UNIVERSITÀ
DI SIENA
1240

Predizione degli Effetti Avversi dei Farmaci Tramite Reti Neurali Artificiali

Carlo Merola

- Dipartimento di Ingegneria dell'Informazione e Scienze Matematiche

SIENA, A.A. 2021-2022

Dedico questo studio a Wally - Luigi -, che si trova in condizioni di salute gravi. E alla sua famiglia, Emma, Edoardo ed Eleonora.
Hanno sempre mostrato tanta forza, propria di poche persone.

Obiettivo

- Sviluppare un sistema per prevedere gli effetti collaterali dei farmaci

Dati di partenza

- Caratteristiche strutturali delle molecole

Utilità

- Problema particolarmente importante nel drug design

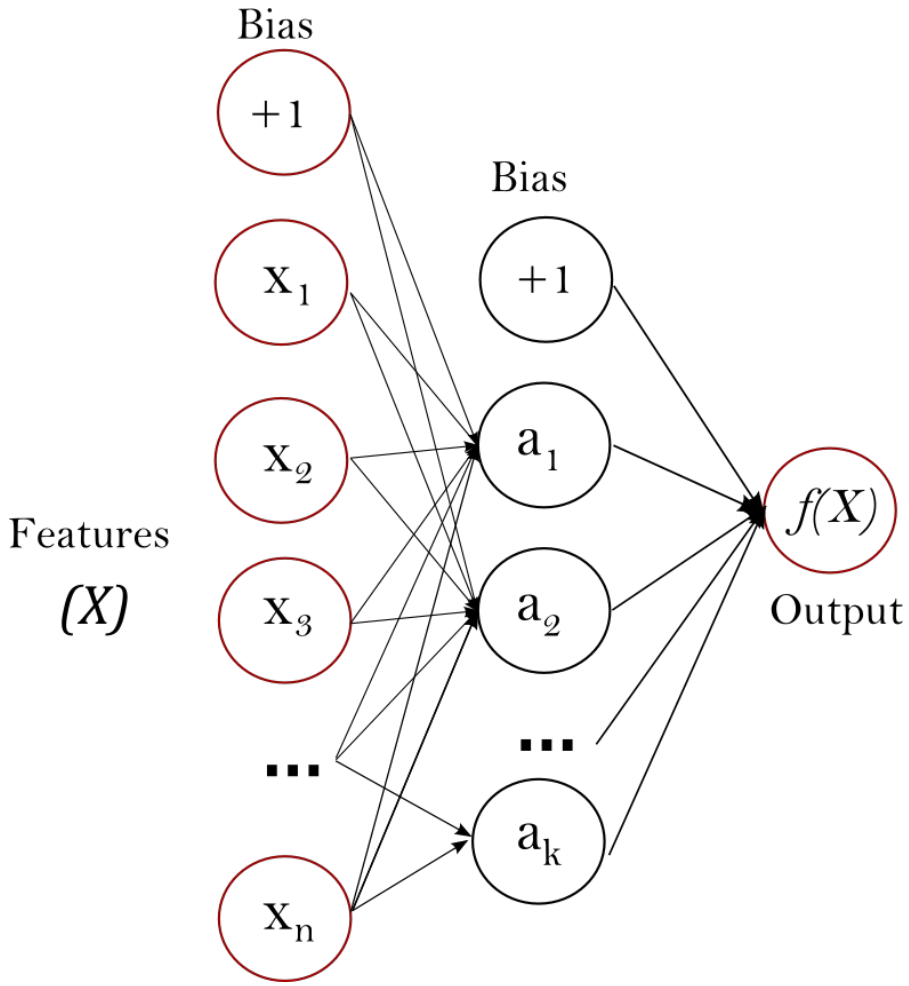
Le molecole possono essere esaminate con metodi computazionali prima di essere sottoposte a studi clinici.

