Statistical learning and deep learning: theoretical background and hands-on sessions

Lezione 1

S. Biffani & F. Biscarini

IBBA/CNR

27 gennaio, 2023

Outline del corso - Monday 30 (S. Biffani)

Morning: 10:00 - 13:00

- Session 1: Statistical Learning
- Inference & Prediction
- Assessing Model Accuracy
- The Bias-Variance Trade-Off (Error Decomposition and Simulation)

Afternoon: 14:00 - 17:00

- Session 2: Cross-Validation
- The Validation Set Approach
- Leave-One-Out Cross-Validation
- k-Fold Cross-Validation (Bias-Variance Trade-Off for k-Fold)
- Exercises (tidymodels)

Bibliografia consigliata, figure ed esempi: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Outline del corso - Tuesday 31 (F. Biscarini)

Morning: 10:00 - 13:00

- Introduction to deep learning [theory]
- Anatomy of neural networks
- Forward and back propagation
- Dense neural networks for classification problems
- Binary and multi-class classification with neural networks [hands-on]

Afternoon: 14:00 - 17:00

- Deep learning for image recognition [theory]
- Convolutional Neural Networks (CNN)
- Overfitting and regularization with deep learning
- A DNN (deep neural network) model for image recognition [hands-on]

Modelli predittivi:

Figura 1: Pubmed Search: machine learning

Pubblicazioni negli ultimi 10 anni:

- 2012: 645
- 2017: 2219
- 2022: 13349

Statistical Learning

Statistical Learning: insieme di strumenti utili a *capire i dati*

- *supervised* : costruisco un modello predittivo per stimare un determinato **output** utilizzando una serie di **inputs**
- *unsupervised*: costruisco un modello **solo** sulla base di alcuni **inputs**. Questo mi permetterà di *scoprire* la struttura dei dati ed i loro eventuali rapporti

Statistical Learning vs Machine Learning

- *Machine Learning* nasce come settore dell'Intelligenza Artificale
- *Statistical Learning* nasce come settore della Statistica
- I due termini si *sovrappongono* anche se:
  - **ML**: maggiore enfasi su applicazioni di **grande scala** sull'**accuratezza** delle predizioni
  - **SL**: maggiore enfasi sui modelli e sulla loro **interpretazione** dei modelli
  - la distinzione tra i due è ora molto *sottile*

Qualche riferimento storico

- inizio '900: *least squares* precursore della ben nota *regressione lineare*
- 1940: regressione logistica (variabili *qualitative*)
- 1970: modelli lineari generalizzati (casi speciali regressioni lineari e logistiche)
- 1980: metodi non lineari (CART, GAM, NN)
- 1990: Support vector Machines
- . . . oggi: Convolutional Neural Network (Deep Learning)

Statistical Learning - Supervised

Un esempio di *supervised* SL:

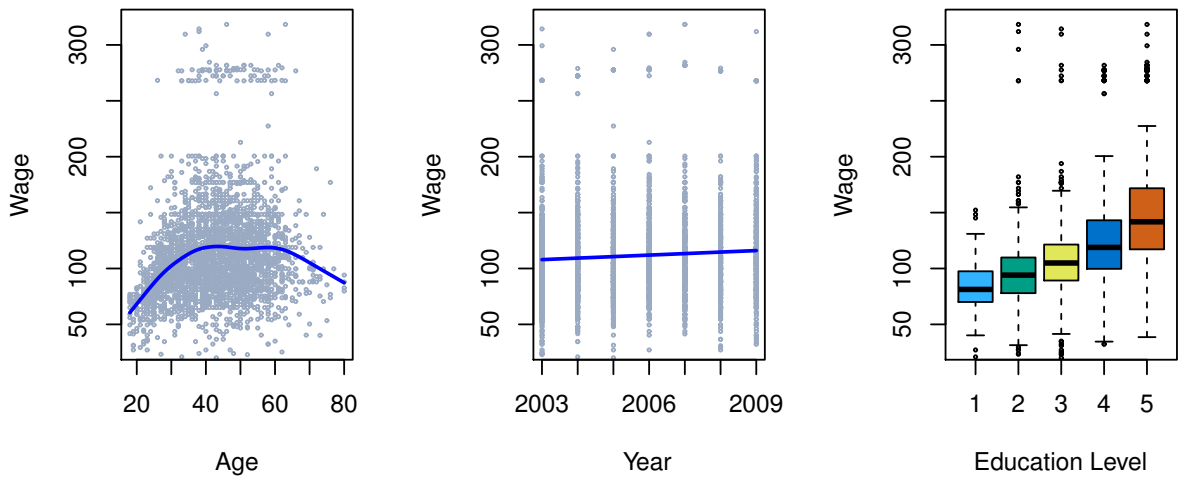


Figura 2: La retribuzione in funzione dell'età, del livello di educazione e dell'anno

Statistical Learning - Unsupervised

Un esempio di *unsupervised* SL:

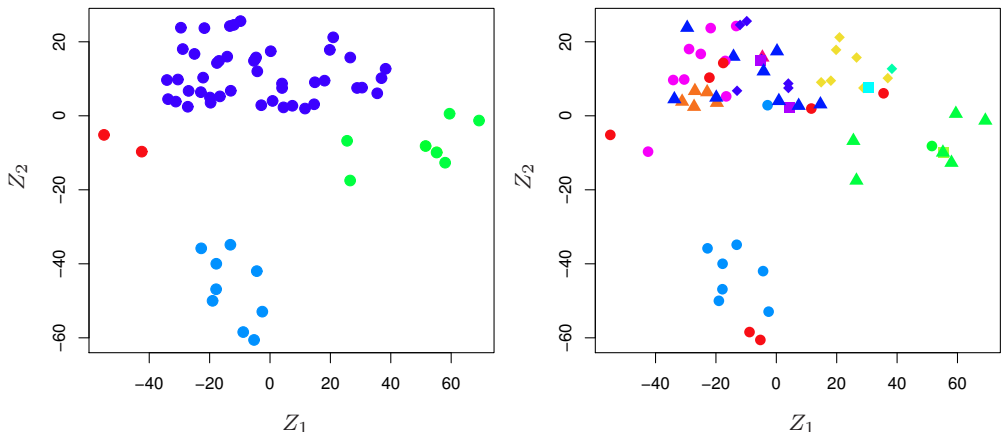


Figura 3: Il rapporto tra diversi tipi di cellule tumorali attraverso dati di espressione genica

a sx le 64 linee diverse di cellule tumorali che si aggregano in 4 *cluster* principali

a dx il rapporto tra linee e tipi di tumori (14)

