.0	owner
75.0	19.6	non-owner
52.8	20.8	non-owner
64.8	17.2	non-owner
43.2	20.4	non-owner
84.0	17.6	non-owner
49.2	17.6	non-owner
59.4	16.0	non-owner
66.0	18.4	non-owner
47.4	16.4	non-owner
33.0	18.8	non-owner
51.0	14.0	non-owner
63.0	14.8	non-owner

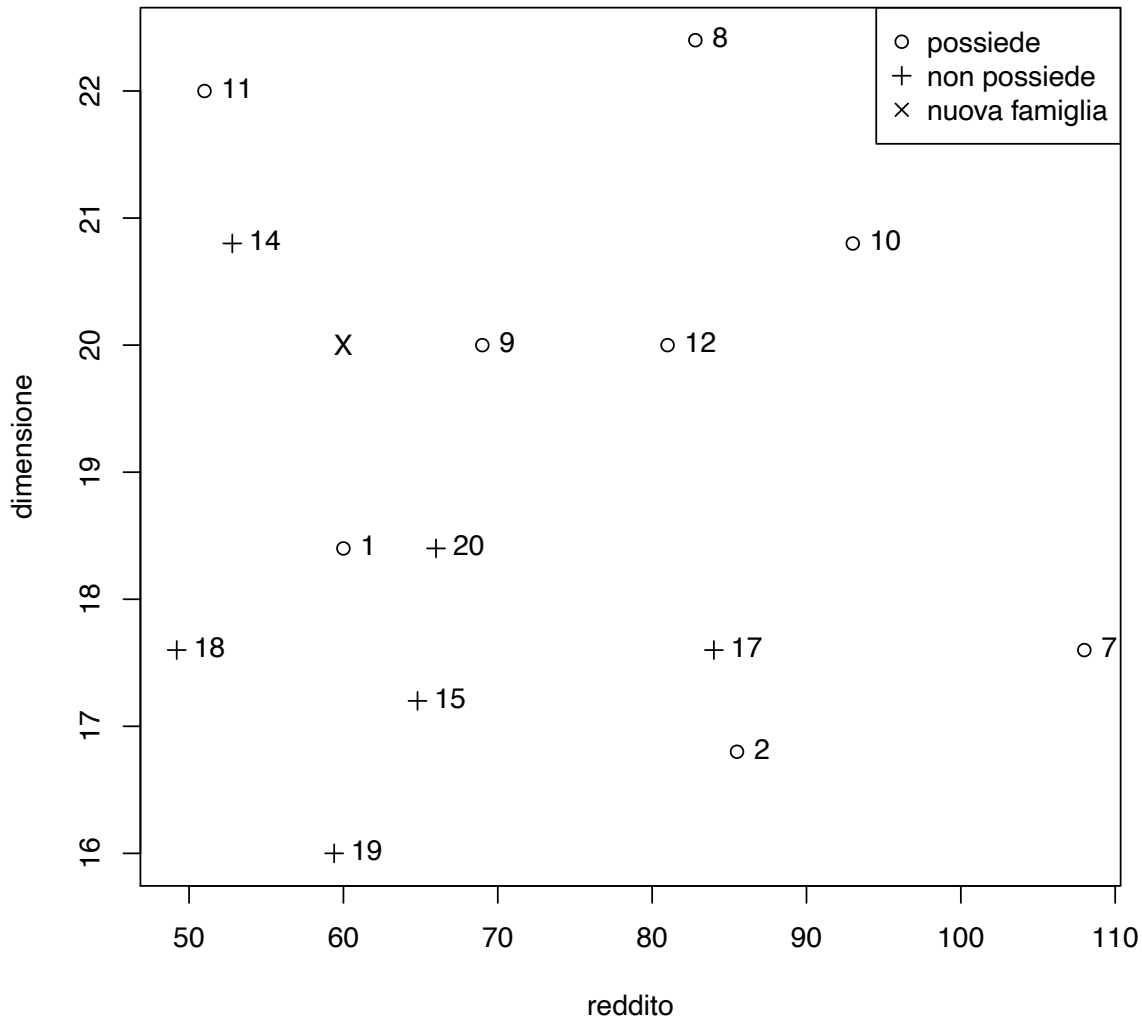
KNN: esempio

Un'azienda che produce tagliaerba vuole classificare famiglie potenzialmente clienti.
Viene selezionato un campione pilota di 12 famiglie che hanno già il tagliaerba (owner) e 12 che non ce l'hanno (nonowner).
Il dataset viene poi suddiviso in training set (14 osservazioni) e test set (10).
I predittori sono il reddito (income) e la dimensione del terreno posseduto dalla famiglia (lot size).

