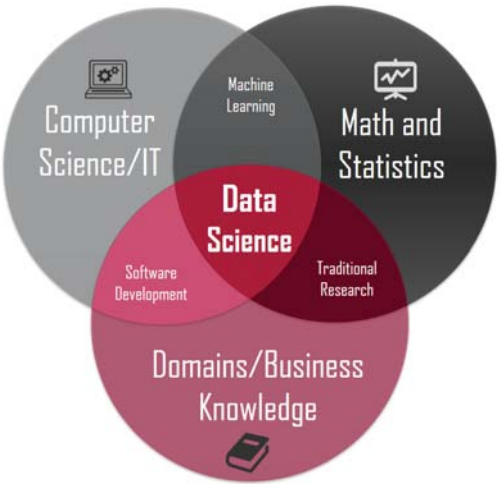
San Pellegrino 03/09/19 – Summer School "La Matematica oggi"

La statistica e la scienza dei dati

Mauro Gasparini Dipartimento di Scienze Matematiche "G.L. Lagrange"
Politecnico di Torino
mauro.gasparini@polito.it
http://calvino.polito.it/~gasparini

Statistica e Data Science

Evoluzione di un diagramma di Venn dovuto a Conway (2013).



https://towardsdatascience.com/

Un problema statistico: la classificazione

Decidere se un oggetto appartiene a una popolazione positiva oppure a una popolazione negativa tramite una serie di rilevazioni sull'oggetto. Esempi:

1. diagnosticare se un paziente sano o malato
2. assegnare a un segnale elettrico un valore binario (vero o falso)
3. giudicare un imputato è colpevole o innocente
4. decidere se un essere vivente è un animale o un vegetale
5. valutare se un debitore sarà insolvente o no

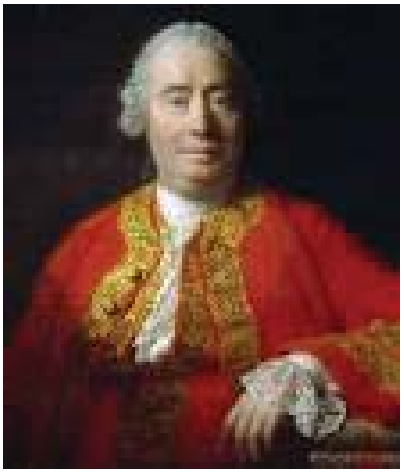
Un problema statistico è un problema di induzione

Cosa vuol dire induzione?

Induzione "scandalo della filosofia".

Aristotele, Bacon, Hume, Mill.

David Hume 1711–1776



Filosofo.
Il problema dell'induzione.
Il concetto di causa.

Cigni neri

"rara avis in terris, nigroque simillima cygno" (Giovenale)
I cigni neri furono effettivamente scoperti in Australia alla fine del Seicento.
Recentemente, l'espressione "cigno nero" ha assunto un significato più specifico di evento imprevedibile e devastante, specialmente in economia e in politica (Taleb).

Il gioco della classificazione dei viventi

Torniamo al problema decidere se un essere vivente un animale o un vegetale.

Poniamoci il problema di costruire delle domande a risposta binaria (per semplicità) sulla base delle quali prendere una decisione (a sua volta binaria, "dico vegetale" oppure "dico animale").

Esempio: "E' verde"?

Idea: usare la probabilità condizionata

La domanda "E' verde" non è perfetta, ma à "buona". Perchè?

Supponiamo di potere quantificare le seguenti probabilità

$$\mathrm{P}(Animale) = 0.45$$
$$\mathrm{P}(Vegetale) = ????$$

$$\mathrm{P}(Verde|Animale) = 0.05$$
$$\mathrm{P}(Verde|Vegetale) = 0.70$$

Che tipo di probabilità sono queste? Cosa significa quella barra verticale?

Riusciamo a vedere un motivo per cui la domanda è "buona"? Vediamo.

