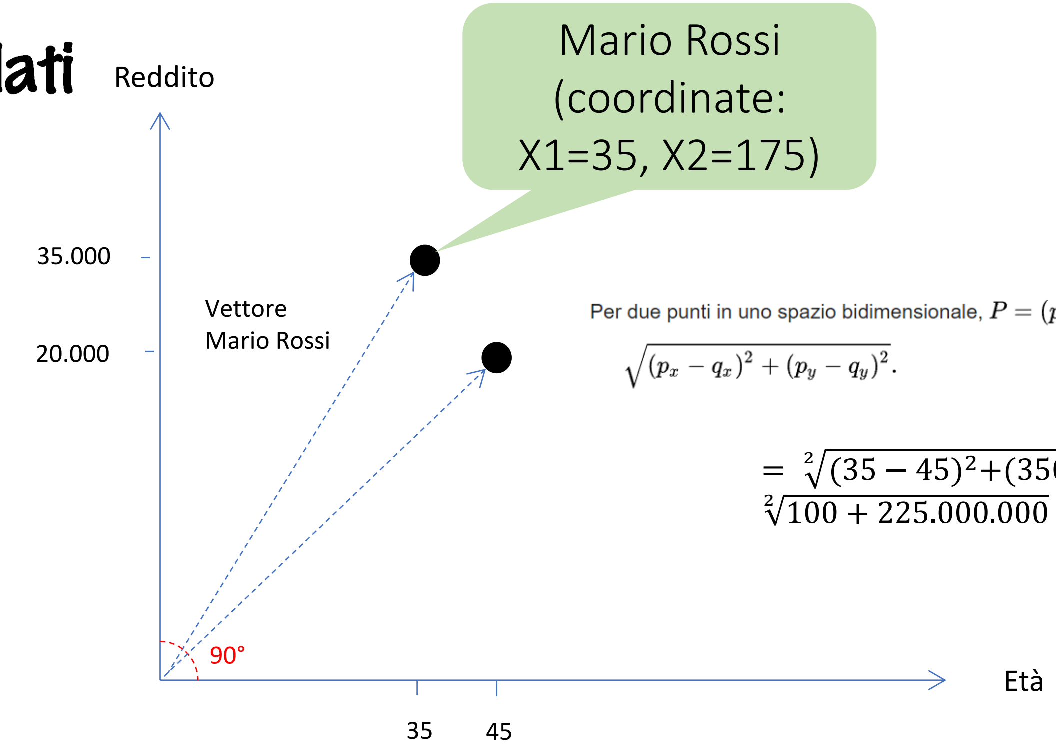
Breve *recap* di algebra lineare per

Machine Learning / Deep Learning

(*per informatici*)

La rappresentazione vettoriale dei dati

Da un punto di vista informatico, Mario Rossi è una riga della tabella, da un punto di vista matematico (algebrico) Mario Rossi è un PUNTO in uno spazio bi-dimensionale ($p = 2$). Tale potente astrazione permette anche di gestire immagini, audio e testi (non-strutturati).



Per due punti in uno spazio bidimensionale, $P = (p_x, p_y)$ e $Q = (q_x, q_y)$, la distanza euclidea è calcolata come: $\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$.

$= \sqrt[2]{(35-45)^2 + (35000-20000)^2} = \sqrt[2]{100 + 225.000.000} = \sqrt[2]{225.000.100} = 15.000$

