Introduzione al Pattern Recognition

Cosa si intende per **pattern recognition**?

"The assignment of a physical object or event to one of several pre-specified categories" (Duda & Hart)

- Un **pattern** è un oggetto, un processo o un evento a cui può essere assegnata una classe.
- Una **classe** (o **categoria**) è un insieme di pattern che condividono caratteristiche ed attributi comuni.
- Per **recognition** (o **classificazione**) si intende il processo che assegna i pattern a classi pre-definite.
- Un **classificatore** è un sistema che effettua un pattern recognition ed eventualmente prende una decisione.

Un esempio

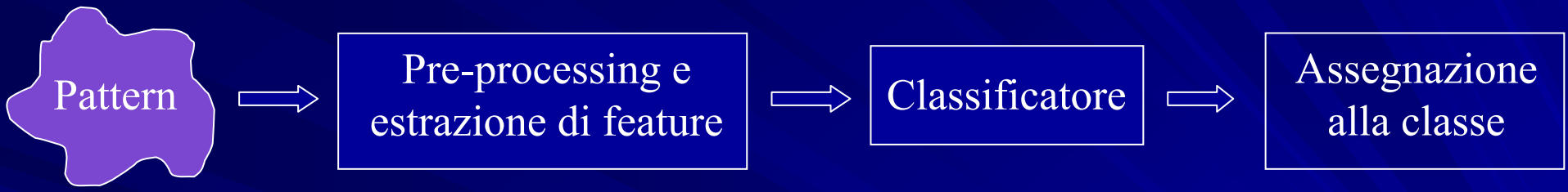
- oggetto: animale
- classe: tigre
- sistema: cervello umano
- decisione: stai alla larga!

- oggetto: animale
- classe: gatto
- sistema: cervello umano
- decisione: stai tranquillo…

Applicazioni

- **Character Recognition:** handwritten (sorting letters by postal code) or printed texts (digitalization of text documents).
- **Biometrics:** face/fingerprints/speech recognition, finger prints recognition.
- **Diagnostic systems:** assisted medical diagnosis (CAD system), DNA sequence identification.
- **Military applications:** automated target recognition, image segmentation and analysis (recognition from aerial/satellite photographs).

Struttura di un sistema di PR



Esempio: classificazione del pesce in base alla specie

Come procedere?

- **Sensing**: fotografare i pesci mediante una camera;
- **Pre-processing**: utilizzare un algoritmo di segmentazione per isolare un pesce dall'altro e/o dal background/rumore
- **Estrazione di caratteristiche (features)** significative per la classificazione delle due specie:
  - lunghezza;
  - peso;
  - larghezza;
  - numero e forma delle alette;
  - posizione della bocca;
- **Classificazione** mediante un sistema che sfrutta l'informazione contenuta nelle feature

Estrazione di feature

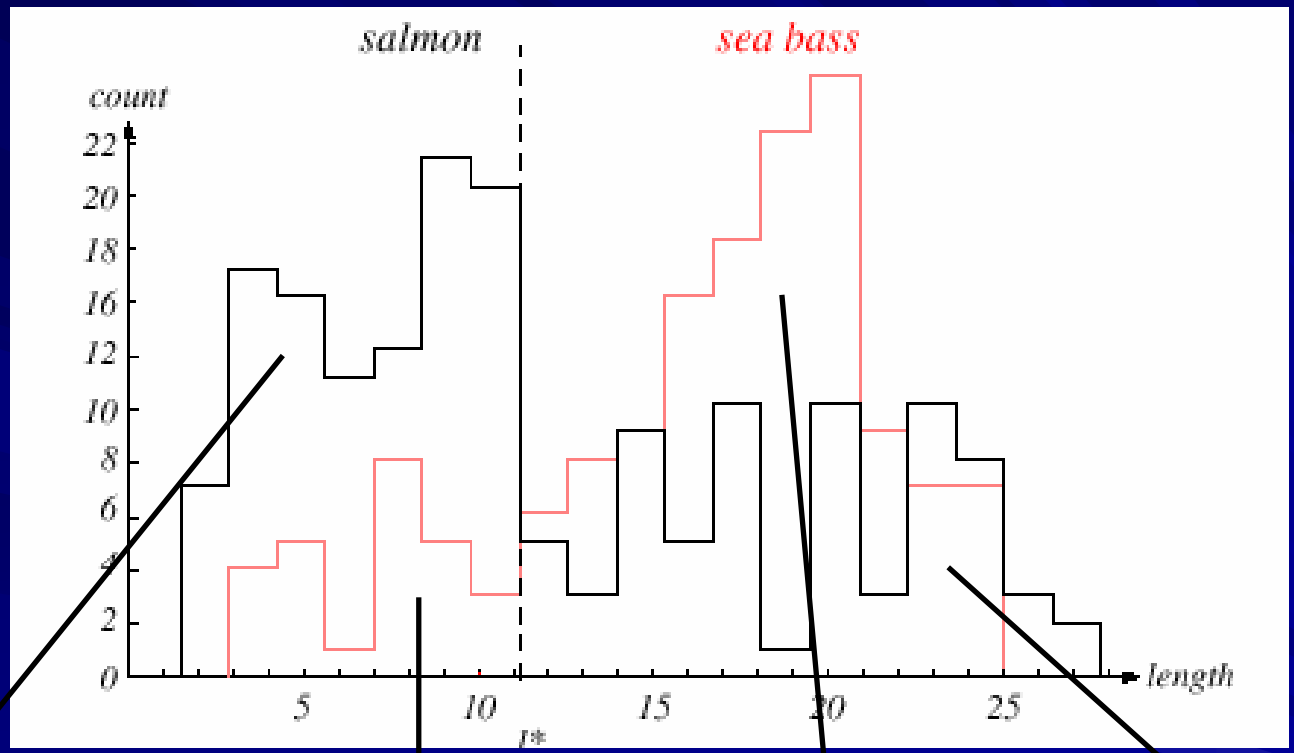
- Conoscenza *a priori:* la triglia è generalmente più lunga del salmone;
- Utilizziamo la lunghezza come caratteristica significativa;
- Il classificatore assegnerà la classe del pesce in base alla seguente regola:

  $L < L_{critica}$ ➡ salmone

  $L > L_{critica}$ ➡ spigola
- Costo della misclassificazione (dipende dall'applicazione):

  è meglio classificare un salmone come triglia o viceversa?
  - se metto i salmoni nel canestro delle triglie ➡ perdo profitto
  - se metto le triglie nel canestro dei salmoni ➡ perdo clienti
- La lunghezza critica viene determinata in modo da minimizzare il costo associato alla classificazione.

