

UNIVERSITÀ DEGLI STUDI DI SALERNO



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

Corso di Laurea Magistrale 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 

Model Based vs. Supervised Regression

- 22/10

Model Based significa stima dei parametri (θ, Y ecc...) con tutti i modelli noti.

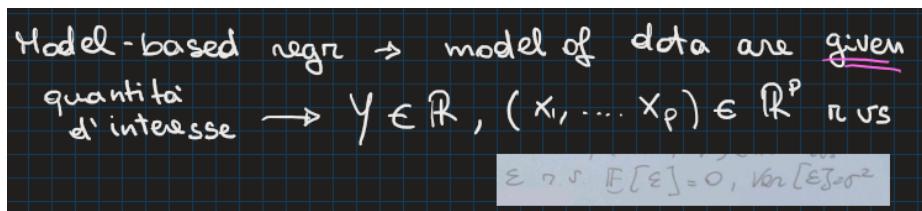
Supervised invece è a partire dai dati.

Cosa succede se usiamo l'approccio model based con la regressione?

Model-based Regression

Model-based → i modelli dei dati sono assegnati.

Anche gli errori sono dati, ϵ è una variabile aleatoria della quale conosciamo media (in questa lezione per la prima volta proveremo a rilassare l'assuzione che sia 0) e varianza.



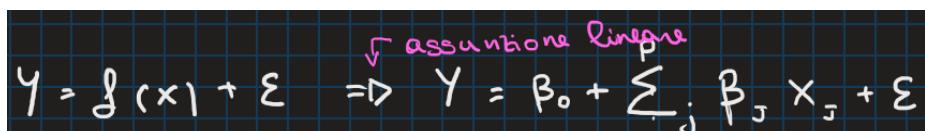
Model-based negr → model of data are given
quantità di interesse → $Y \in \mathbb{R}$, $(x_1, \dots, x_p) \in \mathbb{R}^p$ and ϵ
 $\epsilon \rightsquigarrow E[\epsilon] = 0, \text{Var}[\epsilon] = \sigma^2$

Y è una variabile aleatoria appartenente a \mathbb{R} con modello noto.

I regressori sono random variables a modello noto.

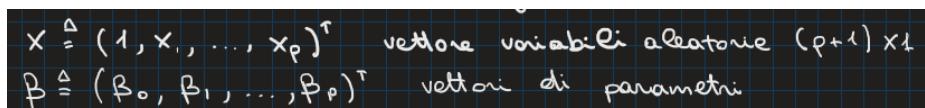
Le ipotesi fin qui sono le ipotesi classiche di Gauss-Markov e della regressione lineare.

Essendo model based costruiamo un modello dal quale partire, che è la stessa cosa che abbiamo fatto in supervised cioè assumiamo una relazione tra Y e i regressori di tipo lineare.



$Y = f(x) + \epsilon$ $\Rightarrow Y = \beta_0 + \sum_j \beta_j x_j + \epsilon$
assunzione lineare

Vediamo come stimare le quantità con un approccio model-based e le confrontiamo con quelle calcolate con un approccio campionario.



$x \triangleq (1, x_1, \dots, x_p)^T$ vettore variabili aleatorie $(p+1) \times 1$
 $\beta \triangleq (\beta_0, \beta_1, \dots, \beta_p)^T$ vettori di parametri

Questa x non è x matrice delle lezioni passate, è solo un vettore con le variabili, è il **random vector dei regressori** che nel nostro caso sono variabili aleatorie.

L'approccio è che nel contesto della regressione vogliamo vedere le cose viste recentemente del tipo "stima della media di una normale", casi in cui avevamo i modelli di probabilità e volevamo stimare parametri, in questo caso i parametri sono le β cioè i coefficienti del modello.

Chiamiamo quindi β il **vettore dei coefficienti dei parametri**.

A questo punto in termini dei vettori che abbiamo trovato riscriviamo la relazione lineare presentata all'inizio.

$$\rightarrow Y = \underbrace{\beta^\top x}_{\text{scalari}} + \varepsilon = \underbrace{x^\top \beta}_{\text{scalari}} + \varepsilon \quad (\star)$$

La relazione lineare si può scrivere in due modi che sono uguali in quanto il risultato è uno scalare e il trasposto di uno scalare è lo scalare stesso, quando si traspone un prodotto si invertono i posti e si traspongono entrambi.