$\boxed{P=2}$

X1 = Età della persona
X2 = Reddito della persona

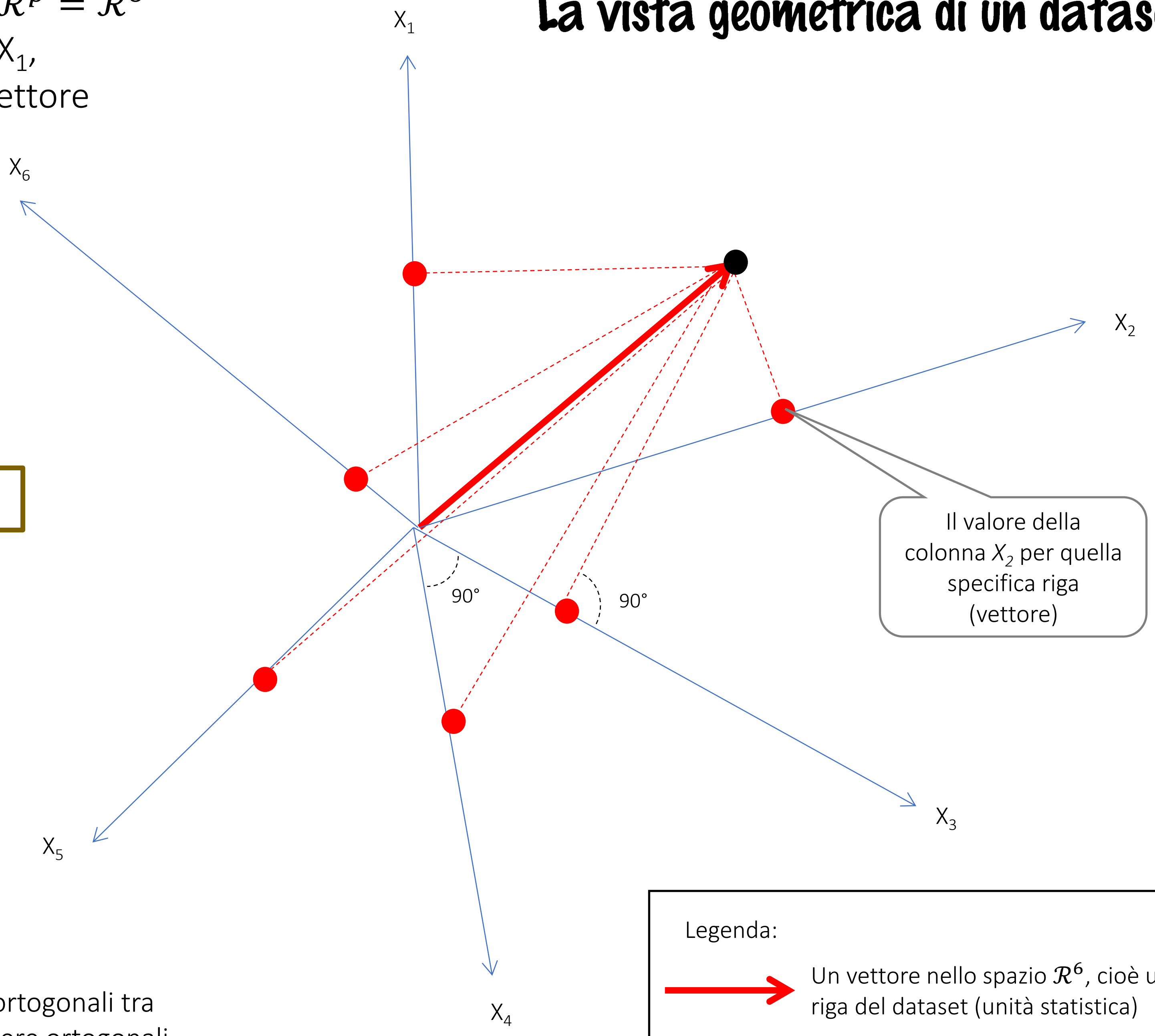
Mario Rossi: età=35, reddito=35.000

	Età	Reddito
Mario Rossi	35	35.000
Giuseppe Verdi	45	20.000

La vista geometrica di un dataset p-dimensionale

Le dimensioni (assi) dello spazio $\mathcal{R}^p = \mathcal{R}^6$ sono qui le colonne del dataset (X_1, X_2,…,X_6), le componenti di ogni vettore sono i valori delle varie righe

