

Introduzione a tecniche multivariate

Laboratorio di Metodi Computazionali e Statistici (2023/2024)

Roberta Cardinale e Fabrizio Parodi

Dipartimento di Fisica - Università di Genova

20 Dicembre 2023

Classificazione eventi

- Immaginiamo di avere un campione di dati con due tipi di eventi
- Supponiamo anche che i due tipi di eventi abbiano delle caratteristiche simili per cui non siamo in grado immediatamente di identificare uno o l'altro tipo
- Ognuno è caratterizzato da n osservabili: $\mathbf{x} = (x_1, \dots, x_n)$
- Classificazione di eventi dal punto di vista dei test statistici
- In un campione di dati in cui sono presenti entrambi i tipi di eventi e non so a priori quale è uno e quale è l'altro, testo l'ipotesi H_0 che l'evento sia di tipo 0
- Se rigetto l'ipotesi H_0 , seleziono candidati di tipo 1 (segnaletico) (candidati perchè sappiamo che ci saranno casi in cui accetto/rigetto H_0 non correttamente)
- Altrimenti, se non rigetto H_0 , seleziono candidati di tipo 0 (fondo)

Efficienza/Reiezione

prob di rigettare H_0 quando è vero
prob accettare eventi di fondo che non lo sono

- Analogalemente a come abbiamo definito gli errori di tipo I (α) e di tipo II (β) e significanza del test (α) e potere del test ($1 - \beta$)
- Probabilità di rigettare l'ipotesi H_0 (fondo) per eventi di tipo H_0 (efficienza di fondo): errore di tipo I

$$\epsilon_{bkg} = \int_{t_{cut}}^{\infty} p(t|H_0) dt = \alpha$$

- Per definizione è la significanza del test
- Probabilità di accettare l'ipotesi di fondo per evento di segnale: errore di tipo II

$$\beta = \int_{-\infty}^{t_{cut}} p(t|H_1) dt$$

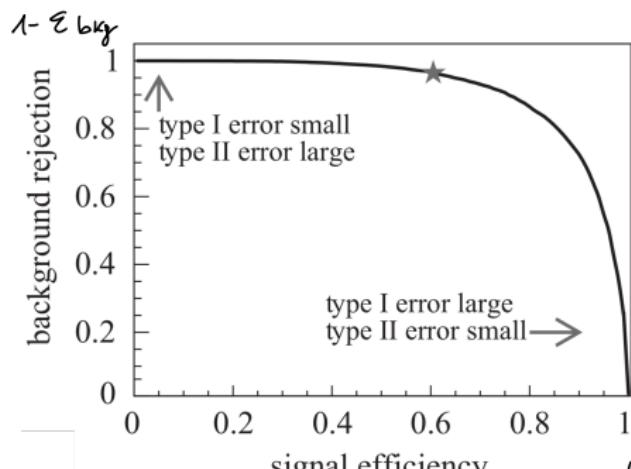
- Probabilità di accettare un evento di segnale come segnale (efficienza di segnale)

$$\epsilon_{sig} = \int_{t_{cut}}^{\infty} p(t|H_1) dt = 1 - \beta$$

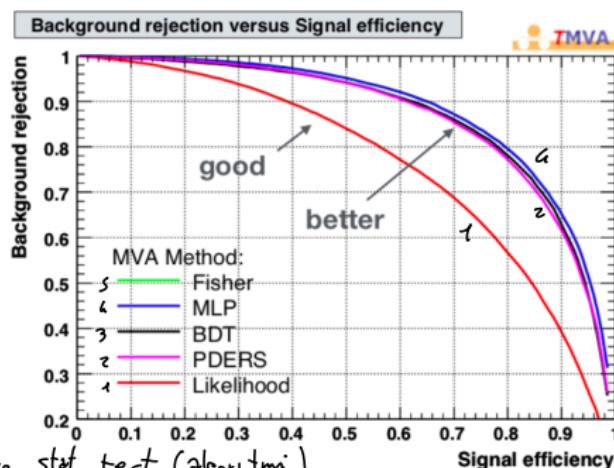
($1 - \beta$ è la potenza del test)

Curva ROC

- La curva ROC (Receiver Operation Characteristics) permette di quantificare la purezza del campione selezionato ($1 - \epsilon_{bkg}$) in funzione dell'efficienza del segnale (ϵ_{sig}).
- Spesso si stima il potere discriminante del metodo tramite il parametro AUC (Area Under the Curve). Maggiore è meglio (il massimo è 1).



Tanto più statistica di segnali distribuiti diverse stat test (algoritmi) negli



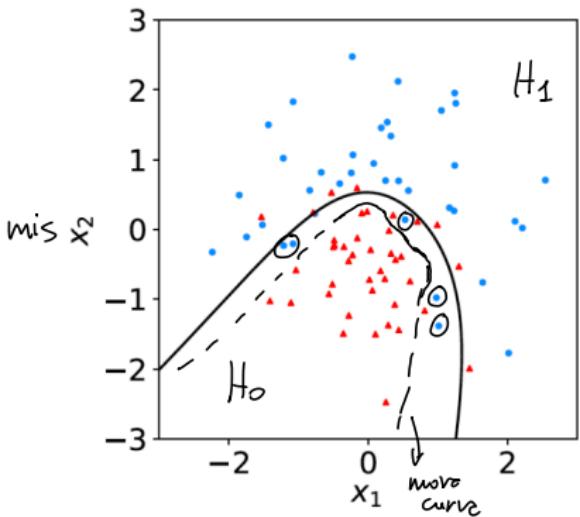
Classificazione di eventi

Cerco curva in 2D che separa eventi diversi \Rightarrow è def di regole critica \mathcal{W}
 \Rightarrow def la significanza curva riferito ad H_0