Come? sulla base di un set di esempi (*training set*) mediante analisi dell'istogramma:



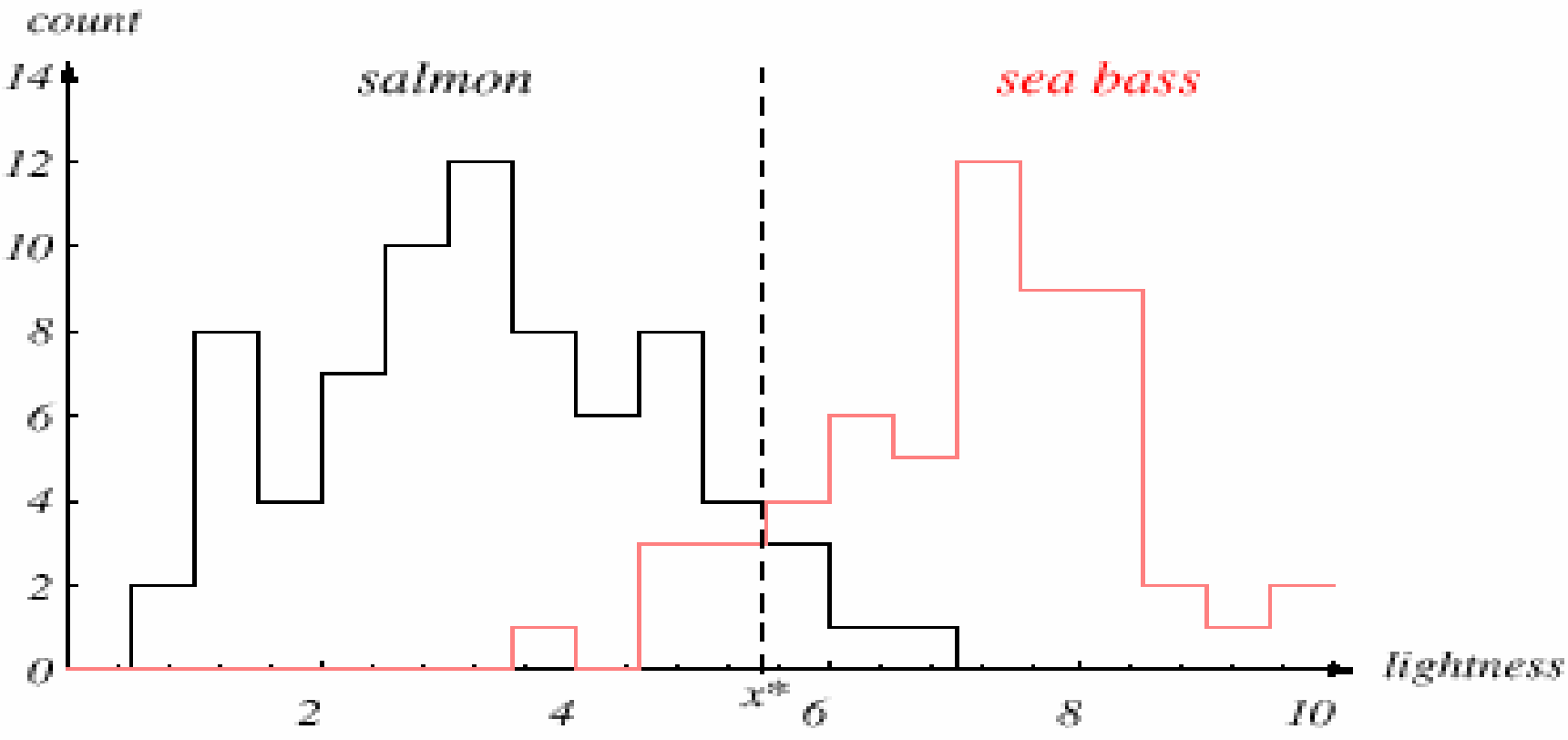
salmone correttamente classificato

spigola classificata come salmone

spigola correttamente classificata

salmone classificato come spigola

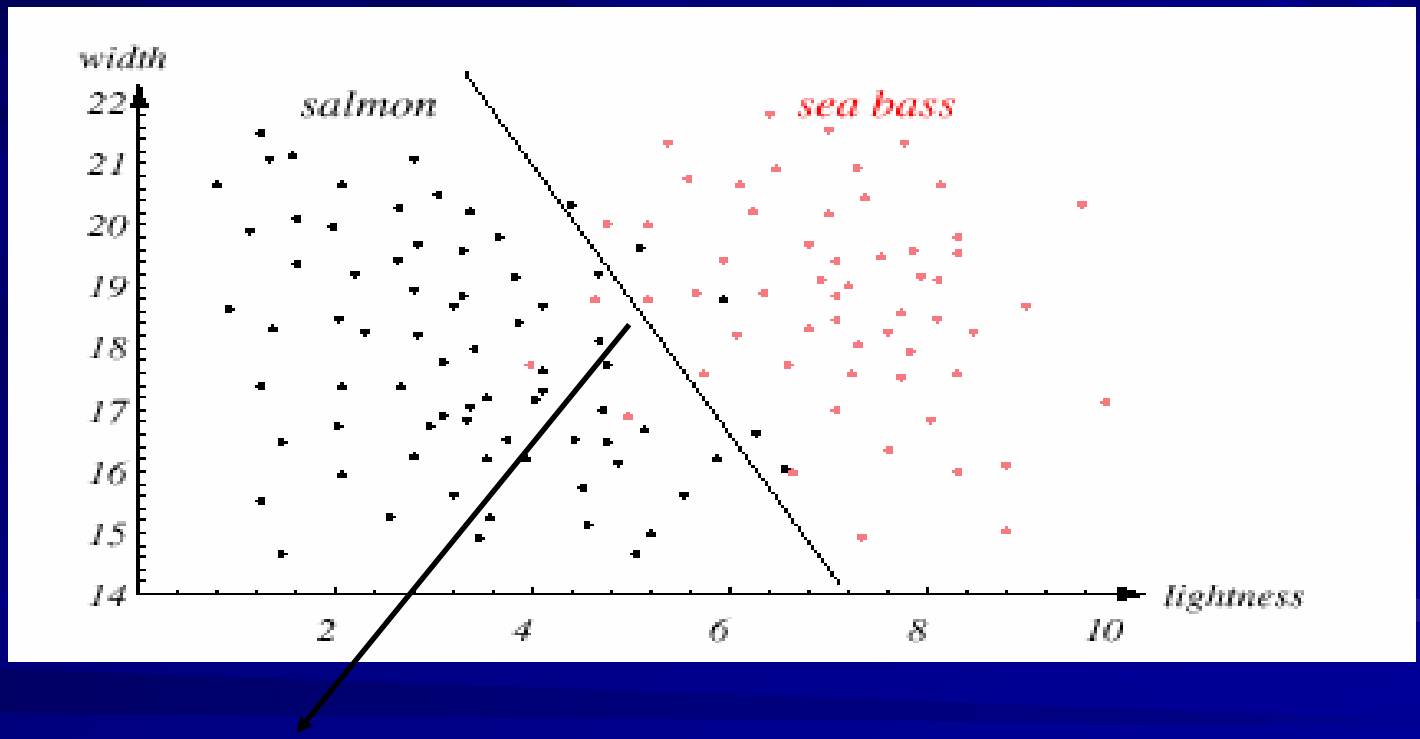
Analisi delle feature: il peso



- Consideriamo 2 features: peso e larghezza; pesce ➡ $\mathbf{x} = [x_1, x_2]$;
- Obiettivo: partizionare lo spazio delle caratteristiche in due regioni:

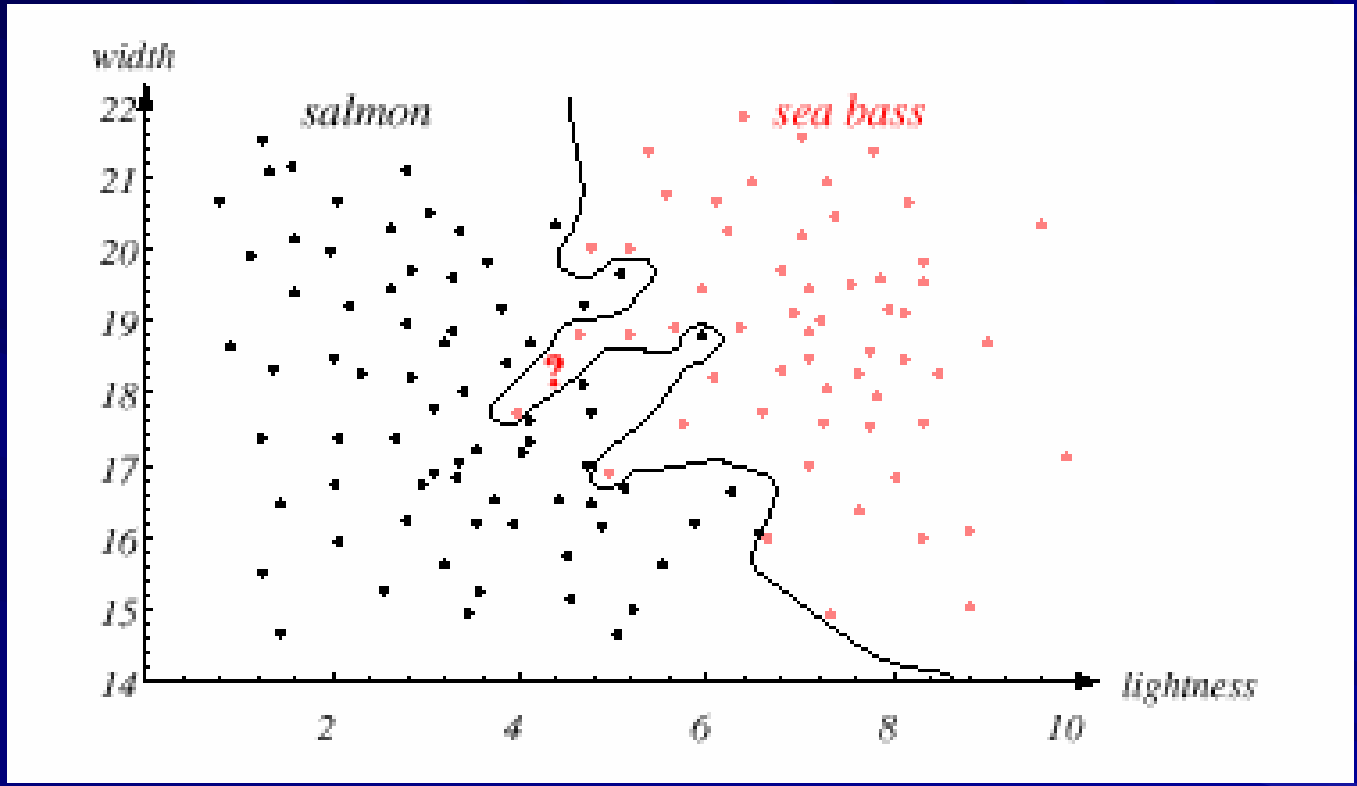
  regione 1 ➡ salmone

  regione 2 ➡ triglia



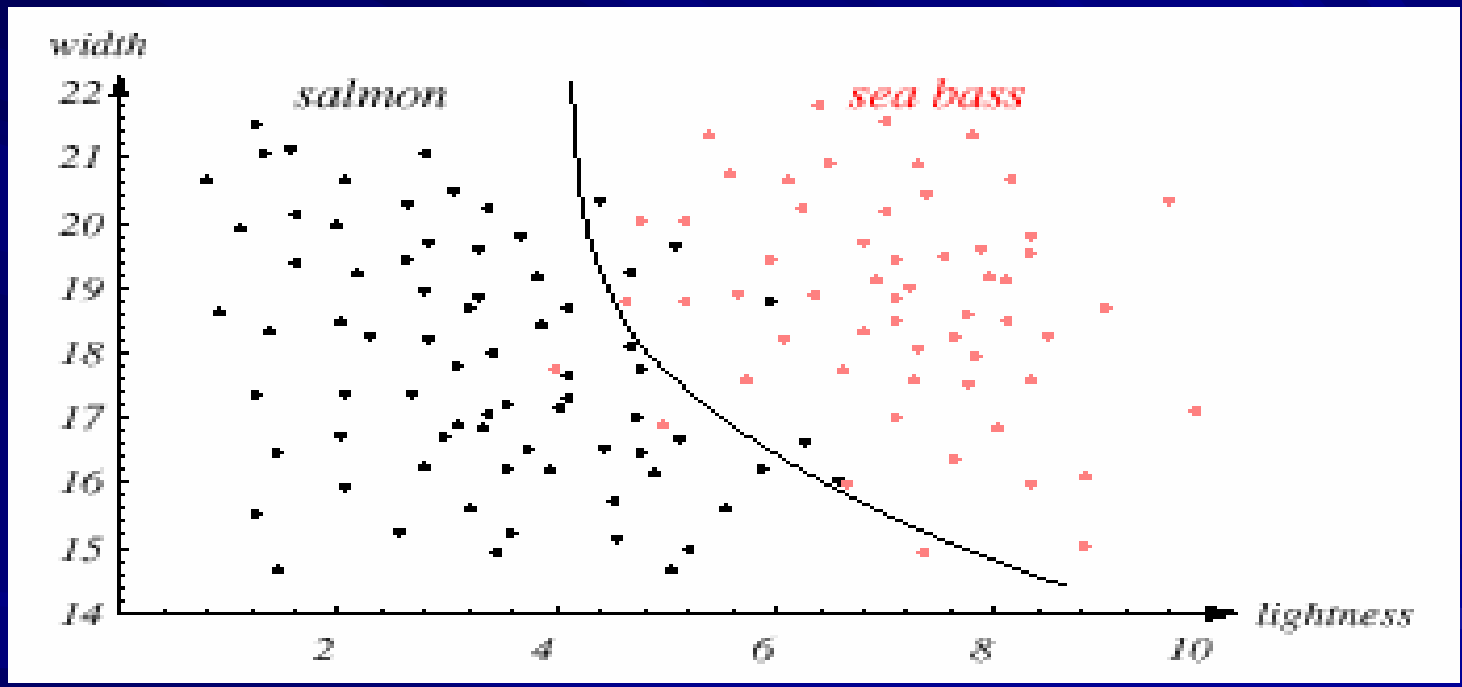
Modello lineare: semplice da implementare, ma generalmente con molte misclassificazioni

- Modelli complessi:

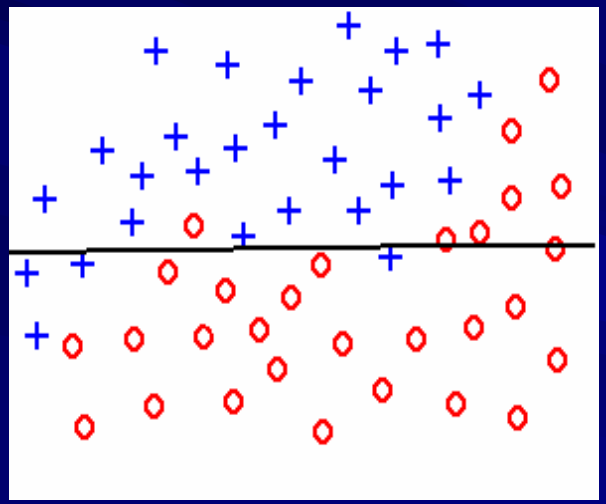


- Possono separare i dati perfettamente ma hanno scarsa capacità di generalizzazione

