Una breve introduzione all'Apprendimento Statistico

(o Machine Learning)

N. Torelli

Settembre 2024

Università di Trieste - Dipartimento di Scienze Economiche Aziendali Matematiche e Statistiche "Bruno de Finetti"

Introduzione: dai modelli statistici all'apprendimento statistico

Alcuni esempi classici in cui si fa ricorso a modelli statistici (in ambito aziendale o assicurativo)

- Valutare la probabilità che un cliente compri un prodotto avendo osservato alcune caratteristiche socio-demografiche
- Classificare i clienti in “fedeli” e “infedeli” sulla base della storia delle caratteristiche e del loro rapporto con l’azienda
- Prevedere la spesa aggiuntiva per l’upgrade di un prodotto o di un servizio
- In ambito assicurativo
    - prevedere il numero di sinistri per classi di assicurati in base alle loro caratteristiche
    - Valutare quali caratteristiche abbiano gli assicurati (o i sinistri) che potrebbero ricorrere a una frode
    - Prevedere la spesa per l’assicurazione di persone di una famiglia

Modelli statistici e apprendimento statistico

- A illustrazione del graduale spostamento dell'interesse per le tecniche di apprendimento statistico (o machine learning) partiremo dall'esempio dell'uso dei modelli statistici in ambito assicurativo. In questo caso, ad esempio, la necessità di classificare un portafoglio di clienti in base al loro profilo di rischio ha condotto all'impiego di modelli di regressione o a loro generalizzazioni
- I modelli di regressione consentono di prevedere una variabile risposta in funzione di fattori di rischio o variabili esplicative note. Ad esempio, valutare la probabilità di un sinistro tenendo conto delle caratteristiche degli assicurati.
- Le informazioni disponibili per valutare i profili di rischio degli assicurati, e quindi prevedere sinistri e perdite connesse, sono diventate col tempo sempre più ricche (fino a caratterizzarsi recentemente come big data)
- Tuttavia le assunzioni standard dei modelli statistici impiegati sono talvolta troppo rigide e poco realistiche
- I modelli statistici hanno quindi dovuto adattarsi e sono divenuti sempre più complessi
- Sono emerse, più recentemente, le nuove tecniche di apprendimento statistico (AS) (o di machine learning - ML) che forniscono una risposta alternativa e a volte più adatta alle nuove esigenze.
- In particolare questo è vero per i cosiddetti problemi di **AS supervisionato** (che sono quelli dei quali tratteremo in questo corso).

Dai modelli statistici all'apprendimento statistico

Obiettivo dei modelli statistici

Gli esempi citati possono essere inquadrati in uno stesso schema che coinvolge:

- la definizione di un fenomeno di interesse: la variabile risposta (variabile dipendente, output, target);
- la disponibilità di informazioni su uno o più variabili concomitanti o esplicative (covariate, caratteristiche o features, inputs, predittori), che si presumono legati al fenomeno di interesse.

Obiettivo dei modelli di regressione

- Comprendere ***se*** e ***come*** il fenomeno di interesse (la variabile risposta) sia legata ai fattori concomitanti al fine di
  - **prevedere**: fornire una previsione del fenomeno di interesse se si conoscono le variabili concomitanti
  - **interpretare**: individuare quali siano le covariate che incidono maggiormente sul fenomeno di interesse e qual è il rapporto tra il fenomeno di interesse e una specifica variabile esplicativa

Formalizzazione di base (per un modello statistico GLM)

- Denoteremo la variabile risposta con Y
- e le variabili esplicative con $x_1, \dots x_p$
- Nei problemi di apprendimento statistico supervisionato si può mantenere la stessa notazione.
- un esempio notevole di modello statistico è il Modello Lineare Generalizzato (GLM) ove per una unità statistica osserviamo il valore y_i che è determinazione della variabile aleatoria Y_i e i valori delle variabili esplicative $x_{1i}, \dots x_{pi}$ vale:

$$g(E(Y_i)) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

I modelli lineari (generalizzati)

- Nei modelli lineari generalizzati (GLM)

$$g(E(Y_i)) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

- e Y_i è una opportuna variabile aleatoria (Bernoulli, Poisson, Normale, Gamma) coerente con la natura della variabile risposta, $E(Y_i)$ è il suo valore atteso
- la funzione g è scelta in modo opportuno in relazione ai valori assunti da $E(Y_i)$
- Si osserva il valore y_i e i valori $x_{1i}, \ldots x_{pi}$ per un campione di n individui

Il più noto modello della classe introdotta è il modello lineare in cui Y_i è gaussiana e la funzione g è la funzione identità

Il modello lineare gaussiano

Nel modello lineare gaussiano quindi possiamo anche scrivere per l'i-esima osservazione

$$y_i = f(x_{i1}, \dots, x_{ip}) + \epsilon_i$$
$$y_i = \beta_0 + \beta_1 x_i + \dots + \beta_p x_{ip} + \epsilon_i$$
$$y_i = \text{componente sistematica } + \text{componente stocastica}$$

- la componente sistematica
  - è una combinazione lineare delle variabili esplicative
  - $\beta_0 + \beta_1 x_i + \dots + \beta_p x_{ip}$
- la componente stocastica:
  - è una variabile aleatoria gaussiana di media nulla e varianza costante $var(\epsilon_i)$ che è indipendente dalle variabili aleatorie relative alle altre ossservazioni
- Si tratta di cercare i valori dei parametri (in particolare i β) e preferire i valori per i quali vi è un buon accostamento fra i valori prodotti dal modello e quelli effettivamente osservati (in questo caso la teoria statistica assicura che utilizzare il criterio dei minimi quadrati è un'ottima scelta)

Modellazione statistica: Alcuni aspetti generali

- Il tratto comune dei modelli statistici è che essi postulano un “meccanismo di generazione dei dati” e i dati osservati sono usati per inferire su di esso.
- Si introducono quindi alcune ipotesi che postulano strutture semplici, più o meno, definite in termini di pochi parametri.
- I dati disponibili vengono utilizzati per stimare questi parametri e quindi il modello aiuta a rispondere a domande di ricerca, a indagare possibili relazioni “causali”, a interpretare i dati e a prevedere i valori futuri della variabile risposta.
- Un modello finale emerge dopo molti controlli e prove ed è spesso diverso dal modello provvisorio iniziale
- Per selezionare il modello si invoca spesso il principio del rasoio di Occam: esso afferma che tra le ipotesi in competizione, dovrebbe essere preferita quella più parsimoniosa se i modelli alternativi si adattano “quasi” ugualmente bene ai dati.

Modellazione statistica: All models are wrong...

- G.E.P. Box (uno dei più grandi statistici del secolo scorso) viene spesso ricordato per questa frase

"Essentially, all models are wrong, but some are useful"

- Qualsiasi modello è una semplificazione della realtà e può essere sbagliato (si spera solo un po' sbagliato), ma può aiutare a spiegare, prevedere, capire. In una parola: a imparare dai dati.
- Anche la frase di John Tukey rende bene l'idea:

"Molto meglio una risposta approssimativa alla domanda giusta, spesso vaga, che una risposta esatta alla domanda sbagliata, che può essere sempre resa più precisa."