Nuova osservazione: una famiglia con 60000 $ di reddito e dimensione del terreno $20000 ft^2$.
Fra tutte le osservazioni del training set, la più vicina (come distanza Euclidea) è la 9.

- per $K = 1 \rightarrow 1NN = 9 \rightarrow$ owner
- per $K = 3 \rightarrow 3NN = 9, 14, 1 \rightarrow$ owner

KNN: esempio

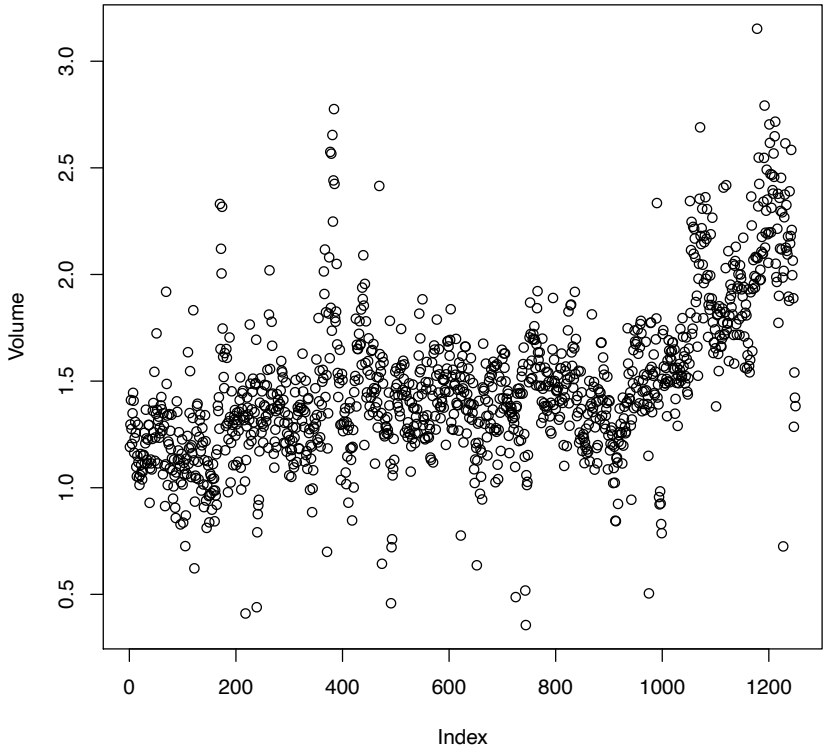

possiede
non possiede
nuova famiglia
dimensione
reddito
16
17
18
19
20
21
22
50
60
70
80
90
100
110
8
11
14
10
X
9
12
1
20
18
17
7
15
2
19