- Obiettivo: determinare un confine in modo da determinare il tipo di evento in base a dove si trova rispetto al confine
- Supponiamo di avere $\mathbf{x} = (x_1, x_2)$
- Rosso: Eventi di tipo 0
- Blu: Eventi di tipo 1
- Estensione a n dimensioni: ipersuperficie

efficienza più alta ma contaminazione peggiore se, con meno dati (più Δ e meno σ) ma tralascio dei Δ mis curva nuova grande



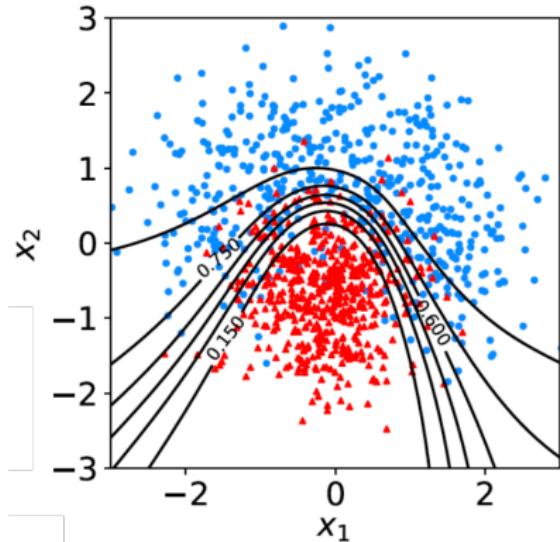
Regione Critica

Quanto mi posso permettere errori di tipo I
Varie significanze

- Una superficie in uno spazio n-dimensionale può essere descritta da una funzione scalare $t(x_1, \dots, x_n)$ con la condizione sia uguale ad una costante t_c

$$t(x_1, \dots, x_n) = t_c$$

- Valori diversi di t_c equivalgono in una famiglia di superfici con diversa significanza α (ottengo una selezione più o meno pura di eventi di tipo 1, rossi)
- Il problema si riduce nel cercare la miglior funzione decisionale (o statistica di test) $t(\mathbf{x})$ (a seconda della funzione scelta e dei parametri scelti ottengo curve diverse)

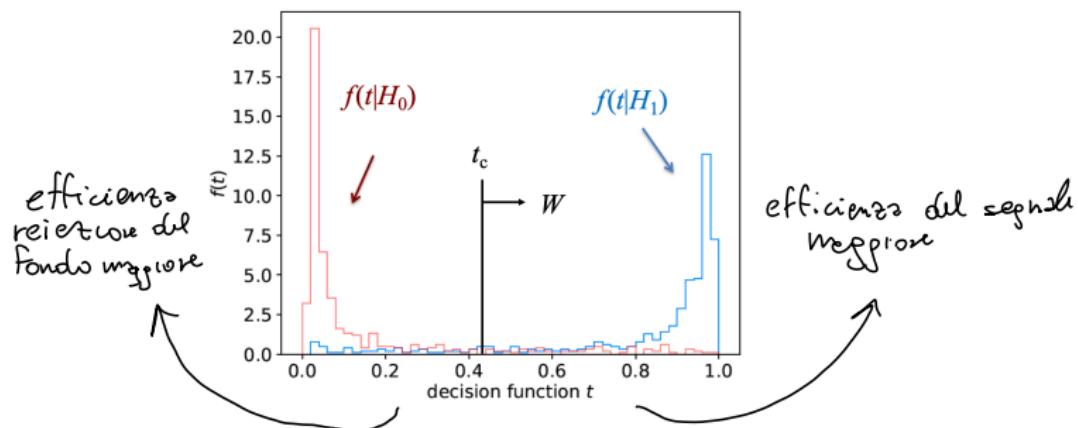


Statistica di test

- Si mappa il vettore di osservabili in una singola variabile “statistica del test”: $\mathbb{R}^n \rightarrow \mathbb{R} : t(\mathbf{x})$

La condizione $t > t_c$ (taglio, in gergo) per selezionare eventi di segnale corrisponde ad una complicata iper-superficie nello spazio delle osservabili (in generale superiore ad una combinazione di tagli nelle singole variabili)

$t > t_c$ individua la regione critica



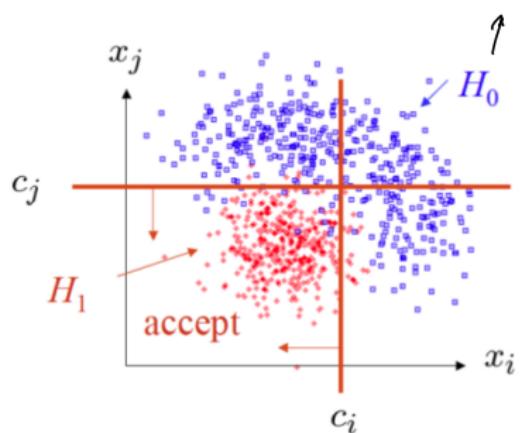
Classificazione eventi (II)

Quale è il miglior confine che definisce la regione critica (o anche quale è la statistica di test ottimale)? *won mi interess*

- Il modo più semplice, intuitivo (ma in generale non ottimale) è usare tagli indipendenti sulle singole variabili (confine rettangolare):