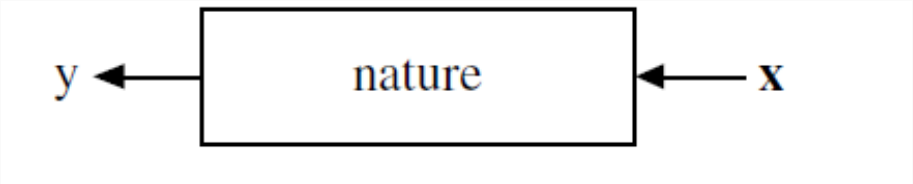
- È però vero che l'uso di alcuni modelli statistici più diffusi e ben consolidati con assunzioni comode (quali Normalità, linearità, indipendenza, stazionarietà) spesso non è in grado di produrre buone previsioni (o condurre a una buona comprensione) per il fenomeno in esame.

“Modelli generatori dei dati” o algoritmi predittivi?

- Abbiamo definito un modello statistico come una congettura su un “meccanismo di generazione di dati” che combina componenti casuali e sistematiche. E i dati sono impiegati per inferire su questo meccanismo.
- La qualità della nostra inferenza dipende fortemente dalle assunzioni fatte e si basa sull’uso di procedure statistiche “buone”.
- Si è anche osservato come i modelli plausibili in competizione possano esser molti e che sia cruciale l’obiettivo di selezionare modelli semplici e interpretabili senza rinunciare a una buona capacità predittiva.
- I modelli statistici quindi, denotando più sinteticamente con $X = (x_1, \dots, x_p)$ il vettore delle variabili esplicative (gli **inputs**), conducono a stimare $f(x_1, \dots, x_p) = f(X)$ con $\hat{f}(X)$ e di ottenere poi la previsione di Y come $\hat{Y} = \hat{f}(X)$.
- La modellazione statistica classica prevede che la funzione f sia nota a meno di un insieme di parametri e la forma di f è spesso rilevante per agevolare l’interpretazione delle relazioni fra le variabili.

“Modelli generatori dei dati” o algoritmi predittivi?

- Tuttavia in molti casi, soprattutto se vi è un maggior interesse sulla previsione, si preferisce mantenere flessibilità nella specificazione di f (con meno enfasi sulla interpretazione del fenomeno almeno nella fase di costruzione del modello)
- quello che conta è quindi costruire algoritmi (procedure di stima) che utilizzino i dati disponibili per ottenere buone approssimazioni di f, immaginata come una scatola nera, e di ottenere quale obiettivo prioritario buone previsioni.



- Si giudicano e si valutano quindi tali algoritmi a partire dalla loro capacità di fornire previsioni accurate

“Modelli generatori dei dati” o algoritmi predittivi?

- Gli algoritmi che si cercano e che tentano di approssimare la scatola nera con $\hat{f}(X)$
  - possono essere molto complessi
  - spesso prevedono relazioni non lineari fra le variabili
  - sono non parametrici
  - in molti casi non vi sono assunzioni formali sulla componente aleatoria (che pure è presente e ineliminabile).
- Per giudicare le prestazioni dei diversi metodi si ricorre prevalentemente al confronto fra le previsioni che l’algoritmo produce e i valori effettivamente realizzati (rinunciando spesso a cercare le proprietà formali del metodo utilizzato).
- Questo approccio da origine alle procedure di apprendimento statistico (AS) o machine learning (ML). Viene enfatizzato l’aspetto della previsione rispetto a quello della spiegazione (o dell’inferenza su un meccanismo generatore dei dati).

Apprendimento supervisionato e non supervisionato

- L'obiettivo enfatizzato finora è quello di imparare come un insieme di caratteristiche osservate siano collegate a un risultato di interesse (la variabile risposta) sia per fare previsioni che per comprendere il fenomeno
- Il processo di apprendimento è guidato dai dati
- Nella presentazione del problema e negli esempi introdotti abbiamo immaginato di disporre di dati che riguardino sia le variabili concomitanti che la variabile risposta.
    - Questa situazione, definisce un problema di apprendimento supervisionato poiché i valori osservati della variabile obiettivo "supervisionano" l'apprendimento
- Di rilevante interesse è anche il caso in cui non viene definita una variabile obiettivo e i dati su un insieme di variabili servono a apprendere sulla presenza di strutture (pattern) caratteristiche (un esempio sono le tecniche di raggruppamento o cluster analysis).
    - In questo caso si parla di apprendimento non supervisionato

Nel corso (anche per questioni di tempo) non andremo ad affrontare il tema dell'apprendimento non supervisionato

Apprendimento supervisionato: previsione e classificazione

Con riferimento all'apprendimento supervisionato conviene distinguere il caso in cui la variabile obiettivo Y è quantitativa da quello in cui è categoriale

- Nel caso in cui Y è **quantitativa** (valore economico di un danno, pressione sanguigna, redditi, ...) si parla di problemi di previsione (o di regressione)
- Nel caso in cui Y è **categoriale** (cliente fedele/infedele, frode/non frode, propenso/non propenso all'acquisto, ...) si parla di problemi di classificazione

Obiettivi dell'apprendimento supervisionato

Come nel caso dei modelli statistici, sulla base dei dati disponibili si vuole

- prevedere con la maggiore accuratezza possibile l'esito se si presentano nuove unità (nuovi dati)
- comprendere quali variabili influenzano il risultato e in quale modo
- disporre di strumenti adeguati per valutare la qualità delle nostre previsioni e deduzioni

Come già osservato, nell'AS l'enfasi è più spesso rivolta all'aspetto predittivo ma sono numerose le applicazioni nelle quali si richiede che vi sia un buon bilanciamento fra obiettivo di previsione e di spiegazione

Stimare f (o allenare -training- l'algoritmo)

Come detto si dispone di dati riguardo la variabile risposta Y e su un vettore di variabili di input $X = (x_1, \dots, x_p)$. La determinazione della stima dela funzione $\hat{f}$ deve essere tale da consentirci di ottenere previsioni $\hat{f}(X_i) = \hat{Y}_i$ che per ogni unità i che siano vicine ai valori osservati.

- La misura comunemente utilizzata nel contesto della regressione (previsione) è la devianza valutata per l'intero insieme di dati ovvero

$$DEV = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2$$

- A tutti è noto che se io utilizzassi una definizione ristretta per f, ad esempio limitandosi alle funzioni lineari degli input X, la ricerca di $\hat{f}$ che minimizza lo scostamento fra valori osservati e previsti definito con DEV conduce alla stima dei minimi quadrati dei parametri che caratterizzano la funzione lineare
- Dobbiamo ora immaginare che nella ricerca di una opportuna funzione $\hat{f}$ si possa scegliere fra versioni molto più flessibili di questa (immaginate ad esempio, se vi fosse una sola x, di poter utilizzare un polinomio di grado s con s libero di variare da 1 a $n - 1$)

Ancora Occam?

