

UNIVERSITÀ DEGLI STUDI DI SALERNO



**Dipartimento di Ingegneria dell'Informazione ed
Elettrica e Matematica applicata**

Corso di Laurea in Ingegneria Informatica

**APPUNTI DI DATA SCIENCE
DI FRANCESCO PIO CIRILLO**

<https://github.com/francescopiocirillo>



"Sii sempre forte"

 Ehi, un attimo prima di iniziare!

Hai appena aperto una raccolta di appunti che ho deciso di condividere **gratuitamente** su GitHub, se ti sono utili fai **una buona azione digitale**:

-  **Lascia una stellina alla repo:** è gratis, indolore e fa super piacere!
-  **Condividerla con amici**, compagni di corso, o chiunque possa averne bisogno.

Insomma, se questi appunti ti salvano anche solo una giornata di studio... fammelo sapere con una **stellina!**

Grazie di cuore 

Intro Data Science - 16/09

Cos'è lo statistical learning?

(Capitolo 2 di ISLR)



L'apprendimento statistico ha lo scopo di usare i dati per imparare

Supponiamo di osservare due tipi di dati Y_i e $X_i = (X_{i1}, \dots, X_{ip})$ per $i = 1, \dots, n$.

Raccogliamo questi dati perché pensiamo che ci sia una relazione tra Y e almeno una delle X.

Costruiamo le relazioni usando un approccio empirico, questo si fa quando non ci possiamo affidare alla fisica che ci da relazioni sicure.

Ad esempio un modello empirico può riguardare rapporti tra persone.

Modelliamo questa relazione che pensiamo possa esserci come:

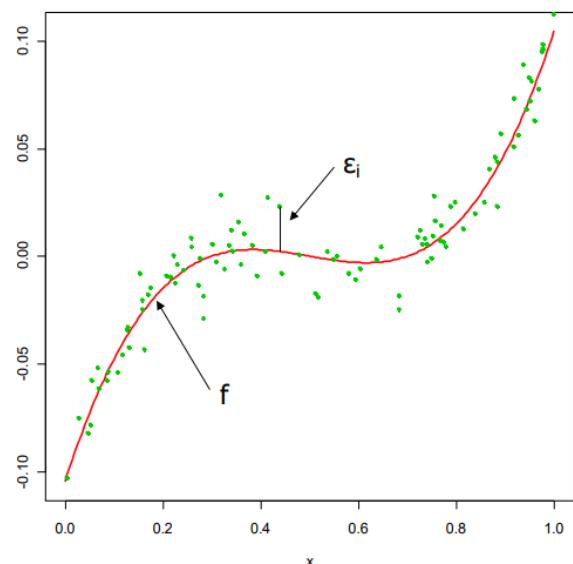
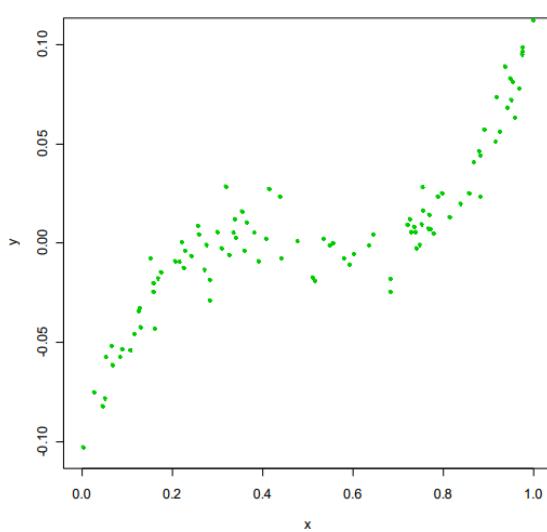
$$Y_i = f(X_i) + \varepsilon_i$$

dove f è una funzione ignota ed ε è un errore casuale con media 0.

Il termine ε_i è riferito al fatto che la funzione f non viene mai stabilita con assoluta precisione, inoltre ad una misura corrisponde sempre incertezza.

La statistica si poggia sulla probabilità.

Un semplice esempio

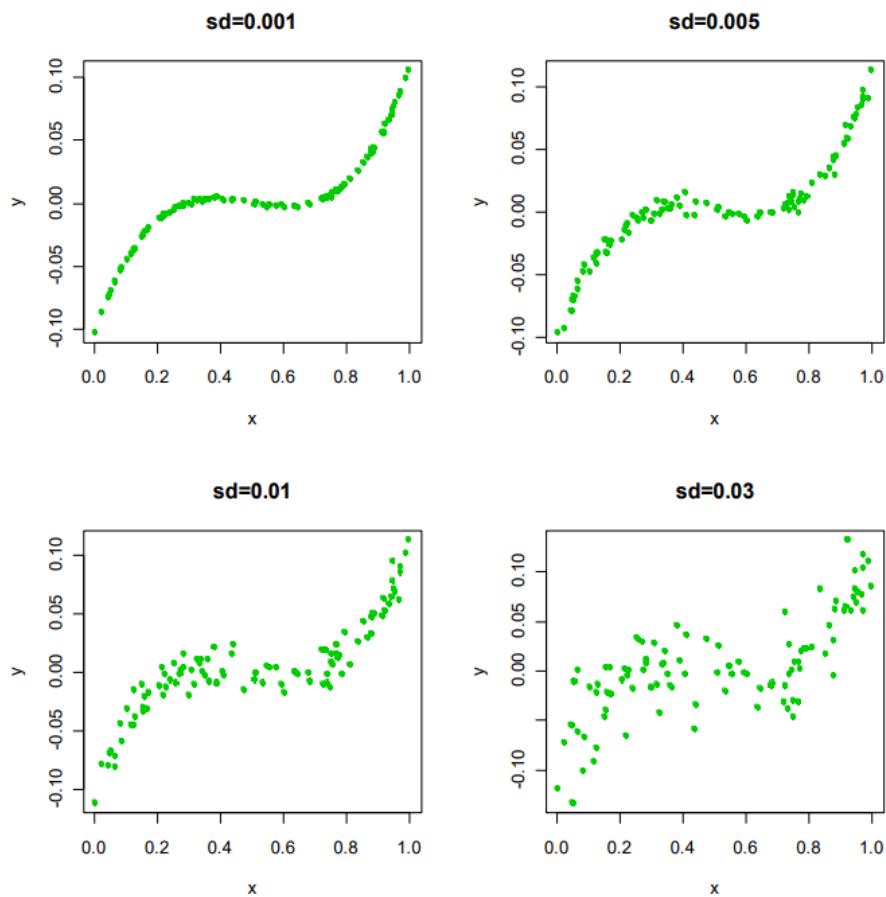


Questo è uno **scatter plot** o grafico di dispersione, nel quale si visualizzano i potenziali legami tra Y e X. Anche se ad occhio non si vede il legame potrebbe comunque esserci e bisogna analizzare meglio.

Sulla base dello scatter plot si definisce la funzione che lega Y ed X.

Bisogna identificare il rapporto di massima tra i dati, la differenza tra un determinato punto e la funzione sarà proprio ε_i .