Quello che abbiamo ottenuto è il modello stocastico che interpreta le nostre assunzioni per i dati, sono modelli aleatori (perché sono aleatori X , Y ed ε). Questo rappresenta sempre la regressione, solo che conosciamo il modello.

Confronteremo poi questo con quanto fatto con i campioni.

Lo stimatore ottimo

Dobbiamo stimare β e stimare quindi la nostra funzione di regressione.

Conosciamo lo stimatore ottimo, cioè quello che minimizza l'errore quadratico (quadratic loss, rischio Bayesiano), vogliamo fare predizione, prevedere Y e questa volta conosciamo anche il modello per le X .

Lo stimatore ottimo è quello che minimizza la distanza tra Y e \hat{Y} al quadrato, per fare questo la \hat{Y} deve essere generata come il valore atteso di Y dato X .

Visto che $E[Y|X]$ non la sapevamo calcolare dovevamo risolvere in maniera campionaria, dovevamo trovare le β per minimizzare le distanze.

$$\begin{aligned} \text{We Know the optimal estimator (minimizziamo la perdita quadratica)} \\ r(x) &= E[Y|x] = E[\beta^\top x + \varepsilon|x] \Rightarrow r(x) = \underbrace{\beta^\top}_{\text{scalare}} \underbrace{E[x|x]}_{x} + \underbrace{E[\varepsilon|x]}_{\text{è media nulla dato } x} \end{aligned}$$

Lo stimatore ottimo minimizza la perdita quadratica, il rischio quadratico.

La funzione di regressione $r(x)$ non è altro che il valore atteso di $Y|X$, è proprio la premessa a tutta la regressione questa, noi questa media non la sappiamo calcolare allora ci costruiamo un modello, poi lo stimiamo dai dati, poi troviamo le β .

Questa volta conosciamo i modelli di Y ed X e quindi abbiamo una marcia in più.

A questo punto diamo una forma ad $r(x)$ sostituendo Y con la sua espressione. Poi dividiamo in somma di due medie perché se è vero che la media è lineare allora lo è anche la media condizionata.

β^T è un vettore ma indipendente dalla media che stiamo calcolando quindi può uscire fuori.

$E[X|x] = x$, perché una volta fissato x allora non c'è più aleatorietà.

$E[\epsilon|x] = 0$, e qui cambia l'assunzione che davamo sempre, invece di dire che la media di ϵ è 0 diciamo che la media di ϵ CONDIZIONATA ad x è 0. Si dice che ϵ is zero mean given X .

$$\Rightarrow \pi(x) = \beta^T x = x^T \beta \quad \text{funzione di regressione ottima}$$

Il ragionamento non è ciclico perché noi abbiamo sostituito il modello beta trasposto per x e abbiamo dimostrato che era anche la funzione di regressione ottima.

Lo stimatore ottimo e il modello sono due cose diverse, noi abbiamo assunto il modello e abbiamo trovato lo stimatore ottimo.

Teorema

Per una generica coppia di punti (x_0, y_0) , dove ovviamente x_0 contiene tutti i regressori, nel problema di regressione generale $y_0 = r(x_0) + \epsilon$, l'errore ϵ è, e deve essere, condizionatamente a media 0.

Theor (x_0, y_0) il problema di regressione generale
 $y_0 = \pi(x_0) + \epsilon$, l'errore ϵ è condizionatamente a media zero
 $E[\epsilon|x_0] = 0$

DIMOSTRAZIONE

$$\begin{aligned} \text{Proof. } \epsilon &= y_0 - \pi(x_0) \Rightarrow E[\epsilon|x_0] = E[y_0 - \pi(x_0)|x_0] \\ &= E[y_0|x_0] - \pi(x_0) = 0 \end{aligned}$$