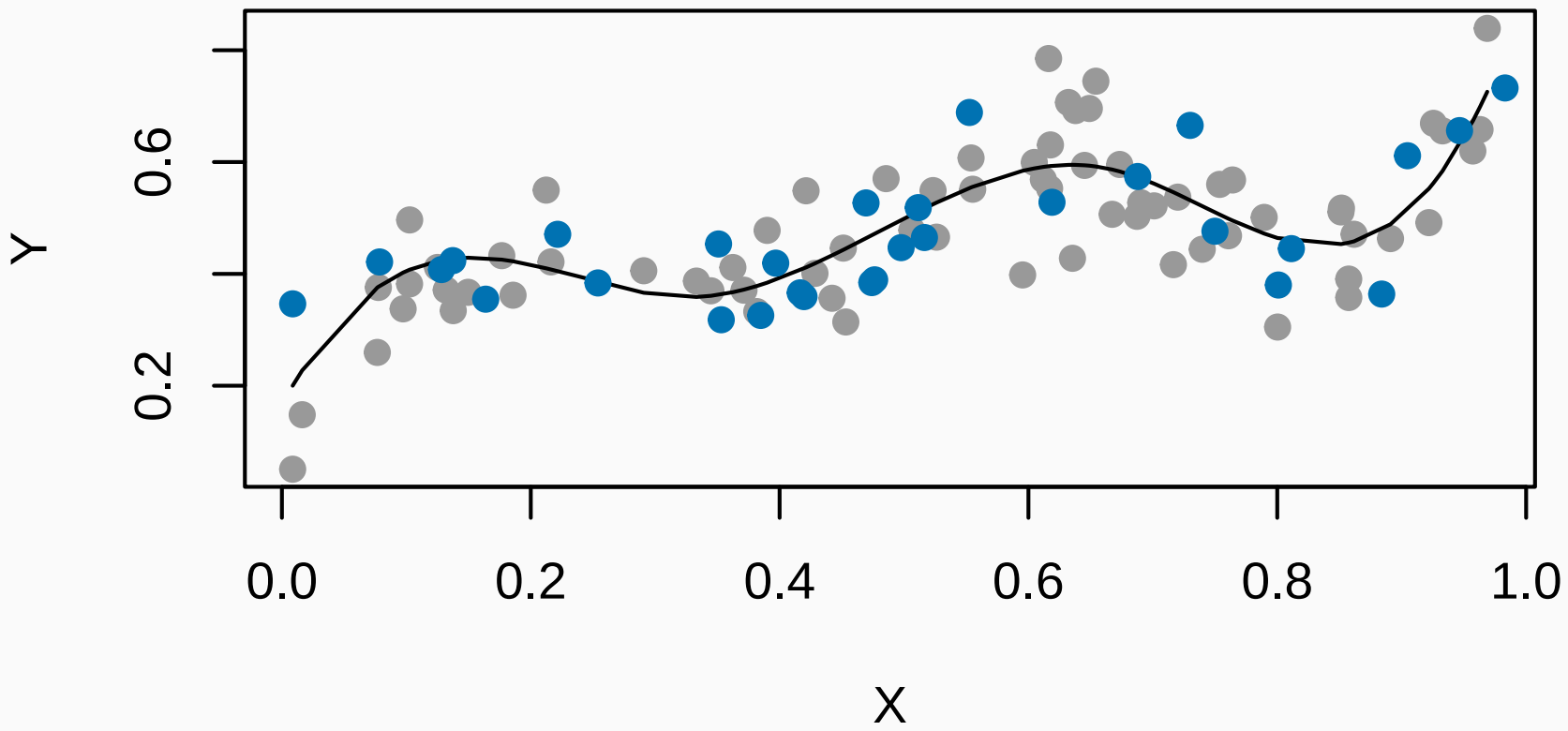
- Se posso scegliere liberamente fra le funzioni f, ad esempio utilizzando metodi non parametrici, è evidente che posso ottenere una $\hat{f}$ che approssimi in modo molto preciso i punti osservati (o al limite li riproduca fedelmente)
- Tuttavia resta l'obiettivo di scegliere una funzione meno complessa che passi abbastanza vicino i punti ma che non sia al contempo troppo contorta (sia la più semplice possibile)
- Se però mi attengo al criterio di cercare quella funzione $\hat{f}$ che produca previsioni che conducano a un valore di DEV piccolo questo porterebbe, all'interno di una ampia famiglia di funzioni, a scegliere sempre il modello più complicato
- Però noi siamo interessati a vedere come la funzione $\hat{f}$ che ottengo riesca a prevedere non i dati che abbiamo utilizzato per determinarla ma dei dati nuovi che l'algoritmo non ha mai visto

Apprendimento statistico supervisionato

Insieme di apprendimento e insieme di valutazione

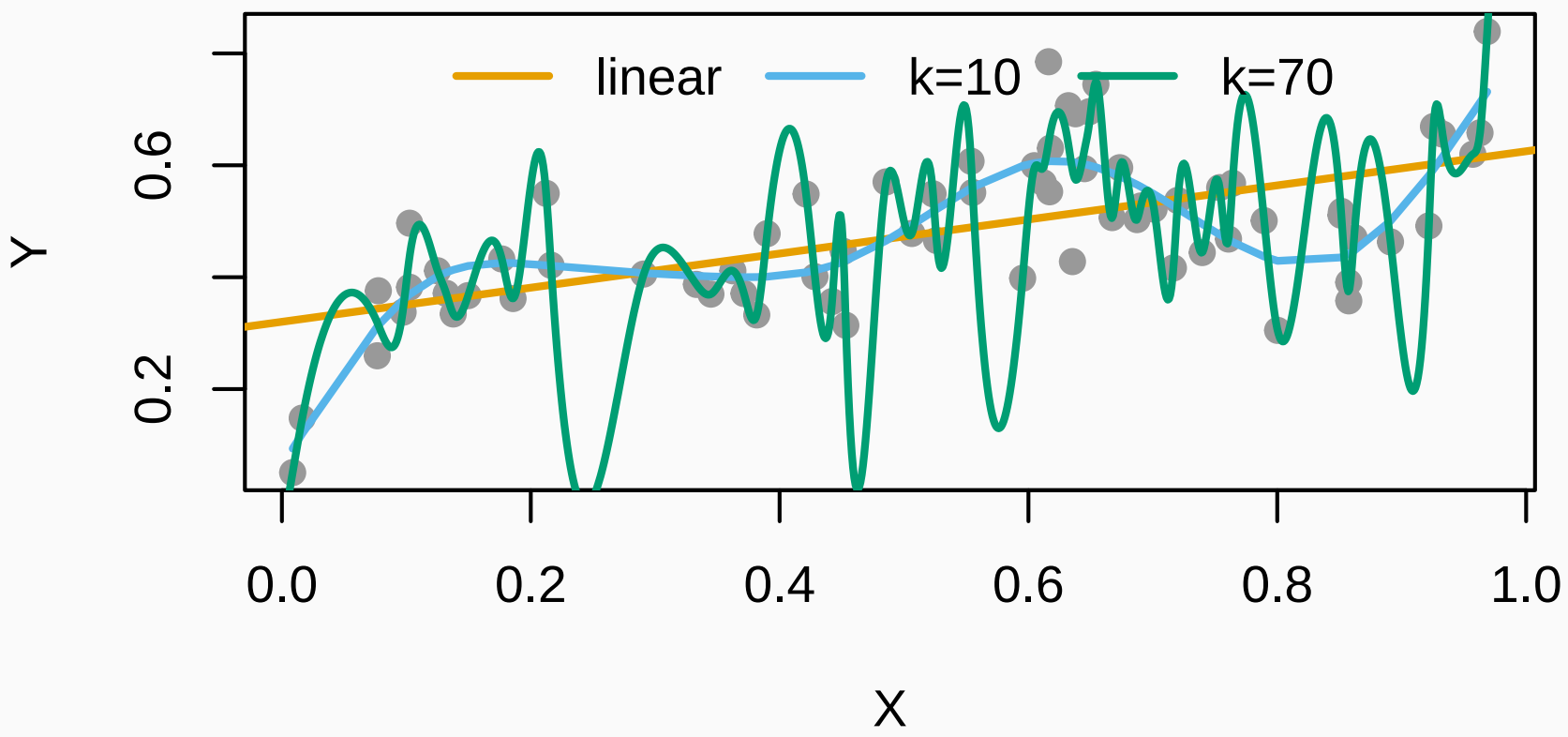
- Se vogliamo valutare modelli alternativi guardando all'errore DEV sia nella fase di determinazione del modello che nella fase di valutazione delle sue performance previsive su nuovi dati una strategia molto semplice è la seguente :
  - separiamo i dati disponibili casualmente in due sottoinsiemi disgiunti
    - l'insieme di apprendimento o allenamento (training set)
    - l'insieme di valutazione (test set)
  - ad esempio, 70% dei dati vengono selezionati per la fase di apprendimento e il restante 30% è invece inserito nel test set
- introdurremo ora un semplice esempio che consenta di vedere come si comporta DEV nei due casi utilizzando modelli via via più flessibili