Differenti deviazioni standard

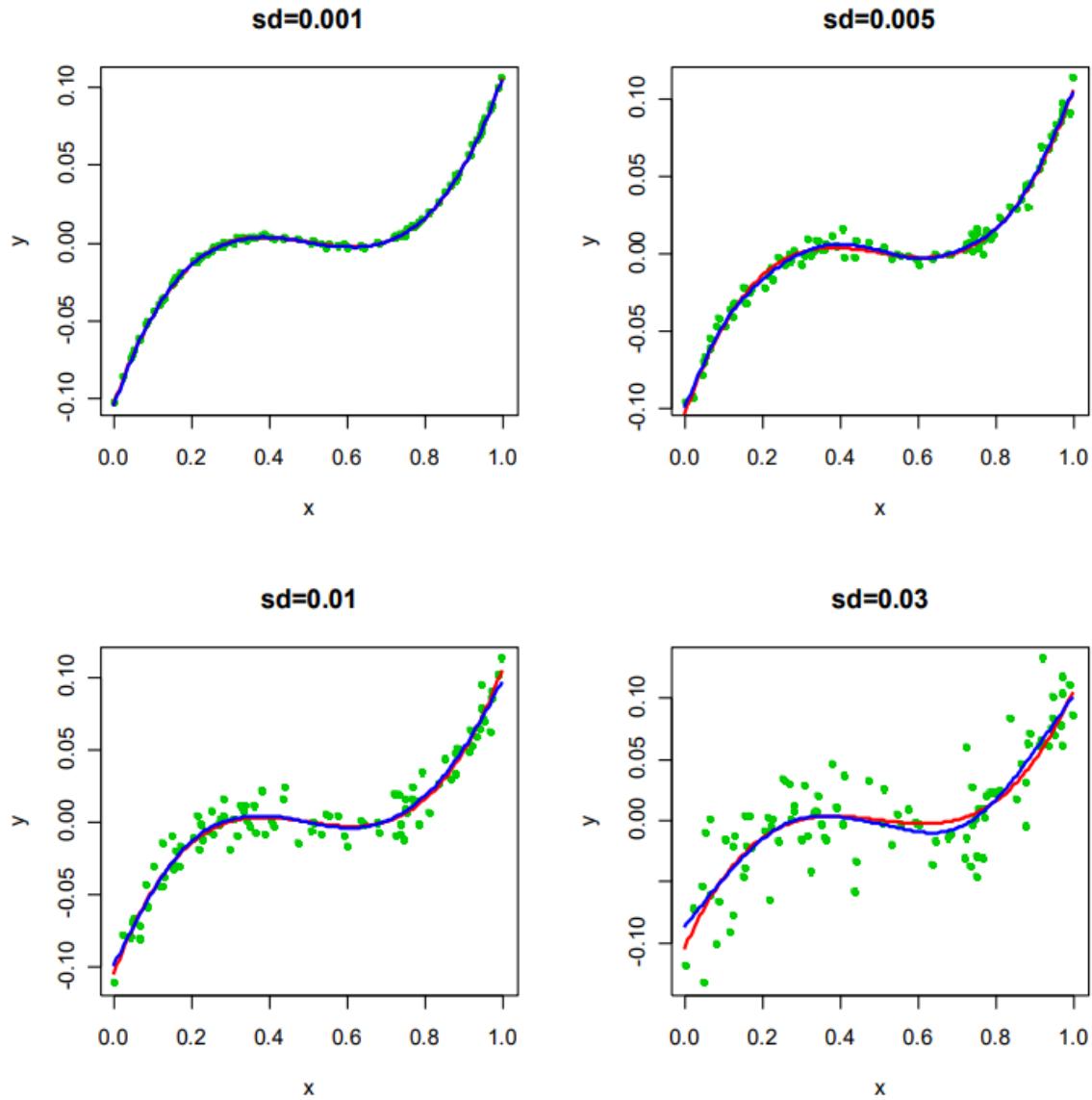


La fluttuazione della y è quello che rende più o meno facile stimare f.

La fluttuazione della y è tanto più grande quanto è grande la deviazione standard della ε .

Il legame funzionale immediato non si vede quasi mai, noi proviamo a costruire dei modelli per meglio approssimare i dati a nostra disposizione.

Diverse stime di f



Una deviazione standard maggiore comporta una funzione f che è meno simile rispetto alla funzione usata per generare i dati (ovviamente in un caso artificiale in cui i dati sono stati creati per esercizio).

Un altro esempio

Supponiamo di essere consulenti statistici assunti da un cliente per fornire consigli su come migliorare le vendite di un determinato prodotto. Il dataset *Advertising* consiste nelle vendite di quel prodotto in 200 mercati diversi, insieme ai budget pubblicitari per il prodotto in ciascuno di questi mercati per tre media differenti: TV, radio e giornali. I dati sono mostrati nella Figura.

Non è possibile per il nostro cliente aumentare direttamente le vendite del prodotto. D'altra parte, essi possono controllare le spese pubblicitarie in ciascuno dei tre media. Il nostro obiettivo è sviluppare un modello accurato che possa essere utilizzato per prevedere le vendite sulla base dei budget destinati ai tre media.

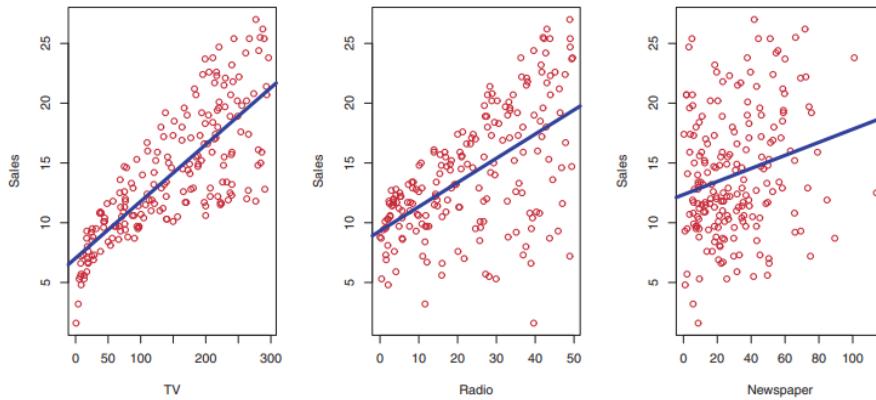
In questo contesto, i budget pubblicitari sono variabili di input, mentre le vendite sono una variabile di output. Le variabili di input sono tipicamente indicate usando il simbolo X, con un pedice per distinguere. Quindi X_1 potrebbe essere il budget per la TV, X_2 il budget per la radio, e X_3 il budget per i giornali.



Le variabili di input possono avere nomi diversi, come **predittori**, variabili indipendenti, **features**, o a volte semplicemente variabili.

La variabile di output — in questo caso, le vendite — è spesso chiamata **response (risposta)** o variabile dipendente, ed è tipicamente indicata con il simbolo Y.

Per partire mettiamo un modello semplice, una retta.



Per noi è importante poter predire (predizione) futuri dati che non ho ancora, per mezzo della serie storica cioè i vecchi dati.

