

Data Mining e Bioimmagini  
**LEZIONE INTRODUTTIVA**

Prof. Pierangelo Veltri – 25/09/2023- Autori: Gulizia, Cassalia - Revisionatori: Gulizia, Cassalia

---

- Email: [pierangelo.veltri@dimes.unical.it](mailto:pierangelo.veltri@dimes.unical.it)
- Studio: DIMES, Cubo 44Z – ricevimento concordabile tranne il mercoledì pomeriggio

**Programma:**

- Introduzione e richiami di statistica
- Esempi di uso e di analisi di dati biomedicali (in R e in Python)
- Esempi applicati in ambito medico (e bioimmagini)
- Introduzione Data Mining
- Richiami di definizione di Elementi di Statistica e motivazioni in ambito medico-clinico

**Testi di riferimento:**

1. *“Introduction to Data Mining”*; Pang-Ning Tan, Michael Steinbach Anuj Karpatne, Vipin Kumar, 2<sup>a</sup> edizione, Pearson
2. *“Fondamenti di Statistica applicata alla biomedica”*; Marc M. Triola, Mario F. Triola, Jason Roy, 2<sup>a</sup> edizione, Pearson
3. *“Machine Learning with PyTorch and Scikit-Learn”*; Sebastian Raschka, Yuxi (Hayden) Liu, Vahid Mirjalili, Packt

**DEFINIZIONE DEI CONCETTI**

**Data mining:** tecniche ed algoritmi di data processing per supportare l'estrazione della conoscenza attraverso delle regole codificate (si rappresentano i dati con la stessa modellistica con la quale vengono forniti).

**Machine Learning:** la macchina apprende e si costruisce una rappresentazione interna del dominio che non è necessariamente e direttamente intellegibile. In quest'ambito, a partire dai dati, si prova a definire un modello.

Alla base di tutto c'è il dato e il suo utilizzo, in quanto se non si ha a disposizione un “meccanismo” con cui gestire la grande quantità di dati a disposizione, il modello non può esser definito. Tale “meccanismo” è generalmente una regola matematica.

- Chat gtp funziona perché apprende utilizzando delle regole;
- La proposta in alternativa di Google funziona più o meno bene e anch'essa utilizza i dati che vengono forniti nel corso del tempo, sia in ambito di web che viene indicizzato e fornito, sia in termini di dati che si continuano a fornire con i meccanismi comportamentali.

**Deep Learning:** la macchina apprende, ma con una complessità del modello di apprendimento molto elevato. In ambito biologico, si ha l'esigenza di essere il più precisi possibile e, quindi, si utilizzano dei meccanismi di apprendimento più sofisticati: le **reti deep** o **reti profonde** o **reti neurali** (la parte delle tecniche che sono in grado di rispondere a determinate richieste utilizzando tecniche avanzate sono quelle che apprendono attraverso le reti profonde). Si definisce un piccolo modello in cui ci sono tante interazioni che imparano a partire da dati reali chiamati **training set**.

**DATA MINING E APPLICAZIONE IN AMBITO MEDICO**

**Prevenzione delle malattie:** il data mining può essere utilizzato per identificare i fattori di rischio

per le malattie e sviluppare programmi di prevenzione, ovvero utilizzare un'analisi del dato che può essere di tipo preventivo. Ad esempio, viene usato per identificare i pazienti che sono a rischio di sviluppare il cancro del colon e fornire loro screening e trattamenti preventivi. (Il progetto Prostat cancer, sviluppato dall'università della Florida, era basato sull'utilizzo di tecniche di analisi di dati per cercare di tirare fuori delle caratteristiche del dataset stesso).

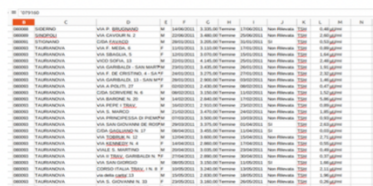
**Diagnosi delle malattie:** il data mining può essere utilizzato per sviluppare nuovi strumenti di diagnosi che possono aiutare i medici a diagnosticare le malattie in modo più accurato e tempestivo. Ad esempio, il data mining può essere utilizzato per sviluppare modelli di intelligenza artificiale che possono analizzare le immagini di esami medici e identificare segni di malattia. Ciò può essere fatto sia utilizzando dei dati alfanumerici, ad esempio analisi dei dati di tipo clinico, delle analisi del sangue, delle analisi biologiche; sia tirando fuori una combinazione delle due cose ovvero una combinazione di analisi biologiche o combinazioni di immagini che possono essere ecografiche o più dettagliate in termini diagnostici.