Un esempio (con dati simulati)



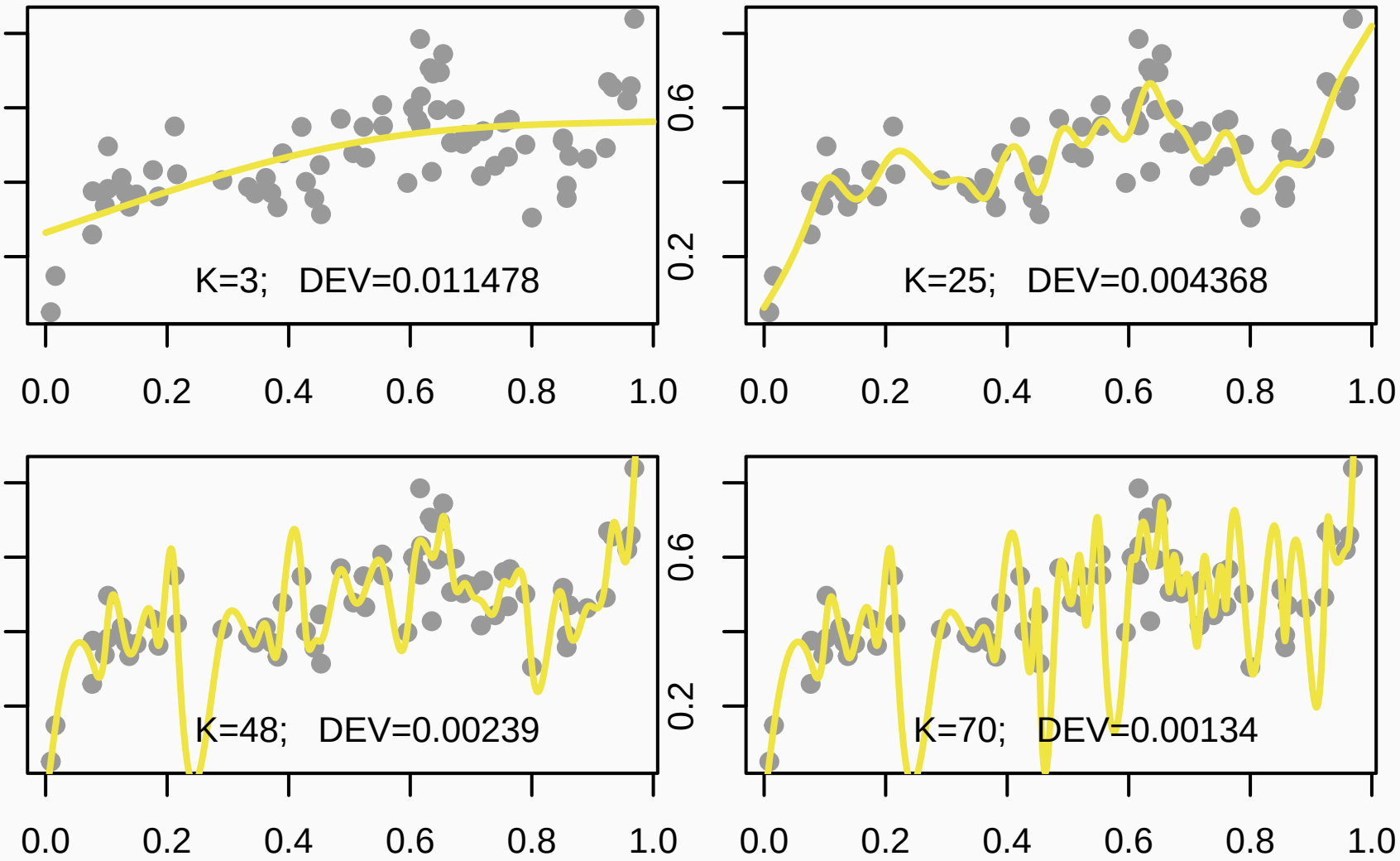
- si sono simulati dati del tipo $Y = f(x) + \epsilon$
- i punti grigi sono quelli del training set (li useremo per cercare il modello), i blu sono quelli del test set (li useremo per vedere se il modello funziona su dati nuovi)