Statistical Learning - Concetti di base

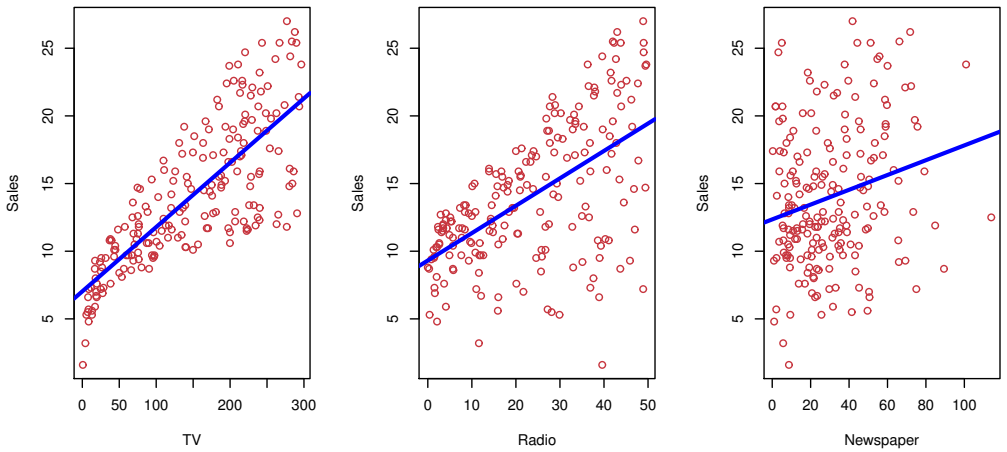


Figura 4: Rapporto tra la vendita di un prodotto (Y) e la spesa sostenuta nei diversi mass media (X) utilizzati

- c'è un legame tra le vendite e le spese sostenute per reclamizzarlo?
- possiamo sviluppare un modello che predica il ricavo sulla base dei 3 budgets di spesa?

Statistical Learning - Concetti di base

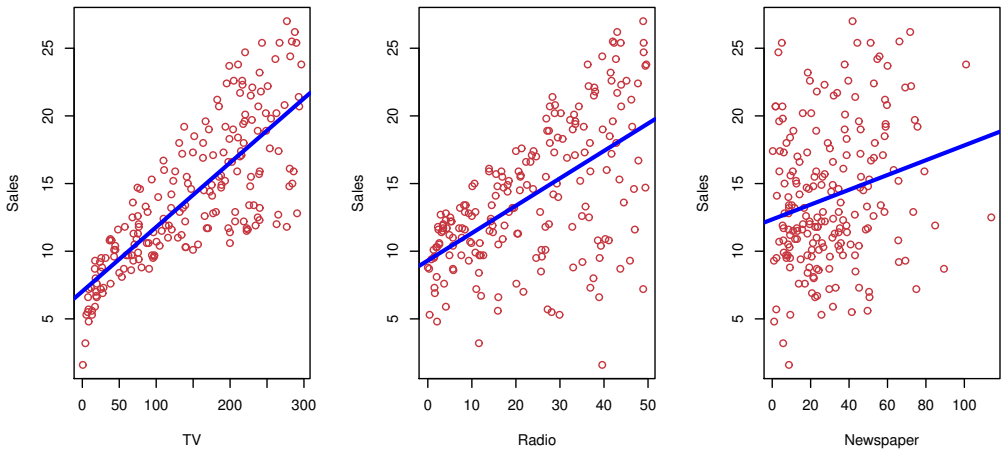


Figura 5: Rapporto tra la vendita di un prodotto (Y) e la spesa sostenuta nei diversi mass media (X) utilizzati

- inputs: Mass Media (X_1, X_2, X_3) - variabile indipendente, predittori, features
- outputs: ricavo (Y) - variabile dipendente o di risposta (**continua** in questo caso)

Statistical Learning - Concetti di base

In termini più generali:

- avendo una variabile di risposta ($\mathbf{Y}$) ed una serie di *predittori* ($\mathbf{X}$), noi assumiamo che esista una relazione tra Y e X_n che può essere cosi scritta:

$$Y = f(X) + \epsilon,$$

dove,

- f = funzione non nota che relaziona $X_1, ..., X_p$
- ϵ = errore casuale con media = 0 e non correlato ad X

f è il mezzo attraverso il quale X fornisce informazioni (stima/predice) Y

Statistical Learning - Concetti di base

Guardiamo la figura a sinistra :

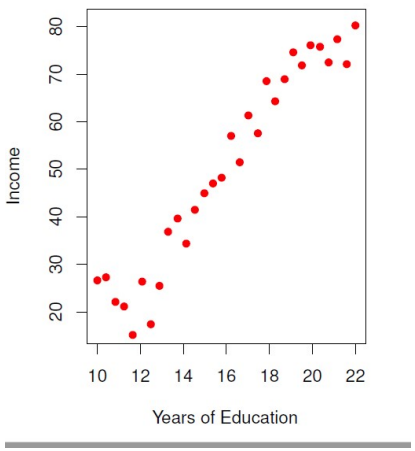


Figura 6: Rapporto tra anni di studio (X) e reddito (Y)

- Cogliamo che esiste una relazione tra gli anni di studio ed il reddito
- ma quale sarà la funzione (inizialmente ignota) che a partire dal reddito osservato (Y) per anno di istruzione (X) lo predice meglio?

Statistical Learning - il ruolo

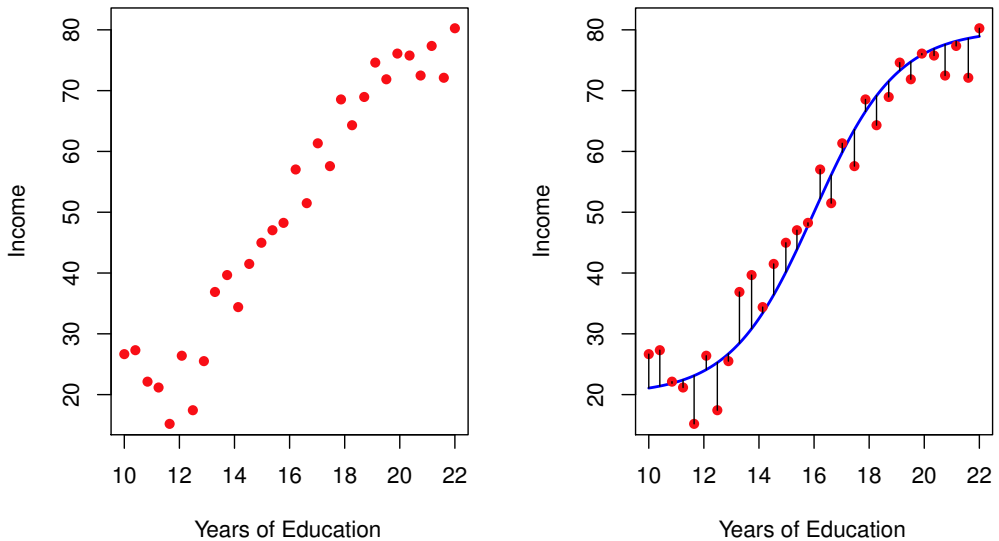


Figura 7: Rapporto tra anni di studio (X) e reddito (Y)

SL si riferisce a tutti quegli approcci (più o meno complessi, parametrici o non parametrici) che mi permettono di stimare la *migliore f*

Inferenza o Predizione ? Inferenza e Predizione

Ci sono due motivi principali per i quali vogliamo stimare f:

- per fare *predizione*
- per fare *inferenza*

Vediamo come e se le possiamo *distinguere*

Fare predizione 1

Situazione comune: abbiamo un insieme di *inputs* disponibili ma l'*output* (Y) non è facilmente ottenibile.