**Trattamento delle malattie:** il data mining può essere utilizzato per sviluppare nuovi trattamenti per le malattie. Ad esempio, il data mining può essere utilizzato per identificare i farmaci più efficaci per il trattamento del cancro e per sviluppare nuovi farmaci che sono più efficaci e meno tossici dei farmaci attualmente in commercio.

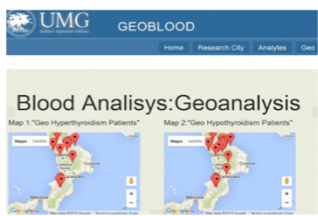
**Personalizzazione delle cure:** il data mining può essere utilizzato per personalizzare le cure mediche per ogni paziente. Ad esempio, il data mining può essere utilizzato per identificare farmaci più efficaci per ogni paziente e per sviluppare programmi di trattamento personalizzati che sono più efficaci e meno costosi delle cure standard.

#### Esempio applicativo: analisi di TSH neonatale

Identificazione di percentili e definizioni di valori soglia.



id	name	age	sex	city	state	country	date	time	value
1	John Doe	1	M	New York	NY	USA	2010-01-01	10:00	1.2
2	Jane Smith	1	F	California	CA	USA	2010-01-01	11:00	1.5
3	Bob Johnson	1	M	Texas	TX	USA	2010-01-01	12:00	1.8
4	Alice Brown	1	F	Florida	FL	USA	2010-01-01	13:00	2.1
5	Charlie Davis	1	M	Illinois	IL	USA	2010-01-01	14:00	2.4



Per **screening neonatale** si intende un insieme di esami volti a identificare patologie genetiche, endocrinologiche, metaboliche ed ematologiche per le quali esista una terapia, durante i primi giorni di vita. In questo caso viene preso in considerazione l'analisi del

TSH. Se il valore analizzato è maggiore di una soglia, il soggetto presenta ipotiroidismo. La **soglia** viene calcolata con il **percentile**.

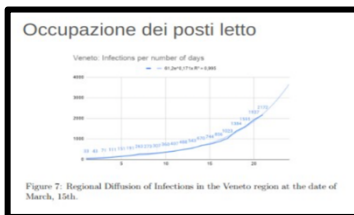
Con l'analisi dei dati si possono inoltre fare dei **test di arricchimento** (enrichment test), ovvero arricchire i dati con nuove informazioni. Ad esempio, i dati possono essere arricchiti con le informazioni geografiche.

Per quanto riguarda le analisi di mercato e le predizioni queste possono prevedere l'inizio e la numerosità di influenza autunnale, basarsi sull'analisi di andamento epidemiologico, analisi del traffico usando sistemi maps, analisi del mercato ed esigenze (ad esempio Amazon conta 1.4 MLD di items che servono per capire la profilazione e l'andamento del mercato).

È il **large scraping**, ovvero si utilizzano delle tecniche di analisi del dato per tirar fuori delle informazioni.

Una direction nell'ambito medico-clinico molto rilevante, iniziata da Obama con *l'Obamacare*, afferma che poiché i centri che fanno health (prestazioni sanitarie) acquisiscono i dati e siccome i centri che fanno cancer data (ricerca) acquisiscono anch'essi dei dati, questi potrebbero essere uniti tutti assieme per avere delle informazioni statistiche su un quantitativo di dati più ampio. Ciò diede inizio agli Open Data ovvero la possibilità di avere dei dati sempre a disposizione per essere utilizzati da chiunque.

## Predizione di necessità ICUs nelle regioni



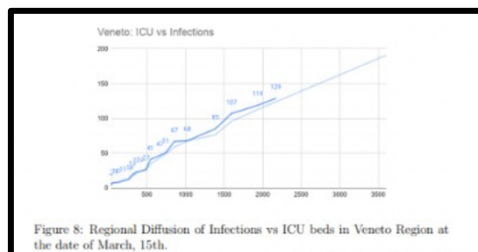
La velocità di diffusione del virus e la velocità di diffusione della malattia erano molto più rapide della possibilità di aumentare i posti letto.

La curva con la quale venivano rappresentati i dati di diffusione del virus era un tipo di curva esponenziale.