Modelli alternativi sul training set



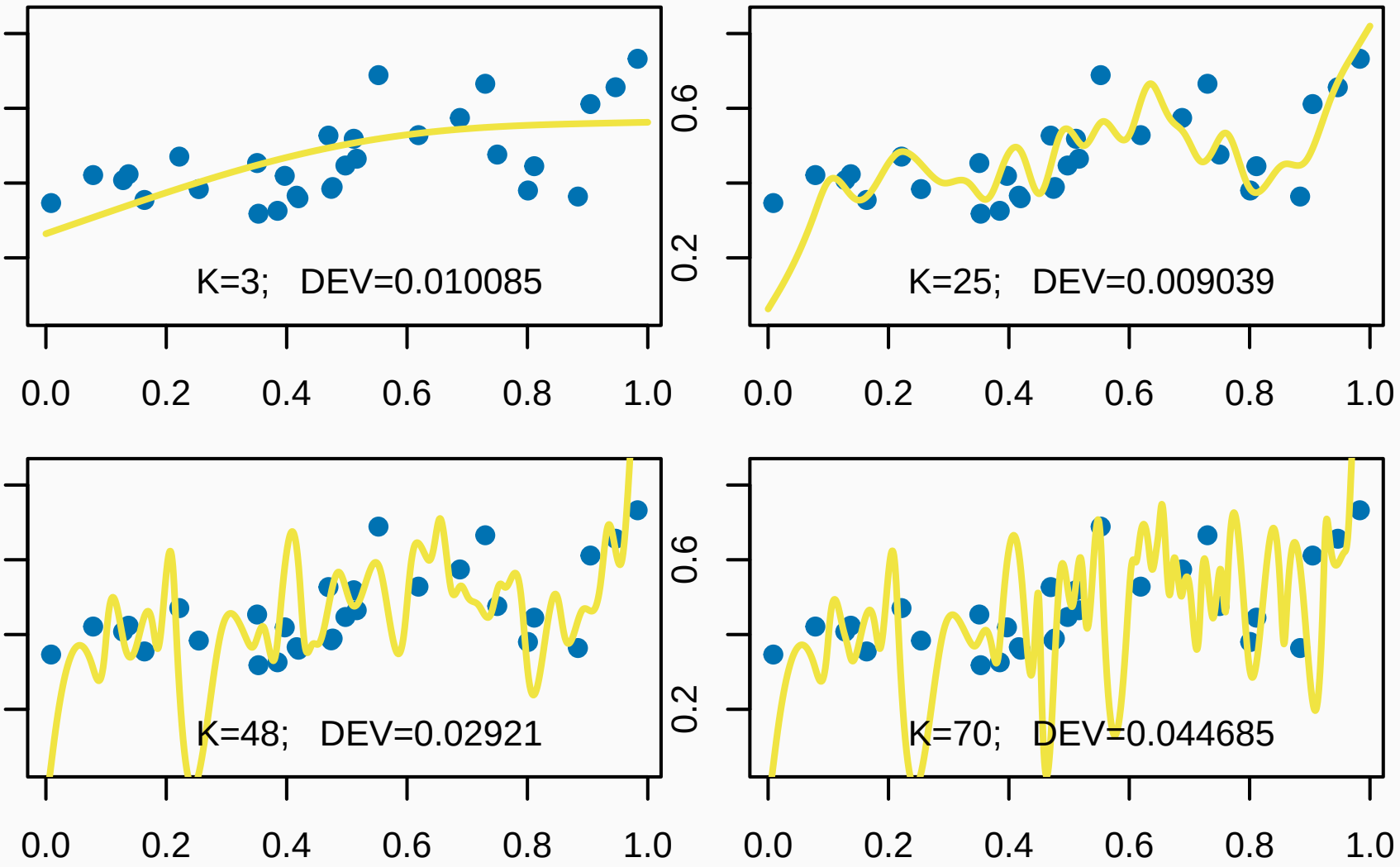
- i modelli alternativi che danno luogo alle curve in figura sono modelli semiparametrici; in essi più elevato è K più il modello sarà flessibile e fornirà previsioni molto vicine ai veri valori.

Qualità dell'adattamento dei modelli (training set)



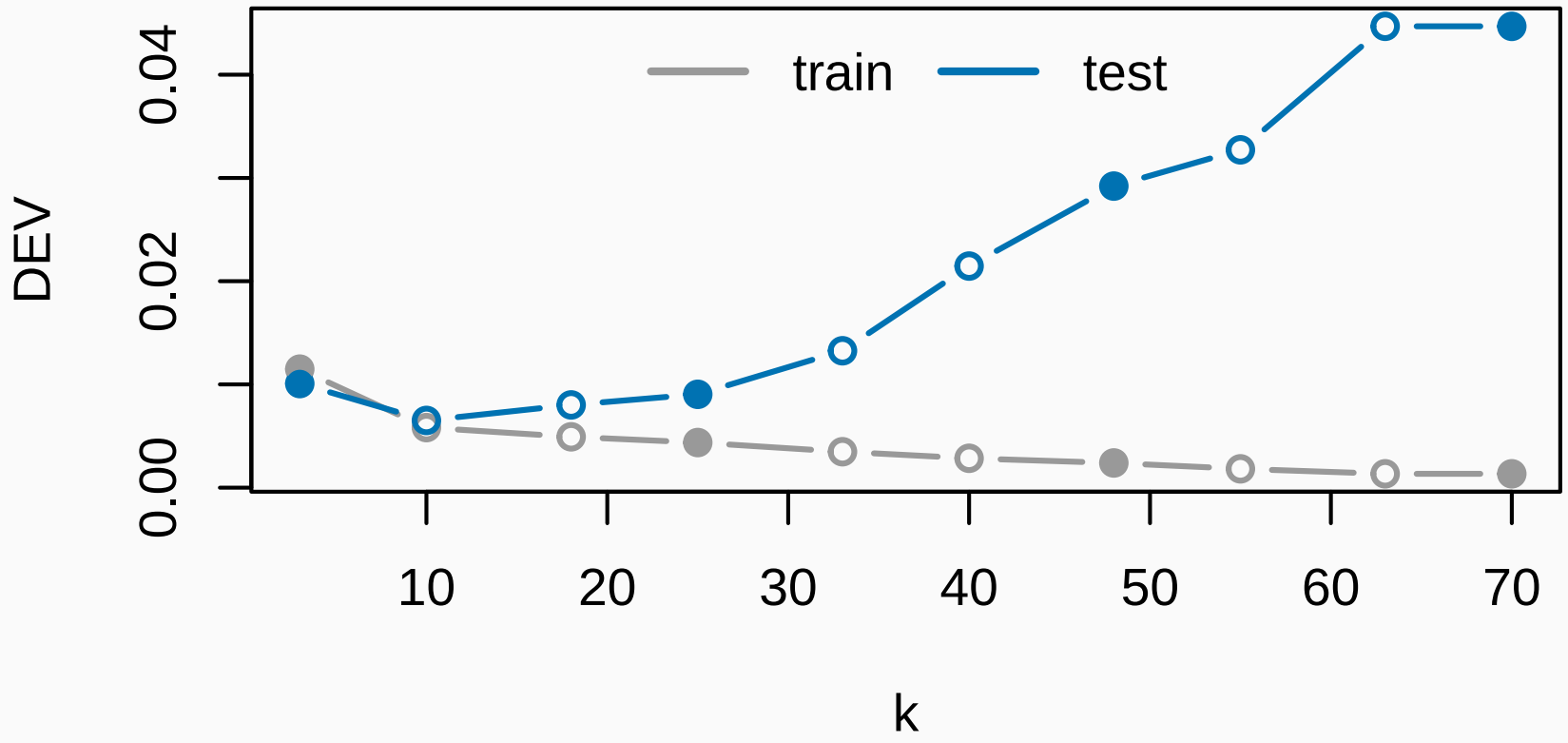
- Si noti che sul training set DEV decresce con k

Qualità dell'adattamento dei modelli (test set)



- Sul test set DEV prima decresce poi torna a crescere

Confrontiamo le performances sui due insiemi (al variare di k)



- Occorre evitare di scegliere modelli troppo flessibili essendo abbagliati dal buon adattamento nel training set
- Il fenomeno per cui modelli eccessivamente "dettagliati" selezionati sul test set poi potrebbero dare pessimi risultati di previsione è detto sovradattamento, in inglese overfitting