In qualunque problema affronteremo, qualunque sia la forma della nostra funzione di regressione ottima $r(x)$, cioè il valore atteso di $Y|X$, gli errori condizionati alla x devono essere 0, altrimenti il modello è sbagliato, c'è un bias nella struttura, non tutto è endogeno, cioè non tutto ciò che stiamo modellando è contenuto nel modello delle x , c'è bisogno di qualcosa di esterno per riportare l'oggetto a media nulla e tornare alla teoria della predizione. Quando non si riesce perché il modello è complicato si usa una variabile detta esogena (modello con variabile strumentale) per provare a tornare all'ambito che abbiamo visto.

$$E[\epsilon|x_0] = \pi(x_0) - \beta^T x_0 \neq 0 \Leftrightarrow \pi(x_0) \neq \beta^T x_0$$

Questa assunzione su ϵ è l'assunzione più ampia che si può fare quando si lavora con i modelli di regressione.

La ricerca del vettore β

Anche se adesso rispetto alla prima volta che abbiamo fatto la regressione lineare stiamo lavorando Model Based comunque l'obiettivo è stimare il vettore β dei miei parametri.

Visto che abbiamo X , Y ed ϵ (condizionatamente a media nulla) capiamo che β è funzione di X ed Y e nello specifico sarà funzione dei momenti di X e di Y (valore atteso di X , X^2 , XY ecc..., insomma i momenti della coppia di variabili aleatorie XY).

Nella regressione lineare multipla

$$Y = X^\top \beta + \epsilon \rightarrow Xy = \underbrace{X^\top \beta}_{\substack{\text{colonna} \\ (\rho+1) \times 1}} + \underbrace{\epsilon}_{\substack{\text{scalare} \\ (\rho+1) \times 1}} = \underbrace{\underbrace{\underbrace{(X^\top \beta)^\top}_{(\rho+1) \times 1} \times \underbrace{\epsilon}_{(\rho+1) \times 1}}_{(\rho+1) \times (\rho+1)}}_{(\rho+1) \times 1}$$

Abbiamo premoltiplicato, sia a destra che a sinistra, per x .

Le dimensioni sono coerenti.

Siamo partiti da due vettori e ci troviamo una matrice (si riferisce solo a XX^T).

Per fare la stima di β calcoliamo il valore atteso di questa espressione.

Valutazione delle medie

Dobbiamo calcolare le tre medie in questa espressione.

$$\begin{aligned} E[\cdot] \\ \Rightarrow E[X^\top Y] &= E[X^\top X^\top] \beta + E[X^\top \epsilon] \quad (\star\star) \\ \text{What is } E[X^\top X^\top] ? & \quad E[(X^\top X^\top)_{ij}] = E[X_i X_j] \\ \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{bmatrix}^\top \begin{bmatrix} 1 & x_1 & \dots & x_p \end{bmatrix} &= \begin{bmatrix} 1 & x_1 & \dots & x_p \\ x_1 & x_1^2 & x_1 x_2 & \dots \\ \vdots & x_2 x_1 & \dots \\ x_p & & & \end{bmatrix} \end{aligned}$$

Il prodotto funziona sempre allo stesso modo, riga per colonna.

Stiamo costruendo una matrice dove sulla diagonale c'è x_1^2, x_2^2 ecc..., poi nei posti off-diagonalabbiamo le varie combinazioni, $x_1 x_2, x_1 x_3$ ecc..., insomma stiamo costruendo una matrice, a meno della media, di covarianza, di correlazione.

▼ Come si calcola la matrice di var-covar

Per calcolare la matrice di covarianza di due variabili aleatorie X e Y , dobbiamo calcolare le covarianze tra le coppie di variabili (X, X) , (X, Y) , (Y, X) e (Y, Y) . La matrice di covarianza per due variabili casuali è una matrice 2×2 definita come:

$$\Sigma = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Var}(Y) \end{pmatrix}$$

dove:

- $\text{Var}(X)$ è la varianza di X ,
- $\text{Var}(Y)$ è la varianza di Y ,
- $\text{Cov}(X, Y)$ è la covarianza tra X e Y ,
- $\text{Cov}(Y, X) = \text{Cov}(X, Y)$ perché la covarianza è simmetrica.

