Lezione 9
Verifica di Ipotesi

Verifica di Ipotesi

- La verifica di ipotesi (detta anche Teoria delle Decisioni) è un altro aspetto fondamentale della Statistica Inferenziale.

- All'interno di un campione di dati (o eventi) capita spesso di dover decidere se l'evento è di un certo tipo (che chiamiamo segnale) oppure se non è di questo tipo e lo chiamiamo fondo.

- Problemi di questo tipo si ritrovano praticamente in ogni attività umana:
  - Decidere se quello che si sta osservando è un evento raro che si sta cercando oppure se è un evento di altro tipo che appare come quello raro che stiamo cercando
  - Decidere se un lotto di un certo materiale prodotto si possa mettere in vendita (in quanto ha i requisiti richiesti) o va trattato diversamente.
  - Un nuovo prodotto eè superiore al precedente oppure no?
  - Una fabbrica va impiantata in Italia, in Brasile oppure in Cina?

Verifica di Ipotesi

- Per poter decidere quale ipotesi è più favorita dalle misure fatte (o più in generale dalle informazioni disponibili) devo fare un test statistico.

- Noi facciamo una certa assunzione che chiamiamo ipotesi. Tradizionalmente questa ipotesi è detta ipotesi nulla H_0 . In genere si fa anche una ipotesi alternativa H_1 ed il test statistico serve a scegliere tra queste due ipotesi

- Se l'ipotesi fatta determina completamente la p.d.f. f(x) di una variabile casuale X, allora l'ipotesi è detta **semplice**

- Se invece la p.d.f. contiene ancora qualche parametro libero θ, f(x; θ), allora l'ipotesi è detta **composta**

- Noi consideriamo solo il caso di ipotesi semplici.

Statistica di Test

- Supponiamo di avere n misure della variabile casuale X: $\mathbf{x} = x_1, x_2, .., x_n$.
  L'ipotesi nulla specifica una p.d.f. congiunta $f(\mathbf{x}; H_0)$ mentre l'ipotesi alternativa specifica una p.d.f. congiunta $f(\mathbf{x}; H_1)$

- Per scegliere tra queste due ipotesi introduco una **statistica di test t(x)**

- Per ogni tipo di ipotesi fatta, la statistica di test avrà una determinata p.d.f. : $g(t; H_0)$ per l'ipotesi nulla e $g(t; H_1)$ per quella alternativa

- La statistica di test t(x) può essere un vettore a più dimensioni:

$$\mathbf{t} = t(t_1, t_2, .. , t_m) \qquad \text{con } m \leq n$$

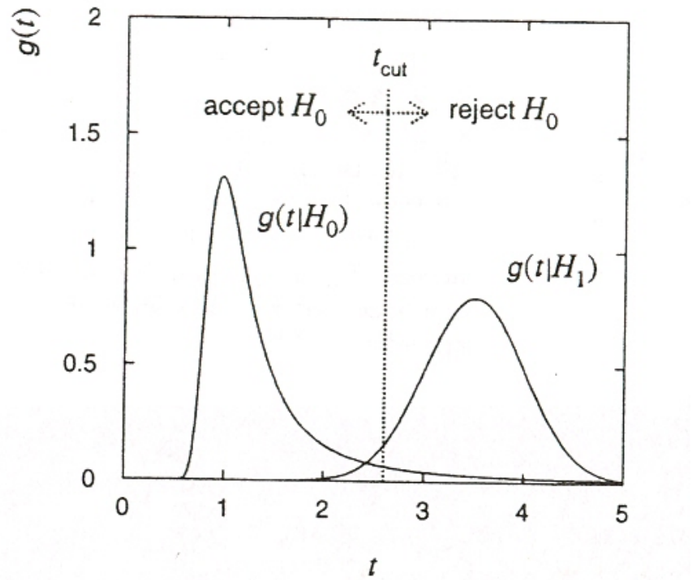
- Noi per semplicità assumiamo che la statistica di test sia una funzione scalare

Statistica di Test

- Le p.d.f. g(t; H_0) e g(t; H_1) della statistica di test le ottengo con eventi MC o direttamente dai dati (quando possibile)

- Definisco un valore di taglio t_{cut} in base al quale decido se l'ipotesi nulla debba essere accettata oppure no