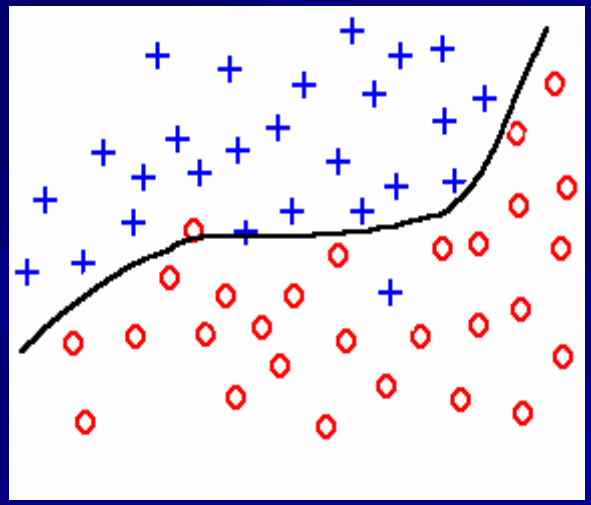
- Il modello ottimale è quello che ha la capacità di classificare al meglio dati mai visti prima (generalizzazione)



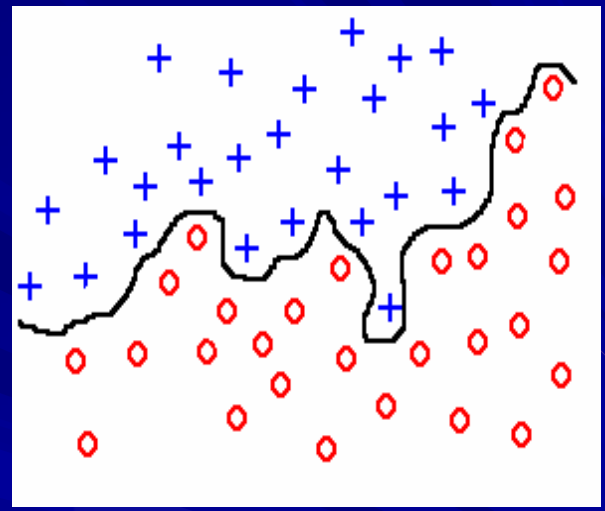
Overfitting e underfitting



underfitting

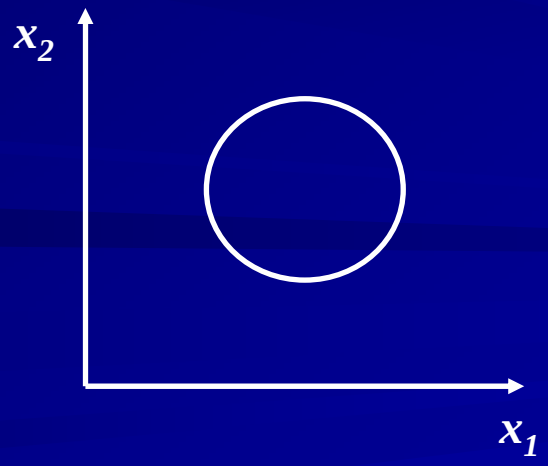


buon fit



overfitting

- Possiamo pensare di aggiungere altre feature che non siano correlate con le precedenti (informazione ridondante) e che non siano "rumorose", altrimenti si ha una riduzione della performance di classificazione.
- Inoltre l'aumento delle feature implica il fenomeno noto come "course graining" (aumento esponenziale dei dati di training necessari)
- Un dataset in *d* dimensioni ha una *dimensionalità intrinseca* $d' \leq d$ se i dati giacciono interamente in un sotto-spazio di dimensione *d'*



Principal Component Analysis

- per eliminare le correlazioni fra le features
- per ridurre la dimensionalità del dataset

Si tratta di una trasformazione che permette di ottenere un nuovo set di variabili dette componenti principali. Queste variabili sono scorrelate fra loro e sono ordinate in base alla frazione di informazione, i.e. varianza dei dati. Di queste variabili se ne possono mantenere un numero minore rispetto al dataset originale, ma tali da contenere la maggior parte dell'informazione significativa del dataset.

Definizione algebrica delle PC

Date N osservazioni in uno spazio a d variabili $\mathbf{x}^n = (x_1, x_2, \ldots, x_d)$, con $n = 1, \ldots N$, vogliamo approssimarle con un set di vettori $\mathbf{z}^n = (z_1, z_2, \ldots, z_M)$, con $M < d$:

$$\vec{x} = \sum_{i=1}^{d} z_i \vec{u}_i \approx \sum_{i=1}^{M} z_i \vec{u}_i + \sum_{i=M+1}^{d} b_i \vec{u}_i \qquad \vec{u}_i^{\,T} \vec{u}_j = \delta_{ij}$$

Determiniamo $\boldsymbol{u}_i$ e b_i in modo da minimizzare l'approssimazione:

$$\vec{x} - \vec{x}_{approx} = \approx \sum_{i=M+1}^{d} (z_i - b_i)\vec{u}_i$$

L'errore da minimizzare è:

$$E = \frac{1}{2} \sum_{n=1}^{N} \left\| \vec{x}^n - \vec{x}^n{}_{approx} \right\|^2 = \frac{1}{2} \sum_{n=1}^{N} \sum_{i=M+1}^{d} (z_i{}^n - b_i)^2$$

Definizione algebrica delle PC

Derivando E_M rispetto alle b_i e ponendo nulle le derivate si ottiene:

$$b_i = \frac{1}{N}\sum_{n=1}^{N} z_i^{\,n} = \vec{u}_i^{\,T} \cdot \langle \vec{x} \rangle$$

Quindi l'errore minimo è:

$$E_{\min} = \frac{1}{2}\sum_{i=M+1}^{d}\sum_{n=1}^{N}\left\{\vec{u}_i^{\,T} \cdot (\vec{x}^{\,n} - \langle \vec{x} \rangle)\right\}^2 = \frac{1}{2}\sum_{i=M+1}^{d} \vec{u}_i^{\,T} \Sigma \vec{u}_i$$

$$\Sigma = \sum_{n=1}^{N} (\vec{x}^{\,n} - \langle \vec{x} \rangle) \cdot (\vec{x}^{\,n} - \langle \vec{x} \rangle)^T$$

Rimane il problema di minimizzare E_{min} rispetto alla base $\boldsymbol{u}_i$. Si mostra che ciò avviene quando la base soddisfa l'equazione