Passi per calcolare la matrice di covarianza

1. Calcola la media di X e Y :

$$\mu_X = \mathbb{E}[X], \quad \mu_Y = \mathbb{E}[Y]$$

2. Calcola le varianze:

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2]$$

$$\text{Var}(Y) = \mathbb{E}[(Y - \mu_Y)^2]$$

3. Calcola la covarianza:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

4. Costruisci la matrice usando i risultati ottenuti.

Se hai a disposizione un campione di dati per X e Y , puoi stimare la matrice di covarianza con la seguente formula:

$$\Sigma = \frac{1}{n-1} \sum_{i=1}^n \begin{pmatrix} X_i - \bar{X} \\ Y_i - \bar{Y} \end{pmatrix} \begin{pmatrix} X_i - \bar{X} & Y_i - \bar{Y} \end{pmatrix}$$

dove n è il numero di osservazioni, e \bar{X} e \bar{Y} sono le medie campionarie di X e Y .

La matrice di autocorrelazione R_x dipende solo dalle x , dai dati, dal modello probabilistico delle x .

$$R_x \triangleq \mathbb{E}[X X^T] \quad \text{"(auto)-correlation matrix between } X \text{ (dato)"}$$

$$R_{xy} \triangleq \mathbb{E}[XY] \quad \text{"(cross) - correlation" è un vettore}$$

R_{xy} è la **cross-correlation**, la correlazione tra x ed y , questa è un vettore e si calcola tra i dati e la risposta.

Queste matrici si trovano spesso in statistical signal processing o in image processing.

Manca da considerare il valore atteso $X\epsilon$, era facile quando erano tutti indipendenti e quindi si spezzava il valore atteso, ora è diverso e usiamo la tower property (total expectation).

If X and ϵ are not independent, tower prop (total expect.)

$$\mathbb{E}[X\epsilon] = \mathbb{E}_x[\mathbb{E}[X\epsilon|_x]] = \mathbb{E}_x[X\mathbb{E}[\epsilon|_x]] = 0$$

$\Rightarrow X$ e ϵ sono incorrelati

Nelle ipotesi classiche della regressione c'è sempre che X ed ϵ devono essere incorrelati, talvolta si dice statistica indipendenza che è anche una cosa più forte visto che indipendenza implica incorrelazione. Incorrelazione l'assunto più generico, come è più generico che sia la media condizionata ad essere uguale a 0 e non la media di ϵ da sola.

Sostituiamo le medie calcolate nell'espressione da cui siamo partiti.

$$(\star\star) \rightarrow R_{xy} = R_x \beta \Rightarrow \beta = R_x^{-1} R_{xy}$$

$$\rightarrow \text{optimal regression function } r(x) = x^T \beta = x^T R_x^{-1} R_{xy}$$

La β calcolata in questo modo (per trovarla è necessario che R_x sia invertibile) è quella che noi usiamo nel nostro modello di regressione ottima, sostituiamo e troviamo $r(x)$.

Tutto a media nulla la matrice di correlazione coincide con la matrice di varianza-covarianza, noi la matrice di varianza-covarianza l'avevamo scritta come l'inversa del prodotto tra la matrice X trasposta e la matrice X stessa, però lì avevamo la matrice X che aveva i dati dentro, era l'approccio empirico alla **stima**, non i valori attesi perché abbiamo potenzialmente la possibilità di calcolarli e di calcolare la matrice di var-covar, come ora. Quello era un approccio basato sui dati, questo è un approccio basato sui modelli.

Invece di fare la minimizzazione per trovare il β che minimizza le distanze, che minimizza l'RSS, si può provare a stimare le matrici di correlazione e stimare la funzione di regressione ottima.

Il legame tra Model-based e Supervised Regression

In model-based negr. $\beta = R_x^{-1} R_{xy}$

In supervised learning, we have $y = f(x) + \epsilon \Rightarrow$

$$\Rightarrow y = \beta_0 + \sum_j^p \beta_j x_j + \epsilon \quad \text{parametric regression}$$

Le matrici R_x e R_{xy} bisogna conoscerle per fare model-based regression, al più si possono stimare ma comunque sono essenziali.