In questa situazione posso ottenere una *stima*/*predizione* di Y:

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

dove $\hat{f}$ è una stima di f e $\hat{Y}$ la predizione di Y

Non mi interessa come è fatta f ma come riesco a predire Y

e.g. la scatola su un mobile

Fare predizione 2

$\hat{Y}$ è una stima della vera Y e la sua accuratezza dipende da due *quantità*:

1. il cosiddetto *reducible error*: conseguenza diretta del fatto che $\hat{f}$ è a sua volta una stima della vera f. Questo errore è *riducibile* identificando la metodologia migliore per stimare f (uno degli obiettivi dello SL)

- cosa uso per raggiungere la scatola sul mobile? un libro, una sedia, un tavolo, una scala

Fare predizione 2

$\hat{Y}$ è una stima della vera Y e la sua accuratezza dipende da due *quantità*:

2. ed il cosiddetto *irreducible error*: l'Y che vogliamo stimare può contenere a sua volta un errore (e.g. variazione nella misurazione causata da evento inatteso e non misurabile o misurato) che però non può essere stimato da X $(Y = f(X) + \epsilon)$

- qualcuno o qualcosa mi sposta la scatola mentre salgo

Fare predizione 3

L'errore che commetto (sostituendo) può essere cosi espresso:

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2$$

ed essendo ϵ una costante:

$$E[f(X) + \epsilon - \hat{f}(X)]^2 = E[f(X) - \hat{f}(X)]^2 + Var(\epsilon)$$

`Reducible` `Irreducible`

L'obiettivo è minimizzare l'*errore riducibile*

Inferenza

In alcune situazioni l'obiettivo può non essere necessariamente predire Y sulla base di $X_1, ..., X_n$ ma piuttosto capire:

1. la relazione tra X e Y
2. come Y cambia in funzione di X

- quale predittore è associato alla risposta (feature selections)?
- che tipo di relazione esiste tra la risposta ed i vari predittori?
- la relazione è lineare o più complessa?

Posso rispondere a queste domande solo se conosco tutti i dettagli di $\hat{f}$

Inferenza e Predizione

In realtà i 2 aspetti sono spesso sovrapponibili e direttamente legati al metodo di valutazione:

- i metodi lineari (più semplici) sono più interpretabili a scapito dell'accuratezza
- i metodo non-lineari raggiungono elevati livelli di accuratezza a scapito dell'interpretabilità

Come stimare f ?

Partiamo sempre da un cosidetto *training data set*

$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ che usiamo per trovare la migliore $\hat{f}$ tale che

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

per stimare $\hat{f}$ possiamo usare metodi

1. parametrici
2. non parametrici

Come stimare f - metodi parametrici

I metodi parametrici si basano su 2 step consecutivi:

1. ipotizziamo una certa forma di f (e.g. funzione lineare)

$$f(X) = \beta_0 + \beta_1 X_1 ... + \beta_p X_p$$

e ne stimiamo i parametri dei quali conosciamo il numero a priori (p+1)

2. quindi testiamo il nostro modello sui *train data* stimando i parametri necessari

$$Y \approx \beta_0 + \beta_1 X_1 ... + \beta_p X_p$$

Come stimare f - metodi parametrici

il termine *parametrico* si riferisce proprio al fatto di stimare f attraverso la stima di una serie di *parametri*

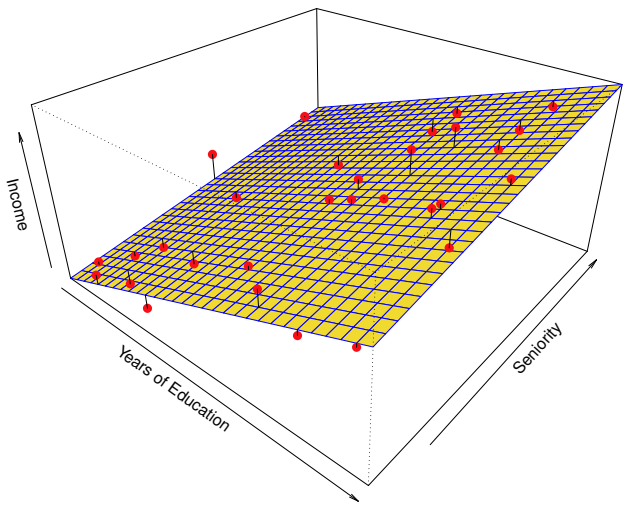


Figura 8: Stima del reddito con un modello lineare

metodo *flessibile*: posso aumentare il numero di parametri

rischio **overfitting**

Come stimare f - metodi non parametrici

I metodi *non parametrici* non fanno assunzioni sulla forma iniziale della f ma usano i dati (...e dovrebbero essere tanti...) per identificarla

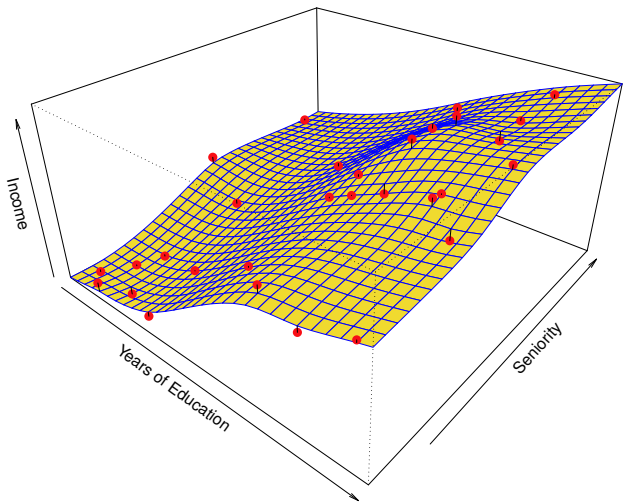


Figura 9: Stima del reddito con un modello non parametrico

overfitting

Quale scegliere?

la risposta dipende da molte variabili ma:

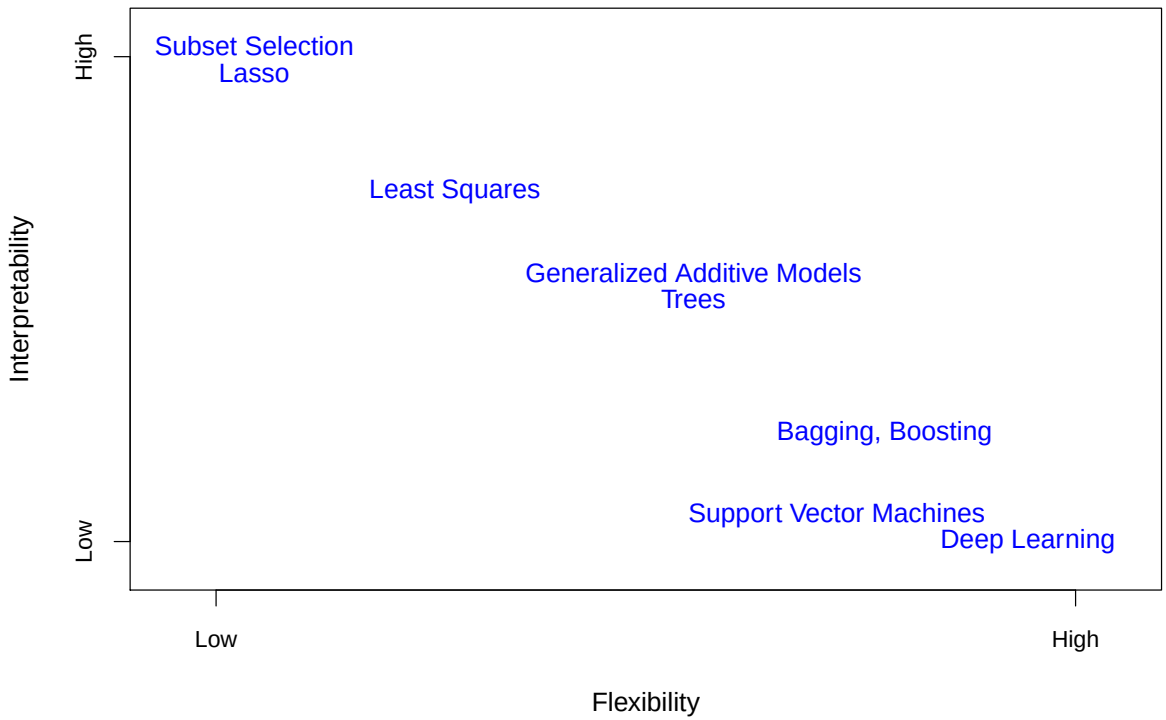


Figura 10: metodi parametrici o metodi non parametrici?