Aggiornamento bayesiano

Se osserviamo "Verde", possiamo usare la probabilità condizionata inversa

$$\begin{aligned}\mathrm{P}(Animale|Verde) &= \frac{\mathrm{P}(Animale \cap Verde)}{\mathrm{P}(Verde)} \\ &= \frac{\mathrm{P}(Animale)\,\mathrm{P}(Verde|Animale)}{\mathrm{P}(Animale)\,\mathrm{P}(Verde|Animale) + \mathrm{P}(Vegetale)\,\mathrm{P}(Verde|Vegetale)} \\ &= \frac{0.45 \times 0.05}{0.45 \times 0.05 + 0.55 \times 0.70} = 0.05521472\end{aligned}$$

e conseguentemente

$$\mathrm{P}(Vegetale|Verde) = 1 - 0.05521472 = 0.9447853$$

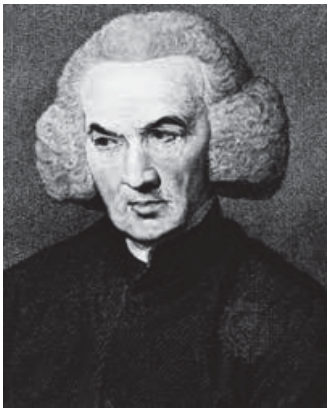
E' un esempio di aggiornamento bayesiano.

Thomas Bayes 1702–1761

Teologo e matematico.
Teorema di Bayes.
Inconsapevole capostipite.
Pastore non conformista
(congregazionista).

Richard Price 1723 –1791

Amico e divulgatore di Bayes.
Contributi alle → Scienze Attuariali.
Pastore unitariano, radicale.
Amico della rivoluzione americana.



Rincariamo la dose.

Per essere sicuri, aggiungiamo un'altra domanda. Quale?

Conosciamo le probabilitá?

Che possiamo conoscere o meno tutte le probabilità coinvolte è stato oggetto di lunghe discussioni negli ultimi due secoli tra statistici bayesiani e non bayesiani.

Un bayesiano moderno: Bruno de Finetti 1906-1985

Fondatore della rinascita neobayesiana.
Scambiabilità.
Teorema di rappresentazione.
Schema della scommessa.
Additività finita.
Soggettivista estemo.
'La probabilità non esiste''.
Fascista in gioventù, poi radicale.
Economia, Didattica della matematica.

Un non-bayesiano: Ronald A. Fisher 1890–1962

Padre della Statistica moderna.
Padre della → Genetica di Popolazione.
Moderna sintesi Darwin-Mendel.
t di Student.
Regressione, ANOVA, DOE.
Randomizzazione.
Analisi multivariata (e.g. discriminante).

Quantificazione delle probabilità?

Per esempio, abbiamo quantificato P($Animale$) = 0.45 ma potrebbe essere molto minore.

Un alieno cosa userebbe?

Si accettano suggerimenti....

Pierre-Simon Laplace 1749-1827

Matematico e astronomo.
Ottenne indipendentemente il teorema di Bayes.
A priori uniforme.
Thorie analytique des probabilits.
Regola di successione.
Minimi quadrati.
Determinismo.
Politicamente, un trasformista.

Quantificazione delle probabilitá condizionate

E per quanto riguarda queste?

$$\mathrm{P}(Verde|Animale) = 0.05$$
$$\mathrm{P}(Verde|Vegetale) = 0.70$$

Le conosciamo?

Come farebbe un alieno a quantificarle?

Se l'alieno avesse a disposizione una serie di osservazioni analoghe, lo potrebbe fare. Come?

Esempio di training set

	verde?	altro?	vero
1	1	1	vegetale
2	0	1	vegetale
3	0	1	vegetale
4	1	0	animale
5	0	0	animale
6	1	1	animale
7	1	1	vegetale
8	0	0	vegetale
9	1	1	animale
10	0	1	animale
11	1	1	animale
12	1	1	vegetale
13	0	1	animale
14	0	1	animale
15	1	1	animale
16	1	0	vegetale
17	1	0	vegetale
18	1	0	animale
19	1	1	vegetale
20	1	0	vegetale
21	1	1	animale
22	0	1	animale
23	0	1	animale
24	1	1	animale
25	1	1	vegetale
26	0	0	vegetale
27	1	0	vegetale
28	1	1	vegetale
29	0	0	animale
30	0	1	animale
31	1	1	animale
32	1	0	animale
33	0	0	animale
34	1	1	animale
35	1	0	vegetale
36	1	0	vegetale
37	1	1	animale
38	0	1	vegetale
39	0	1	vegetale
40	0	0	vegetale
41	0	0	vegetale
42	1	1	animale
43	1	1	animale
44	0	0	animale
45	1	0	animale
46	1	1	animale
47	0	0	vegetale
48	0	1	vegetale
49	1	0	animale
50	1	1	animale