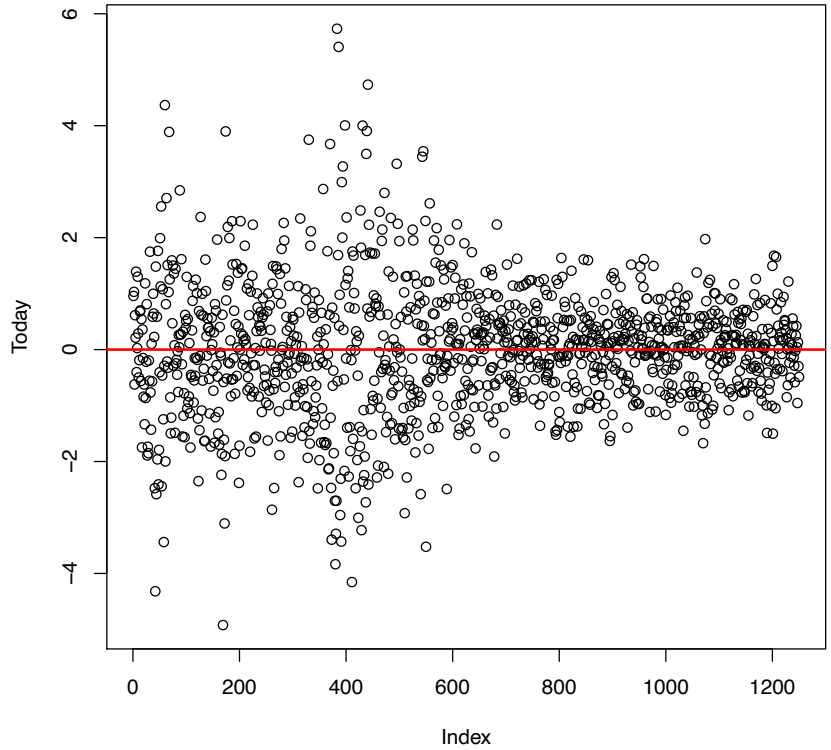
KNN: esempio

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
1	2001.00	0.38	-0.19	-2.62	-1.05	5.01	1.19	0.96	Up
2	2001.00	0.96	0.38	-0.19	-2.62	-1.05	1.30	1.03	Up
3	2001.00	1.03	0.96	0.38	-0.19	-2.62	1.41	-0.62	Down
4	2001.00	-0.62	1.03	0.96	0.38	-0.19	1.28	0.61	Up
5	2001.00	0.61	-0.62	1.03	0.96	0.38	1.21	0.21	Up
6	2001.00	0.21	0.61	-0.62	1.03	0.96	1.35	1.39	Up
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Rendimenti di indice S&P per 1250 giorni, dal 2001 al 2005.
Si intende stabilire se il rendimento sale o scende (Up, Down) in funzione dell'andamento del mercato
Usiamo solo i rendimenti dei due giorni precedenti (Lag1, Lag2)
$Down : 602 \qquad Up : 648$