p=6

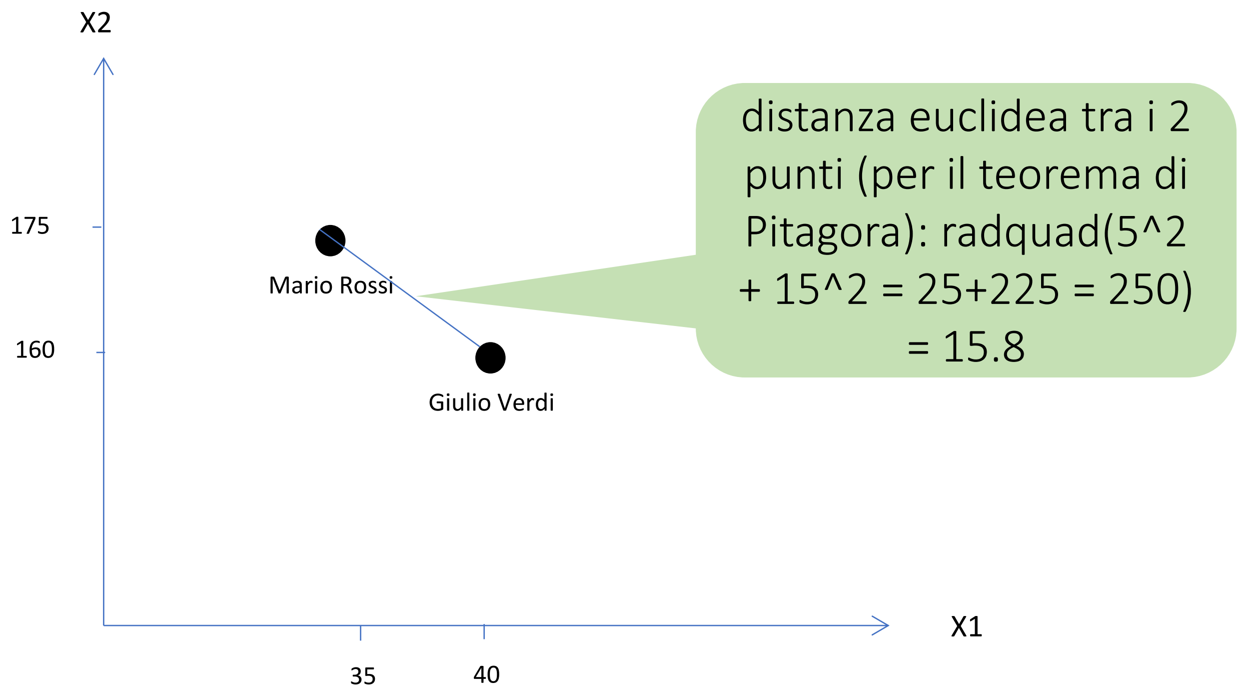


Disclaimer: gli assi dovrebbero essere ortogonali tra loro, tutte le proiezioni dovrebbero essere ortogonali agli assi (non facile da disegnare in uno spazio $\mathcal{R}^6$!)

Legenda:

Un vettore nello spazio $\mathcal{R}^6$, cioè una riga del dataset (unità statistica)

La componente del vettore rispetto alle varie dimensioni.


X2
175
160
Mario Rossi
Giulio Verdi
X1
35
40
distanza euclidea tra i 2 punti (per il teorema di Pitagora): radquad(5^2 + 15^2 = 25+225 = 250) = 15.8

Segnale vs rumore

In questo fenomeno, ovvero nel **processo generativo dei dati** che è ad esso sotteso, e che in genere NON conosciamo, è predominante la parte sistematica (il segnale) o la parte casuale (il rumore)?

$$Y = f(\underline{X}) + \varepsilon$$

Il rumore è un catch-all, che include 2 componenti principali: **lack-of-fit** ed **errore puro**. I residui sono la miglior stima del rumore; per distinguere tra lack-of fit ed errore puro occorrono delle repliche nel campione dati (cioè stesso $\underline{X}$, varie y)

$$\hat{Y} = \hat{f}(\underline{X}^*)$$

Fittare («adattare») un modello (predittivo, in questo caso) ai dati significa stimare dai dati la funzione $\hat{f}$ (il modello).

X-bar = vettore delle vere variabili esplicative del fenomeno
X-bar* = vettore delle variabili esplicative disponibili (eliminarne parti per evitare overfitting, cioè modelli troppo complessi, oppure aggiungerne, se si hanno le conoscenze di business).
Abbiamo a disposizione un campione di training che riporta le vere Y e il vettore X-bar campionato

I due spazi del Machine Learning

$\mathcal{R}^p$: lo spazio delle variabili

(ML non-supervisionato)