Il training set come database statistico

Il *training set* è un semplice database statistico regolare (rettangolare):
– sulle righe ci sono le unità statistiche, o oggetti di interesse
– sulle colonne ci sono le variabili, o caratteristiche, rilevate sugli oggetti.

Ho generato quello nella pagina precedente con il seguente codice scritto nel linguaggio R:

```
  data <- cbind(
        as.data.frame(matrix(rbinom(100,1,.6),50,2)),
sample(c("animale","vegetale"),50,replace=T)
        )
colnames(data) <- c("verde?","altro?","vero")
```


Apprendimento supervisionato e non

Questo è un esempio di apprendimento supervisionato, cioè arricchito da un insieme di addestramento, o training set (a maggior ragione, anche conoscere le probabilitaà si pu considerare una forma di apprendimento supervisionato).

Un problema piú difficile si presenta quando disponiamo di un database parziale, senza vero. Allora dobbiamo essere noi a decidere come partizionarlo in due gruppi, un esempio di apprendimento non supervisionato.

	verde?	altro?
1	1	1
2	0	1
3	0	1
4	1	0
5	0	0
6	1	1
7	1	1
8	0	0
9	1	1
10	0	1

	verde?	altro?
11	0	0
12	1	0
13	1	1
14	0	0
15	0	1
16	1	1
17	1	0
18	0	0
19	1	1
20	1	0

La maledizione della dimensionalità

Due domande spesso non sono sufficienti.

Si presenta, crudele, un problema di dimensionalità.

Quante sono le possibili assegnazioni di questi $n = 20$ elementi a due popolazioni (partizioni di n)?

Quanti sono gli m possibili profili con d domande?

Quanti sono le possibili assegnazioni dei profili?

La maledizione della dimensionalità: esempio

Esercizio: scrivere tutti i possibili profili di risposta a tre domande binarie.

Calcolare tutte le possibili assegnazioni di questi profili.

Quale metodo usare per stimare le probabilità ?

Ammesso che le domande e quindi i profili siano un numero maneggevole, rimane il problema di come stimare le probabilità dei profili.

Una assunzione semplificatoria, che abbiamo fatto anche per il gioco della classificazione dei viventi, è la indipendenza locale delle risposte.

Una volta assunta questa indipendenza, si possono stimare le probabilità dei profili tramite il metodo della massima verosimiglianza, dovuto a R.A. Fisher. E' uno dei concetti più importanti della Statistica nonbayesiana moderna.

Metodi basati sulle distanze

Se invece la maledizione della dimensionalità è insuperabile, ci arrendiamo e ricorriamo a metodi più euristici fondati sulla distanza tra unità statistiche (o, specularmente, sulla loro vicinanza, o similarità).

Una volta scelta una distanza tra due unitá statistiche, cerchiamo di individuare due classi tali che:

1. entro le classi, le distanze tra le unità statistiche siano, in media, basse;

2.tra le classi, le distanze tra le unità statistiche siano, in media, alte.

Questa è una impostazione algoritmica più vicina al Machine Learning.

Buone distanze

La distanza tra due unità con lo stesso profilo dovrebbe essere nulla.

Qual è una buona distanza tra due unità con diversi profili?

Che distanza conoscete dalla scuola superiore?

Distanza euclidea

La distanza euclidea tra due vettori $x = (x_1, \ldots, x_n)$ e $y = (y_1, \ldots, y_n)$ è

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Cos'è in una dimensione $n = 1$?

In due dimensioni $n = 2$?

Per x e y binari?

Una tabella per x e y binari

Raccogliamo in una tabella di contingenza il numero di concordanze e discordanze delle diverse possibili paia

x \ y	0	1	
0	n_{00}	n_{01}	n_{1+} zeri in x
1	n_{10}	n_{11}	n_{2+} uni in x
	n_{+1} zeri in y	n_{+2} uni in y	n oggetti

Varie distanze per x e y binari

La distanza euclidea è la radice del numero di concordanze.