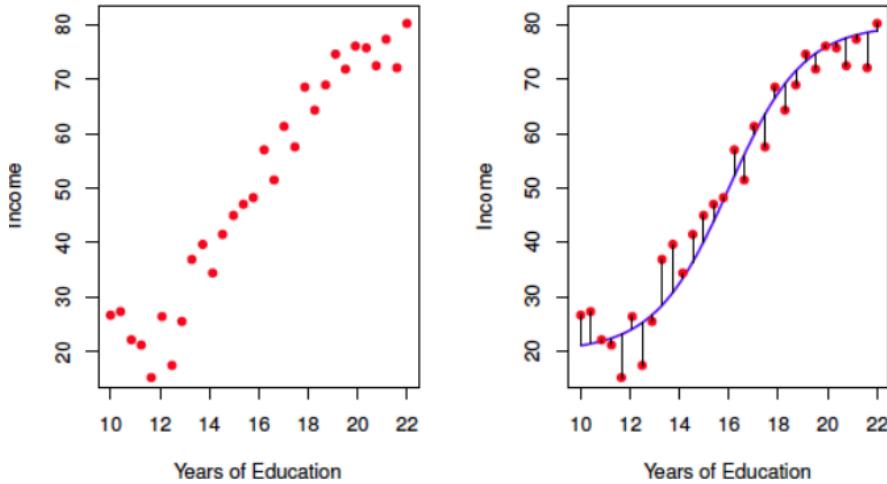
$X = (X_1, X_2, X_3)^T$ è il vettore di input.



Lo statistical learning (apprendimento statistico) si riferisce ad un insieme di approcci per la stima di f.

Se ci sono tanti parametri rispetto ai dati un problema è detto ad "Alta dimensionalità" ma noi faremo problemi più semplici.

X può essere decisa a tavolino o talvolta può essere aleatoria.



Supponiamo di voler predire l'income usando gli anni di istruzione basandoci sul data set presentato nell'immagine.

Questo è un dataset simulato, la linea blu è la vera funzione f e le linee verticali rappresentano il termine di errore ϵ , che ha all'incirca media 0.

Perché stimiamo f ?

Lo statistical learning si riferisce ad un insieme di approcci per la stima di f per mezzo dei dati. Ci interessa stimare f per due ragioni:

- fare predizioni;
- fare inferenza, cioè dedurre il legame tra Y ed X .

Predizioni

Se produciamo una buona stima di f (e la varianza di ϵ non è troppo grande) possiamo fare predizioni accurate della risposta Y relativa a nuovi punti $X = x$.

Visto che il termine di errore è a media 0 possiamo predire Y usando:

$$\hat{Y} = \hat{f}(X)$$

dove f cappello rappresenta la nostra stima di f (la vera f , cioè il cosiddetto legame di natura, sarà sempre sconosciuta) e \hat{Y} rappresenta la predizione di Y . Quella che otteniamo è una stima intervallare, perché invece di un punto abbiamo un intervallo a causa della ϵ .

Costruire f a partire dai dati è un approccio black box, non consideriamo relazioni fisiche ma ci basiamo solo su ingressi e uscite.

Inferenza

Alternativamente, potremmo essere interessati al tipo di relazione tra l'output Y e gli input X .

Ad esempio:

- quali specifici predittori hanno davvero effetto sulla risposta? Se abbiamo 50 variabili X non è detto che tutte e 50 partecipano significativamente nel determinare Y.
- quali componenti del vettore $X = (X_1, \dots, X_p)^T$ sono importanti per spiegare Y e quali sono rilevanti?
- la relazione è positiva o negativa?
- la relazione è lineare o più complessa?

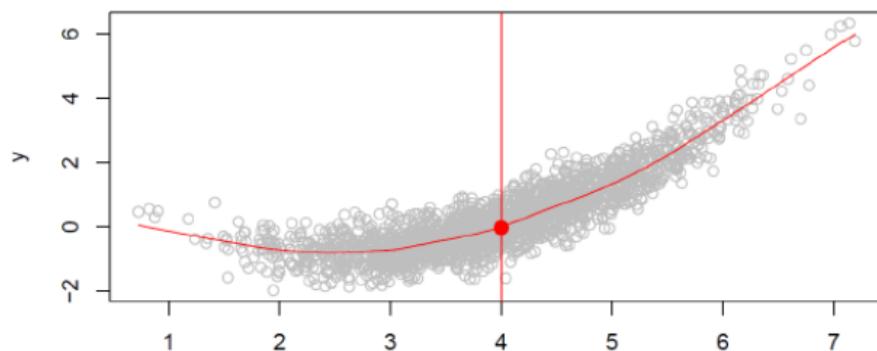
Un modello costruito per inferenza poi può essere usato per predizione.

Inferenza, esempio: abitazioni

Desiderare di prevedere il prezzo mediano delle case in base a diverse variabili, come il tasso di criminalità, la distanza dal mare, la qualità dell'aria, le scuole, il livello di reddito della comunità, la dimensione delle case e così via.

- Potrebbe essere interessante capire quali fattori hanno il maggiore impatto sulla risposta e quanto è grande l'effetto.
- Ad esempio, quanto impatto ha una vista sul mare sul valore della casa, ecc.
- In alternativa, si potrebbe voler prevedere il valore di una casa date le sue caratteristiche. Questo è un problema di **previsione**.

Qual è $f(X)$



Qual è un buon valore di $f(X)$ per ogni specifico valore X, ad esempio $X=4$? Possono esserci molti valori Y corrispondenti ad $X=4$. Un buon valore è:

$$f(4) = E(Y|X=4)$$

$E(Y | X = 4)$ è il **valore atteso** di Y dato $X=4$.



La funzione $f(x) = E(Y | X=x)$ è chiamata **funzione di regressione**, è anche definita per il vettore X , è la funzione data dal valore atteso di Y dato un certo valore di X .



La funzione $f(x)$ è il **predittore ideale o ottimale di Y** per quanto riguarda l'errore quadratico atteso di predizione (Expected squared Prediction Error), infatti $f(x) = E(Y | X=x)$ è la funzione che minimizza

$$EPE(g) = E[(Y - g(X))^2 | X = x]$$

per tutte le funzioni g in tutti i punti $X=x$.

Y è il valore vero e $g(X)$ è la stima calcolata da noi

Esiste un **errore irriducibile** $\epsilon = Y - f(x)$, anche se conoscessimo $f(x)$ faremmo comunque errori di predizione, visto che per ogni $X = x$ c'è tipicamente una distribuzione di possibili valori di Y (stima intervallare?).

Definiamo \hat{f} come una stima di f dati i dati:

$$\mathcal{D} = \{(y_i, x_i)\}_{i=1}^n$$

Poi per ogni dato test data point (y_0, x_0) , abbiamo:

$$E[(y_0 - \hat{f}(x_0))^2] = \underbrace{E[f(x_0) - \hat{f}(x_0)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}$$

Come si stima f ?

Assumiamo di aver osservato un insieme di dati di training:

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

Vogliamo utilizzare il training data set e il metodo statistico per stimare f .

Ci sono **due tipi** importanti di metodi di statistical learning:

- **Metodi parametrici:** fanno assunzioni sulla forma funzionale di f ;
- **Metodi non-parametrici:** non fanno esplicite assunzioni sulla forma funzionale di f .

Metodi parametrici

Si riduce il problema della stima di f al problema della stima di un insieme di parametri.

Sono basati su un approccio basato sul modello in due step:

- STEP 1:

Si fanno assunzioni sulla forma funzionale di f , cioè si **identifica un modello**.