Valutare l'accuratezza di un modello 1

Necessità di un parametro che quantifichi la *distanza* tra la mia stima ed il vero valore (...che ricordiamo può contenere un errore non stimabile...)

Nel caso più semplice della regressione lineare:

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

che viene di solito stimato sulla base dei nostri *train data*

Valutare l'accuratezza di un modello 2

domanda: mi serve stimare un errore su qualcosa che conosco già? (e.g. il prezzo del gas nei mesi scorsi)

oppure mi interessa di più stimare il prezzo del gas nei mesi che verranno e quindi:

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

dove y_0 e x_0 sono output e input futuri e $\hat{f}$ è la funzione che ho stimato sui *train data*

Se ho dei dati di test posso farlo...ma conosco già il prezzo del gas futuro e ho dati relativi a questo??

Valutare l'accuratezza di un modello 3

. . . no, non li ho, quindi posso usare solo i *train data* per stimare il mio *Mean Square Error*

Sembra logico ma. . . non è detto che il *train MSE* minimizzi anche il *test MSE*

Vediamo perchè:

Valutare l'accuratezza di un modello 4

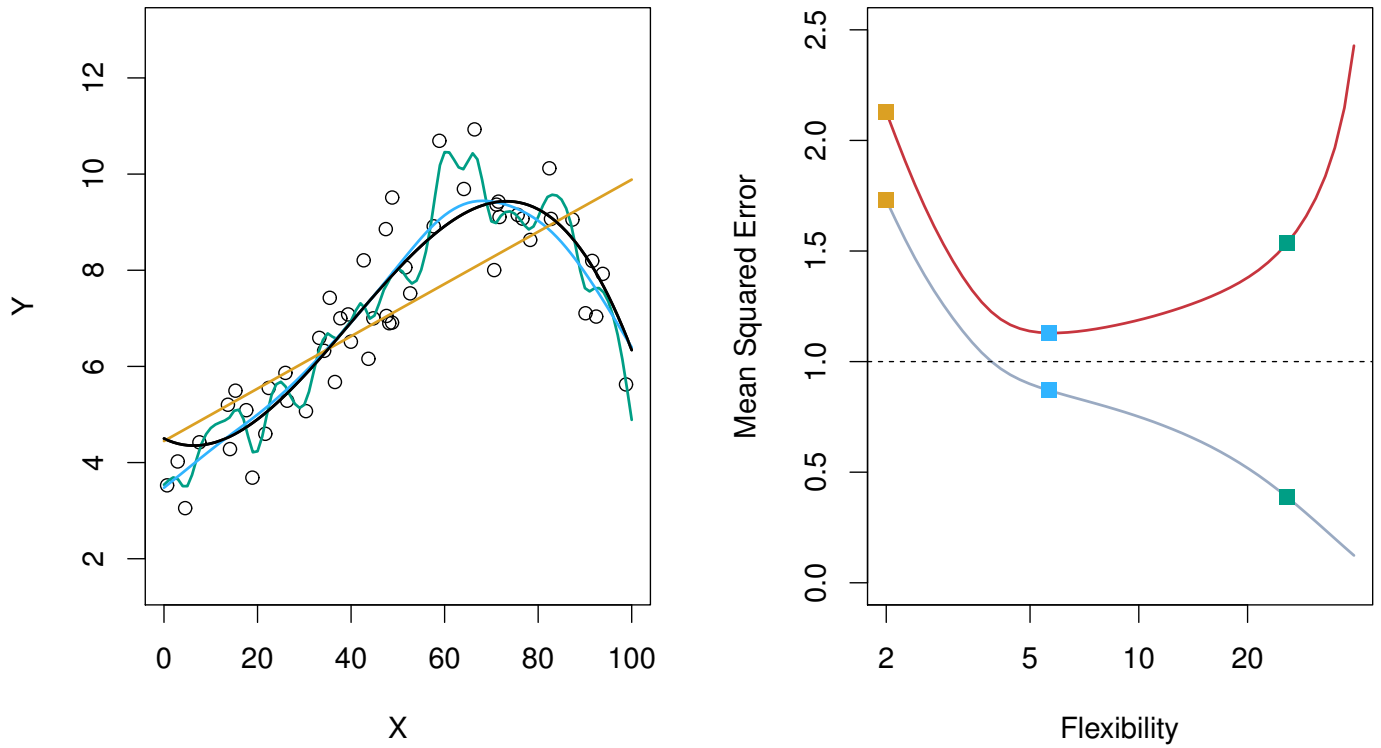


Figura 11: SINISTRA: Dati Simulati con f (nero). regressione lineare (arancio), 2 smoothing splines (Blu e Verde), DESTRA: Training MSE (grigio), test MSE (rosso)