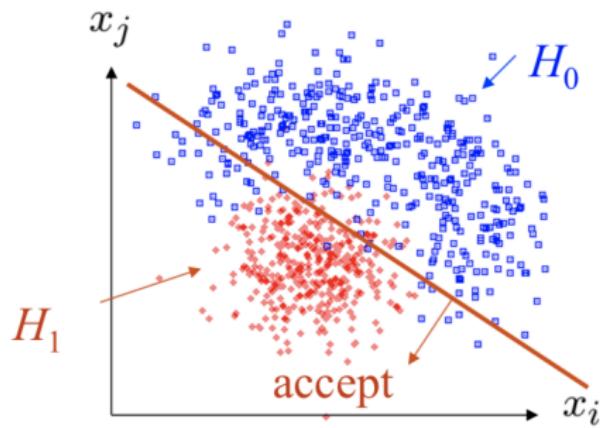
$$x_i < c_i$$

$$x_j < c_j$$

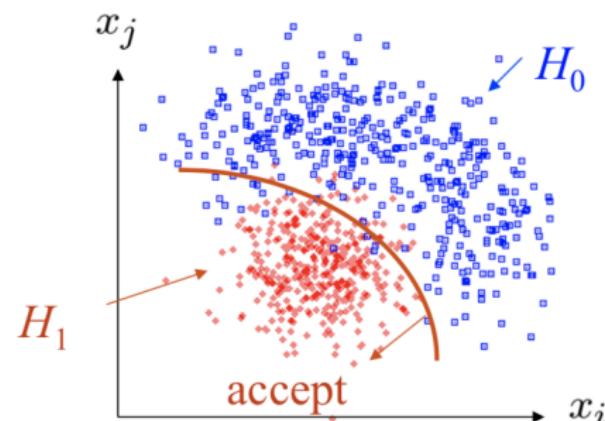


Classificazione eventi (III)

Algoritmi multivariabili creano ipersuperficie per separare eventi



lineare



non lineare

C'è un modo ottimale per decidere la linea di separazione tra i due tipi di eventi?

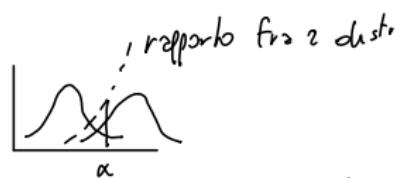
Neyman-Pearson lemma

C'è un modo ottimale per scegliere la regione critica ?

Neyman-Pearson Lemma

Per ottenere la più elevata reiezione di eventi di tipo 0 (fondo) data un'efficienza di segnale (massima potenza per un dato livello di significanza) si deve scegliere la regione critica tale che:

$$\Lambda(x) = \frac{f(x|H_1)}{f(x|H_0)} \geq c_\alpha$$



dove c_α è una costante scelta per una certa efficienza di segnale

Equivalentemente, il test statistico ottimale è

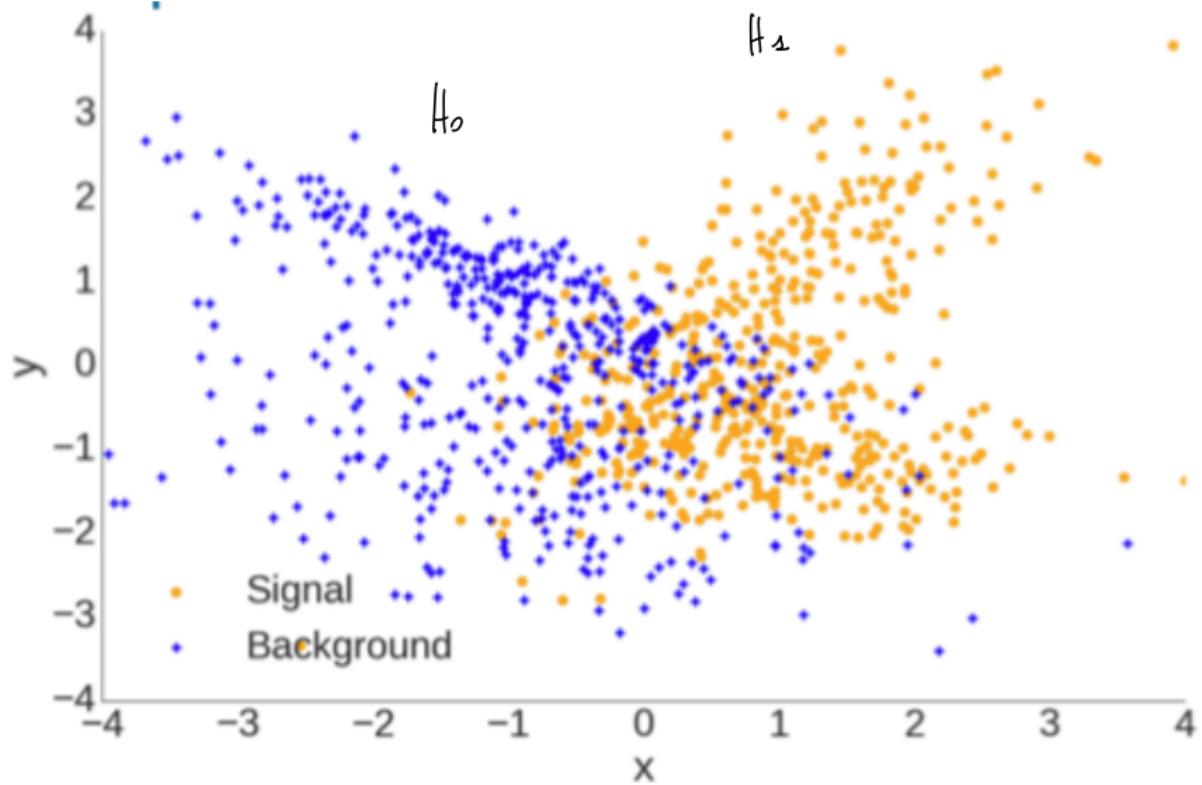
$$t(x) = \frac{f(x|H_1)}{f(x|H_0)}$$

taglio a dx → più efficacia
ma perdi effenz
giulli

taglio A sx → più efficacia
selez giulli
ma più fondo
(vedi avanti)