Nell'approccio supervisionato ipotizziamo una relazione lineare tra X ed Y.

Per risolvere il problema della stima di β abbiamo un training set.

Training set (x, y) and we want to learn β to make prediction,
 i.e. minimizing $(y - \hat{y})^2$ error

Vogliamo creare un ponte da regression a model based e iniziamo scrivendo il nostro training set come una matrice \tilde{x} che è praticamente la trasposta di come facemmo in regressione lineare multipla, dove avevamo un regressore per colonna e ogni riga rappresentava un campionamento, ora invece ogni colonna è il vettore di ingresso corrispondente ad una certa uscita \tilde{y} .

In supervised setting, we use training set \tilde{x}

$$\tilde{x} = \begin{bmatrix} x_1 & & x_{1n} \\ \vdots & & \\ x_{(p+1)1} & & x_{(p+1)n} \end{bmatrix} = [\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n]$$

\tilde{x}_l vector column l of training matrix $l = 1, \dots, n$
 ↑ data size

$$\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_n)$$

la prima riga in realtà dovrebbe essere tutti 1, neanche il prof lo ha scritto comunque, sarebbe la colonna relativa a β_0

Non è un problema che abbiamo ragionato in maniera trasposta rispetto alla scorsa volta, semplicemente dobbiamo tenerlo a mente, ci usciranno le espressioni che sono il trasposto di quelle trovate in passato.

Per tradizione i dati si mettono per colonne che penso sia come è stato fatto qui ma ogni modo è corretto.

Anche \tilde{y} è parte del training set visto che seguiamo un approccio supervisionato.

Ora abbiamo i dati ma non abbiamo le matrici di correlazione quindi non possiamo trovare β ottimo che abbiamo dimostrato prima ma invece dobbiamo stimarlo e trovare $\hat{\beta}$ sulla base dei dati, in particolare sulla base del training data set.

From data, we estimate β as $\hat{\beta}$ based on (\tilde{x}, \tilde{y})

$$\hat{\beta} = (\hat{R}_x)^{-1} \hat{R}_{xy} \quad \text{where} \quad (\hat{R}_x)_{ij} = \frac{1}{n} \sum_{m=1}^n \tilde{x}_{im} \tilde{x}_{mj}$$

$$= \frac{1}{n} [\tilde{x} \tilde{x}^\top]_{ij}$$

$$(\hat{R}_{xy})_{i\cdot} = \frac{1}{n} \sum_{m=1}^n \tilde{x}_{im} \tilde{y}_m = \frac{1}{n} [\tilde{x} \tilde{y}^\top]_{i\cdot}$$

Invece di usare le medie vere per calcolare le vere matrici R_x e R_{xy} noi facciamo la versione empirica e invece dell'integrale per calcolare la media ci basta la media campionaria.

Notiamo la corrispondenza tra la matrice R_x e la matrice che trovammo in regressione $X^T X$ (qui è al contrario, quella trasposta è la seconda, perché la nostra X è trasposta).

Trovate le matrici approssimate (che invece dei valori attesi hanno le stime) abbiamo la formula per $\hat{\beta}$.

$$\Rightarrow \hat{\beta} = n (\tilde{x} \tilde{x}^\top)^{-1} \cdot \frac{1}{n} \tilde{x} \tilde{y}^\top = (\tilde{x} \tilde{x}^\top)^{-1} \tilde{x} \tilde{y}^\top$$

La somiglianza con questa espressione è significativa. Se avessimo usato una matrice \tilde{x} trasposta, in modo da averla coerente a quella che usammo per la regressione lineare multipla, e concordantemente la \tilde{y} non deve essere più un vettore riga ma diventa un vettore colonna. Avremmo trovato con questa differenza proprio lo stimatore $\hat{\beta}$ che avevamo trovato in regressione lineare multipla.

Quindi abbiamo trovato essenzialmente la stessa cosa ma abbiamo scritto le matrici in modo diverso, in un modo più simile a come facemmo con le variabili aleatorie in model based.

Richiamo alla strategia classica per trovare $\hat{\beta}$

Come delineammo $\hat{\beta}$? Minimizzando l'RSS.