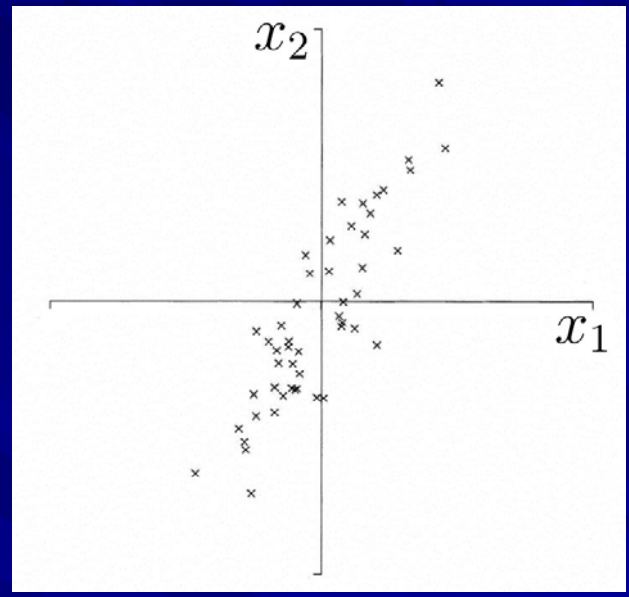
$$\Sigma \vec{u}_i = \lambda_i \vec{u}_i \Rightarrow E_{\min} = \frac{1}{2}\sum_{i=M+1}^{d} \lambda_i$$

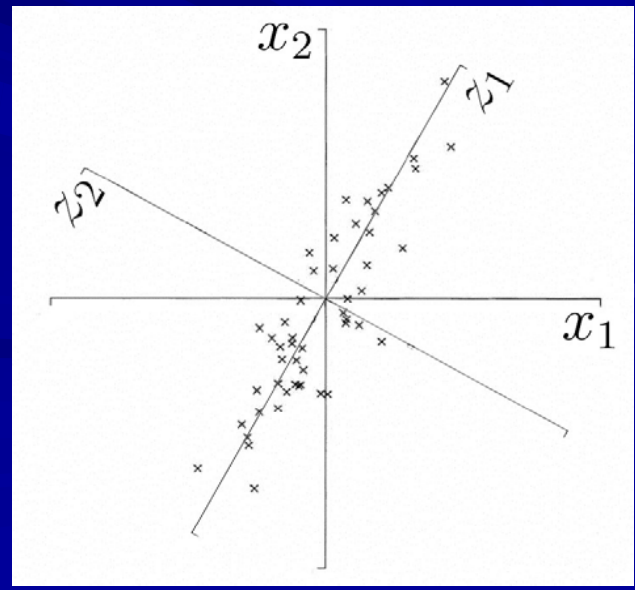
Spiegazione geometrica

- Dato il campione di *N* dati in uno spazio 2D: $\mathbf{x} = (x_1, x_2)$, l'algoritmo si calcola la matrice di covarianza e determina autovettori e autovalori.
- Si proiettano i vettori $\boldsymbol{x}^n$ sugli autovettori relativi agli *M* autovalori maggiori, ottenendo le nuove variabili:

$$z_i = \vec{u}_i^{\,T} \cdot \vec{x}$$

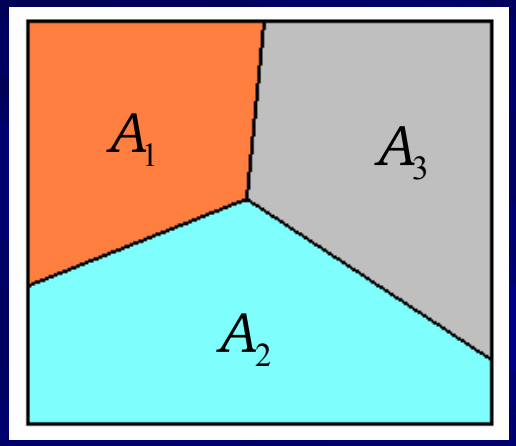
- Ciò corrisponde ad effettuare una rotazione degli assi: la 1ª PC z_1 è la direzione di massima varianza dei dati; la 2ª PC z_2 è la direzione ortogonale alla prima e così via.





Classificatori

- Un classificatore partiziona lo spazio A dei dati di input in K subset disgiunti $\{A_1, ..., A_K\}$



$$A = \bigcup_{i=1}^{K} A_i \qquad \bigcap_{i=1}^{K} A_i = \{0\}$$

in modo che se l'oggetto $\mathbf{X} = (X_1, ..., X_n) \in A_k$, allora appartiene alla classe k-esima:

$$C : \mathbf{X} \in A \rightarrow \{1,2, ... ,K\} \Leftrightarrow C(\mathbf{X}) = k \Leftrightarrow \mathbf{X} \in A_k$$

Tipologia di classificatori

- Classificatori non supervisionati: le classi sono sconosciute, l'algoritmo deve scoprire da sé correlazioni e similitudini fra i dati. Vengono usati quando non si ha nessuna conoscenza *a priori* sul dataset.
- Classificatori supervisionti: le classi sono note; questa conoscenza *a priori* viene utilizzata per "addestrare" (*training* o *learning*) il classificatore per future osservazioni.

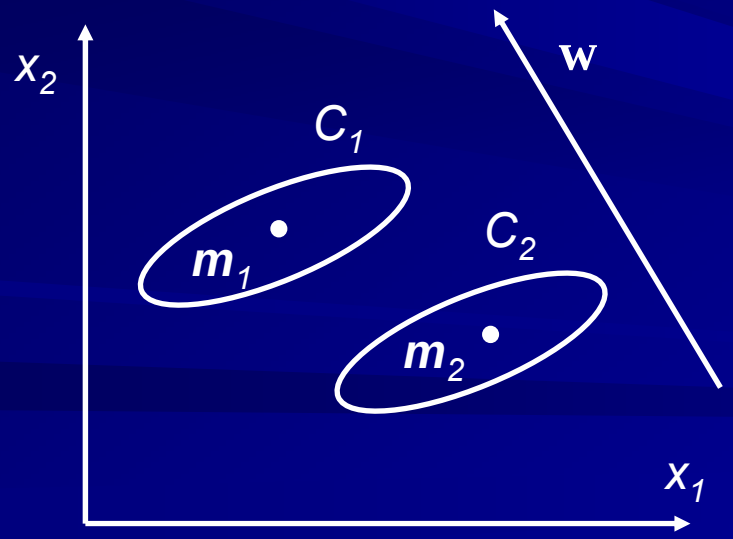
Esempi di metodi di classificazione

- Discriminante Lineare di Fisher.
- Regressione lineare e logistica.
- Reti neurali.
- Algoritmi di clustering.

Discriminante lineare di Fisher

R.A.Fisher. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179-188, 1936.

- Quando? in un problema di classificazione a 2 classi.
- Scopo: trovare un vettore **w** tale che la proiezione dei dati su questo vettore massimizza la separazione fra le classi.



Supponiamo di avere N_1 oggetti di classe C_1 e N_2 di classe C_2 e siano $\boldsymbol{m}_1$ e $\boldsymbol{m}_2$ i vettori medi:

$$m_1 = \frac{1}{N_1} \sum_{x \in C_1} x$$

$$m_2 = \frac{1}{N_2} \sum_{x \in C_2} x$$

Discriminante lineare di Fisher

- Si determina il vettore **w** in modo da massimizzare la separazione fra i vettori medi proiettati e minimizzare la covarianza totale *intra*-classe dei dati proiettati:

$$J(w) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{S}_1 + \tilde{S}_2}$$

- $\tilde{m}_i = \frac{1}{N_i} \sum_{x \in C_i} w' x = w' m_i$ vettore medio dei punti della classe C_i proiettati su **w**;

- $\tilde{S}_i = \sum_{x \in C_i} (w' x - w' m_i)^2 = w' S_i w$ varianza dei punti della classe C_i proiettati su **w**

$$S_i = \sum_{x \in C_i} (x - m_i)^2$$

Discriminante Lineare di Fisher

- Con un po' di algebra: $J(w) = \dfrac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{S}_1^2 + \tilde{S}_2^2} = \dfrac{w' S_B w}{w' S_w w}$
- $S_B = (m_2 - m_1) \cdot (m_2 - m_1)'$ matrice di covarianza *between-class*
- $S_W = S_1 + S_2$ matrice di covarianza *within-class* totale
- Massimizzando $J(\mathbf{w})$:

$$\frac{\partial J(w)}{\partial w} = 0 \Rightarrow w \propto S_W^{-1}(m_2 - m_1)$$

Regressione logistica

- Regressione lineare semplice e multipla.
- Regression logistica semplice:
  - La funzione logistica
  - Stima dei parametri
- Regressione logistica multipla.
- Logit e Probit.