L'errore quadratico medio

- Quando k (l'indicatore di complessità del modello) aumenta, l'adattamento migliora sul training set, ma questo non è vero per il test set.
- L'errore di adattamento diminuisce con k mentre l'errore di previsione (valutato sul test set) prima diminuisce poi torna ad aumentare
- Quando k aumenta troppo quindi "sovradattiamo" i dati
- Per capire perchè accada questo negli algoritmi di AS (ML) che cercano di ottenere un buon $\hat{f}$, conviene introdurre la seguente quantità che andrà valutata su un test set

$$MSE_0 = E(Y_0 - \hat{f}(X_0))^2$$

Si tratta dell'Errore Quadratico Medio e il valore atteso è fatto rispetto al valore Y_0, ovvero si valuta cosa succederebbe in media se potessi osservare diversi valori di Y_0 ripetendo l'osservazione. Si può calcolare poi l'MSE complessivo facendo la media su tutti i possibili valori di X_0 nel test set
$MSE = \frac{1}{n} \sum_{i=1}^{n} E(Y_i - \hat{f}(X_i))^2$

Il trade-off distorsione/varianza

Il bilanciamento (trade-off) tra distorsione e varianza

La quantità definita da MSE può essere scomposta come segue:

$$E(Y_0 - \hat{f}(X_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon)$$

- Quando noi scegliamo un metodo di AS $\hat{f}$ vorremo quindi che abbia simultaneamente **bassa varianza** e **basso bias**.
- la quota di MSE legata alla varianza della componente casuale $\text{Var}(\epsilon)$ non può essere ridotta (è un errore casuale che avremmo anche se conoscessimo la vera f)
- Bias e varianza dipendono quindi dalla funzione (dall'algoritmo) $\hat{f}$ che utilizziamo come approssimazione di un'ipotetica vera funzione

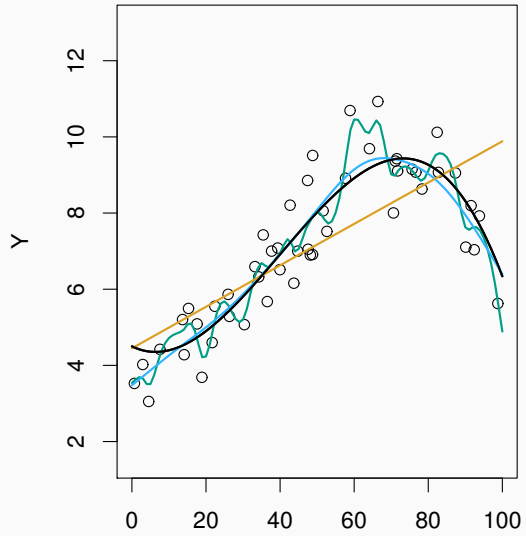
Il trade-off tra distorsione e varianza

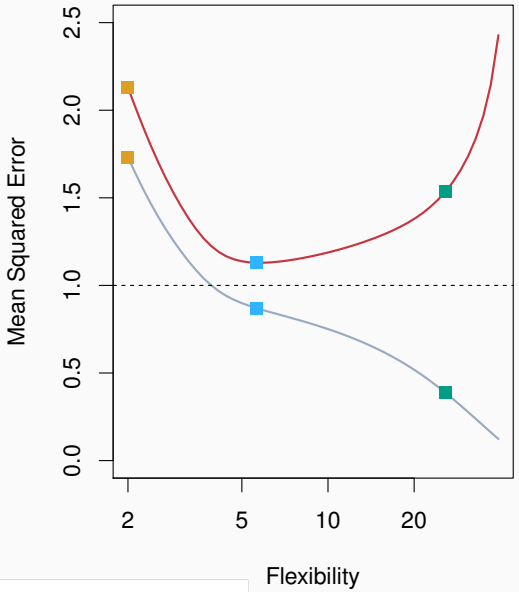
- la varianza $\text{Var}(\hat{f}(x_0))$ attiene a quanto la funzione stimata $\hat{f}(x)$ cambia se noi usiamo un training set diverso (un nuovo campione di dati)
    - se un metodo ha alta varianza otterremo stime molto diverse (e quindi forme di $\hat{f}(x)$ che ballano molto) al variare dei dati di training
    - in generale più flessibile è la funzione che utilizziamo più elevata sarà questa componente
- il bias $\text{Bias}(\hat{f}(x_0))$ attiene all'errore che commettiamo quando approssimiamo la vera funzione f con una funzione più semplice
    - ad esempio un modello di regressione lineare assume un comportamemnto semplificato di quanto accade nella realtà, tuttavia la sua forma è rigida e non varierà tanto al variare dei dati
    - qualunque sia il nuovo insieme di dati la forma lineare non potrà mai approssimare la vera f e quindi avrà un alto Bias
        - ovviamente essa dipende anche dalla forma della vera f

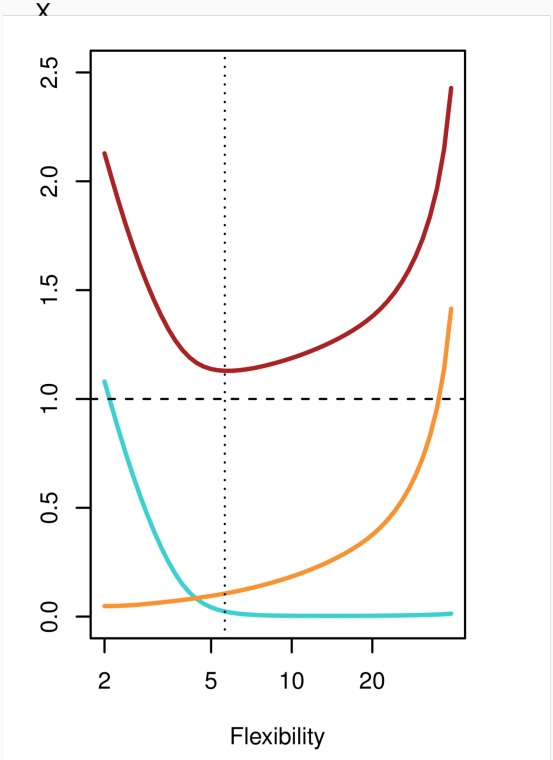
Il trade-off tra distorsione e varianza

- Se la complessità dell'algoritmo (del modello) è bassa avrò quindi molta distorsione (bias) ma poca varianza mentre al crescere della complessità il bias tenderà ad annullarsi e dominerà la varianza.
- Questo spiega
    - l'andamento della devianza visto nell'esempio precedente
    - la necessità di calcolare l'errore di previsione su dati diversi da quelli che abbiamo utilizzato per determinare (allenare) l'algoritmo $\widehat{f}$

Il trade-off tra distorsione e varianza







Il trade-off tra distorsione e varianza: conseguenze

- Uno degli aspetti di maggior delicatezza nella costruzione di algoritmo di ML è quello di trovare il giusto compromesso e evitare il sovraadattamento (**overfitting**)
- Quindi non bisogna valutare un modello utilizzando gli stessi dati utilizzati per cercare (stimare, allenare) la funzione
- Inoltre bisogna che per i metodi di AS siano tali da cercare un compromesso tra adattamento e flessibilità
- Ciò conduce alla suddivisione dei compiti (si cerca un buon adattamento sul training set e si valutano le performance sul test set)

Come selezionare i modelli (gli algoritmi)

