27 Settembre 2017

Regressione Lineare - REGRESSIONE LINEARE

Ho 2 grandezze, occorre soffermarsi sulla dipendenza lineare, + semplice che cmq mi permette di capire le dipendenze funzionali.

Richiami:

X, Y

$$Cov(x,y) = E[(X - E(x))(Y - E(y))] = E[XY] - E[X]\,E[Y]$$

$E[X] = \mu_X$

$E[Y] = \mu_Y$

$$\rho(x,y) = \frac{Cov(x,y)}{sd(x)\cdot sd(y)} = \rho_{xy}$$

σ

$$sd(x) = \sqrt{var(x)}$$

$$\sigma_X = \sqrt{var(x)} = \sqrt{\sigma_X^2}$$

↑ deviazione standard

σ^2 = varianza

$$var(x) = Cov(X,X)$$

Proprietà:

1. $Cov(X,Y) = Cov(Y,X)$
2. $Cov(aX, bY) = ab\,Cov(X,Y)$
3. $Cov(aX + bY + c, Z) = a\,Cov(X,Z) + b\,Cov(Y,Z)$
4. $\rho(aX, bY) = \rho(X,Y)$
5. $|\rho(x,y)| \leq 1$
6. $\rho(X,Y) = \pm 1$ solo se $Y = aX + b$ (e viceversa)

inoltre $a, \rho(x,y)$ sono concordi

7. X, Y indipendenti, allora $Cov(x,y) = \rho(x,y) = 0$

(ma non il viceversa)

a livello empirico

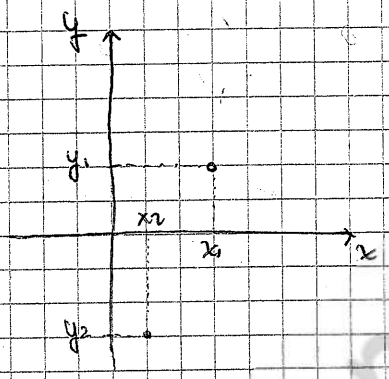
x_1	y_1
⋮	⋮
x_n	y_n

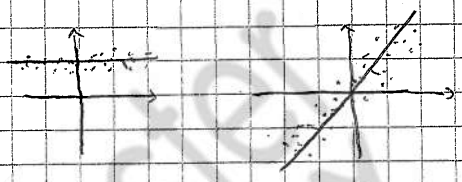
$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

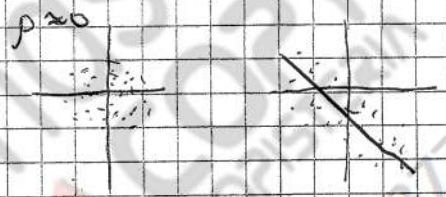
$$\widehat{cov}(x,y) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

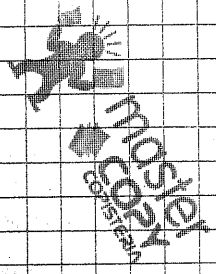
$$\hat{\sigma}_x^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 \qquad \hat{\sigma}_y^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2$$

$$\hat{\rho}(x,y) = \frac{\widehat{cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y}$$









Modello di regressione

$$Y = \underbrace{aX + b}_{} + \underbrace{\varepsilon}_{\text{errore}}$$

Se è risolta BENE la nostra valutazione sarà non buona

ipotesi
- $\mathbb{E}[\varepsilon] = 0$
- $cov(x, \varepsilon) = 0$ (il nostro modello sarebbe OTTIMO)

- Trovare a e b

$$cov(x,y) = cov(x, aX + b + \varepsilon) \quad \text{(per le proprietà viste nella scorsa puntata)}$$
$$= a\,Var(x) + \underbrace{cov(x,\varepsilon)}_{0}$$
$$= a\,var(x)$$

ovvero $a = \dfrac{cov(x,y)}{var(x)}$

$$\mathbb{E}Y = \mathbb{E}(aX + b + \varepsilon) = a\mathbb{E} + b$$
$$b = \mathbb{E}Y - a\mathbb{E}X$$

+ ε ha varianza bassa e + $aX + b$ tende a spiegare meglio il fattore di uscita Y

- Trovare la varianza di ϵ

$$Var(y) = Var(aX + b + \epsilon)$$

$$= a^2 Var(x) + Var(\epsilon)$$

$$Var(\epsilon) = Var(y) - a^2 Var(x)$$

Abbiamo la nostra tabella dati e vogliamo calcolare i valori attesi di qst tabella...