L'esempio di modello più comune è il modello lineare, specificato da $p+1$ parametri $\beta_0, \beta_1, \dots, \beta_p$:

$$f(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}.$$

Molti modelli potrebbero essere scelti per f , più è flessibile e complesso è il modello scelto e più vicina la f sarà al data set.

- STEP 2:

Si usano i dati di training per fare il fit del modello. Cioè per adattare il modello.

Per stimare f si stimano (equivalentemente) i parametri ignoti come $\beta_0, \beta_1, \dots, \beta_p$.

L'approccio più comune per la stima dei parametri in un modello lineare è l'**ordinary least squares (OLS)**, cioè metodo dei minimi quadrati, ci sono ovviamente altri approcci.

Un modello lineare spesso fornisce una buona, e facilmente interpretabile, prima approssimazione della vera funzione ignota.

Esempio: la stima di una regressione lineare

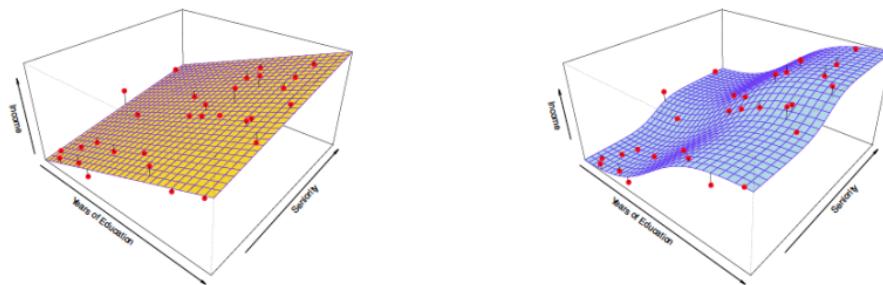
I punti rossi sono valori simulati di Income da

$$\text{Income} = f(\text{Education}, \text{Seniority}) + \epsilon$$

dove f è la superficie blu.

Il piano giallo è il modello lineare adattato (fit) da

$$f = \beta_0 + \beta_1 \times \text{Education} + \beta_2 \times \text{Seniority}$$



L'adattamento lineare (giallo) sembra essere abbastanza ragionevole nel catturare la relazione positiva, ma fallisce nel catturare alcune curvature della vera f (blu).

Metodi non-parametrici

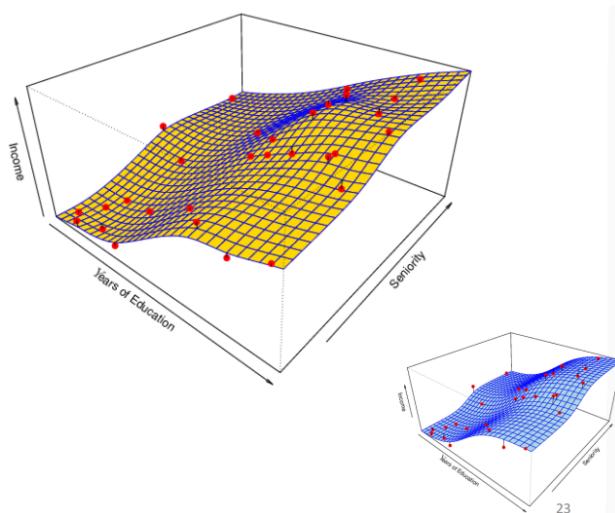
Non fanno assunzioni esplicite sulla forma funzionale di f .

Vantaggi: si adattano (fit) accuratamente ad una gamma più ampia di possibili forme di f .

Svantaggi: un numero molto grande di osservazioni sono necessarie per ottenere una stima accurata di f , in altre parole occorre un dataset molto più grande.

Esempio: una stima Thin-Plate Spline

Una **Thin-Plate Spline (TPS)** è uno strumento matematico molto utilizzato per l'interpolazione smooth e il warping ad alta dimensionalità. È particolarmente utile per trasformare un insieme di punti in un altro, minimizzando la distorsione.



I metodi di regressione non lineare sono più flessibili e possono potenzialmente fornire stime più accurate.

Trade-off tra accuratezza della predizione e interpretabilità del modello

Perché non usare sempre un metodo più flessibile se è più realistico?

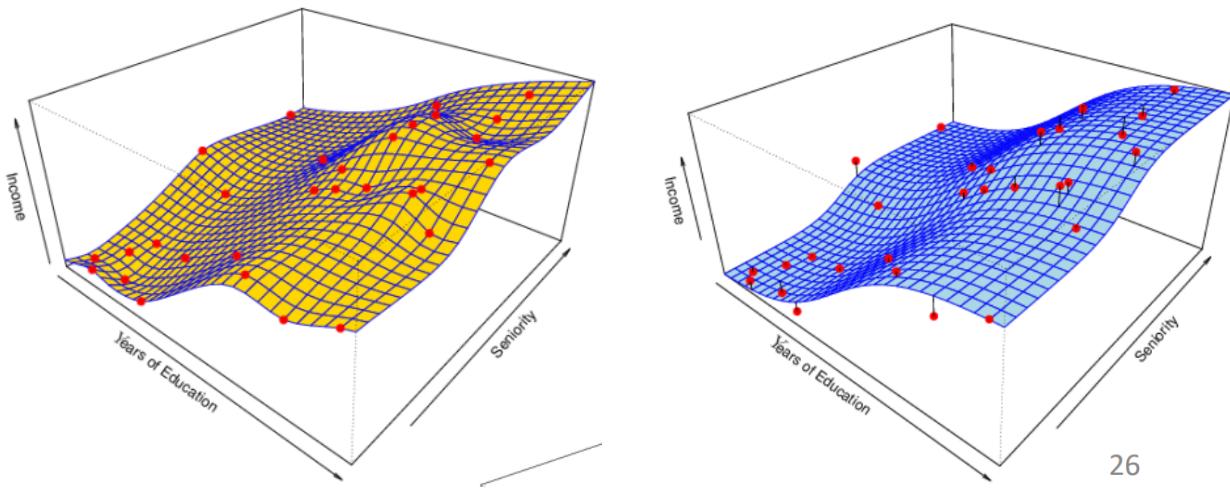
1. Un metodo semplice come una regressione lineare produce un modello che è molto più facile da comprendere (la parte di Inferenza è migliore).

Ad esempio, in un modello lineare

β_j è l'aumento medio di $f(X)$ per un aumento di una unità in X_j mantenendo tutte le altre variabili costanti.

2. Anche se si è solo interessati alla predizione, e quindi il primo punto è irrilevante, è spesso possibile ottenere predizioni più accurate con un modello semplice piuttosto che con uno complesso.