KNN: esempio





Divido i dati in

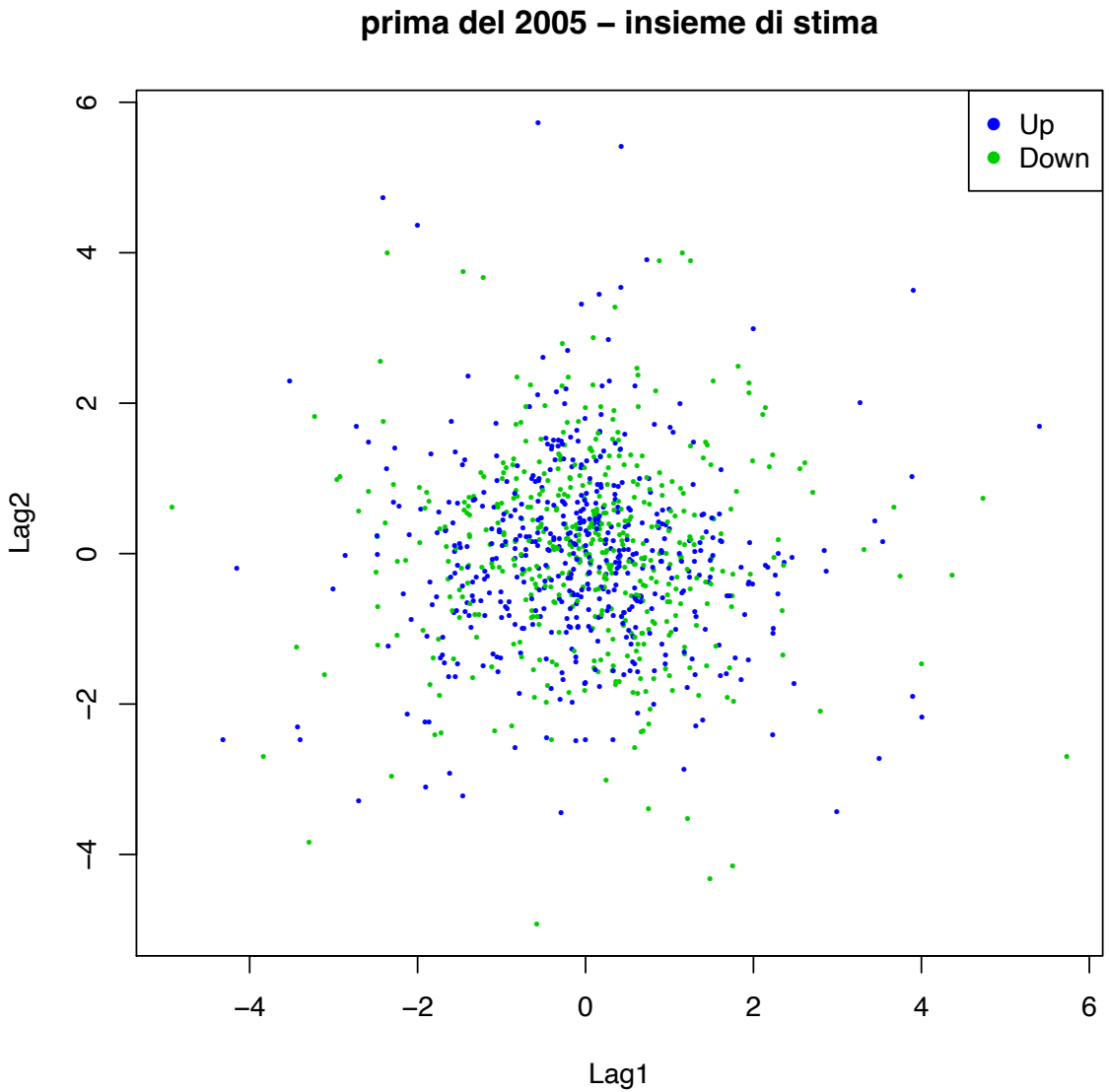
insieme di stima: dal 2001 al 2004

insieme di verifica: dal 2005

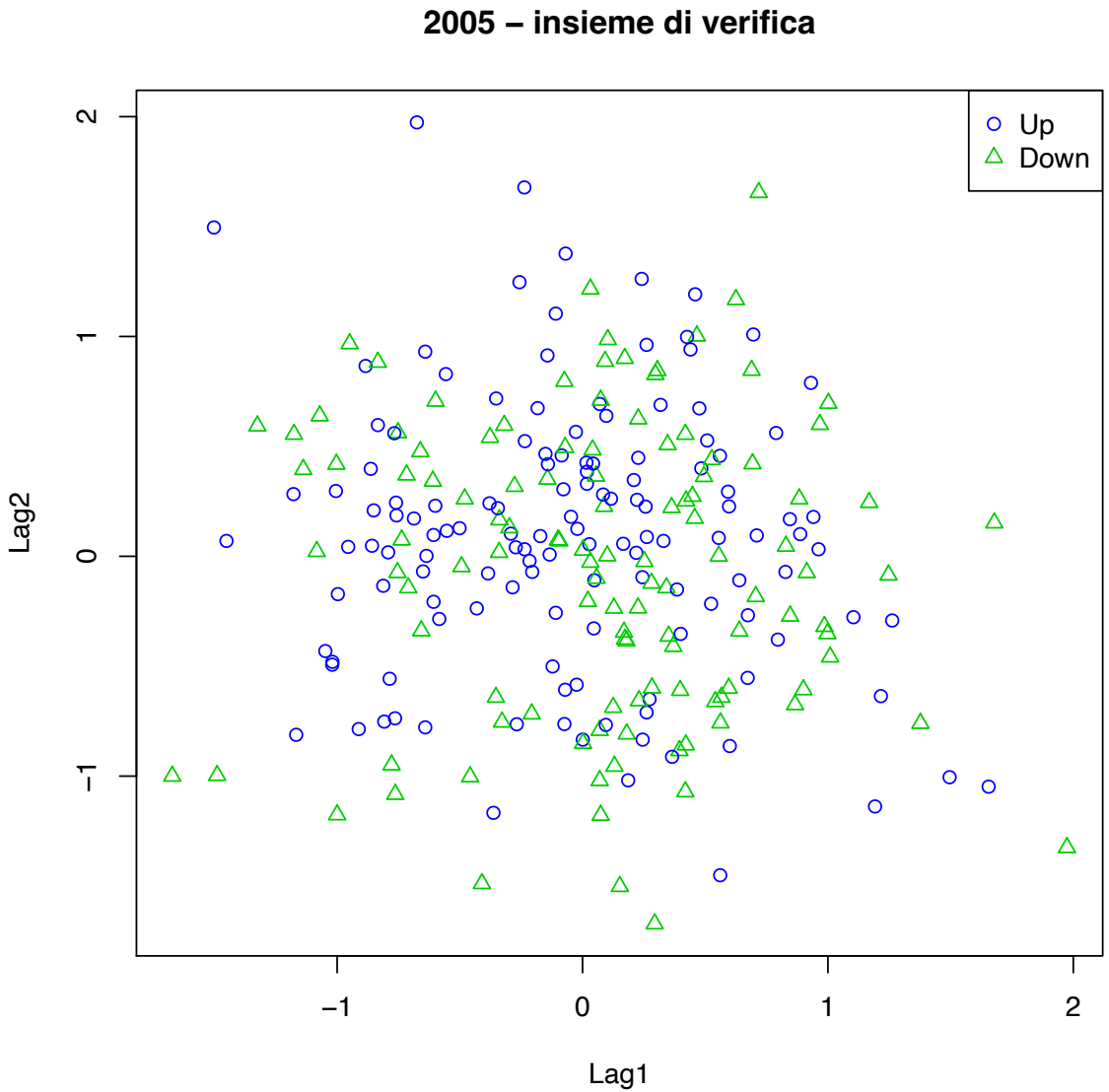
Nota: *non* abbiamo diviso in maniera casuale!