- Per i valori di $t > t_{cut}$ io respingo l'ipotesi nulla



- La regione dei valori in cui l'ipotesi nulla è respinta si dice **regione critica**

- La regione complementare a quella critica è detta **regione di accettanza** (dell'ipotesi nulla)

Statistica di Test

- Calcoliamo ora l'integrale dell p.d.f. della statistica di test nell'ipotesi nulla H_0 estesa a tutta la regione critica:

$$\alpha = \int_{t_{cut}}^{\infty} g(t \mid H_0) dt$$

- α è detto **livello di significanza del test** o anche **misura del test**. Eventi veri dell'ipotesi H_0 per i quali $t > t_{cut}$ vengono rigettati come falsi . α misura la probabilità di rigettare l'ipotesi nulla H_0 quando questa è vera

- L'errore che si commette rigettando l'ipotesi H_0 quando è vera si dice **errore di prima specie** o **errore di tipo I**

- È possibile che nella regione di accettanza ($t \leq t_{cut}$) l'ipotesi accettata come vera non sia H_0 ma l'ipotesi alternativa H_1. La probabilità β che ciò succeda è data da

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

Statistica di Test

- Questo tipo di errore si dice **errore di seconda specie** o di **tipo II**

- **1- β** è la probabilità di rigettare l'ipotesi nulla H_0 quando questa ipotesi è falsa (quindi di rigettare l'ipotesi alternativa). 1 – β è detta **potere del test**

- La caratteristica del test è data dall'insieme (α, β)

- Nel caso di statistica di test monodimensionale (come stiamo supponendo ora) il taglio t_{cut} fissa automaticamente i due tipi di errore e quindi sia l'efficienza della selezione che la purezza del campione selezionato. Variando il taglio all'aumentare di una diminuisce l'altra.

- In talune situazioni ho bisogno di maggiore efficienza (ad esempio ricerca di eventi rari). In altre situazioni ho bisogno di maggiore purezza (selezione di campioni di controllo per calibrare un rivelatore ad esempio). Scelgo quindi il taglio di volta in volta più opportuno

Test Più Potenti

- Per statistiche di test multidimensionali la scelta della regione critica e della regione di accettanza non è ovvia nè semplice da trovare

- Si possono avere diverse regioni critiche ω_α con la stessa misura α del test. Tra queste regioni critiche scegliamo quella che, fissato una misura α, fornisce il valore massimo per la probabilità $(1 - \beta)$

- Queste regioni critiche si chiamano regioni critiche migliori (BCR) e i test che che si basano su queste regioni si chiamano test più potenti (MP).

- Il test MP assicura per un fissato valore di α il valore massimo per la probabilità $(1 - \beta)$

- L'esistenza e l'individuazione del test più potente per la verifica di due **ipotesi semplici** tra loro in alternativa sono garantite dal **Lemma di Neyman-Pearson.**

Lemma di Neyman-Pearson

- Si abbiano due ipotesi semplici ed in alternativa tra di loro H_0 e H_1 ed una statistica di test multidimensionale **t** = t(t_1, t_2, .. t_m)

- Come facciamo a costruire la regione critica migliore che per una fissata efficienza (misura del test α) dia il massimo di purezza (massimo potere del test (1 – β)) ?

- La risposta viene dal **lemma di Neyman-Pearson** (1933):
  La regione di accettanza con la più elevata purezza per una fissata efficienza è data dalla regione nello spazio **t** nella quale si ha:

$$\frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)} > c$$

  dove c è una costante che dipende dalla efficienza richiesta

- Questo rapporto è detto rapporto di massima verosimiglianza (likelihood ratio)

Identificazione di Particelle

- Vediamo un caso interessante di verifica di due ipotesi, considerando la identificazione delle particelle in Fisica Subnucleare

- In un esperimento di alte energie ad un acceleratore è possibile produrre e studiare particelle (a vita media breve) che decadono in altra particelle (elettroni, pioni, kaoni, ecc). Per esempio si può studiare se è prodotto e con quale tasso decade un mesone B in η' K. Questo è un decadimento raro (B decade cosi 65 volte su 10^6). Oltre a questo decadimento c'è anche B in η' π (che ha un tasso di decadimento molto più elevato !)