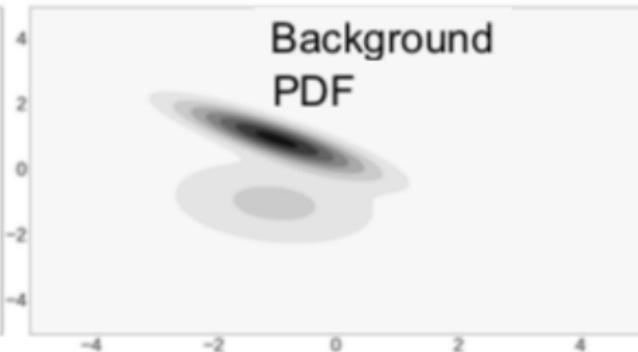
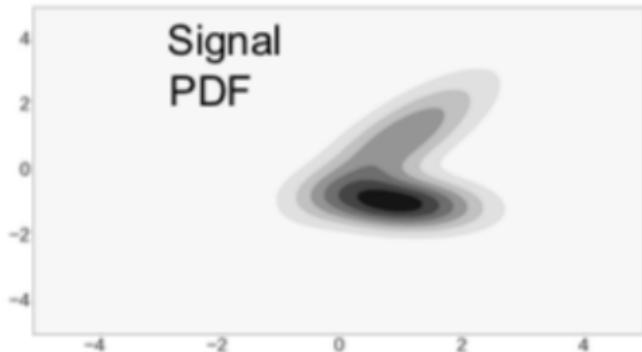
per cui c_α corrisponde al valore di taglio Una qualunque trasformazione monotona del rapporto è anche ok

Esempio

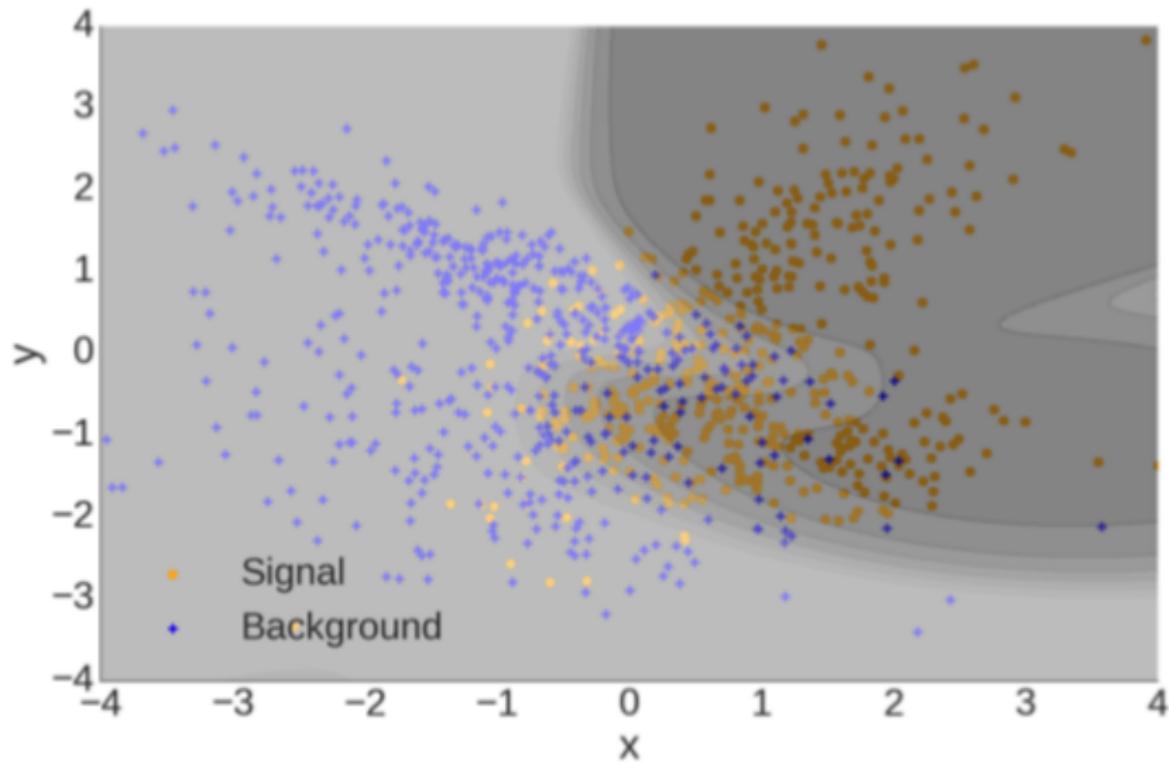


Esempio

densità eventi



Esempio

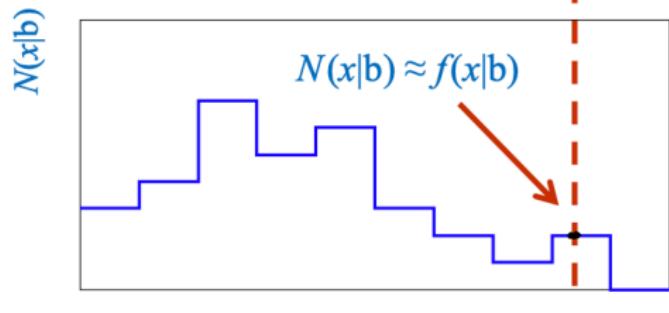
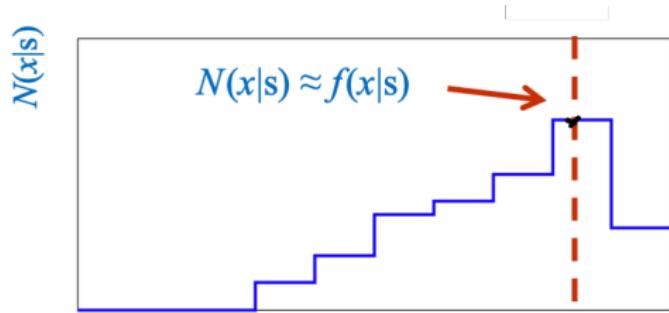


NP Lemma?