$\rightarrow$ serie storica. . . interessa prevedere il futuro. . .

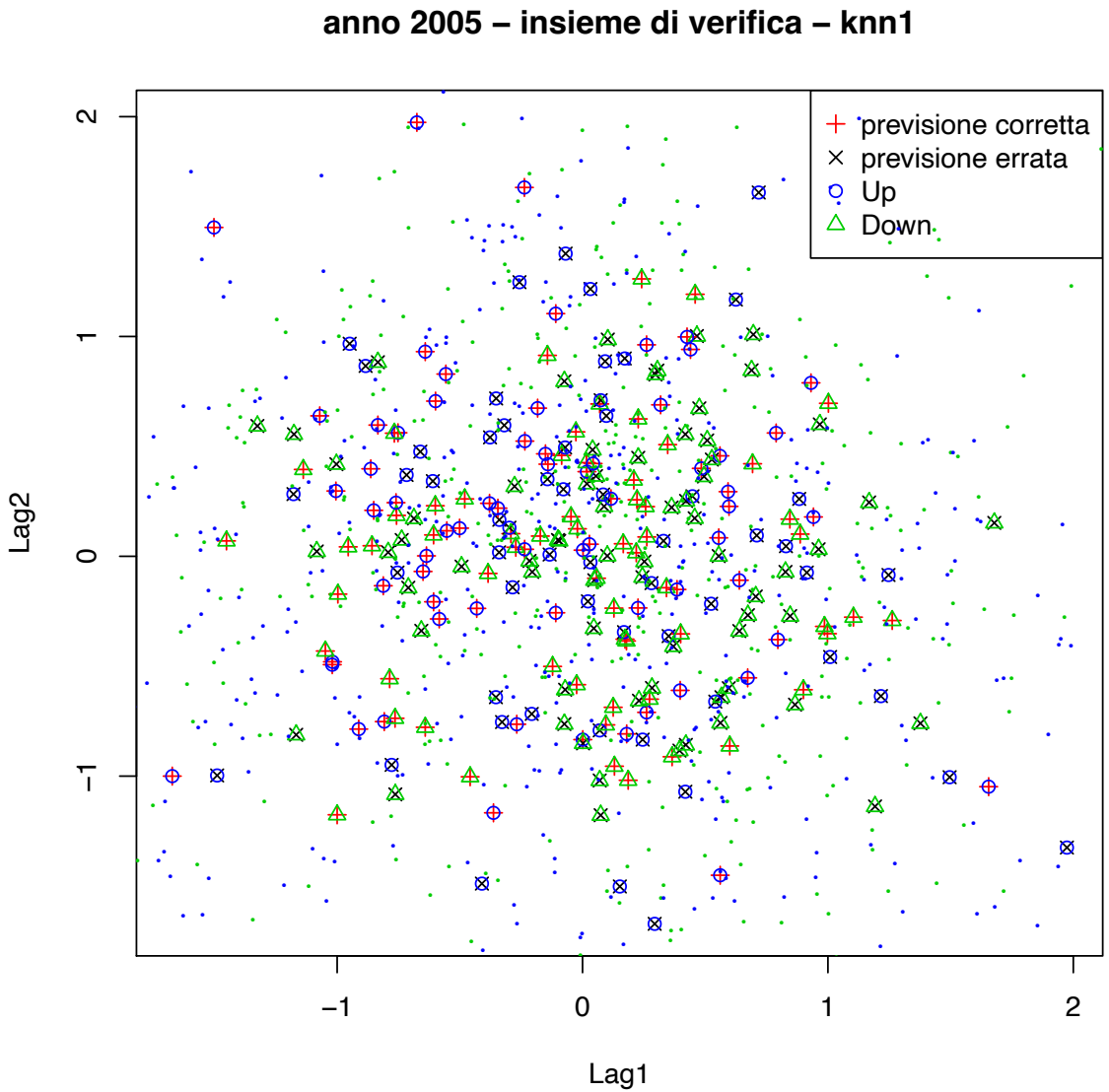
KNN: esempio


prima del 2005 – insieme di stima
Up
Down
Lag2
Lag1
6
4
2
0
−2
−4
−4
−2
0
2
4
6

KNN: esempio



KNN: esempio


anno 2005 – insieme di verifica – knn1
previsione corretta
previsione errata
Up
Down
Lag2
Lag1
2
1
0
−1
−1
0
1
2