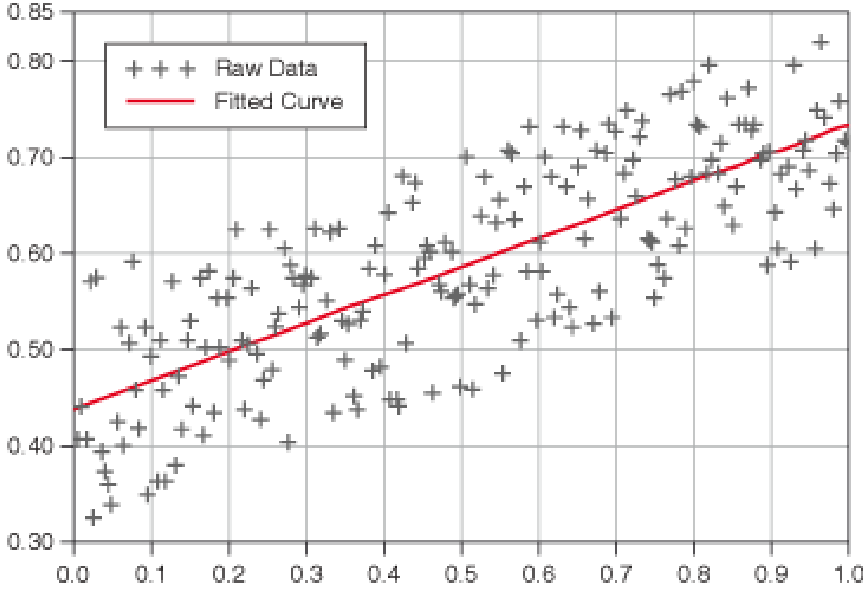
$\mathcal{R}^{p+1}$: lo spazio completo

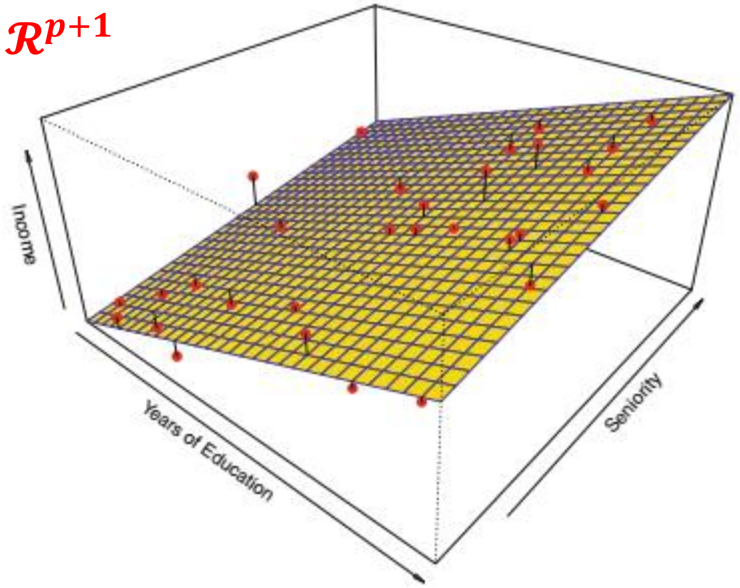
(ML supervisionato)

in questo spazio si trova $f(\underline{X})$, ma non la soluzione $\underline{\theta}^* \in \mathcal{R}^\theta$ del modello di ottimizzazione - vedi dopo

Il fitting lineare dei dati nel ML supervisionato in $\mathcal{R}^{p+1}$

Approssimazione numerica

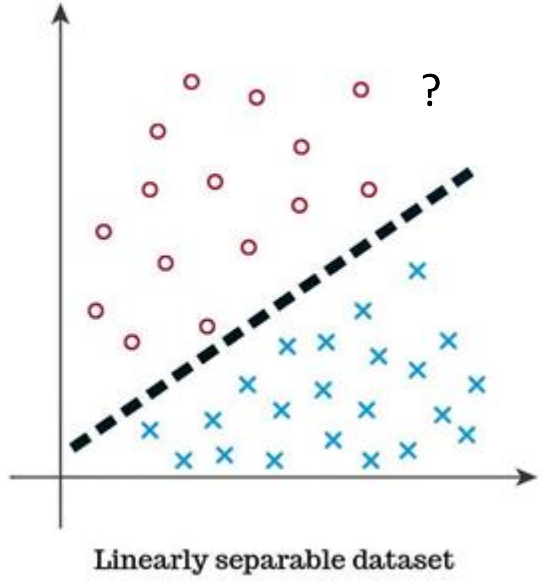


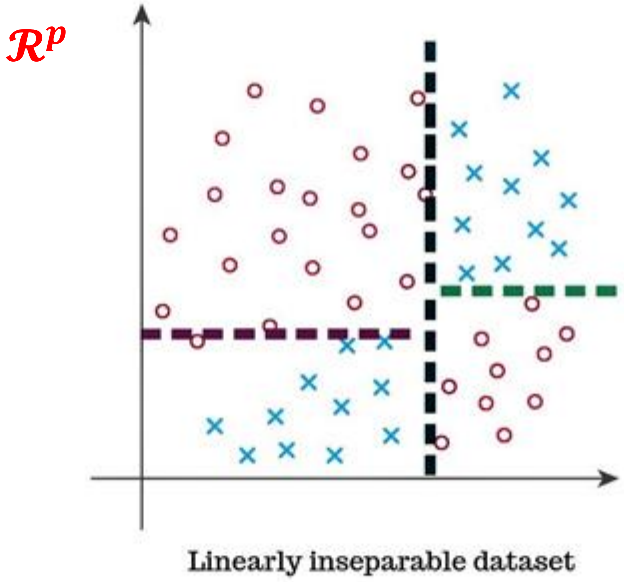


Fitting lineare con:

- Algoritmi lineari (regressione, SVM)
- NN con funzione di attivazione lineare

Decision Boundary

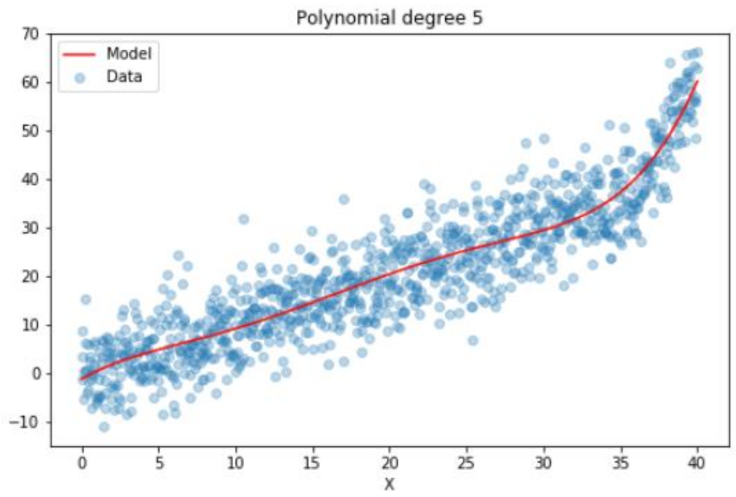


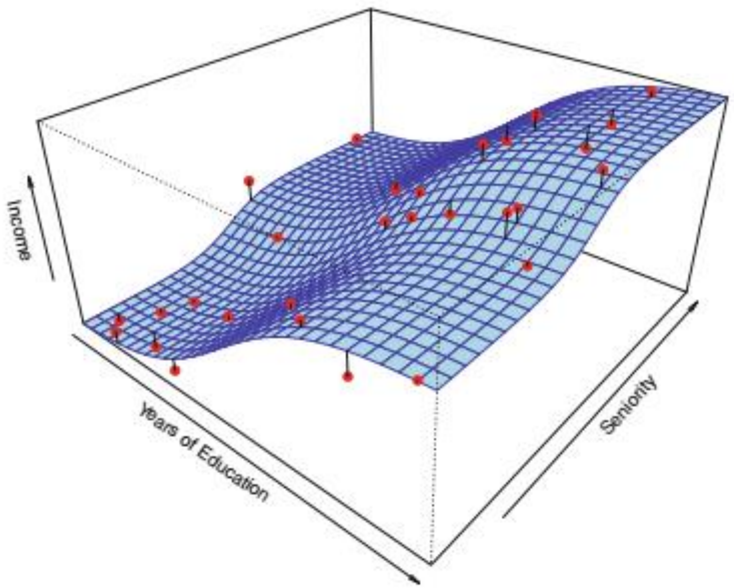


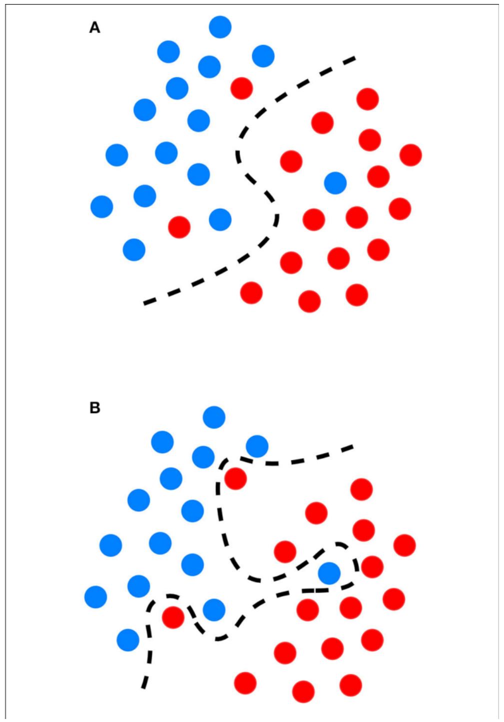
Uno spazio lineare (retta, piano, iperpiano) è definito da una combinazione lineare delle variabili: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. I coefficienti del modello (reale e stimato) sono in genere espressi tramite lettere greche.