Se si prendessero i punti e i dati relativi alla diffusione del virus e all'occupazione del posto, ovvero ICU vs infection, si otterrebbe un grafico come riportato a lato.

Nei grafici è possibile tracciare una retta di correlazione (una retta ideale che cerca di associare le informazioni): considerando i punti e le coordinate, si identificano in un grafico e in seguito viene tracciata la retta. Più i punti si avvicinano alla retta, più il modello considerato è valido, idealmente si riesce a rappresentare la correlazione.

Se si ha la funzione esponenziale, si conosce il punto di partenza e di conseguenza si è in grado di sapere il valore delle infezioni in una settimana, oppure si può predire la necessità di posti in terapia intensiva in dieci giorni. Anche la ricollocazione dei posti letto può essere studiata e rappresentata con un grafico di questo tipo.



Gli esercizi e le attività di analisi possono essere fatti sia utilizzando **R** (un linguaggio di programmazione specifico per la statistica e la grafica computazionali) sia **RStudio** (è un ambiente di sviluppo integrato per R). R è a linea di comando ed RStudio che da una parte mi dà la directory di lavoro e una parte è dedicata alla copiatura del codice.

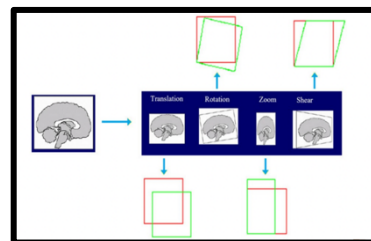
## Esempio di test statistici

Sono prese in esempio le PET cerebrali e si effettua un processo di **preprocessing**, ovvero quando si confrontano dei dati spesso devono essere ripuliti e resi confrontabili tra di loro.

Nell'ambito delle bioimmagini si mappano le informazioni sui template, si riposizionano le immagini rispetto ad un formato comune.

Tutto ciò si fa per preparare i dati al processamento. Si è poi cercato di organizzare un meccanismo di classificazione tra sani e malati prendendo in considerazione delle regioni di interesse.

Successivamente si cerca di capire perché prendendo due classi, una sana e una malata, siano tra loro statisticamente significative: se si considerano come elementi diversi la probabilità che si faccia un errore rimane bassa. Ciò si misura con un concetto di *verifica dell'ipotesi nulla*, un test statistico.



La statistica è una metodologia per disegnare esperimenti, ricavare i dati ed estrarre (graficamente) risultati che rappresentano un fenomeno.

Termini:

- **Popolazione**: tutti gli elementi di uno studio
- **Censimento**: raccolta dati che riguardano ogni membro della popolazione
- **Dati**: osservazioni che si fanno e si conservano per successivo studio (misure di un sottoinsieme di popolazione) presi con metodi specifici

## Significatività statistica

Misurare se un dato e/o un risultato è significativo ed utilizzabile. Uno studio statistico ha una significatività statistica se è molto improbabile che il risultato ottenuto sia stato ottenuto per caso (per esempio minore del 5%)

Esempio → Registrare 98 neonati su 100 di sesso F è statisticamente significativo perché tale evento è molto improbabile secondo leggi casuali. Invece, registrare 52 su 100 non è statisticamente significativo perché molto probabile per le leggi del caso.

La modalità con la quale si acquisiscono i dati è la parte più difficile; ad esempio, l'associazione tra QI e misure del cervello è un processo che si fa in modo tabellare. Lo studio si effettua per capire se tra le due informazioni ci può essere una correlazione significativa, indicando che il campionamento e lo studio del campione risultano essere rilevanti.

La scelta degli individui del campione può essere fatta tramite:

- Selezione in base alle storie medico/cliniche
- Selezione casuale
- Selezione di campione retribuito (volontari pagati, ...)

Una volta composto il campione è necessario verificarne la significatività statistica (guardare esempio dei neonati sopra), così come è necessario misurare la significatività del risultato ottenuto.

La misura della significatività statistica dei risultati prevede le seguenti fasi:

- Ipotesi su popolazione sana e malata (*Si ipotizza la presenza di due classi diversificate*)
- Identificare ipotesi nulla  $H_0$  (*Ossia che la popolazione sia o solo sana o solo malata*)
- Calcolare la probabilità di errore
- Se la probabilità di sbagliare è più piccola di una soglia prestabilita (in ambito scientifico generalmente è  $<0.05$ ):
  - Si rifiuta l'ipotesi nulla di sovrapposizione tra le due classi
  - I risultati significativi dei campioni sani e malati si distinguono
  - Il test è robusto e l'identificazione di pattern sono utili

Esempio di applicazione innovativa dell'algoritmo di Goertzel sull'analisi vocale.