L'RSS è ben definito essendo tutto campionario.

Come otteniamo $\hat{\beta}$? minimizzando RSS = $\frac{1}{n} \sum_{m=1}^n (\tilde{y}_m - \beta^\top \tilde{x}_m)^2$

$\xrightarrow{n \rightarrow \infty} \#[(Y - \beta^\top x)^2]$

RISCHIO SUPERVISIONATO RISCHIO EMPIRICO

\tilde{x}_m è l'ennesima colonna della matrice dei dati. Serve l'indice perché la matrice contiene tanti campionamenti e noi li analizziamo uno per volta.

Questo calcolo converge per n che va a infinito al valore atteso del quadrato della differenza tra y e $\beta^\top x$, cioè quello che abbiamo trovato noi è la versione campionaria.



Quello che noi minimizziamo è il **Rischio Empirico**, perché è calcolato a partire dai dati, il **Rischio Bayesiano** è il valore atteso invece.

Valutazione delle prestazioni, approccio Model based vs. Approccio supervisionato

Noi usiamo il supervisionato perché NON abbiamo i modelli, abbiamo i dati e ci industriamo con quelli, con n estremamente grande però si converge al valore vero che è quello che si avrebbe con la conoscenza del modello.

Performance comparison
training set (\tilde{x}, \tilde{y}) or $D_{tr} = \{(x_i, y_i)\}_{i=1}^n$ prediction on new obs.
 (x_0, y_0) with the same distribution (\tilde{x}, \tilde{y})

in alto a destra è random samples

Il nostro obiettivo è fare prediction e misurare l'errore su un data set di test, facciamo tutto semplice per non dover fare calcoli e quindi valutiamo un test data set fatto da un solo punto (x_0, y_0) che avrà sempre la stessa distribuzione dei dati di training ovviamente, non è che addestriamo su dei dati e testiamo su altri dati che non c'entrano nulla.

Model based approach : opt. est. for $y_0 \rightarrow \pi(x_0)$ given MMSE($\mathbb{E}[Y|X]$)
MMSE è la media di $Y|X$

Lo stimatore ottimo model based è il nostro r valutato in x_0 , r lo abbiamo trovato tale da minimizzare l'MMSE, cioè il valore atteso di $Y|X$.

Supervised approach : $\hat{\pi}(x_0)$ given minimizing empirical risk

Lo stimatore per l'approccio supervisionato sarà \hat{r} calcolato in x_0 , r nel model based lo abbiamo stimato con la stessa filosofia, ma qui non abbiamo le matrici di varianza-covarianza, non conosciamo β^T , non conosciamo tutto, qui dobbiamo stimare β . Qui abbiamo minimizzato il rischio empirico per trovare \hat{r} perché abbiamo solo le medie campionarie calcolabili a partire dai dati.

Teorema sull'errore della funzione di regressione

Theor The error regress. funct. $\hat{\pi}(x_0)$ (suboptimal) is:
$$\mathbb{E}[(\hat{\pi}(x_0) - y_0)^2] = \text{MMSE} + \mathbb{E}\{[(\hat{\pi}(x_0) - \pi(x_0))^2]\}$$

where $\text{MMSE} = \mathbb{E}\{[\pi(x_0) - Y_0]^2\}$

Ovviamente andrà meglio la stima con il modello, quella che possiamo fare con i dati è suboptimal, l'errore della funzione stimata a partire dai dati è presentato in immagine.

Essenzialmente oltre al MMSE che c'è anche nel caso con i modelli si aggiunge un termine di errore aggiuntivo.