Metodi basati sulla devianza o misure similari (con penalizzazione)

- La devianza in realtà potrebbe rivelarsi un indicatore non affidabile della qualità del modello essendo troppo ottimistico nel valutare l'errore di previsione.
- Si potrebbe quindi penalizzare la *devianza*

$$DEV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  o una sua trasformazione monotona, ad esempio, $ n log(DEV) + (costant)$ con un'opportuna quantità che quantifichi la complessità del modello
- In un modello lineare gaussiano, ad esempio, la quantità vista equivarrebbe a $-2logL$ ovvero a -2 volte la log-verosimiglianza
- Vi sono diversi criteri ispirati alla logica della penalizzazione per decidere fra modelli alternativi. Tornando all'esempio dei modelli lineari (o loro varianti), la funzione da minimizzare potrebbe essere del tipo

$$IC(p) = -2logL + \text{penalità}(p)$$

ove p è il numero di parametri del modello - Il criterio cui si arriva dipende dalla funzione di penalità scelta. I noti criteri tipo AIC o BIC seguono questa logica

Training set e test set

La suddivisione dell'insieme di dati fra training set e test set è una scelta piuttosto comune quando si dispone di molti dati (spesso è questo il caso) e come già visto:

- il training set viene utilizzato per cercare fra i modelli candidati quelli che danno un buon adattamento
- il test set (a volte chiamato set di validazione) viene utilizzato per valutare le prestazioni dei modelli disponibili e quindi scegliere quello che fornisce previsioni più accurate.
- Spesso si utilizzano diverse tecniche e diversi modelli e si confrontano i risultati ottenuti sul test set
- Ovviamente questo riduce i dati utilizzati nella fase di adattamento (se n è grande questo però non è un problema)
- Le proporzioni utilizzate per i due insiemi sono di solito sbilanciate a favore del training set (circa il 75% o 70% dei casi)

Validazione incrociata (cross-validation a K vie)

Quando n non è molto grande, possiamo modificare lo schema:

- se utilizziamo solo il 75% dei dati per la fase di training (la ricerca del modello) la stima potrebbe essere meno affidabile (alta variabilità), mentre vorremmo sfruttare al meglio tutte informazioni disponibili.
- una tecnica alternativa potrebbe essere quella dividere i dati in, diciamo, K parti uguali e poi utilizzare $K-1$ porzioni a rotazione per la fase di training (costruzione) del modello e la parte rimanente per la fase di test.
- a turno ogni gruppo di dati viene usato nella parte di adattamento e nella parte di previsione
- questo schema richiede K iterazioni della fase di training e del calcolo degli errori di previsione
- Si considera poi la media (o qualche altra combinazione) delle K performance di previsione

Tale tecnica è detta di validazione incrociata (cross-validation a K vie)

Validazione incrociata del tipo lascia-fuori-uno (Leave-one-out)

La tecnica illustrata sopra è la validazione incrociata (cross-validation a K vie)

- Si può arrivare all'estremo di porre $K = n$. Si stimano quindi n modelli con $n - 1$ dati e si usa quello rimasto per valutare la previsione. In questo caso si parla di *leave-one-out cross validation*.
- Tale schema è computazionalmente più dispendioso ma in alcuni casi (nei modelli lineari ad esempio) non è necessario iterare i calcoli.

Apprendimento statistico supervisionato: problemi di classificazione

La misura della qualità di un modello (algoritmo di classificazione)

- Molti dei concetti che abbiamo incontrato, sono applicabili ai problemi di classificazione. In particolare anche in questo caso vi è da affrontare il problema del trade-off bias-varianza
- La variabile Y è in questo caso categoriale e il problema di classificazione consiste nel prevedere in quale categoria si collocherà una unità con caratteristiche X
- Un equivalente della devianza (come scostamento fra classe osservata e classe prevista) potrebbe essere

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

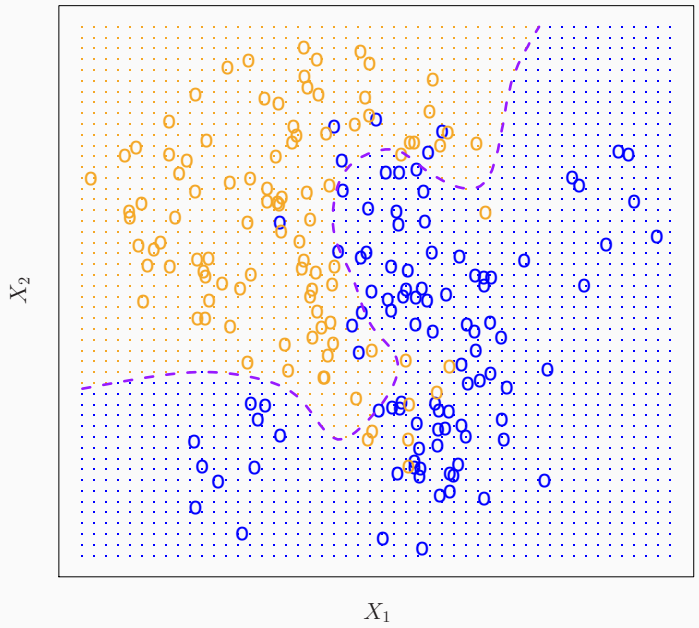
  dove $\hat{y}_i$ è stavolta la classe assegnata dall'algoritmo di classificazione e la quantità appena vista è la proporzione di errori commessi.
- nel training set si cerca di minimizzare tale quantità e l'algoritmo che ne risulta è poi valutato calcolando la stessa proporzione però su un test set.
- Tuttavia come vedremo durante il corso la misura dell'accuratezza come proporzione di corrette classificazioni è molto grezza e ne introdurremo molte altre per tenere conto delle specificità dei problemi applicativi

Problemi di classificazione

Nei problemi di classificazione la regola che si adotta per assegnare un elemento a una classe è banalmente quella per cui se il valore dei predittori è x_0 la soluzione ottima è quella di assegnare l'osservazione alla classe j per cui è massima

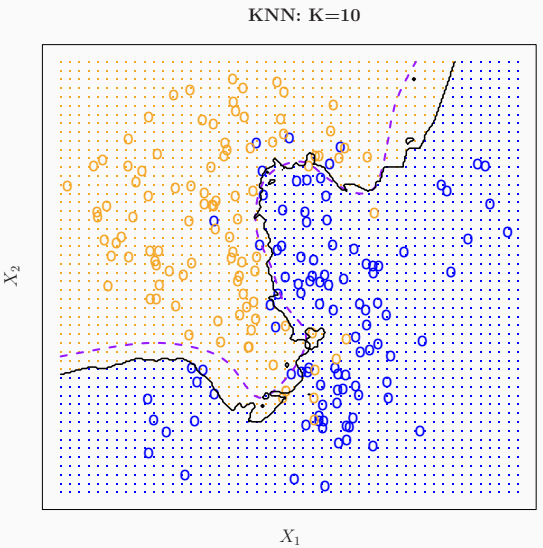
$$Pr(Y = j|X = x_0)$$

- questo è detto classificatore di Bayes. Vediamo un esempio di *confine di Bayes* in un problema con 2 classi e due predittori