Machine Learning e Reti Neurali Artificiali

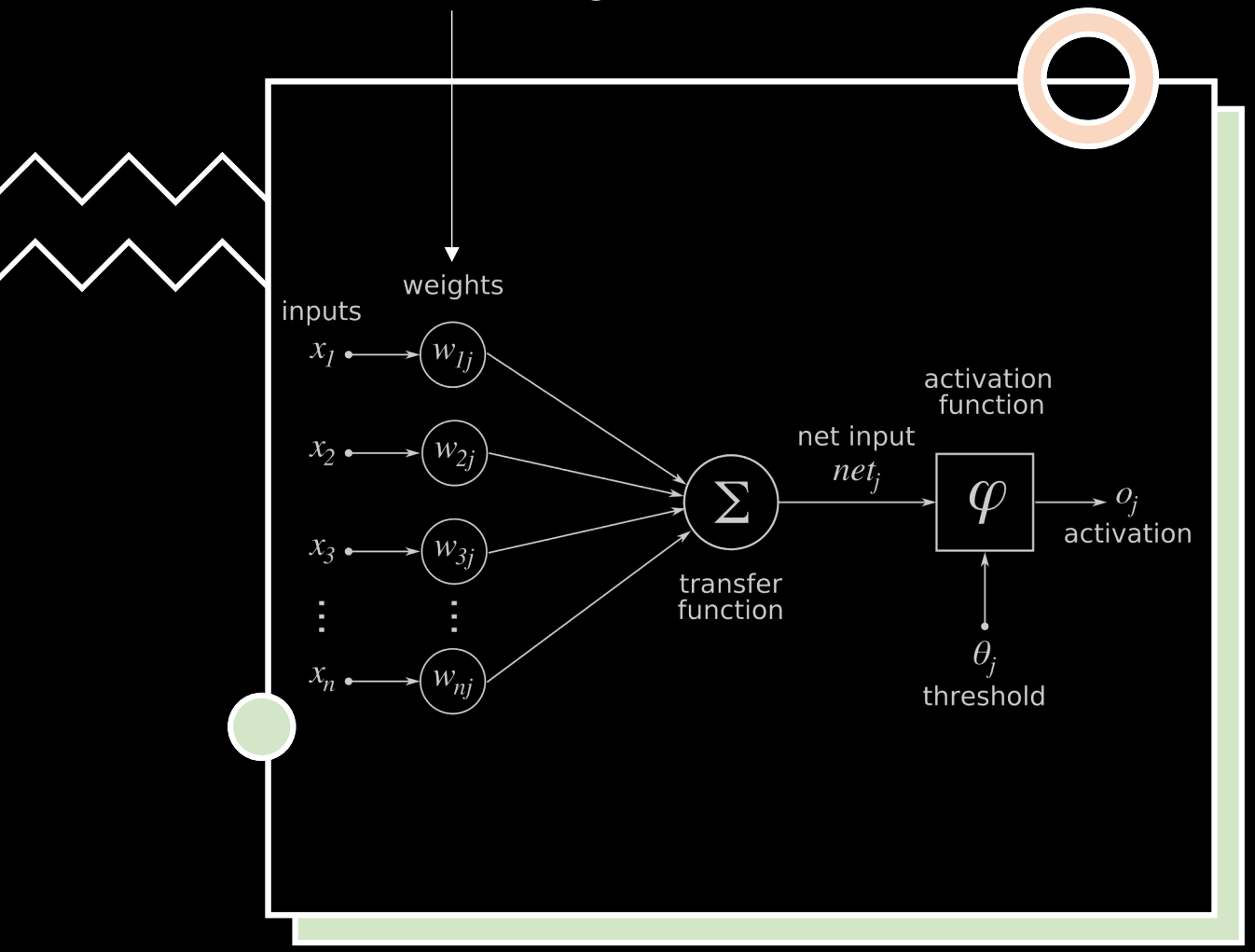
- La capacità di generalizzazione, e dunque **predizione**, delle Reti Neurali Artificiali può aiutare in questo processo.

- Attraverso l'apprendimento supervisionato il modello impara una **funzione** sui dati.

- Differisce dall'algoritmo classico, dove la logica viene descritta in forma esplicita. Quest'ultimo può essere inadatto in contesti di elevata **variabilità** dei dati.