Un esempio di stima di bassa qualità



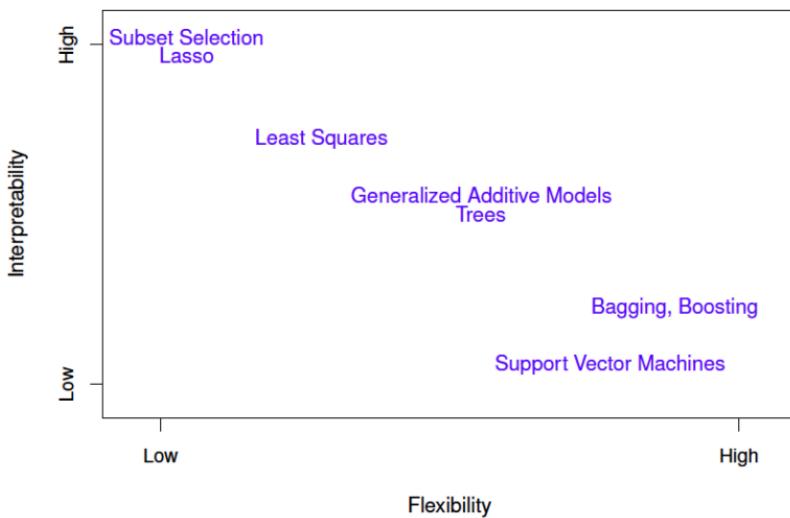
26

Metodi di regressione non lineari possono produrre stime di f di bassa qualità, se il fit ha un livello più basso di smoothness (levigatezza) rispetto alla vera funzione f , visivamente questo si vede perché il fit è molto più variabile della funzione f , questo è noto come over-fitting.

Iper addestrare (**over-fitting**) rispetto al proprio data-set rischia di portare a performance peggiori su un test-set.

Bisogna avere errore piccolo sulla generalizzazione del problema non sullo specifico problema alla mano.

Per ottenere questo l'idea è dividere tutti i dati in nostro possesso in data set e test set e allenare sul data set fino ad ottenere la versione che performa meglio sul test set.



Un modello lineare nel quale sono stati tolti molti regressori inutili è facile da interpretare ma è meno performante. Viceversa un modello più performante risulta meno comprensibile.

Supervised vs. Unsupervised Learning

Possiamo dividere tutti i problemi di apprendimento in situazioni supervisionate e non supervisionate.

APPRENDIMENTO SUPERVISIONATO:

- l'apprendimento supervisionato è quando sia i predittori X_i che la risposta Y_i sono osservati;
- desideriamo predire in maniera accurata dei test cases sconosciuti, comprendere quali input condizionano l'outcome e come, e infine stabilire la qualità delle nostre predizioni e inferenze.

APPRENDIMENTO NON SUPERVISIONATO:

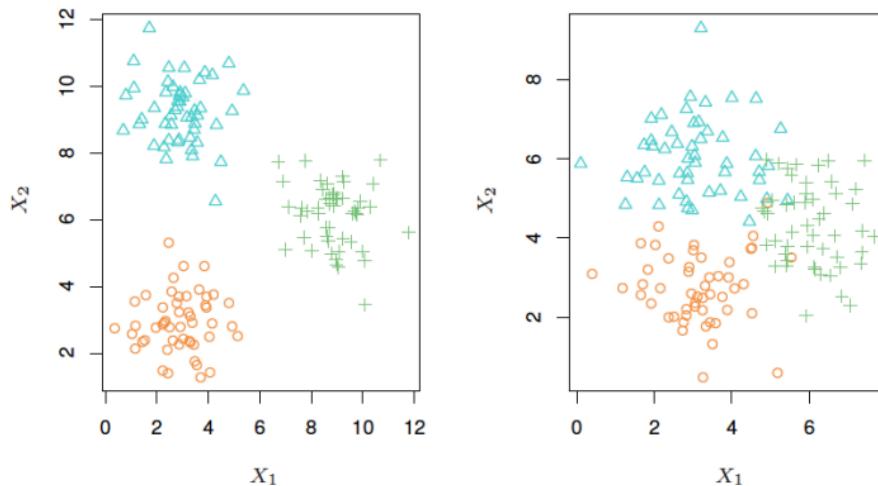
- in questa situazione solo le X_i sono osservate;
- dobbiamo usare le X_i per dedurre quale sarebbe stata la Y se l'avessimo osservata e costruire un modello da lì. (Esempio: segmentazione del mercato, dividere potenziale clienti in gruppi);
- a volte è utile fare unsupervised learning come passo di pre-elaborazione (pre-processing) per il supervised learning;
- alcuni esempi sono il **clustering** e la **principal component analysis** (PCA).

SEMI-SUPERVISED LEARNING:

Date n osservazioni, per $m < n$ osservazioni abbiamo sia misurazioni dei predittori che misurazioni della risposta, mentre per le restanti $n-m$ osservazioni abbiamo solo misurazioni dei predittori. In questi casi sono necessari specifici metodi che non approfondiremo.

Uno scenario di questo tipo può sorgere se i predittori possono essere misurati in maniera molto economica rispetto alle corrispondenti risposte.

Un semplice esempio di Clustering



I tre colori rappresentano il fatto che ci sono 3 leggi diverse che generano quei punti, ma noi all'inizio magari non lo sappiamo, si chiama clustering il processo di identificare questi cluster, questo diventa più difficile quanto più sono sparse le nuvolette. Non c'è Y perché è unsupervised.

Regression vs. Classification

I problemi di apprendimento **supervisionato** possono essere ulteriormente divisi in problemi di regressione e problemi di classificazione.

Le variabili possono essere caratterizzate come quantitative o qualitative (note anche come categorical). Le variabili quantitative assumono valori numerici.

Esempi includono l'età di una persona, l'altezza o il reddito, il valore di una casa e il prezzo di un'azione. Al contrario, le variabili qualitative assumono valori in una delle K diverse classi, o categorie. Esempi di variabili qualitative includono il genere di una persona (maschio o femmina), la marca di un prodotto acquistato (marca A, B o C), se una persona è inadempiente su un debito (sì o no), o una diagnosi di cancro (Leucemia Mieloide Acuta, Leucemia Linfoblastica Acuta, o nessuna leucemia).



Tendiamo a riferirci ai problemi con una risposta quantitativa come **problem di regressione**, mentre quelli che coinvolgono una risposta qualitativa sono spesso chiamati **problem di classificazione**.

Tuttavia, la distinzione non è sempre così netta. La *least squares linear regression* viene utilizzata con una risposta quantitativa, mentre la *logistic regression* è tipicamente usata con una risposta qualitativa (a due classi, o binaria); per questo motivo, viene spesso usata come metodo di classificazione.

Ma poiché stima le probabilità di classe, la logistic regression può anche essere considerata un metodo di *regression*.

Alcuni metodi statistici, come *K-nearest neighbors* e *boosting*, possono essere utilizzati sia nel caso di risposte quantitative che qualitative.

Tendiamo a selezionare i metodi di *statistical learning* sulla base del fatto che la *response* sia quantitativa o qualitativa; cioè, potremmo usare la *linear regression* quando la risposta è quantitativa e la *logistic regression* quando è qualitativa. Tuttavia, se i predittori sono qualitativi o quantitativi è generalmente considerato meno importante.

La **Regressione** riguarda situazioni nelle quali Y è continua/numerica:

- predire il valore di una metrica di performance di un server nell'istante successivo di tempo, raccogliendo diverse misurazioni di features SW o HW;
- predire il valore di una certa casa sulla base di vari input.

La **Classificazione** riguarda situazioni nelle quali Y è **categorical**

Ex: $Y \in \{0,1\}$, $Y \in \{0,1, \dots, M\}$,

$Y \in \{\text{no disease}, \text{disease}\}$, $Y \in \{\text{Up}, \text{Down}\}$

Spesso sono chiamati problemi di decisione:

- il server sarà Up o Down tra 2 ore?