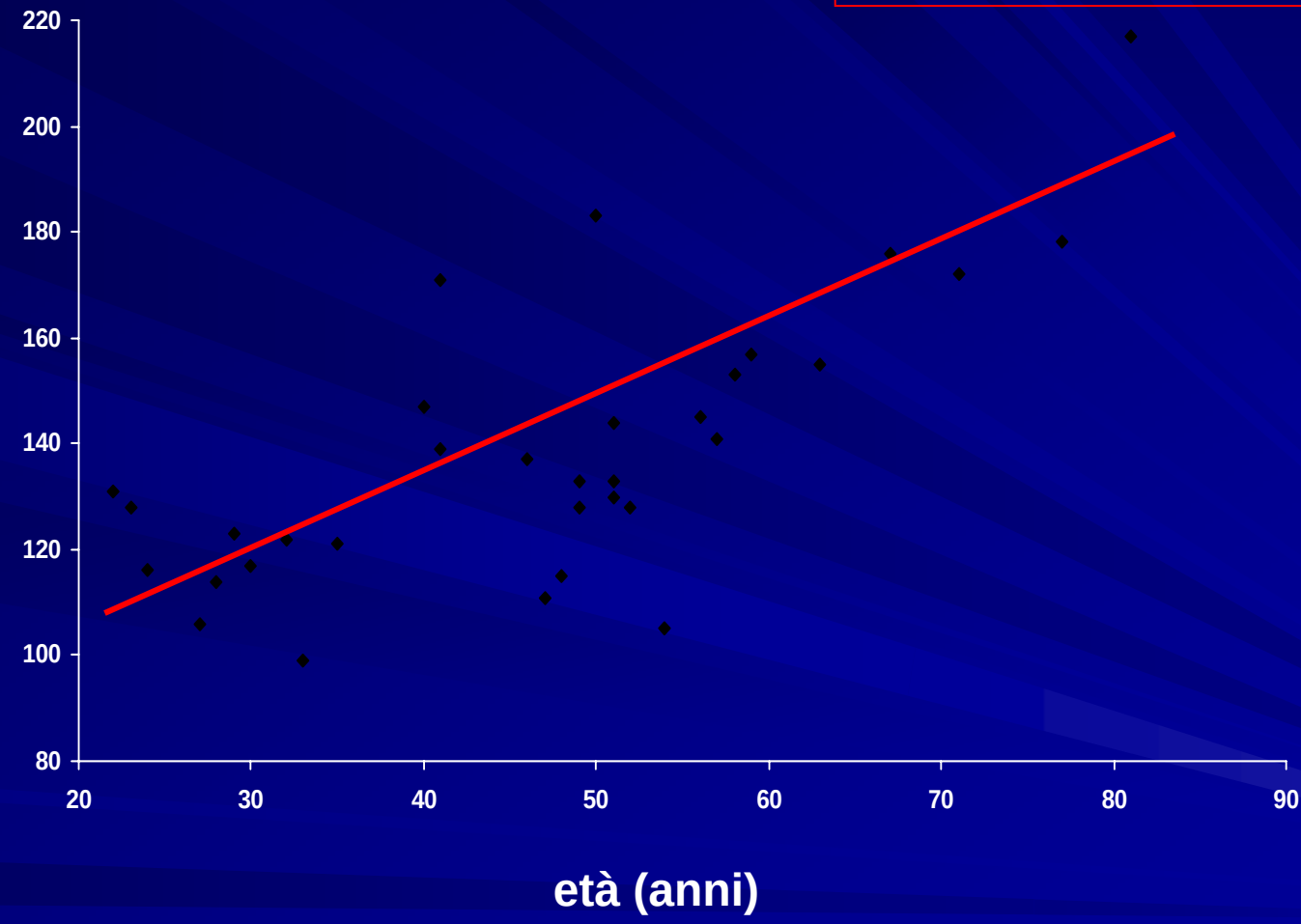
Regressione lineare semplice

- Modella una relazione lineare fra una variabile indipendente *x* continua e una variabile dipendente *y* continua

Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

Tabella 1: età e pressione sistolica del sangue in 33 donne adulte

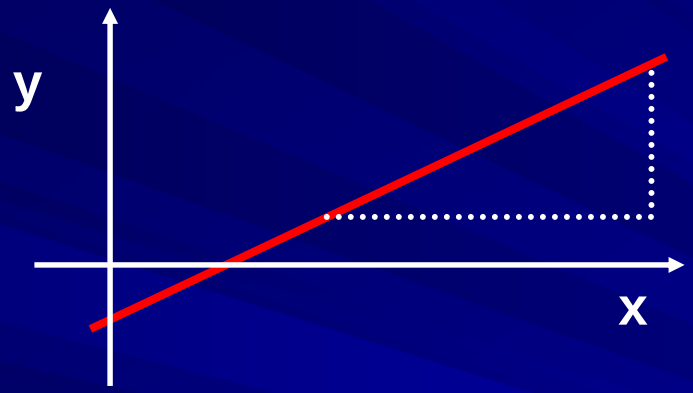
SBP = 81.54 + 1.222 x età



T. Colton, Statistics in Medicine. Boston: Little Brown, 1974

Regressione lineare semplice

- Relazione fra 2 variabili continue (SBP e età)



$$y = \alpha + \beta x$$

- I coefficienti di regressione α e β vengono stimati con il metodo dei minimi quadrati:

$$E = \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i \right)^2$$

Stima dei coefficienti di regressione

$$\begin{cases} \dfrac{\partial E}{\partial \alpha} = 0 \\ \dfrac{\partial E}{\partial \beta} = 0 \end{cases} \Rightarrow \begin{cases} \alpha = \langle y \rangle - \beta \langle x \rangle \\ \beta = \dfrac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} \end{cases}$$

dove:

$$\mathrm{cov}(x, y) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$$

$$\mathrm{var}(x) = \mathrm{cov}(x, x)$$

Regressione lineare multipla

- Relazione fra una variabile continua e un set di variabili continue:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

- Esempio:
  - SBP in funzione di età, peso, altezza, etc

Regressione Logistica

- Modella la relazione fra una variabile *x* dicotomica (o binaria) e/o continua e una variabile dicotomica *y*

- *y* variabile dicotomica $\Leftrightarrow$ y = 0,1 dove:

  – y = 1 indica il verificarsi di un evento di interesse, chiamato “successo”;