- Normalmente non abbiamo a disposizione le pdfs esplicite per $f(x|H_0)$ e $f(x|H_1)$
- Non possiamo utilizzare il lemma di NP perchè non possiamo calcolare $t(\mathbf{x}) = \frac{f(\mathbf{x}|S)}{f(\mathbf{x}|S')}$
- Spesso sono disponibili modelli che ci permettono di generare eventi \mathbf{x} secondo i due tipi di eventi: eventi di tipo 0 o eventi di tipo 1
- Questo ci fornisce un campione di dati con eventi noti (di cui conosciamo il tipo di eventi)

NP Lemma?

- Possiamo utilizzare questi campioni Monte Carlo di segnale e fondo per riempire istogrammi per eventi di segnale e di fondo
- Possiamo usare i valori degli istogrammi (normalizzati) per approssimare il rapporto delle likelihood:
$$t(x) \sim \frac{N(x|S)}{N(x|B)}$$
- Nel caso unidimensionale funziona bene



NP Lemma?

- Se passiamo al caso n -dimensionale
- Dobbiamo riempire istogrammi n -dimensionali (tanti quante la variabili) ciascuno con M bins per un totale di celle M^n
- Raramente la statistica MC è sufficiente per parametrizzare le pdf (anche per M e n non troppo grandi)

NP Lemma?

- Due soluzioni
 - Utilizzando delle assunzioni si stimano delle pdf approssimate
 - Si determinano i confini delle ipotesi senza approssimare le pdf

Metodi di classificazione

1 Metodi che approssimano le pdf multidimensionali del lemma di NP

- **Naive Bayesian Classifier**
- Kernel density estimator
- ...

2 Metodi che combinano le variabili (o le selezioni di esse) in modo da approssimare il risultato del lemma di NP

- **Discriminante lineare (Fisher)** linear
- (Boosted) Decision Trees
- Support Vector Machine
- **Neural Networks** non lineare
- ...

Machine Learning

Da el PC campioni di interesse e non interesse \Rightarrow quando do mno campione oppure quale è di interesse

- “Machine learning” è l’ambito dell’informatica che si occupa di fornire ai calcolatori l’abilità di imparare senza essere esplicitamente programmati.
- Si riferisce in questo caso alla determinazione automatica del confine che suddivide lo spazio nelle due classi di eventi secondo un determinato algoritmo
- Si effettua “supervised learning”: il calcolatore stabilisce a quale categoria appartiene un evento basandosi su campioni di apprendimento in cui gli eventi sono correttamente categorizzati
- Questo processo è definito il training del classificatore

Procedura di training e test

- Si divide il campione simulato in un campione di “training” e uno di “test”;
- Si effettua la procedura di training in cui l’algoritmo impara a distinguere tra campioni di categorie diverse
- È importante fare attenzione a non avere overtraining: fluttuazioni statistiche del campione di dati sono interpretate come informazioni significative
- Dopo il training, si applica la selezione ad un campione di test indipendente (che ci permette di verificare tra le altre cose se si ha overtraining)
due copie quegli fluttuazioni statistiche per escluderle

Likelihood (Naive Bayes Classifier)

- Si assume che la PDF a n dimensioni è semplicemente per ogni evento i , il prodotto delle pdf unidimensionali: $p_{s(b)}(x_i)$ (tralasciando possibili correlazioni)
- La pdf unidimensionale può essere stimata dalla proiezione unidimensionale degli eventi simulati che si usano per il training
- Per cui il test statistico $t(\mathbf{x})$ (usando il lemma di NP e una trasformazione monotona) diventa

$$\begin{aligned} t(\mathbf{x}) &= \frac{s(x_1, x_2, \dots, x_n)}{s(x_1, x_2, \dots, x_n) + b(x_1, x_2, \dots, x_n)} \\ &= \frac{\prod_i s_i(x_i)}{\prod_i s_i(x_i) + \prod_i b_i(x_i)} \\ &= \frac{\prod_i s_i(x_i)}{\prod_i s_i(x_i) + \prod_i b_i(x_i)} \end{aligned}$$

- Performance non ottimale se le PDF non fattorizzano

Test statistici lineari: Fisher

- Discriminante lineare: classificatore corrisponde ad un hyper-piano (nello spazio a n dimensioni)
- Si considera la variabile $t(\mathbf{x})$ come una combinazione lineare delle componenti di \mathbf{x} :

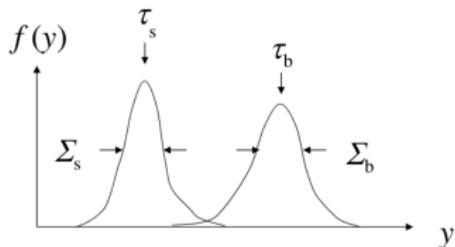
$$t(\mathbf{x}) = \sum_{i=1}^n w_i x_i$$

adatta i pesi per essere
efficiente

- dove \mathbf{w} è il vettore dei coefficienti (chiamati anche pesi)
- Si sceglie il set di parametri \mathbf{w} che dia la massima separazione tra le distribuzioni di t per segnale e fondo
- Il numero di parametri che bisogna determinare sono n