KNN: esempio

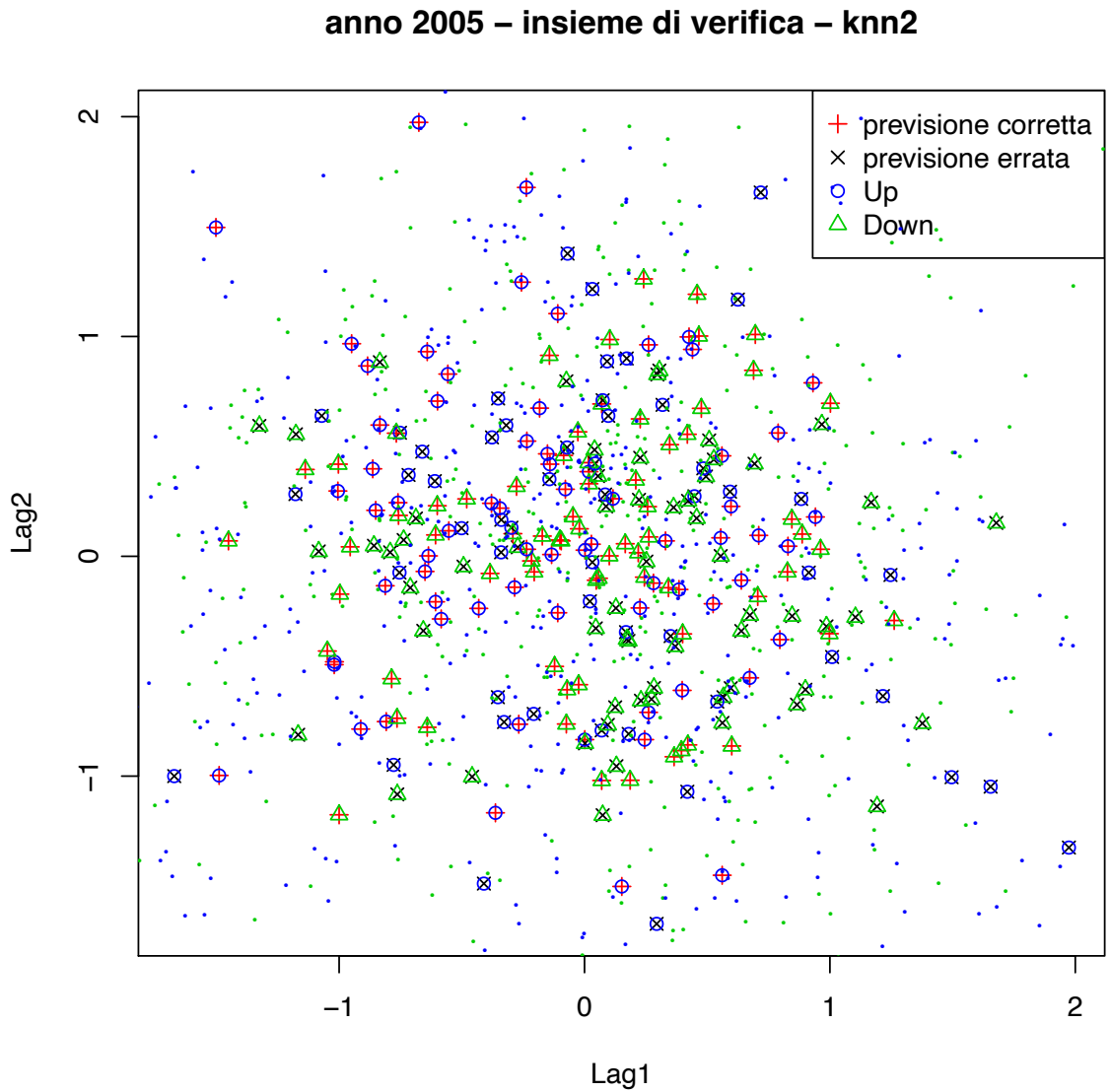
KNN con $k = 1$

	Down	Up
Down	43	58
Up	68	83

Tasso di corretta classificazione: $\frac{43 + 83}{43 + 58 + 68 + 83} = 50\%$

Nota: Tasso di corretta classificazione$= \frac{1}{n} \sum_{i=1}^{n} I(y_i = \hat{y}_i)$