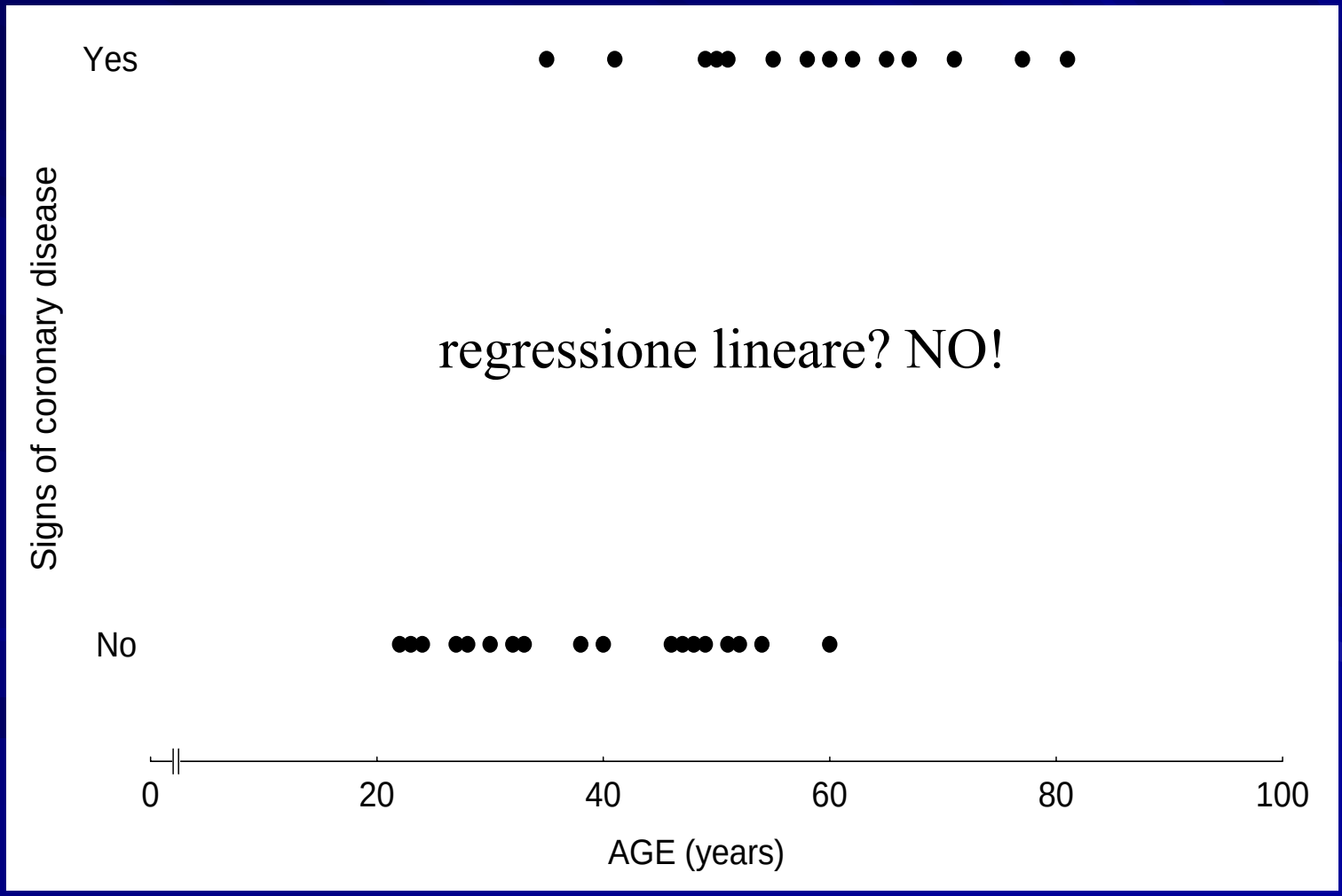
  – y = 0 per l’evento complementare.

Esempio

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

Tabella 2: età e sintomi di infarto coronarico (CD)

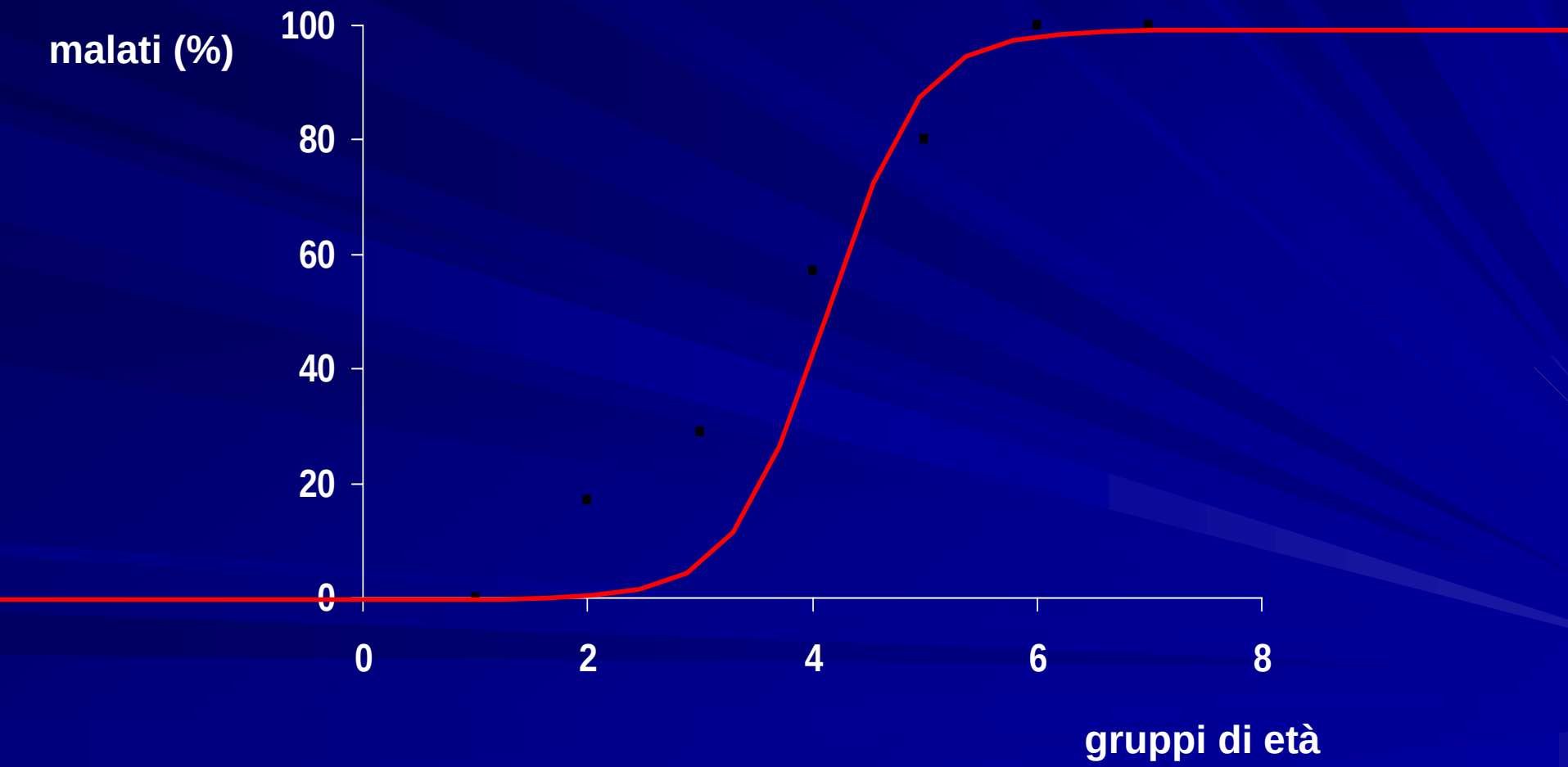
Come analizzare questi dati?



