CdL in Scienze Statistiche ed Economiche - Università degli Studi di Milano-Bicocca

Lezione : La matrice dei dati

Docente: Aldo Solari

1 La matrice X

$$\underset{n\times p}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

- Media per la j-sima variabile

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \qquad j = 1, \ldots, p$$

- Varianza per la j-sima variabile

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \qquad j = 1, \ldots, p$$

- Covarianza tra la j-sima e la k-sima variabile

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \qquad j = 1, \ldots, p,\ k = 1, \ldots, p$$

  Si noti che $s_{jk} = s_{kj}$ e che $s_{jj} = s_j^2$

- Correlazione tra la j-sima e la k-sima variabile

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}, \qquad j = 1, \ldots, p,\ k = 1, \ldots, p$$

  Si noti che $-1 \leq r_{jk} \leq 1$

2 Vettore delle medie e matrice di varianze/covarianze e di correlazione

- Vettore delle medie

$$\underset{p\times 1}{\bar{x}} = \begin{bmatrix} \bar{x}_1 \\ \cdots \\ \bar{x}_j \\ \cdots \\ \bar{x}_p \end{bmatrix}$$

- Matrice di varianze/covarianze

$$\underset{p\times p}{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1j} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2j} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ s_{j1} & s_{j2} & \cdots & s_{jj} & \cdots & s_{jp} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pj} & \cdots & s_{pp} \end{bmatrix}$$

- Matrice di correlazione

$$\underset{p\times p}{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1j} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2j} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ r_{j1} & r_{j2} & \cdots & 1 & \cdots & r_{jp} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pj} & \cdots & 1 \end{bmatrix}$$

2.1 Esempio

Variabile 1 (prezzo in Dollari per libro):	42	52	48	58
Variabile 2 (numero di libri venduti):	4	5	4	3

$$\underset{4\times 2}{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

$$\underset{2\times 1}{\bar{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

$$\underset{2\times 2}{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix}$$

$$\underset{2\times 2}{R} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} = \begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$$

3 Diagramma di dispersione

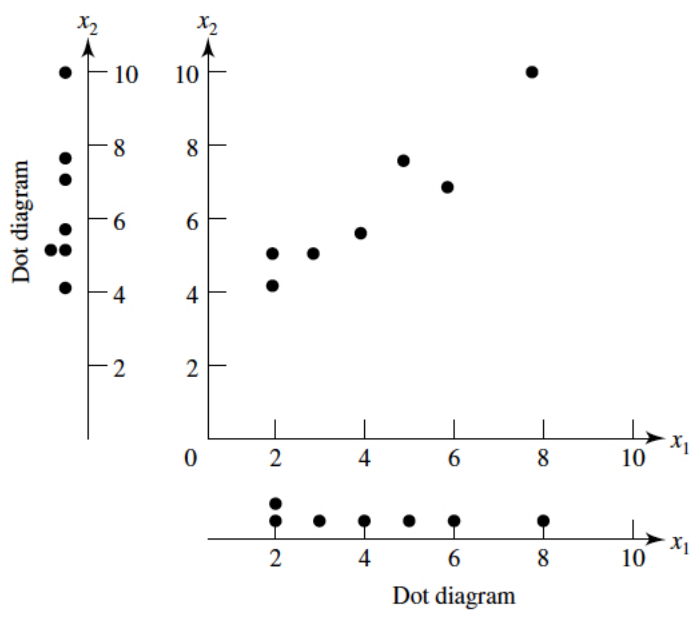
	u_1	u_2	u_3	u_4	u_5	u_6	u_7
x_1	3	4	2	6	8	2	5
x_2	5	5.5	4	7	10	5	7.5

Medie: $\bar{x}_1 = 4.2$, $\bar{x}_2 = 6.2$

Varianze: $s_{11} = 4.2$, $s_{22} = 0.56$

Covarianza: $s_{12} = 3.70$

Correlazione: $r_{12} = 0.95$



Cosa succede se mescolo a caso (permutazione) i valori della prima riga della tabella?

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
x_1	5	4	6	2	2	8	3
x_2	5	5.5	4	7	10	5	7.5

Medie: $\bar{x}_1 = 4.2$, $\bar{x}_2 = 6.2$

Varianze: $s_{11} = 4.20$, $s_{22} = 0.56$

Covarianza $s_{12} = -3.01$

Correlazione $r_{12} = -0.78$