- la mail appena ricevuta è spam o ham (buona)?

Statistical learning vs. Machine learning

Il Machine learning nasce come sottoinsieme dell'intelligenza artificiale.

Lo Statistical learning nasce come sottoinsieme della statistica.

C'è una certa sovrapposizione tra i due, entrambi si concentrano su problemi supervisionati e non ed entrambi beneficiano dei progressi dell'altro (cross-fertilization), tuttavia ci sono differenze:

- il Machine learning è maggiormente concentrato su applicazioni su larga scala e sull'accuratezza delle previsioni, in generale ci si concentra più sulla realizzabilità e gli algoritmi;
- lo Statistical learning enfatizza i modelli e la loro interpretabilità precisione e incertezza, in generale ci si concentra più sulle definizioni matematiche.

Constatare (assessing) l'accuratezza dei modelli

In statistica nessun metodo domina sugli altri su tutti i possibili data set.

Su un particolare data set uno specifico metodo può funzionare meglio degli altri ma altri metodi potrebbero funzionare meglio su un data set simile ma differente, di conseguenza è importante decidere per ogni data set quale metodo può produrre i risultati migliori.

Scegliere l'approccio ottimale è una delle parti più difficili dello statistical learning in pratica.

Misurare la qualità del fit

Supponiamo di aver adattato un modello di regressione a dei dati di training

$$\mathcal{D}_{tr} = \{(x_i, y_i)\}_{i=1}^n$$

e desideriamo misurare le performance del modello adattato.

La misura più comunemente utilizzata a questo scopo è il Mean Squared Error (MSE), cioè l'errore quadratico medio:

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

dove \hat{f} cappello è la stima di f .

Visto che è calcolato usando i dati di training è di solito chiamato training MSE.

Il MSE è spesso utilizzato allo scopo di stimare i parametri del modello, ad esempio nei casi di regressione lineare scegliamo la linea tale da minimizzare l'MSE.

Un problema

L'MSE di training potrebbe non funzionare bene allo scopo di misurare la qualità del fit rispetto a nuovi dati.

Quando compariamo modelli non è importante quanto i modelli funzionano sui dati di training, invece è fondamentale l'accuratezza delle previsioni generate quando i modelli sono applicati a nuovi dati, i dati di test o test data.

Dati i dati di test

$$\mathcal{D}_{te} = \{(x_i, y_i)\}_{i=1}^m$$

l'MSE di test si calcola:

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{f}(x_i))^2$$

Non c'è garanzia che il modello con l'MSE di training più piccolo abbia anche il più piccolo MSE di test.

In generale quanto più un metodo è flessibile (ha più parametri), tanto più basso è l'MSE di training (si dice che si adatta, fit, o che "spiega" i dati di training molto bene).

Training vs Test MSE

I metodi più flessibili (ad esempio le splines) possono generare una gamma più ampia di possibili forme per stimare f rispetto ad altri metodi più restrittivi e meno flessibili (ad esempio la regressione lineare).

I metodi meno flessibili comportano però modelli più facili da interpretare (trade-off tra flessibilità e interpretabilità del modello).

In alcuni problemi la strategia di minimizzare il training MSE è sufficiente, questo quando l'obiettivo è stimare empiricamente i parametri del modello in maniera più accurata possibile.

Se facciamo inferenza possiamo non partizionare i dati in data set e test set e avere tutti i dati come data set sul quale a quel punto possiamo fare over fitting perché tanto non dobbiamo fare previsioni, dobbiamo però ricordarci questa scelta.

In molte applicazioni moderne però non ci importa quanto il metodo funzioni sui dati di training ma ci importa invece l'accuratezza delle previsioni che otteniamo applicando il nostro metodo a dati precedentemente ignoti.

Ad esempio:

- Abbiamo misurazioni cliniche (ad esempio peso, pressione sanguigna, altezza, età, storia familiare di malattie) per un certo numero di pazienti, così come informazioni su se ciascun paziente ha il diabete. Possiamo utilizzare questi pazienti per addestrare un metodo di *statistical learning* al fine di prevedere il rischio di diabete basato sulle misurazioni cliniche.

- In pratica, vogliamo che questo metodo predica accuratamente il rischio di diabete per futuri pazienti, basandosi sulle loro misurazioni cliniche.
- Non siamo molto interessati a sapere se il metodo predice correttamente il rischio di diabete per i pazienti usati per addestrare il modello, poiché sappiamo già quali di questi pazienti hanno il diabete.

Il nostro obiettivo

Supponiamo di adattare, fit, il nostro metodo di statistical learning alle nostre osservazioni di training

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

e otteniamo la stima \hat{f} cappello.

Possiamo calcolare

$$\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_n)$$

e se questi risultano approssimativamente uguali a y_1, y_2, \dots, y_n allora l'MSE di training è piccolo.