x_1	y_1
...	...
x_N	y_N

$$y = aX + b + \epsilon$$

$$\hat{a} = \frac{\hat{cov}(x,y)}{\sigma_x^2}$$

$$\hat{b} = \bar{y} - a\bar{x}$$

$$\sigma_\epsilon^2 = \sigma_y^2 - \hat{a}\sigma_x^2 = \sigma_y - \frac{\hat{cov}(x,y)}{\sigma_x^4}\sigma_x^2$$

$$= \sigma_y^2 - \frac{\hat{cov}(x,y)}{\sigma_x^2}$$

$y = \hat{a}x + b$ retta di regressione

la variabilità dell'errore ϵ sarà data dalla variabilità di y meno qualcosa influenzato da x e y

$$\sigma_\epsilon^2 = \sigma_y^2 \left(1 - \frac{\hat{cov}(x,y)^2}{\sigma_x^2 \sigma_y^2}\right)$$

$$= \sigma_y^2 (1 - \hat{\rho}(x,y)^2)$$

qst formula ci da una stima della precisione del modello di regressione

ricordiamo che il coefficiente empirico è dato da

$$\rho(x,y) = \frac{\hat{cov}(x,y)}{\sigma_x \sigma_y}$$

per l'errore la cosa importante è la sua varianza

ci aspettiamo che $y = \epsilon$ condividano la natura qualitativa

master cozy consistenza

(proporzione) di varianza spiegata dal modello

$$R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2}$$

$\frac{\text{varianza/errore}}{\text{varianza del fattore di uscita}}$

poiché

$$\sigma_\varepsilon^2 = \sigma_y (1 - \hat{\rho}(x,y))^2$$

se $\sigma_\varepsilon \neq 0 \rightarrow$ il modello non spiega nulla e quindi il modello determina in nessuna parte il fattore di uscita

quando $a = \hat{a}$, $b = \hat{b}$, si ha

$$R^2 = \hat{\rho}(x,y)^2$$

↑ quadrato del coefficiente di correlaz.

se $\sigma_\varepsilon^2 \ll \sigma_y^2$

↓

R^2 si avvicina a 1 e quindi il modello determina in gran parte il fattore di uscita y

Noi possiamo scrivere i modelli lineari scegliendo a e b

il coeff di correlaz. diventa una buona misura della dipendenza

www.mastercopy.it 050/831211/26 388/1958774

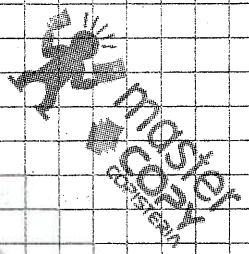
Metodo dei minimi quadrati METODO DEI MINIMI QUADRATI

x_1	y_1
⋮	⋮
x_n	y_n

Immaginiamo che valga una dipendenza lineare di qst tipo:

$$y = ax + b + \varepsilon \qquad \varepsilon = y - ax - b$$

$$y_1 - ax_1 - b$$
$$y_2 - ax_2 - b$$
$$\vdots$$
$$y_n - ax_n - b$$



Scarto quadratico dell'errore

$$\mathcal{E}(a,b) = \frac{1}{n}\sum_{k=1}^{n}(y_k - ax_k - b)^2$$

Il nostro obbiettivo è quello di trovare a e b in modo tale che $\mathcal{E}(a,b)$ sia il minimo possibile, ovvero minimizzare qst funzione

$$\min \mathcal{E}(a,b)$$

$$\begin{cases} \frac{\partial \mathcal{E}}{\partial a} = 0 \\ \frac{\partial \mathcal{E}}{\partial b} = 0 \end{cases}$$ derivata di $\mathcal{E}$ rispetto ad a

$$\begin{cases} \frac{1}{n}\sum_{k=1}^{n} x_k(y_k - ax_k - b) = 0 \\ \frac{1}{n}\sum_{k=1}^{n}(y_k - ax_k - b) = 0 \end{cases} \qquad \begin{cases} \hat{cov}(x,y) - a\sigma_x^2 = 0 \\ \bar{y} - a\bar{x} - b = 0 \end{cases}$$

che ha
- supponendo che la x spieghi bene la y
- supponendo di minimizzare l'errore

abbiamo trovato gli stessi valori

$$\hat{a} = \frac{\hat{cov}(x,y)}{\sigma_x^2}$$
$$b = \bar{y} - a\bar{x}$$

Stessa conclusione