Test statistici lineari: Fisher

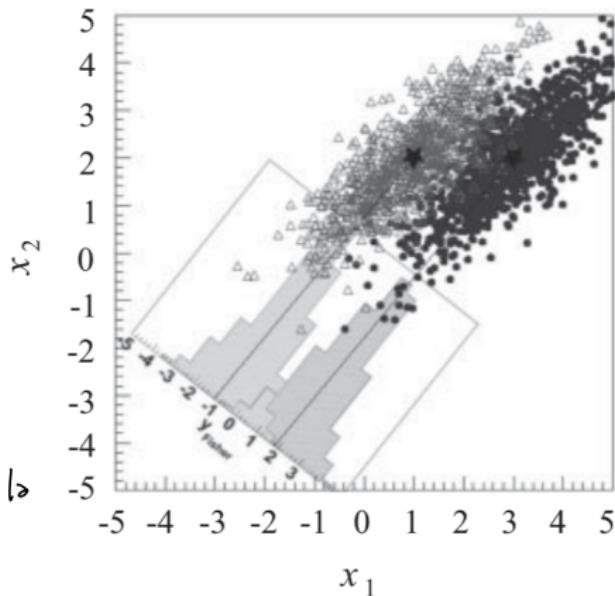
- Per una data scelta di \mathbf{w} determino $f(y|S)$ e $f(y|B)$ sotto le due diverse ipotesi: S e B
- Voglio scegliere i coefficienti \mathbf{w} in cui queste due distribuzioni sono separate il più possibile
- Il discriminante di Fisher massimizza $J(\mathbf{w}) = \frac{(\tau_s - \tau_b)^2}{\sigma_s^2 + \sigma_b^2}$
- dove $\tau_{s(b)} = E_{s(b)}[t(\mathbf{x})]$ è il valore atteso della proiezione della pdf del segnale (fondo) su \mathbf{w} e $\sigma_{s(b)}^2$ sono le rispettive varianze



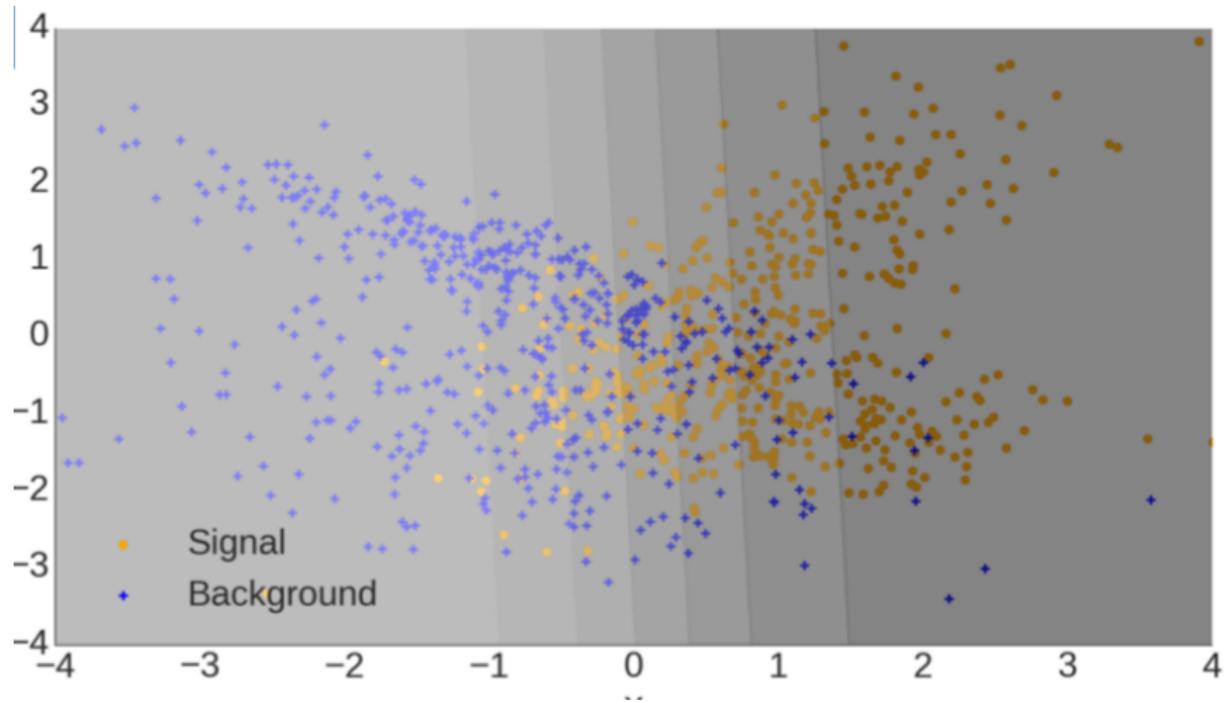
Test statistici lineari: Fisher

- Il vettore peso w può essere interpretato come la direzione di un asse sul quale sono proiettati gli eventi
- La direzione dell'asse sul quale proiettare che massimizza la separazione tra i due tipi di eventi è quella in cui la differenza tra i valori medi delle distribuzioni è massima e lo spread delle distribuzioni è minimo