Valutare l'accuratezza di un modello 5

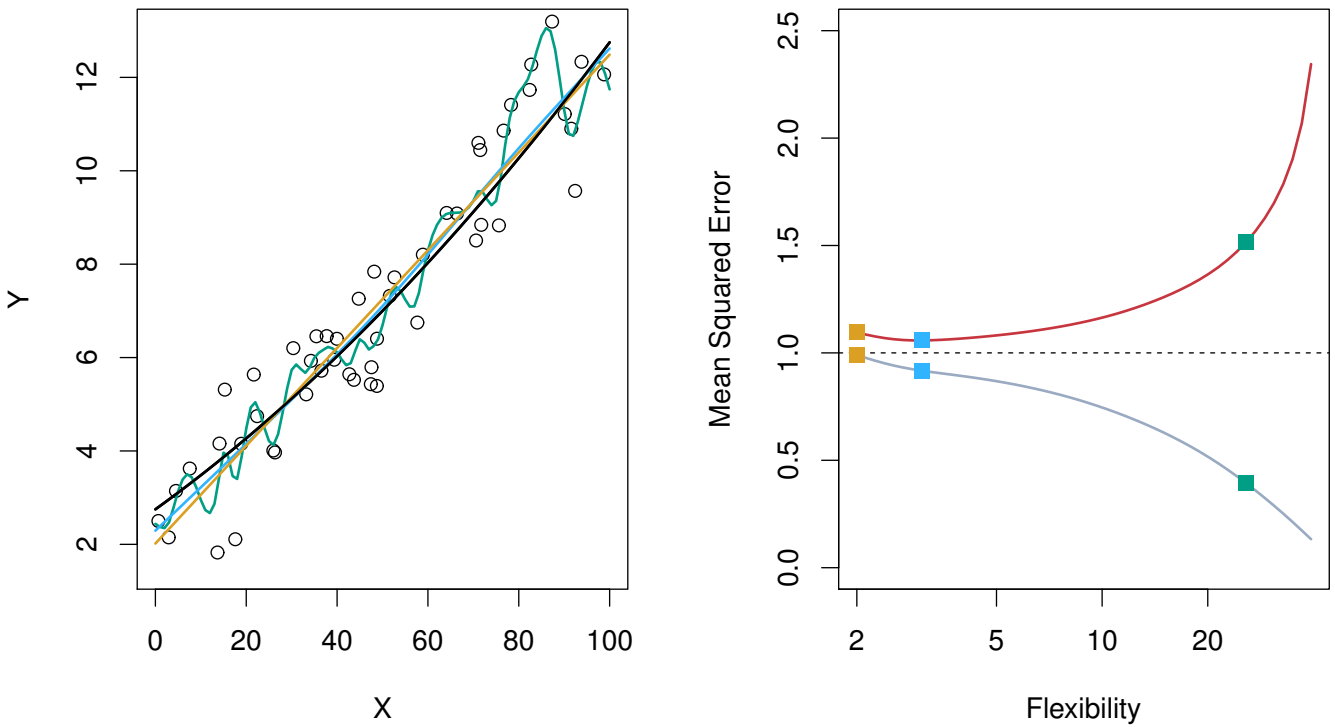


Figura 12: SINISTRA: Dati Simulati con una f più semplice

Valutare l'accuratezza di un modello 6

Considerazioni importanti:

- Un modello più flessibile può portare a valori di *training MSE* molto ridotti, ma lo stesso non vale per il *test MSE*
- la **U-shape** del *test MSE* (fig 11) è una proprietà tipica del Statistical Learning
- basso *training MSE* + elevato *test MSE* = **overfitting**
- come stimare il *test MSE* senza avere dei *veri* dati di testing?

$$cross - validation$$

Il compromesso tra bias e varianza 1

da dove si origina **U-shape** del *test MSE* ?

il *test MSE* dipende sempre da 3 componenti:

1. la *varianza* della mia stima $Var(\hat{y}_0) = Var(\hat{f}(x_0))$
2. il quadrato dell'errore o bias che commetto usando $\hat{f}(x_0)$
3. la varianza dell'errore *irriducibile* $Var(\epsilon)$

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

Cosa ci dice questa formula?

Il compromesso tra bias e varianza 1

Cosa intendiamo per *varianza* e *bias*?

Ricordiamoci che ci sono 2 fattori che influenzano l'accuratezza della mia predizione:

1. i dati usati per stimare il mio modello (creano *varianza*)

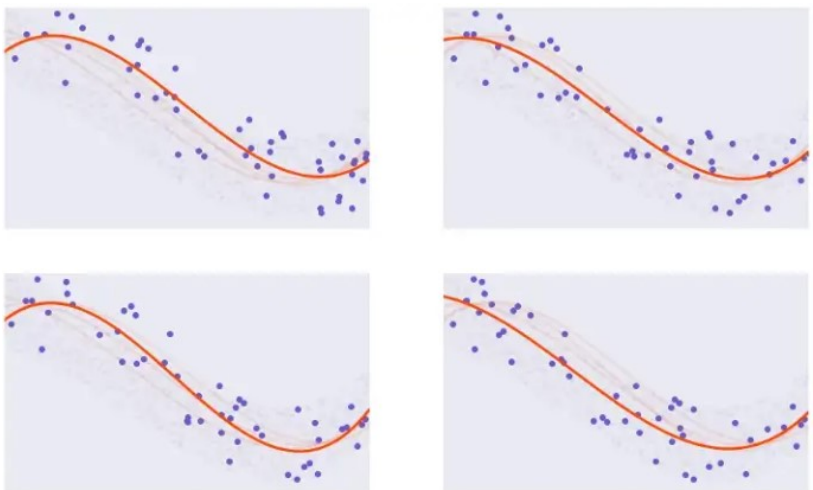


Figura 13: Stesso modello ma dati diversi

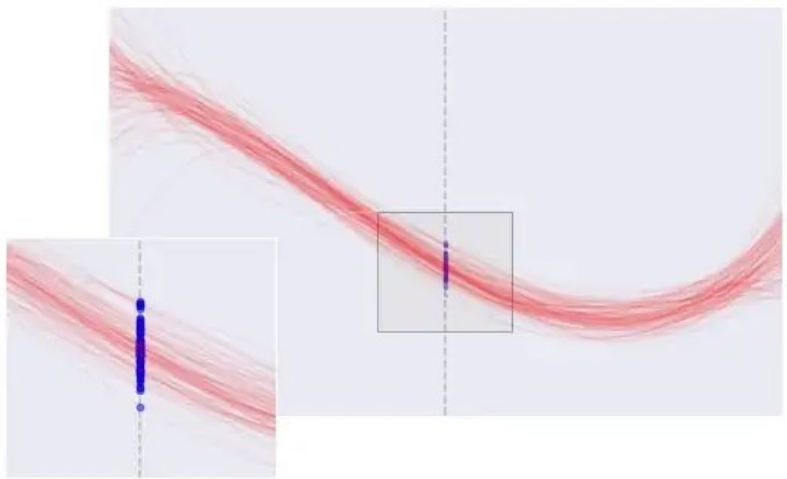


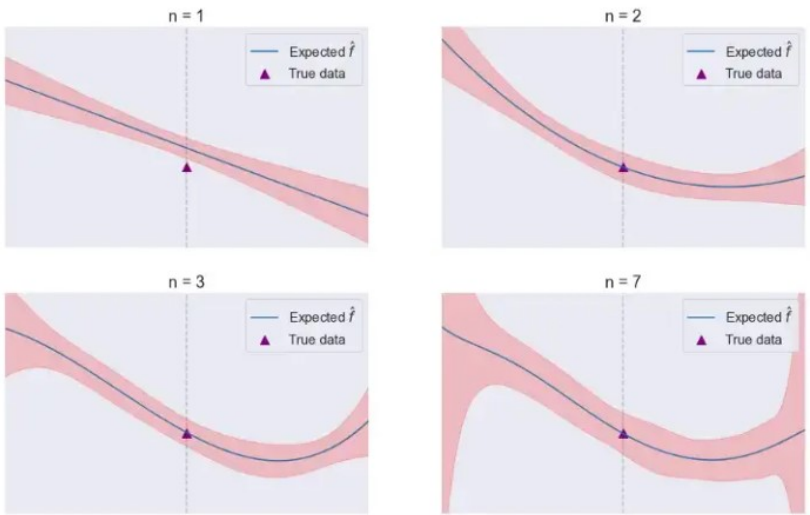
Figura 14: Stime ottenute dai diversi dataset

Il compromesso tra bias e varianza 2

Cosa intendiamo per *varianza* e *bias*?