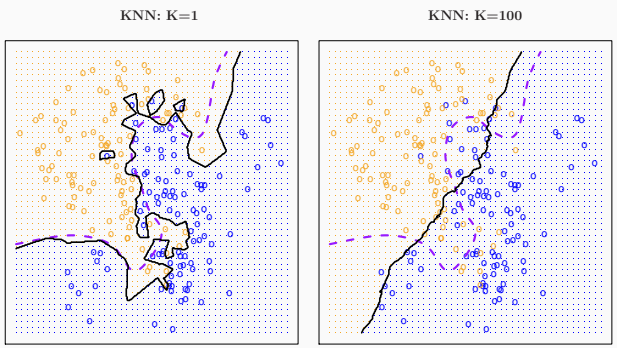
Classificazione con i K più vicini (K-Nearest-Neighbour- KNN)

- Nella realtà noi non conosciamo la probabilità condizionata indicata sopra e cerchiamo di approssimarla con algoritmi opportuni (di classificazione).
- Un metodo semplice è quello di stimare $Pr(Y = j|X = x_0)$ sulla base dei dati con $\hat{Pr}(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in L_0}^{n} I(y_i = j)$
- K è il numero di punti più vicini ad x_0 e L_0 è l'insieme dei K punti più vicini a x_0.
- si classifica nella classe che osserviamo con maggiore ferquenza per i K punti più vicini a x_0. Con $K = 10$ per i dati che abbiamo illustrato nella figura precedente si ottiene

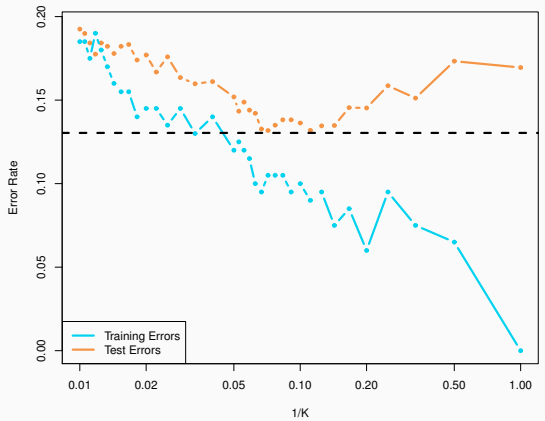


Ancora il trade-off e l'overfitting

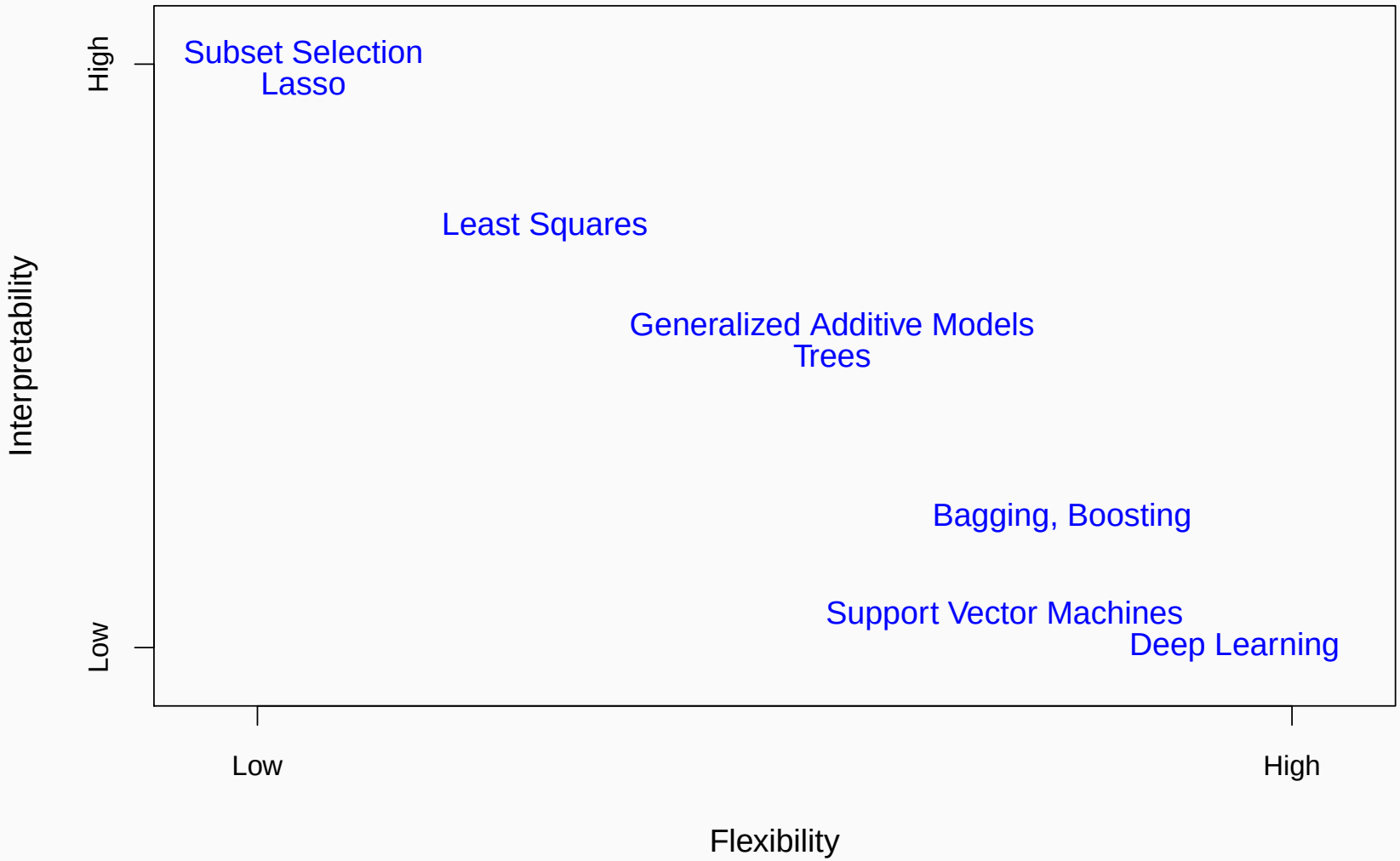
- Al variare di K ottengo rappresentazioni più o meno flessibili per il confine di decisione



- con K=1 esso è troppo flessibile con K=100 è è troppo liscio.
- Nel primo caso vi è il rischio di overfitting
- Se si calcola l'errore di classificazione su un test set e poi applico tale classificatore (per vari valori di K) su un nuovo insieme di dati ottengo un grafico analogo a quellio visto sopra per la previsione (si noti che in ascissa è riportato $1/K$)



Il bilanciamento (trade-off) fra capacità predittiva e interpretabilità



Cose importanti da ricordare

Cose importanti da ricordare

- Le procedure inferenziali tipiche non si applicano
- E' necessario cercare una compensazione tra varianza e bias
- Occorre evitare il sovradattamento (**overfitting**)
- Ci sono diversi modi per affrontare questo problema
    1. Dividere i dati in insieme di adattamento e insieme di valutazione (training e test set)
    2. Usare la validazione incrociata a K-vie
    3. Utilizzare criteri di adattamento che contengano sia termini di errore che termini di complessità del modello