Il neurone artificiale

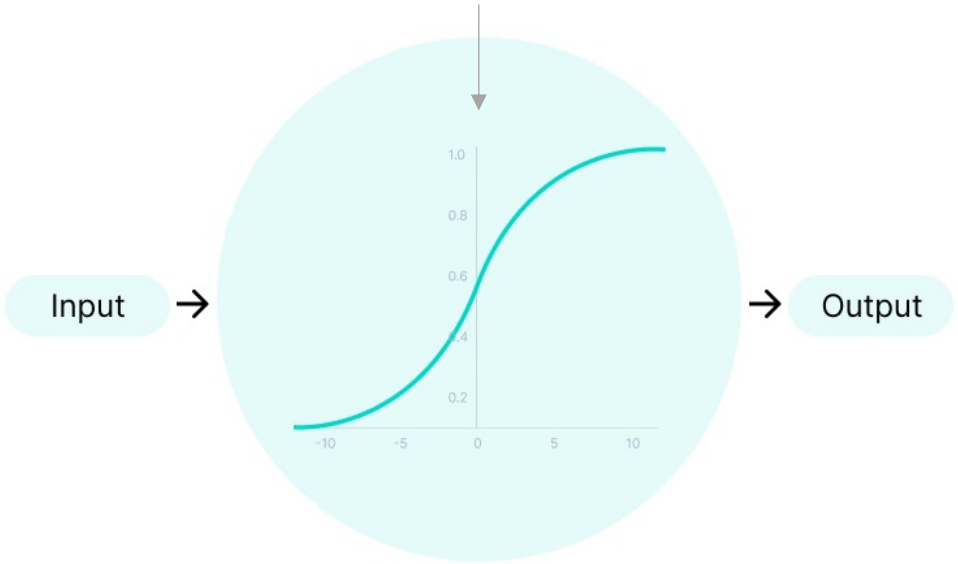


- Metafora neurobiologica a livello della singola unità elementare.
- Ogni connessione può trasmettere un segnale ad altri neuroni densamente interconnessi.
- Elabora l’output attraverso una funzione non lineare della somma **pesata** dei suoi input.

Classificazione multi-label

- Ad ogni campione vengono assegnate *m* **etichette** da *n* possibili classi, dove *m* è compreso tra *0* ed *n* effetti collaterali univoci.
- Le proprietà del campione sono **indipendenti** l'una dall'altra.
- Il vettore binario rappresenta gli **effetti collaterali** propri del farmaco.

Funzione di attivazione logistica.
Accetta qualsiasi valore reale come input e genera valori nell'intervallo da 0 a 1.

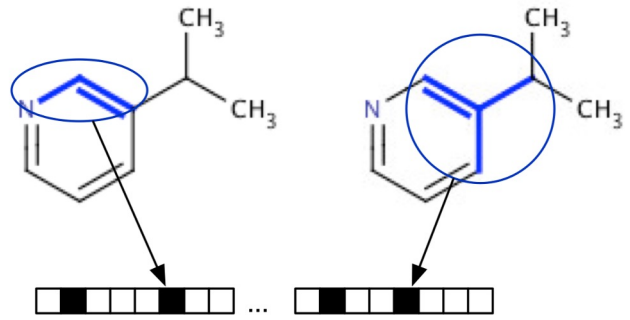


$$\varphi(z) = \frac{1}{1 + e^{-z}}$$

Dati di apprendimento

- «**Features**» ottenute dal database PubChem, liberamente accessibile. Messo a disposizione dal National Institutes of Health (NIH).

- Dati «**target**» prelevati da SIDER, contenente le reazioni avverse dei medicinali attualmente in commercio.



Funzione di hashing di Rdkit. Ogni bit corrisponde ad un frammento della molecola.

Dalle stringhe SMILES (Simplified Molecular Input Line Entry System) delle molecole sono stati ricavati i «**fingerprint**», vettori binari altamente discriminanti, che descrivono le caratteristiche strutturali della molecola.

Il problema delle classi sbilanciate

- Un problema di classificazione **sbilanciato** si ha quando la distribuzione degli esempi tra le classi è disomogenea.
- L’assenza di un effetto collaterale è molto più frequente della sua presenza.
- Scarse prestazioni predittive per la classe meno rappresentata, talvolta la più importante.

Soluzione adottata: vettore di pesi proporzionale al numero di etichette di un farmaco.

Metodo di valutazione

- Sui dataset sbilanciati l'**accuratezza** può non rappresentare la modalità migliore per la misura delle prestazioni.

Potrebbe capitare di avere prestazione elevate, anche nel caso di incapacità di predizione verso la classe minoritaria.

- Un insight più preciso si ha con la **matrice di confusione**.

	Valore **predetto negativo**	Valore **predetto positivo**
Valore **reale negativo**	TN	FP
Valore **reale positivo**	FN	TP

- **Precisione** = $\frac{TP}{TP+FP}$

Misura la capacità di un classificatore binario di identificare solo gli esempi appartenenti ad una data classe (identificata come positiva).