In ogni caso non è importante se è vero che

$$\hat{f}(x_i) \approx y_i$$

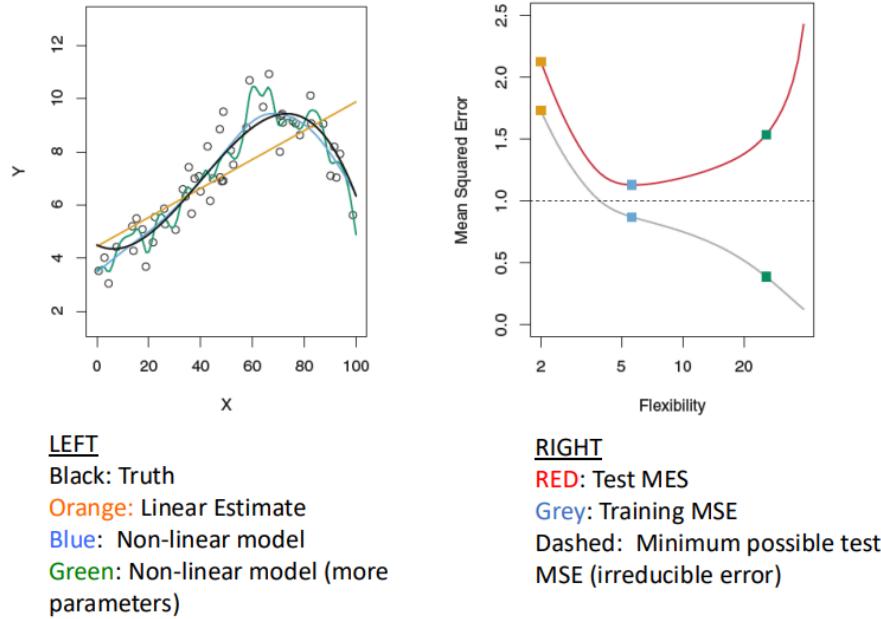
quello che invece è importante è se

$$\hat{f}(x_0) \approx y_0$$

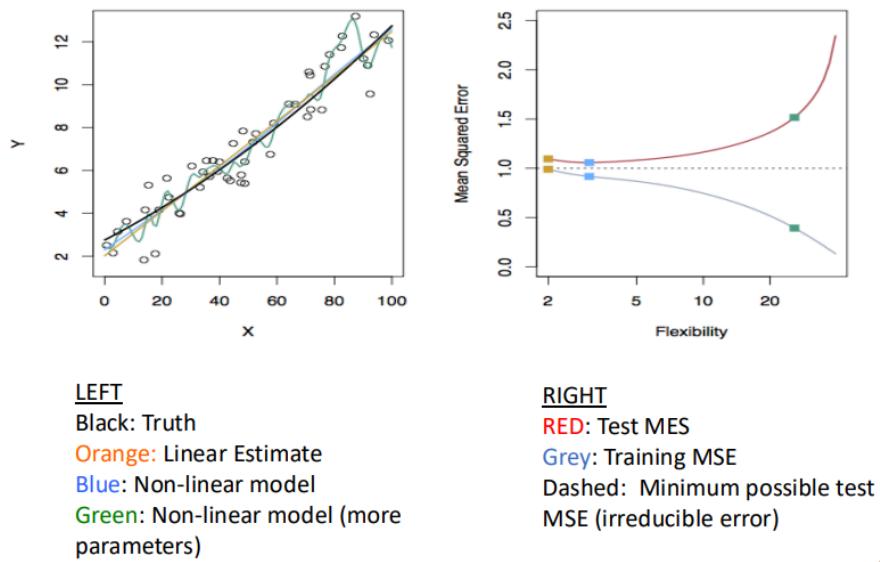
dove (x_0, y_0) è una osservazione di test precedentemente sconosciuta e non usata per fare il training del metodo di statistical learning.

Il nostro obiettivo è **selezione il metodo con il valore minimo di test MSE**.

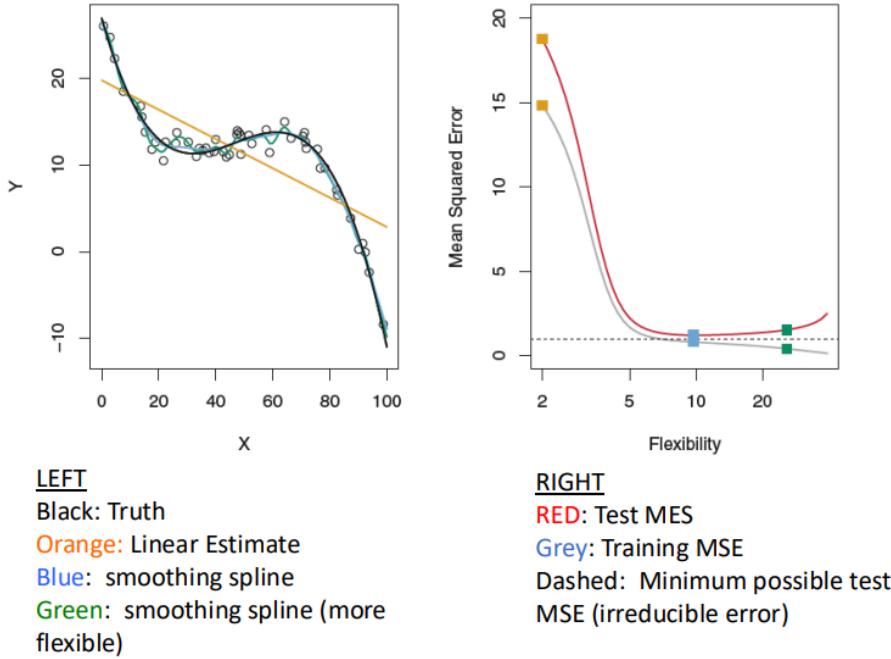
Esempio 1 - Diversi livelli di flessibilità



Esempio 2 - Diversi livelli di flessibilità



Esempio 3 - Diversi livelli di flessibilità



Concludiamo che con l'aumentare della flessibilità del metodo di apprendimento statistico si osserva un decrescere monotono dell'MSE di training e invece una forma ad U nell'MSE di test.

Questa è una proprietà fondamentale al di là dello specifico data set e al di là del metodo statistico utilizzato.

Quando un metodo specifico comporta un piccolo MSE di training ma un grande MSE di test si parla di **overfitting** dei dati.

In pratica, calcolare l'MSE di training è abbastanza semplice ma stimare l'MSE di test è considerevolmente più complesso perché i dati di test non si devono usare per creare il proprio modello.

Un metodo interessante è la cross-validation, che permette la stima dell'MSE di test usando i dati di training.

Il livello di flessibilità corrispondente al modello con l'MSE di test minimo può variare considerevolmente in relazione al data set specifico usato.

Il trade-off tra Bias e Varianza

I grafici precedenti che paragonano l'MSE di training a quello di test illustrano un importante compromesso che governa la scelta di metodi di statistical learning.

Le curve ad U degli MSE di test sono il risultato di due proprietà contrastanti in termini di capacità di stima/predizione di un qualsiasi metodo di learning, queste due proprietà sono chiamate **bias** e **varianza**.

Supponiamo di aver adattato, fit, un modello

$$\hat{f}(x)$$

a dei dati di training D_{tr} , e supponiamo che (x_0, y_0) sia una osservazione di test presa dalla popolazione (in data science, la popolazione è l'insieme completo di dati o elementi di interesse su cui si vogliono fare analisi.).

Se il vero modello è

$$Y = f(X) + \epsilon, \quad E(\epsilon) = 0$$

allora

$$E(y_0 - \hat{f}(x_0))^2 = \underbrace{[\text{Bias}(\hat{f}(x_0))]^2}_{\text{reducible}} + \text{Var}(\hat{f}(x_0)) + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible}},$$

dove l'aspettativa (penso intenda la media) è relativa a y_0 e y_1, \dots, y_n ed è vero che

$$\text{Bias}(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0)$$

L'MSE di test previsto è definito come:

$$E(y_0 - \hat{f}(x_0))^2$$

Il compromesso significa che se un metodo diventa più flessibile il bias diminuisce e la varianza aumenta ma l'MSE di test atteso potrebbe salire o scendere.

Tendenzialmente l'idea è che se il bias viene ridotto aumenta la varianza e viceversa.

Bias dei metodi di learning



Il bias è riferito all'errore che è introdotto dalla modellazione di problemi della vita reale, che sono solitamente molto complessi, tramite l'adozione di modelli molto più semplici.

Tanto più un metodo è flessibile e complesso tanto più sarà piccolo il bias.

Il bias è la misura di quanto differisce il modello scelto dal modello reale.

Varianza dei metodi di learning



La varianza si riferisce a quanto la stima di f , cioè \hat{f} , cambierebbe se si stesse usando un diverso data set per il training.

Visto che i dati di training sono usati per adattare, fit, il metodo di statistical learning, diversi data set di training risulteranno in differenti \hat{f} .

Un metodo più flessibile avrà più varianza, cambiare uno qualsiasi dei punti potrebbe causare un cambiamento considerevole della \hat{f} .

In contrasto, la least squares line (retta dei minimi quadrati, è un metodo per adattare un modello lineare ai dati minimizzando la somma dei quadrati delle differenze tra i valori osservati e i valori previsti dalla retta) è abbastanza inflessibile e ha una varianza bassa, infatti cambiare una osservazione comporterà probabilmente un cambiamento piccolo nella posizione della retta.