I gruppi presi in considerazione sono campioni di voci sane e campioni di voci patologiche. L'informazione è misurata in termini di significatività statistica, se la distinzione tra pazienti sani e malati è affidabile utilizzando due algoritmi:

**Goertzel e la trasformata di Fourier (FFT).** Qualora la significatività statistica tra sani e malati sia molto bassa, l'ipotesi di distinzione tra i gruppi utilizzando lo screening vocale risulta molto efficace.

Un'analisi di questo tipo è un ottimo risultato poiché si sostituisce a dei metodi invasivi.

Subjects	Num. Subj.	Comparison	p-value GA	p-value FFT	Test type
All	56	Healthy vs Pathological	<b>0.01185 *</b>	0.7046	GA, FFT: Wilcoxon
All	52	Healthy vs Multiple Sclerosis	<b>3.182e-10 *</b>	<b>0.03352 *</b>	GA, FFT: Wilcoxon
All	56	Healthy vs Hyperkinetic Dysphonia	<b>0.01256 *</b>	0.3451	GA: Wilcoxon, FFT: t-test
All	36	Healthy vs Hypokinetic Dysphonia	0.2299	0.685	GA: Wilcoxon, FFT: t-test
All	38	Healthy vs Reflux Laryngitis	0.1014	0.6711	GA: Wilcoxon, FFT: t-test
Female	36	Healthy vs Pathological	<b>0.01672 *</b>	0.7887	GA: Wilcoxon, FFT: t-test
Female	36	Healthy vs Multiple Sclerosis	<b>0.00007827 *</b>	<b>0.0002677 *</b>	GA, FFT: Wilcoxon
Female	36	Healthy vs Hyperkinetic Dysphonia	<b>0.009745 *</b>	0.28	GA: Wilcoxon, FFT: t-test
Female	32	Healthy vs Hypokinetic Dysphonia	0.1674	0.7537	GA: Wilcoxon, FFT: t-test
Female	19	Healthy vs Reflux Laryngitis	0.1038	0.2252	GA: Wilcoxon, FFT: t-test
Male	20	Healthy vs Pathological	<b>0.002781 *</b>	0.1577	GA: Wilcoxon, FFT: t-test
Male	16	Healthy vs Multiple Sclerosis	<b>0.00002708 *</b>	0.08889	GA: Wilcoxon, FFT: t-test
Male	20	Healthy vs Hyperkinetic Dysphonia	0.4429	<b>0.001513 *</b>	GA: Wilcoxon, FFT: t-test
Male	4	Healthy vs Hypokinetic Dysphonia	0.6857	0.5953	GA: Wilcoxon, FFT: t-test
Male	19	Healthy vs Reflux Laryngitis	0.1017	0.5876	GA: Wilcoxon, FFT: t-test

**Table 5.** T-test and Wilcoxon test results for GA and FFT algorithm applications. Goertzel succeeds (p-values lines in bold) in most of the performed tests w.r.t. FFT at significance level  $\alpha = 0.05$ . Note that each row reports two test results, one for GA and one for FFT (see columns *p-value GA* and *p-value FFT*). In two comparison cases (p-value in gray), both GA and FFT were unable to reject the null hypothesis, i.e. the test failed, hence the features extracted by both algorithms were not significantly different between the considered groups.

Nella raccolta dei dati alcuni possono mancare ed avere, quindi, dataset incompleti. Si può avere una mancanza casuale o non casuale; in quest'ultimo caso si considerano i dati come non affidabili per cui il campionamento non è corretto. La completezza di un dato, dunque, è fondamentale perché, ad esempio, nel campo medico legale, molte operazioni non sono ripetibili e quindi bisogna essere il più precisi possibili. Gli esperimenti sono anch'essi rilevanti e possono essere:

- A posteriori: ad esempio, lo studio degli effetti del vaccino per il covid;
- Esperimenti: si somministrano i test e si misurano gli effetti (soggetti che partecipano);
- Studio osservazionale: si misurano caratteristiche senza intervenire sui soggetti.