Ricordiamoci che ci sono 2 fattori che influenzano l'accuratezza della mia predizione:

2. il modello usato sui dati (genera *bias*)



Distanza tra il vero valore (triangolo) e la stima (linea blu). L'area rosa identifica la varianza delle stime

Un fattore influenza l'altro

Il compromesso tra bias e varianza 3

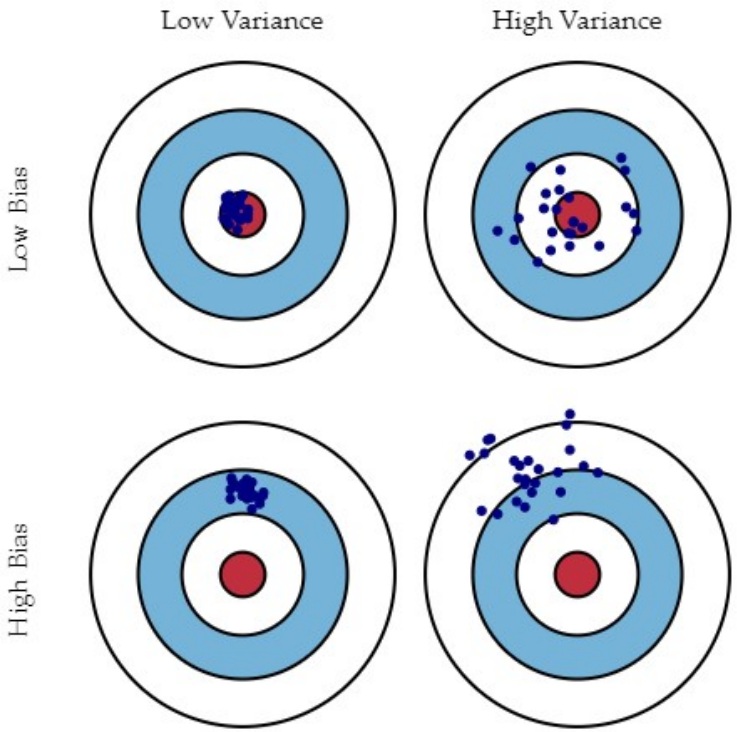


Figura 15: Illustrazione grafica del compromesso Bias-Varianza

tratto da http://scott.fortmann-roe.com/docs/BiasVariance.html

Il compromesso tra bias e varianza 4

Ricapitolando:

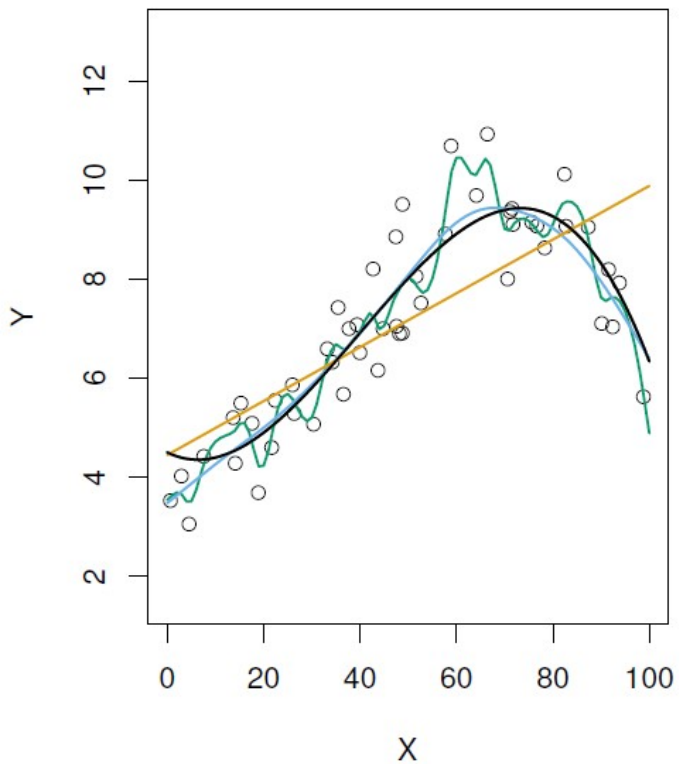


Figura 16: Varianza

- La *varianza* mi dice di quanto $\hat{f}$ cambia in funzione dei dati.
- *Modelli più flessibili hanno maggiore varianza*
- la curva verde ha molta *varianza* (cambia un punto e cambia la curva)
- la linea arancione ha poca varianza (al cambiare dei punti rimane stabile)

Il compromesso tra bias e varianza 5

Ricapitolando:

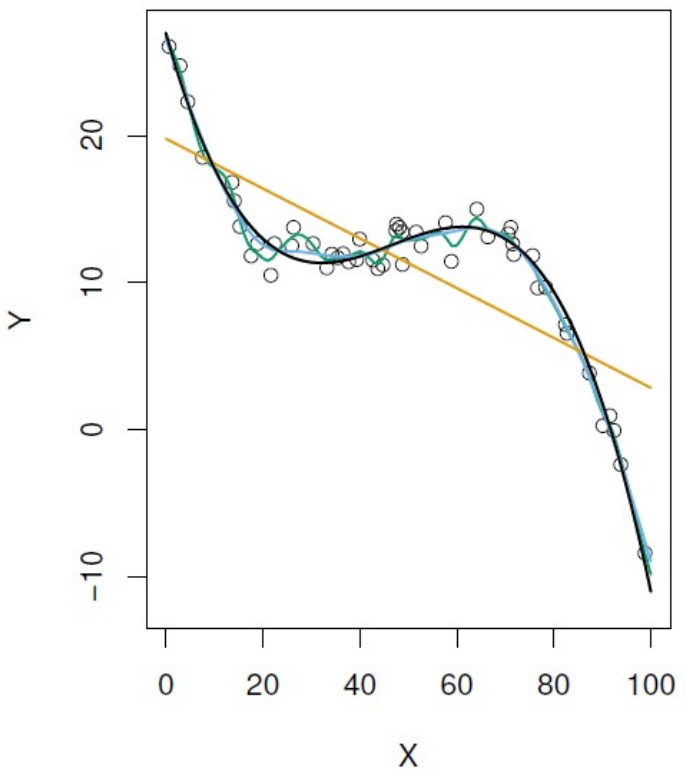


Figura 17: Bias

- Il *bias* mi dice quale è l'errore medio che commetto nell'approssimare la realtà con un'approssimazione ($\hat{f}$)
- il vero andamento dei dati è sostanzialmente non lineare (curva nera)
- la regressione lineare (linea arancione) non avrà mai una buona predizione, neanche aumetando i dati di *training*