Dimostrazione del teorema - 24/10

Proof.

$$\mathbb{E}[(\hat{r}(x_0) - y_0)^2] = \mathbb{E}\{[\hat{r}(x_0) - r(x_0) + r(x_0) - y_0]^2\} =$$

\uparrow
 $\hat{r}(x_0)$

Sommiamo e sottraiamo $r(x_0)$.

$$\mathbb{E}\{[(\hat{r}(x_0) - r(x_0)) + (r(x_0) - y_0)]^2\} =$$

Raggruppiamo a due a due i termini interni alla media da ottenere il quadrato di un binomio.

$$= \mathbb{E}\{[(\hat{r}(x_0) - r(x_0))^2 + (r(x_0) - y_0)^2 + 2(\hat{r}(x_0) - r(x_0))(r(x_0) - y_0)]\} \stackrel{\mathbb{E}[.] \text{ lin.}}{\leq}$$

Svolgiamo il quadrato del binomio.

$$= \underbrace{\mathbb{E}\{[(\hat{r}(x_0) - r(x_0))^2]\}}_{\substack{\text{fun di neg. tra la} \\ \text{ottima e l'empirica} \\ \text{tanto è più grande quanto} \\ \text{sono diverse}}} + \underbrace{\mathbb{E}\{(r(x_0) - y_0)^2\}}_{\substack{\text{MMSE per} \\ \text{definizione}}} + \underbrace{2\mathbb{E}\{[(\hat{r}(x_0) - r(x_0))(r(x_0) - y_0)]\}}_{\substack{\text{errore dello stimatore} \\ (*)}}$$

Dividiamo la media in tre medie per linearità in modo da dividere i tre termini.



Il primo valore atteso è la distanza quadratica tra la funzione di regressione ottima e quella empirica. Più è grande questo termine quanto più \hat{r} è lontano dalla vera r .



Il secondo valore atteso è l'MMSE.

Il terzo valore atteso per essere compreso richiede ulteriore elaborazione, comunque scomparirà. Usiamo la Tower Property.

Dimostrazione che il terzo termine vale 0

Immaginiamo che è la media del prodotto di due funzioni f e g . Il valore atteso continuo diventa l'integrale di f per g^* (complesso coniugato), questo integrale se questi fossero stati segnali di energia sarebbe stato corrispondente proprio al prodotto scalare tra f e g . E infatti il valore atteso del prodotto di due quantità aleatorie è il prodotto scalare.

▼ Prodotto scalare tra segnali di energia

Il prodotto scalare di due segnali di energia $f(t)$ e $g(t)$ è definito come l'integrale del prodotto dei due segnali nel dominio del tempo. Se i segnali sono definiti su tutto l'asse dei tempi, il prodotto scalare è dato da:

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t) g^*(t) dt$$

dove $g^*(t)$ indica il complesso coniugato di $g(t)$. Nel caso di segnali reali, il complesso coniugato può essere omesso, quindi l'espressione diventa:

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t) g(t) dt$$

Questa quantità rappresenta una misura di somiglianza tra i segnali f e g : se il prodotto scalare è grande, significa che i segnali sono molto "simili" (in fase e forma); se è zero, i segnali sono ortogonali e non hanno correlazione energetica.

Questa definizione si estende anche a segnali discreti, dove l'integrale è sostituito dalla somma:

$$\langle f, g \rangle = \sum_{n=-\infty}^{+\infty} f[n] g^*[n]$$

Questa formula è utile, ad esempio, nell'analisi di segnali e nel calcolo della correlazione.

▼ Prodotto scalare tra quantità aleatorie

Se consideriamo due variabili aleatorie X e Y con valori attesi (o medie) dati da $\mathbb{E}[X]$ e $\mathbb{E}[Y]$, e variabili centrate (ovvero con media zero) ottenute come $X - \mathbb{E}[X]$ e $Y - \mathbb{E}[Y]$, allora il valore atteso del prodotto di queste variabili centrate (detto covarianza) può essere visto come un prodotto scalare nello spazio delle variabili aleatorie. Questo è vero in quanto la covarianza è una misura di come due variabili si "proiettano" l'una sull'altra, analogamente a un prodotto scalare in uno spazio vettoriale.