Il fitting non-lineare dei dati nel ML supervisionato $\mathcal{R}^{p+1}$

Approssimazione numerica







Decision Boundary

Sintomo tipico di fit lineare non adeguato è un errore di previsione sul test (RMSE oppure l'accuratezza) alto.
Fitting non lineare con:

- Algoritmi non-lineari (KNN, ecc)
- Cambiando la base, cioè elevare al quadrato od al cubo alcuni predittori (parabole, sinusoide, interpolazioni)
- NN con funzione di attivazione non-lineare

L'inferenza

Nel ML supervisionato, la funzione reale, quella che genera la popolazione e che si identifica con $f(\underline{X})$, è stimata (approssimata) tramite i dati disponibili, sui quali si fitta una funzione $\hat{f}(\underline{X})$.

Questo processo è detto inferenziale, perché «induce» (aristotelicamente) una conoscenza generale da una particolare (cioè dal campione alla popolazione).

La funzione f è la parte deterministica del modello, il quale è formato anche da una parte stocastica (l'errore ε).

Lo spazio θ dei parametri

Dietro molti problemi di ML, supervisionati oppure no, c'è un problema di ottimizzazione:

$$\underline{\theta}^* = \underset{\underline{\theta}}{argmin}\, \mathcal{L}(\underline{x}, \theta)$$

dove $\mathcal{L}(\underline{x}, \underline{\theta})$ è la funzione di costo (*loss function*)

Lo spazio nel quale si opera è $\mathcal{R}^{\theta}$, lo spazio dei parametri del modello, cioè le variabili della loss function da minimizzare.

Anche le reti neurali profonde sono dentro questa framework concettuale.

Due rappresentazioni geometriche del dataset **X**

(da Johnson & Wichern)

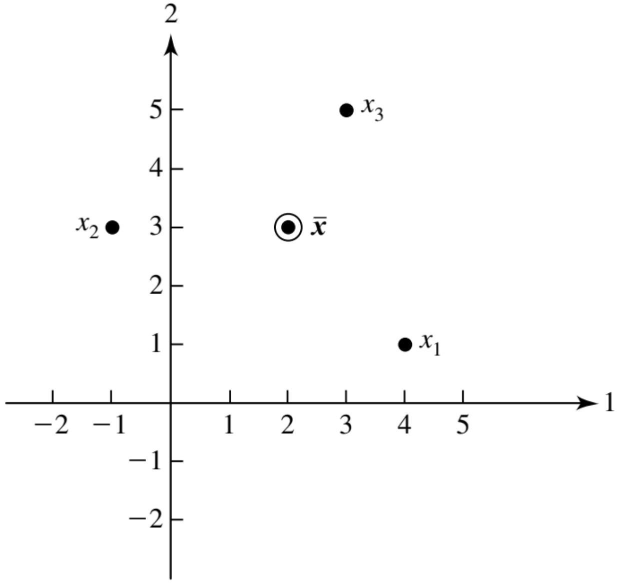


Figure 2.1 A plot of the data matrix **X** as $n = 3$ points in $p = 2$ space.

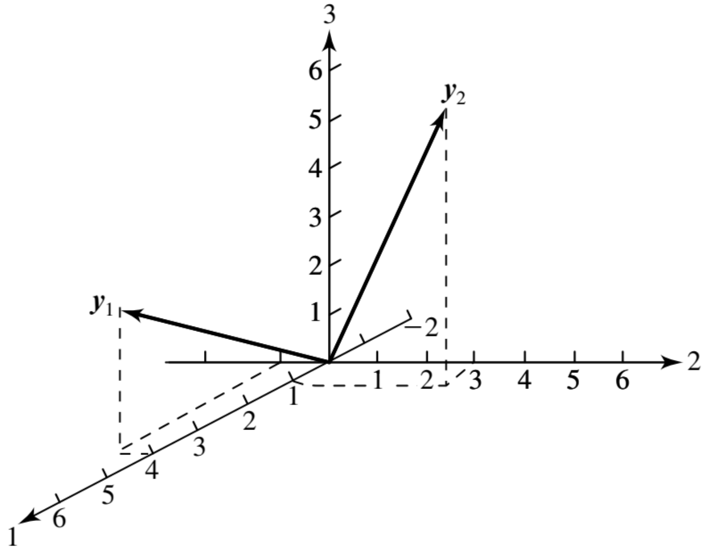


Figure 2.2 A plot of the data matrix **X** as $p = 2$ vectors in $n = 3$ space.