Il compromesso tra bias e varianza 6

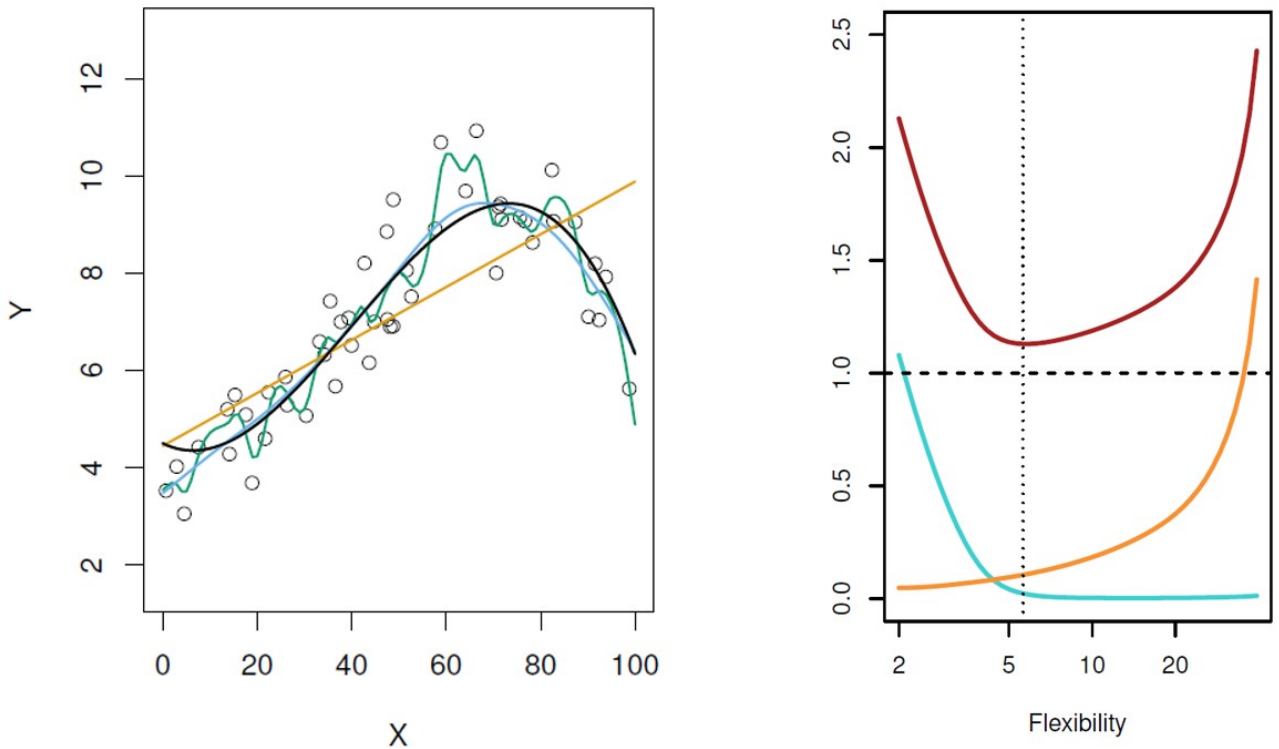


Figura 18: Predizione (sinistra) e bias, varianza e test MSE (destra)

Il compromesso tra bias e varianza 7

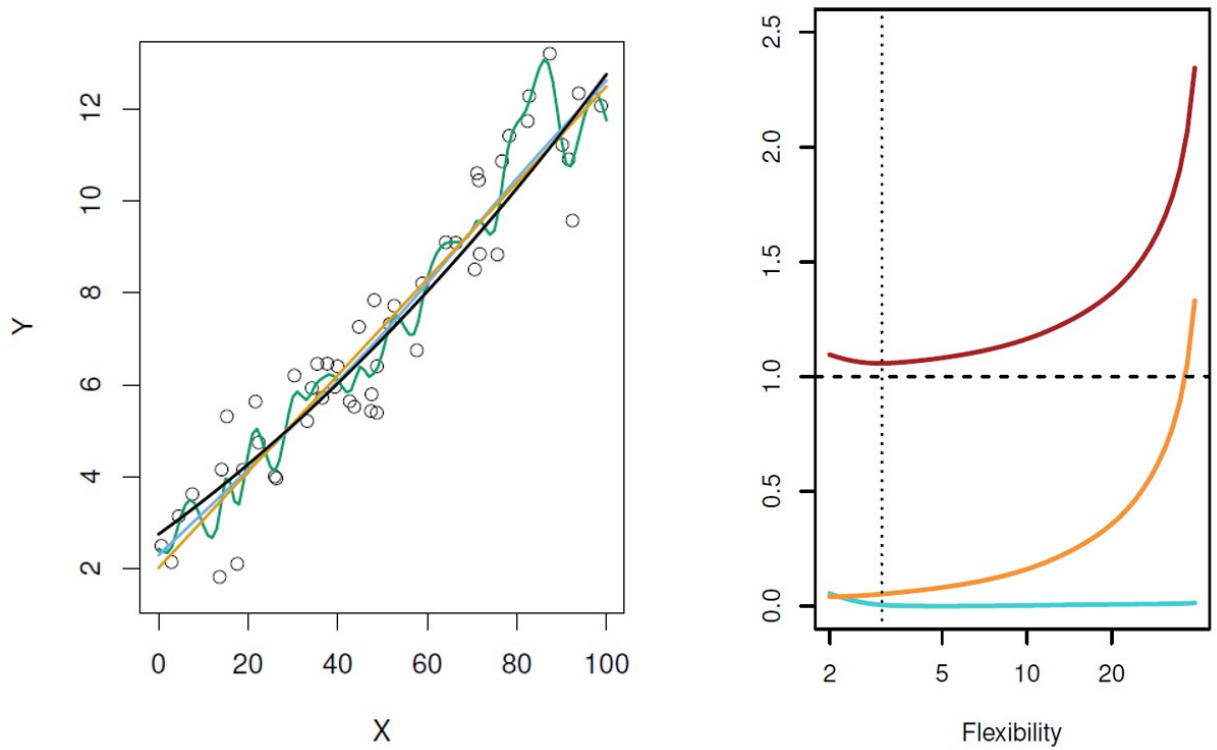


Figura 19: Predizione (sinistra) e bias, varianza e test MSE (destra)