Bias vs. Variance

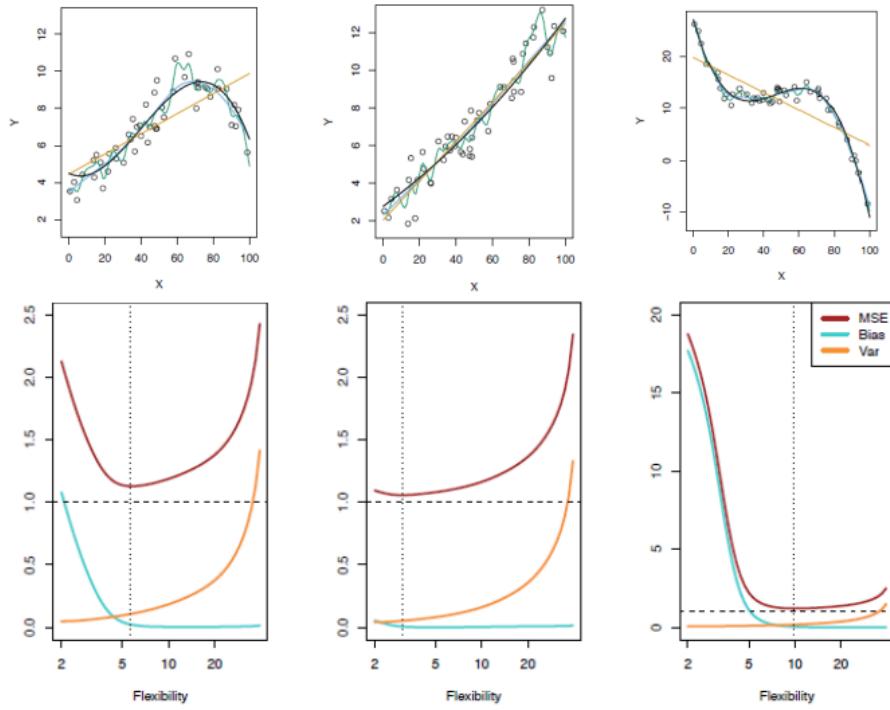
$$E(y_0 - \hat{f}(x_0))^2 = \underbrace{\left[\text{Bias}(\hat{f}(x_0)) \right]^2}_{\text{Reducible}} + \text{Var}(\hat{f}(x_0)) + \text{Var}(\epsilon)$$

Come regola, questa espressione significa che se la complessità del modello aumenta allora il bias si ridurrà ma la varianza aumenterà.

Al fine di minimizzare l'expected test error (l'errore di test atteso), è necessario selezionare metodi di statistical learning che ottengano simultaneamente varianza e bias bassi.

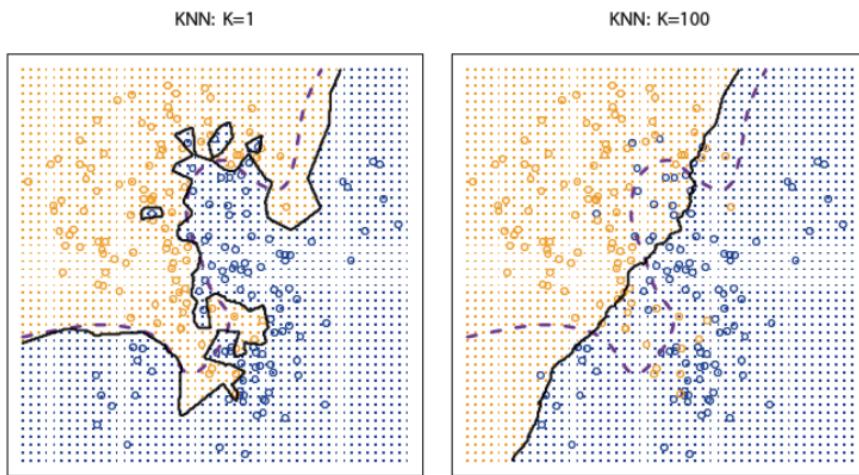
Si nota che la varianza è, innatamente, non negativa e anche il bias al quadrato è non negativo, di conseguenza l'MSE di test atteso non può mai scendere al di sotto di $\text{Var}(\epsilon)$ che è quindi detto errore irriducibile.

Test MSE, Bias e Varianza

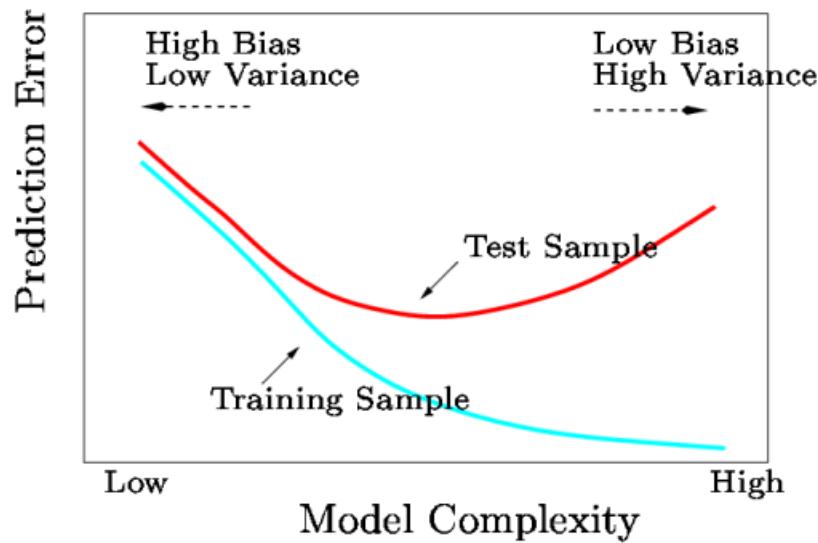


Il compromesso tra bias è varianza esiste per tutti i metodi di learning.

Un discorso simile è valido per i problemi di classificazione:



Un grafico fondamentale



In generale gli errori di training si riducono sempre mentre gli errori di test calano all'inizio (quando la riduzione del Bias è il fattore dominante) ma poi iniziano a salire di nuovo (quando l'aumento della varianza diventa il fattore dominante).

DISCLAIMER

Questi appunti sono stati realizzati a scopo puramente educativo e di condivisione della conoscenza. Non hanno alcun fine commerciale e non intendono violare alcun diritto d'autore o di proprietà intellettuale.

I contenuti di questo documento sono una rielaborazione personale di lezioni universitarie, materiali di studio e concetti appresi, espressi in modo originale ove possibile. Tuttavia, potrebbero includere riferimenti a fonti esterne, concetti accademici o traduzioni di materiale didattico fornito dai docenti o presente in libri di testo.

Se ritieni che questo documento contenga materiale di tua proprietà intellettuale e desideri richiederne la modifica o la rimozione, ti invito a contattarmi. Sarò disponibile a risolvere la questione nel minor tempo possibile.

In quanto autore di questi appunti non posso garantire l'accuratezza, la completezza o l'aggiornamento dei contenuti e non mi assumo alcuna responsabilità per eventuali errori, omissioni o danni derivanti dall'uso di queste informazioni. L'uso di questo materiale è a totale discrezione e responsabilità dell'utente.