w è direz dell'asse che massimizza la separazione

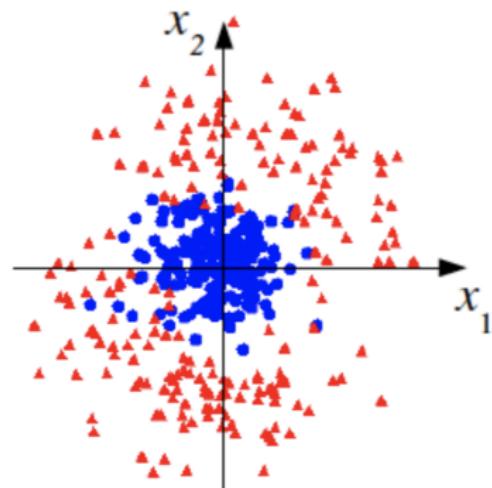
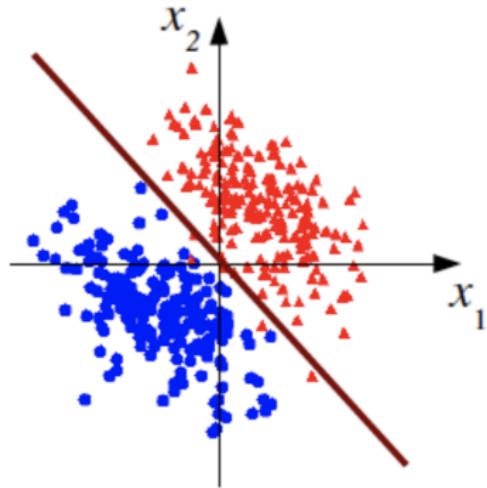


Test statistici lineari: Fisher



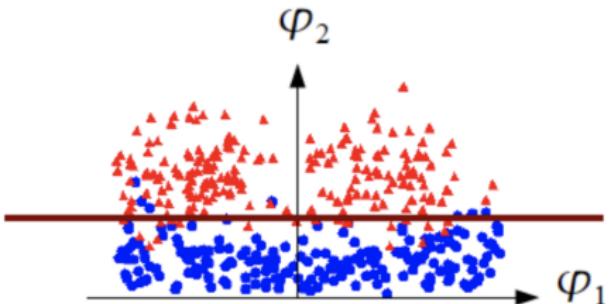
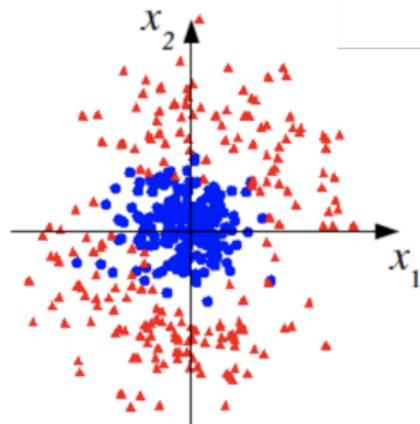
Test statistici lineari

Il concetto di un discriminatore lineare consiste nel costruire una funzione lineare $f(\mathbf{x})$ delle osservabili \mathbf{x} che è usata per fissare il campione di dati di training in modo tale che l'iperpiano di questa funzione, dato da valori costanti di f separa le regioni che sono dominate da eventi di segnale o fondo.



Test statistici non lineari

- L'idea è di trasformare le variabili originali x in un altro set di variabili ϕ :
$$\phi_1 = \tan^{-1}(x_2/x_1)$$
 e $\phi_2 = \sqrt{x_1^2 + x_2^2}$



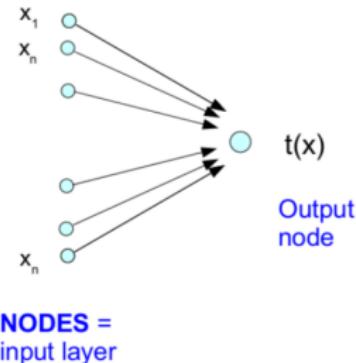
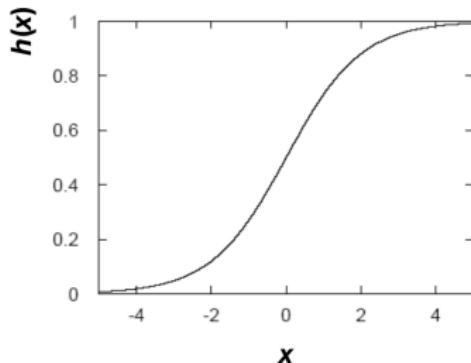
- In questo particolare esempio, abbiamo provato ad indovinare la trasformazione e non avevamo parametri da determinare

Test statistici non lineari: Reti Neurali

- Il concetto di discriminatore lineare può essere esteso ad un caso non-lineare
- Hanno origine dal tentativo di modellizzare i processi neurali ma ora sono utilizzate in numerosi campi
- Metodo specifico di parametrizzare le funzioni che trasformano le componenti di b_{fx}
- Si costruisce $t(x)$ con generiche funzioni non lineari

$$t(x) = h(w_0 + \sum_i w_i x_i)$$

- dove h è una funzione detta di attivazione: sigmoide $h(t) = 1/(1 + e^{-t})$ o tangente iperbolica $h(t) = (e^t - e^{-t})/(e^t + e^{-t})$ (storicamente)