Il compromesso tra bias e varianza 8

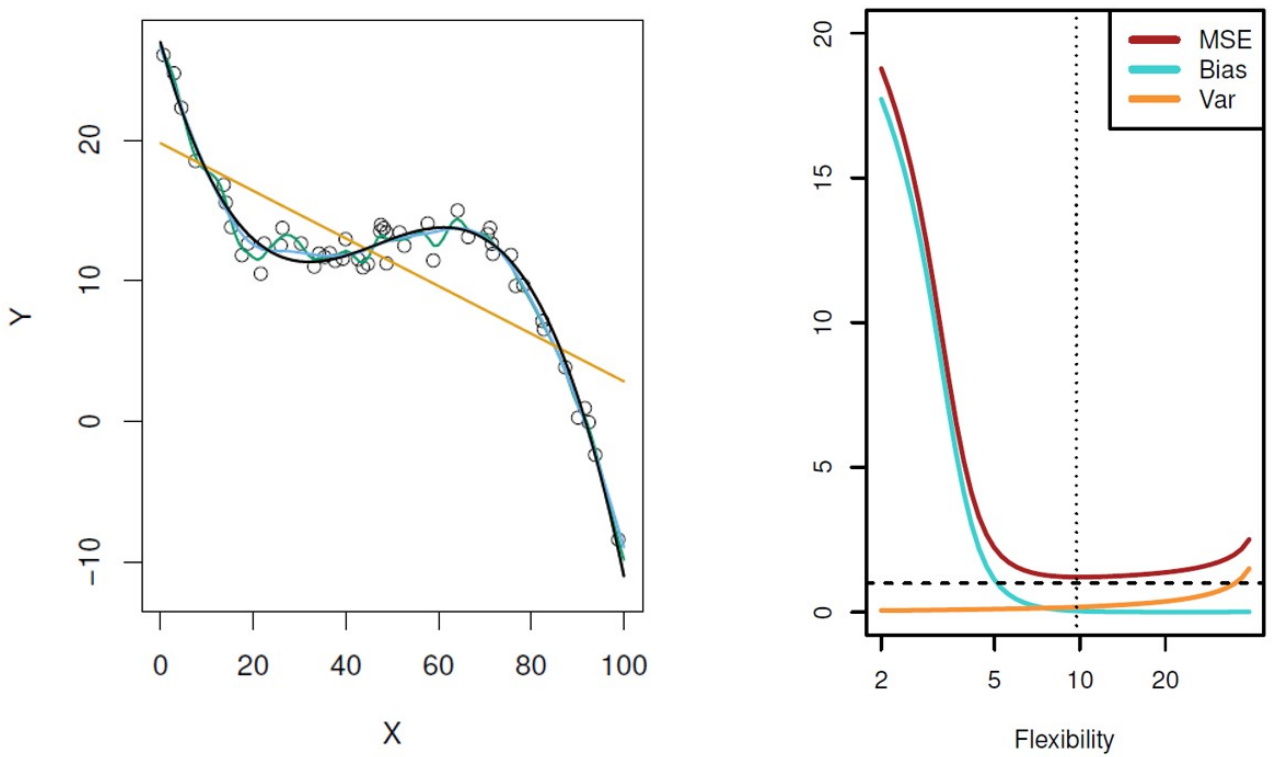


Figura 20: Predizione (sinistra) e bias, varianza e test MSE (destra)

Valutare l'accuratezza di un modello - Variabili qualitative 1

- Nel caso di variabili non quantitative (e.g. sano/malato, alto/basso) non possiamo calcolare *MSE*
- useremo una misura diversa: l'**error rate**

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

dove

$$I(y_i \neq \hat{y}_i) = 1, \text{ se } y_i \neq \hat{y}_i$$

oppure

$$I(y_i = \hat{y}_i) = 0, \text{ se } y_i = \hat{y}_i$$

Valutare l'accuratezza di un modello - Variabili qualitative 2

a noi ovviamente interesserà il *test error rate*:

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

Valutare l'accuratezza di un modello - Variabili qualitative 3

Nel caso delle variabili qualitative la valutazione è fatta *contando* quante volte il nostro modello **identifica** correttamente le diverse classi.

Possiamo individuare 2 tipi di performances:

- Predizioni Corrette
  - True Positive (TP)
  - True Negative (TN)
- Errori di Classificazione
  - False Positive (FP)
  - False Negative (FN)

Valutare l'accuratezza di un modello - Variabili qualitative 4

La *Confusion Matrix*

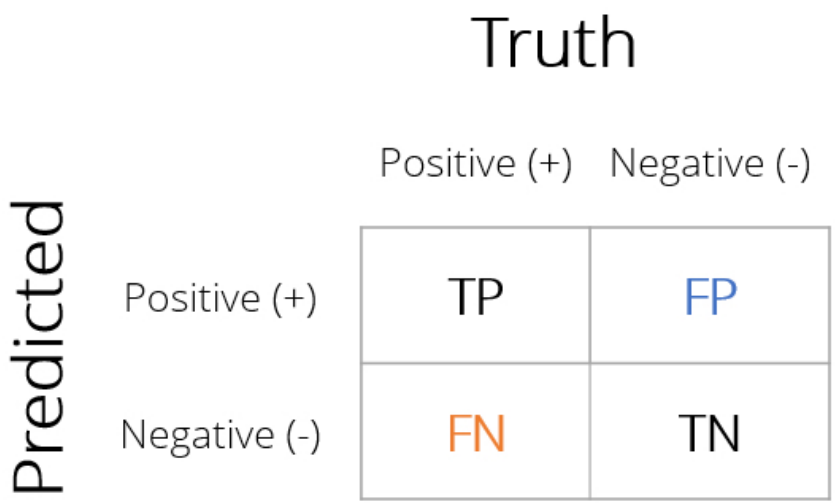


Figura 21: Confusion Matrix

Valutare l'accuratezza di un modello - Variabili qualitative 4

I principali parametri di valutazione dei risultati

1. **Accuracy** : $\frac{TP+TN}{TP+FP+TN+FN}$
2. **Sensitivity**: $\frac{TP}{TP+FP}$, proporzione dei casi positivi classificati correttamente
3. **Specificity**: $\frac{TN}{TN+FN}$, proporzione dei casi negativi classificati correttamente
4. **False positive rate** (FPR): $1 - \textit{specificity}$, proporzione dei falsi positivi tra i **veri** negativi.

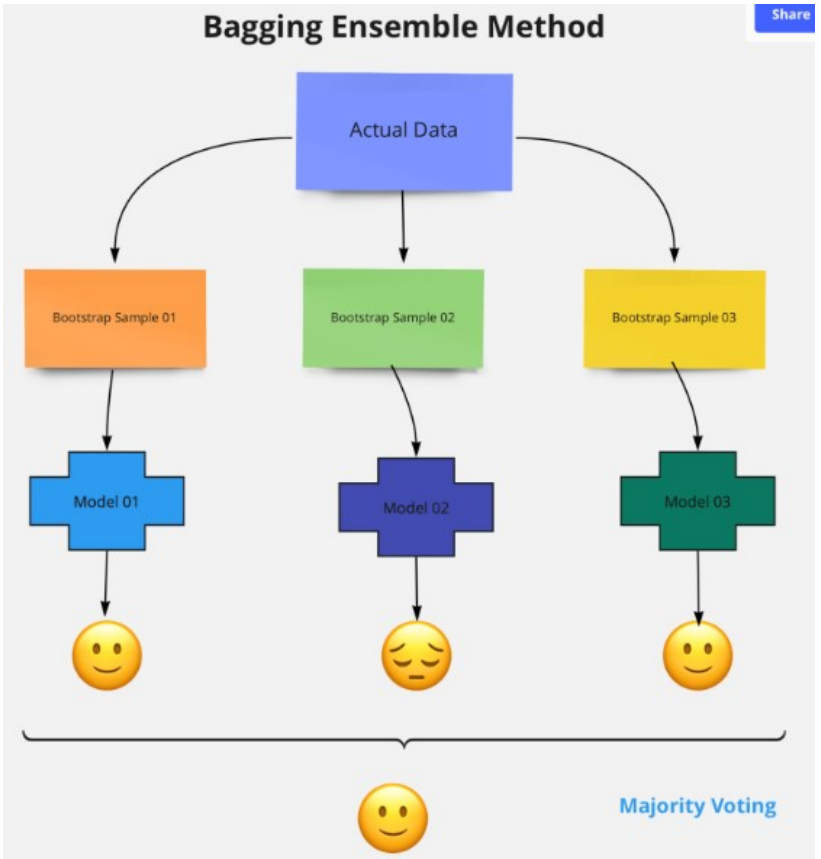
Underfitting o Overfitting? Cosa fare? 1

Sintomi :

1. *Train MSE* molto basso (overfitting)

Possibili Rimedi:

- aggiungere più dati di training
- ridurre la complessità del modello
- *bagging* (Bootstrap aggregation): crea copie multiple a partire dai *train data*



Underfitting o Overfitting? Cosa fare? 2

Sintomi :

1. *Train MSE* alto (underfitting)

Possibili Rimedi:

- aumentare la complessità del modello (modelli non lineari)
- aggiungere predittori
- *boosting* (modelli sequenziali):