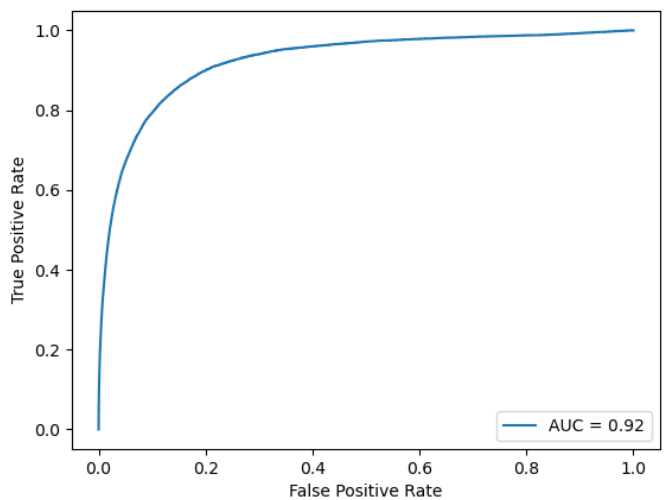
- **Richiamo** = $\frac{TP}{TP+FN}$

Quando un modello è sensibile per una classe, la prevede ogni volta che si verifica.

- **ROC** (Receiver Operating Characteristic)

Grafico che visualizza il compromesso tra tasso di veri positivi e tasso di falsi positivi, calcolato e tracciato per **ogni soglia**.

- **AUC** (Area Under the Curve)

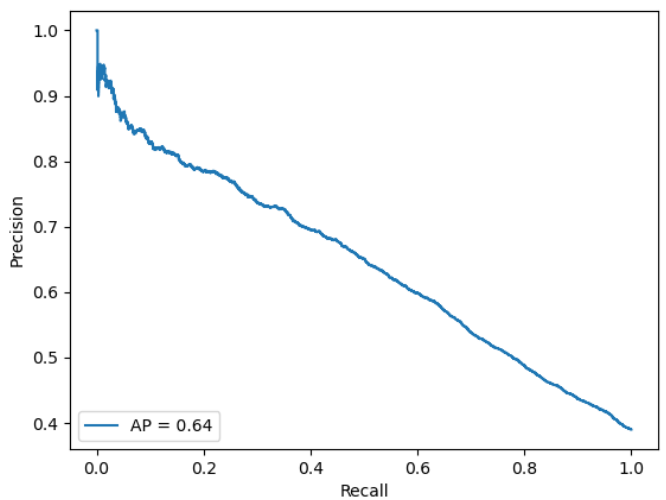


- **Curva di precisione- richiamo**

Combina precisione e richiamo in un'unica misura, calcolata per **ogni soglia**.

- **Average Precision** = $\sum_n (R_n - R_{n-1})\, P_n$

È più sensibile ai miglioramenti predittivi che riguardano l'identificazione della classe come positiva.



Prestazioni del modello predittivo

Peso dei Campioni	Min Occ.	Max Occ.	Unità Nascoste	Epoche	AUC	AP
no	nessun filtro	nessun filtro	1000	300	0.9125	0.3593
sì	nessun filtro	nessun filtro	700	500	0.9181	0.3602
no	10	nessun filtro	300	500	0.8395	0.3765
sì	10	nessun filtro	200	600	0.8450	0.3853
no	370	nessun filtro	100	300	0.6849	0.5910
sì	370	nessun filtro	100	300	0.6865	0.5931

- Numero di effetti collaterali univoci senza il filtro sulle occorrenze: **4251**
  Taglio percentuale: **0%**

- Numero di effetti collaterali univoci riscontrati su un minimo di 10 farmaci: **1389**
  Taglio percentuale: **67,32%**

- Numero di effetti collaterali univoci applicando un filtro di minimo 370 occorrenze: **100**
  Taglio percentuale: **97,65%**

Adattamento della soglia di discriminazione

- Solitamente le predizioni vengono calcolate discriminando se la classe appartiene o meno al campione in base alla **soglia predefinita** di 0.5
- Abbassando la soglia **aumenteremo** la probabilità che il modello applichi l'etichetta dell'effetto collaterale al farmaco.
- Aumento del punteggio di **richiamo** e conseguente diminuzione nel punteggio di precisione.
- Può risultare un **buon compromesso** nel caso l'obiettivo risulti predire il quantitativo maggiore di effetti collaterali del farmaco, prima di un successivo studio clinico.

↓

Per l'ultima configurazione in tabella:
soglia = 0.5 → precisione = 0.5839, richiamo = 0.4940
soglia = 0.15 → precisione = 0.4749, richiamo = 0.7874

Grazie per l’attenzione

Candidato: Carlo Merola

Relatore: Monica Bianchini

Co-relatore: Pietro Bongini