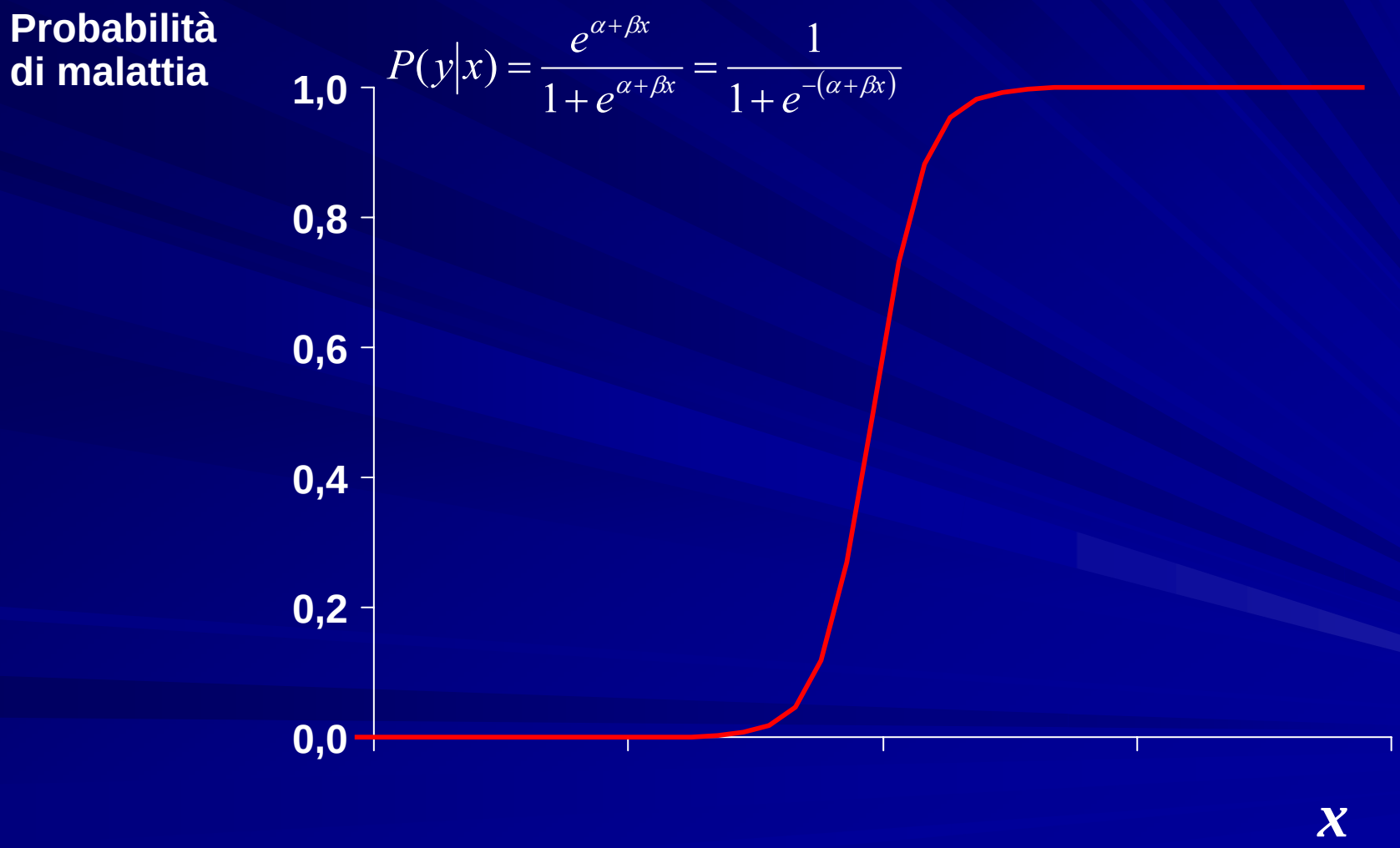
		Diseased	
Age group	# in group	#	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100

Tabella 3: percentuale di sintomi di CD secondo gruppi di età

Plot dei dati della tabella 3



Funzione logistica



Modello logistico

- Un modello di regressione logistica definisce la probabilità di successo di un evento $P(y|x)$ data l'osservazione x.

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \Rightarrow \underbrace{\ln\left[\frac{P(y|x)}{1-P(y|x)}\right]}_{\log it(P(y|x))} = \alpha + \beta x$$

$$\log it(P(successo)) = \log\left[\frac{P(successo)}{P(insuccesso)}\right]$$

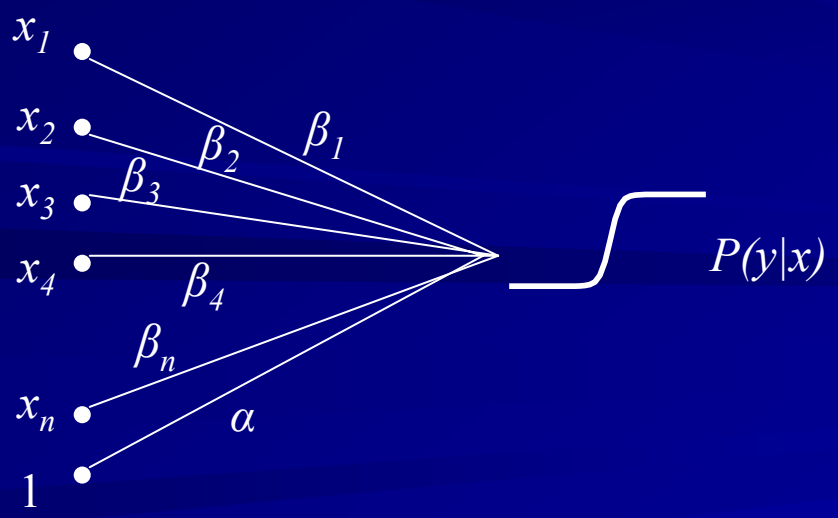
- Probit: $P(y|x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{t^2}{2}} dt$

Regressione logistica multipla

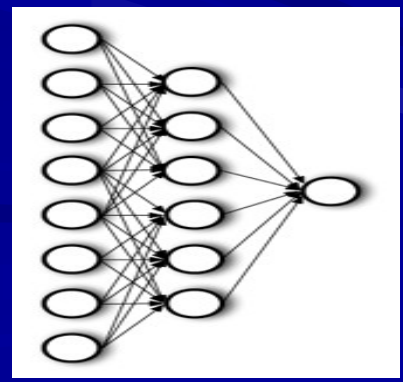
- Più di una variabile indipendente:

$$\ln\left[\frac{P}{1-P}\right] = \alpha + \sum_{i=1}^{n} \beta_i x_i$$

- Schematizzazione:



generalizzando



Rete neurale

References

Books

Duda, Heart: Pattern Classification and Scene Analysis. J. Wiley & Sons, New York, 1982. (2nd edition 2000).

Fukunaga: Introduction to Statistical Pattern Recognition. Academic Press, 1990.

Bishop: Neural Networks for Pattern Recognition. Claredon Press, Oxford, 1997.

Giudici: Data Mining. McGraw-Hill, 2003.