KNN: esempio


anno 2005 – insieme di verifica – knn2
previsione corretta
previsione errata
Up
Down
Lag2
Lag1
2
1
0
−1
−1
0
1
2

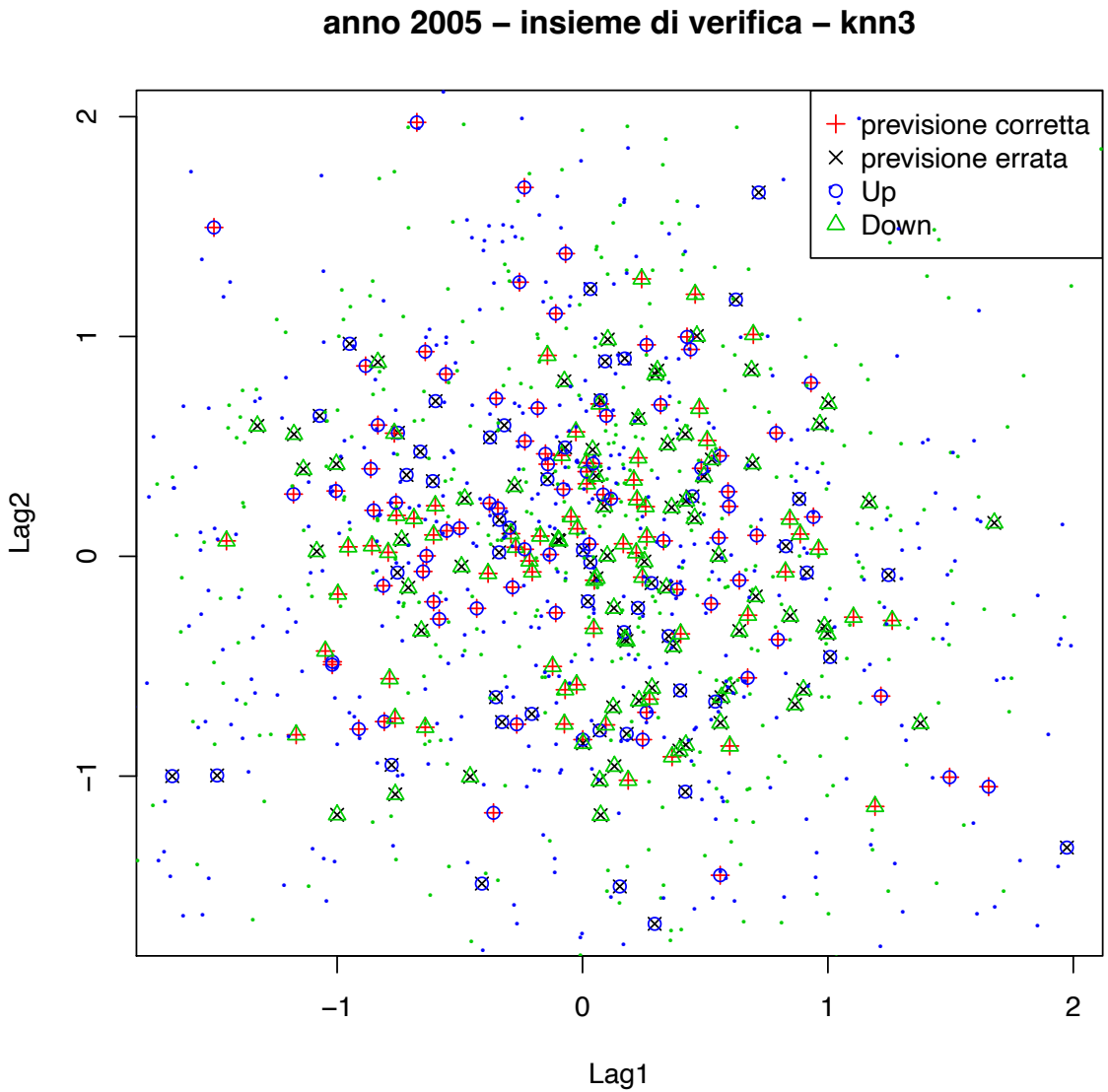
KNN: esempio

KNN con $k = 2$

	Down	Up
Down	52	54
Up	59	87

Tasso di corretta classificazione: 55.15%

KNN: esempio


anno 2005 – insieme di verifica – knn3
previsione corretta
previsione errata
Up
Down
Lag2
Lag1
2
1
0
−1
−1
0
1
2