Reti neurali

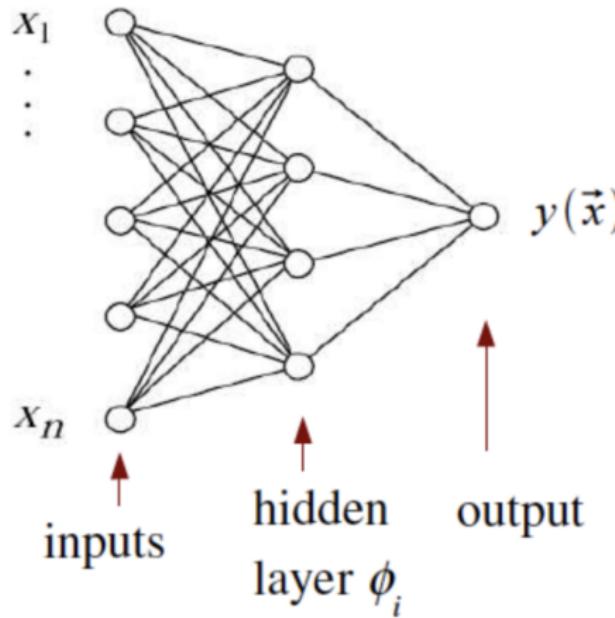
- Riapplichiamo la stessa idea non solo all'output ma anche ad un set di input trasformati ϕ_i che formano uno strato nascosto "hidden layer"

$$\phi_i(\mathbf{x}) = h(w_{i0}^{(1)} + \sum_j^n w_{ij}^{(1)} x_j)$$

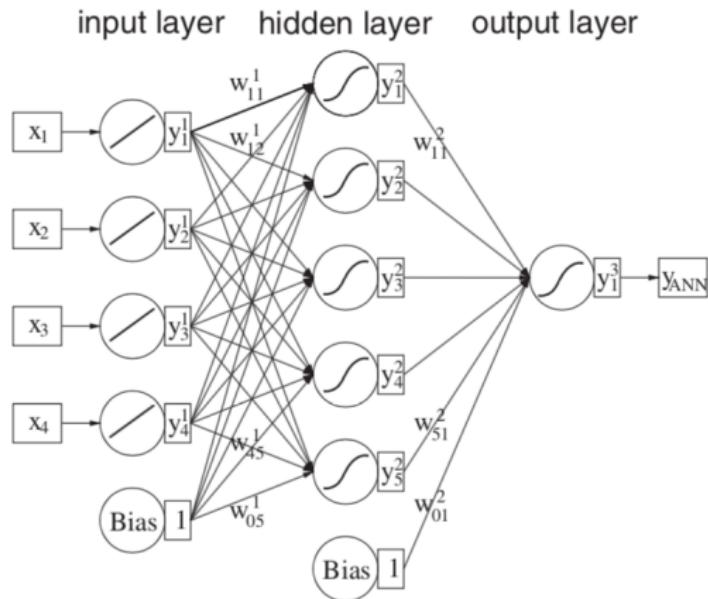
$$y(\mathbf{x}) = h(w_{10}^{(2)} + \sum_j^n w_{1j}^{(2)} \phi_j(\mathbf{x}))$$

L'apice (1) e (2) indica il numero del layer

- Questo è il modello base di una rete neurale
- Il tutto è facilmente estendibile a più di un layer nascosto
- Il numero di parametri è piuttosto elevato (numero di linee del grafico + offsets)



Feed-forward Neural Network



- Gli input vengono solo dai nodi che precedono
- Vengono chiamate feed-forward networks o (multi-layer) perceptrons (MLPs).

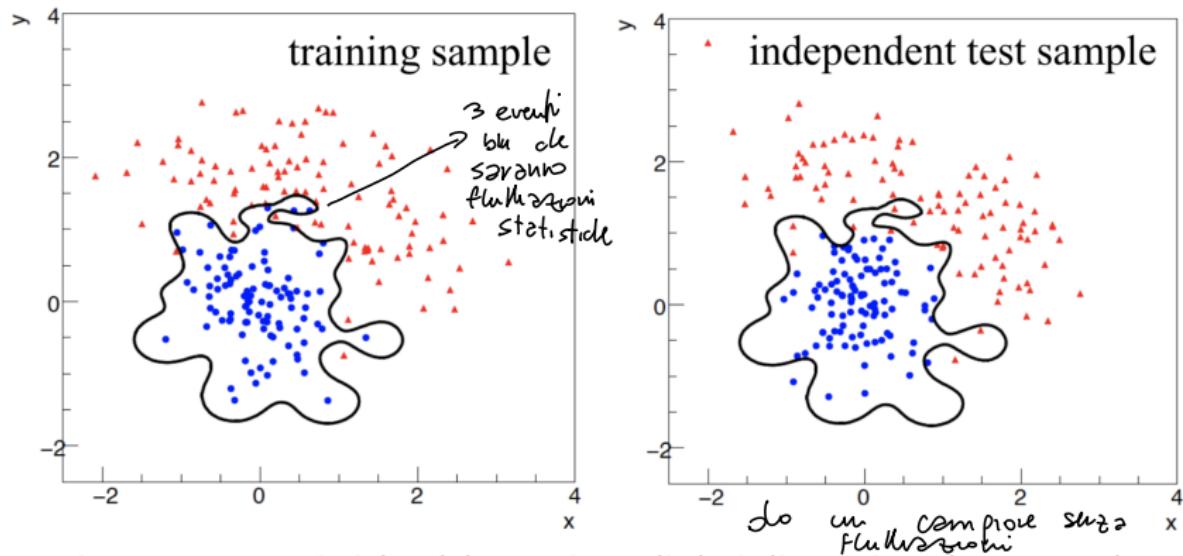
Architettura della rete neurale

Teorema: *una rete neurale con un singolo layer “nascosto” ed un numero sufficientemente grande di nodi può approssimare arbitrariamente la separazione ottimale tra i due campioni (quello che si otterrebbe dal lemma di Neyman-Pearson).*

Solitamente si sceglie come architettura una rete neurale con un singolo layer nascosto e si aumenta il numero di nodi fino a quando non si trova più nessun miglioramento nella performance. Possibile utilizzare più di un singolo layer nascosto: “Deep Neural Networks”

Overtraining

- Neural Network sono soggette a overtraining (numero di parametri grande)



La rete impara caratteristiche del campione di dati di training che sono solamente fluttuazioni statistiche.

Reti neurali