In termini matematici, il valore atteso del prodotto di X e Y , cioè $\mathbb{E}[XY]$, può essere visto come una forma di prodotto scalare in uno spazio di probabilità se si definisce lo spazio con funzioni centrate rispetto alla loro media.

$$\mathbb{E}_{x_0} \left\{ \mathbb{E}_{y_0} \left\{ [\hat{r}(x_0) - r(x_0)] [r(x_0) - y_0] \mid X_0 \right\} \right\} = \mathbb{E}_{x_0} \left\{ g(x_0) \mathbb{E}_{y_0} \left\{ r(x_0) - y_0 \mid X_0 \right\} \right\} =$$

$\underbrace{g(x_0)}_{\text{è una costante}} \mid X_0$

Non abbiamo potuto dividere il valore atteso come prodotto delle medie in quanto le due quantità non sono indipendenti, x_0 e y_0 sono legati tra loro. Abbiamo quindi usato la Tower Property (il valore atteso iterato, la total expectation) condizionando rispetto ad x_0 .

Il primo termine della media interna è essenzialmente una funzione di x_0 quindi lo rinominiamo $g(x_0)$, una volta fissato x_0 non c'è aleatorietà, è una costante, un numero, lo facciamo quindi uscire dalla media interna.

$$\begin{aligned} &= \mathbb{E}_{x_0} \left\{ g(x_0) \left\{ \mathbb{E}_{y_0} [\hat{r}(x_0) \mid X_0] - \mathbb{E}_{y_0} [y_0 \mid X_0] \right\} \right\} = \mathbb{E}_{x_0} \left\{ g(x_0) \left[\underbrace{\mathbb{E}_{y_0} [\hat{r}(x_0) - r(x_0)]}_{=0} \mid X_0 \right] \right\} \\ &= \mathbb{E} \left\{ g(x_0) \cdot 0 \right\} = 0 \end{aligned}$$

\hookrightarrow per non essere ridondante ha compattato

La media interna può essere divisa in due parti per la linearità della media. La prima media interna è $r(x_0)$ che una volta fissato x_0 è una costante e quindi possiamo liberarla dalla media e scrivere solo $r(x_0)$, per la seconda media ricordiamo che media di y_0 dato x_0 è proprio la definizione di $r(x_0)$.

Quindi abbiamo mostrato che il terzo termine vale 0.

Visto che questo termine vale 0 e abbiamo detto che la media di f moltiplicato g equivale al prodotto scalare tra f e g , possiamo dire che le quantità $\hat{r}(x_0) - r(x_0)$ e $r(x_0) - y_0$ sono **ortogonali**.

Avremmo potuto bypassare questi calcoli in quanto c'è un teorema più generale chiamato principio di ortogonalità.

Principio di Ortogonalità

Questo principio vale per l'MSE Bayesiano.

PRINCIPIO DI ORTOGONALITÀ (Fundamental result of MSE)

the error of the optimal estimator is orthogonal to any (integrable) function of data $g(x_0)$

i.e. $\mathbb{E} \{ g(x_0) [\hat{r}(x_0) - y_0] \} = 0 \quad \forall g(x_0)$

Quindi visto che esiste questo principio non serviva fare i calcoli che abbiamo fatto per il terzo elemento. Già capitò in regressione quando parlammo di compromesso bias-varianza che assumemmo ortogonalità dell'errore, ora sappiamo perché.

Continuo dimostrazione del teorema

$$(k) \Rightarrow E\{[\hat{r}(x_0) - r_0]^2\} = MMSE + E\{[\hat{r}(x_0) - \bar{r}(x_0)]^2\}$$

errore col rischio empirico \hookrightarrow distanza quadratica tra \hat{r} e \bar{r}
 termine di penalità

Dell'espressione di partenza restano solo i primi due contributi che ci portano al risultato che intendavamo dimostrare.



Il rischio empirico è pari al rischio Bayesiano (MMSE) più un termine di penalità che è rappresentativo della distanza tra la nostra r stimata e la r vera.

Ricordiamo che la linearità è una cosa che abbiamo imposto noi e non per forza la relazione doveva essere lineare. Il termine di penalità dipende da \hat{r} e proveremo a trovare funzioni non più lineari che provano a minimizzare questo \hat{r} in modo da avvicinare rischio Bayesiano e rischio empirico.

non è detto che dobbiamo usare la regressione lineare, quindi proveremo a minimizzare l'errore con altre tecniche (dopo la pausa didattica)

QUANTILE REGRESSION \rightarrow sostituire la media con la mediana per aumentare la robustezza

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.