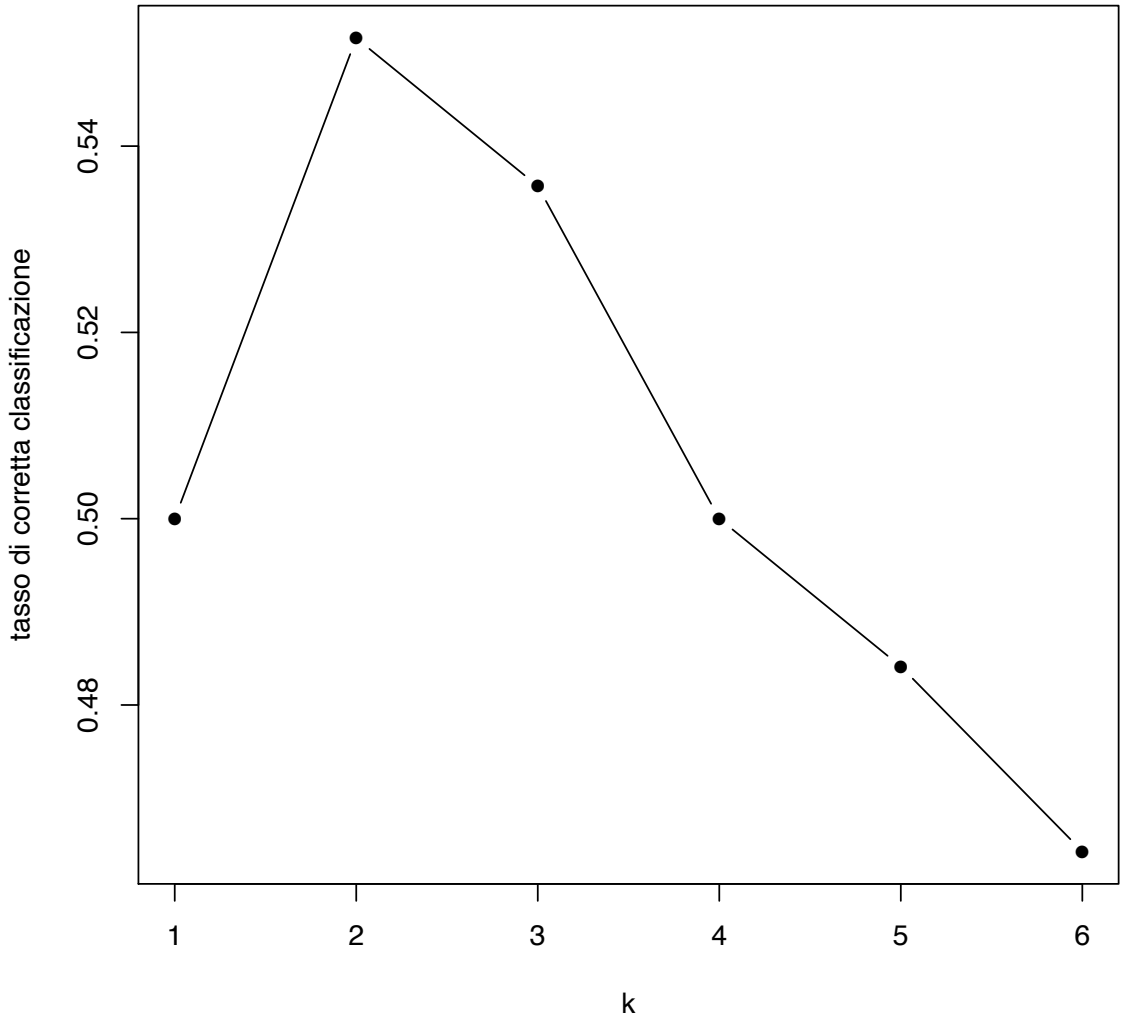
KNN: esempio

KNN con $k = 3$

	Down	Up
Down	48	54
Up	63	87

Tasso di corretta classificazione: 53.57%

KNN: esempio



KNN: esempio