Il coefficiente di concordanza è

$$\frac{n_{01}+n_{10}}{n} = \frac{n_{00}+n_{11}}{n} = \text{proporzione di concordanze.}$$

Il coefficiente di Jaccard è invece

$$\frac{n_{01}+n_{10}}{n_{01}+n_{10}+n_{11}}.$$

Ce ne sono altri.

Esempio ecologico

Due zone ecologiche x e y, di uguale ampiezza (*quadrats*) vengono confrontate per vedere se hanno o meno la specie 1, la specie 2, ..., la specie n.

Le distanze viste misurano il grado di diversità biologica delle due zone.

Riassumendo

Ho cercato di far capire, tramite il *fil rouge* del gioco della classificazione dei viventi, somiglianze e differenze tra Statistica e Machine Learning.

La Statistica ha dei fondamenti teorici più profondi, e meglio formulati dal punto di vista matematico.

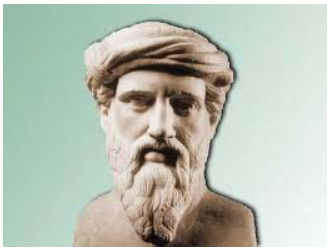
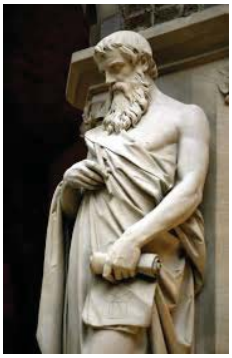
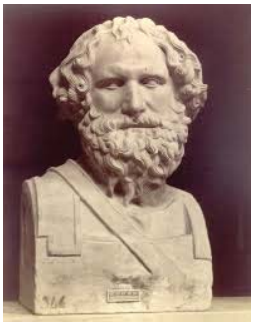
Il Machine Learning è più praticone ma può essere utile per quei casi in cui la Statistica fallisce (*Big Data*).

Entrambe costituiscono il cuore metolodogico della *Data Science*.

Ne ho approfittato per menzionare diversi concetti matematici e non: probabilità, probabilità condizionata, induzione, database, classificazione supervisionata, classificazione non supervisionata, distanza.

Immagini di coda, se rimane tempo

Il più grande matematico dell'antichità.

... se la giocano.

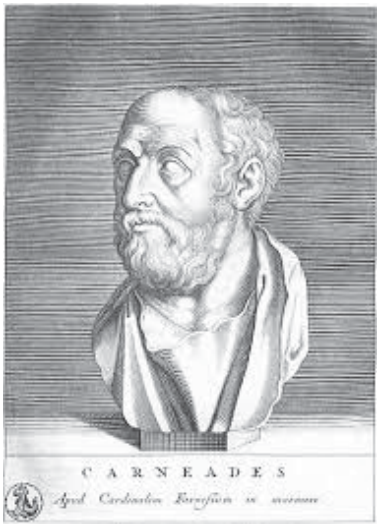
"Il più grande matematico della modernità."

E' abbastanza universalmente riconosciuto. Chi è?

Carl Friedrich Gauss 1777 –1855

Algebra e Analisi.
Astronomia.
Probabilità.
Minimi quadrati.
$\rightarrow$ Metrologia.

Carneade 214aC–129aC



Scettico greco.
Chi era costui?
Usò il termine "probabile".

Cardano 1501–1576

Liber de ludo aleae, postumo.
Regole elementari della probabilità.
Applicazioni nel gioco dei dadi.
Coefficiente binomiale.
→ Combinatorica.

Pierre de Fermat 1601–1665

Corrispondenza con Pascal.
Matematico di primaria importanza.
Quello dell'Ultimo Teorema.
→ Teoria dei numeri.

Blaise Pascal 1623-1662

Corrispondenza con Fermat.
Geometra di primaria importanza.
Triangolo di Pascal (o di Tartaglia).
Apologeta cristiano.

Gottfried Achenwall 1719–1772

Giurista e filosofo.
Coni il termine *Statistica*.