- È chiaro che l'apparato quando una particella lo attraverso deve avere elevata potenza nel discriminare un π da un K !!

- L'apparato sperimentale nel passaggio della particella deve misurare opportune quantità fisiche che permettano di scegliere tra l'ipotesi π e l'ipotesi K

Risposta di un Rivelatore: p.d.f. e LF

- La risposta di un rivelatore al passaggio di una particella è data dalla p.d.f. P(x; p, H) che descrive la densità probabilità che una particelle di tipo H (per esempio e, p, π, K, …) e di quantità di moto p rilasci nel rivelatore un segnale x (perdita di energia, luce Cherenkov, ecc)

- P(x; p, H) dx è la probabilità che una particella di tipo H e quantità di moto p rilasci nel rivelatore un segnale compreso tra x e x + dx

- La p.d.f. P(x; p, H) di risposta del rivelatore viene determinata o da campioni di dati controllo oppure da eventi Monte Carlo

- La likelihood per l'ipotesi di una particella di tipo H che con quantità di moto p rilascia un segnale x è definita da :

$$L(H;p,x) \equiv P(x;p,H)$$

Risposta di un Rivelatore: p.d.f. e LF

- Si noti che la LF è una funzione dei tipo di ipotesi (tipo di particella) H per dato impulso p e segnale rilasciato x mentre la p.d.f. è una funzione del segnale x per una data quantità di moto p e una data ipotesi (tipo di particella) H

- Confronto di ipotesi alternative (π o K ?) su una particella può essere fatto mediante il rapporto delle likelihood. Per esempio per discriminare tra un pione positivo π^+ e un kaone positivo K^+ utilizzo il rapporto:

$$L(K^+; p_{oss}, x_{oss}) / L(\pi^+; p_{oss}, x_{oss})$$

con p_{oss} e x_{oss} valori della quantita' di moto misurato dall'apparato sperimentale e segnale rilasciato

- Determino una costante c che mi permette di avere una efficienza di identificazione fissata e quindi considero K tutte le particelle per le quali il rapporto delle likelihood e' maggiore di c .

Consistenza e Livello di Significanza

- Un test statistico di consistenza non è un test che permette di scegliere tra due ipotesi concorrenti. Esso permette di stabilire quanto bene un una misura si accorda con quanto aspettato nell'ipotesi che la particella sia di tipo H

- Si pone la seguente domanda: Qual è la frazione di tracce vere di tipo H che sembrerebbero meno vere di questa traccia ?

- Sia P(x|H) la p.d.f. della variabile X misurata per l'ipotesi H. Il livello di significanza (SL) o consistenza di una misura x_{oss} data l'ipotesi H è data da:

$$SL(x_{oss} \mid H) = 1 - \int_{P(x|H)>P(x_{oss}|H)} P(x \mid H)dx$$

- O anche equivalentemente da: $SL(x_{oss} \mid H) = \int_{P(x|H)<P(x_{oss}|H)} P(x \mid H)dx$

Consistenza e Livello di Significanza

- Supponiamo che una certa quantità X sia misurata da un rivelatore con una p.d.f. gaussiana:

$$P(x \mid H) = \frac{1}{\sqrt{2\pi}\sigma(H)} \exp\left[-\frac{1}{2}\left(\frac{(x-\mu(H))}{\sigma(H)}\right)^2\right]$$

- Il livello di significanza per una misura x_{oss} data l'ipotesi H è definito da

$$SL(x_{oss} \mid H) = 1 - \int_{\mu(H)-x_{oss}}^{\mu(H)+x_{oss}} P(x;H)dx$$

- Questo è un test di consistenza a due lati. Per p.d.f. non simmetriche si posso fare test da un lato integrando da x_{oss} a $+\infty$ oppure da $-\infty$ a x_{oss}

- Questo test può essere utilizzato per eliminare tracce inconsistenti con l'ipotesi H fatta.

- È anche possibile fare un confronto tra due ipotesi confrontando il livello di significanza per le due ipotesi.

Probabilità a Posteriori

- In alcuni casi le probabilità $P_A(H)$ a priori (cioè prima che si faccia la misura) delle due ipotesi competitive sono note. Per esempio posso sapere che in un fascio di particelle su sette pioni c'è un kaone.

- In questo caso la probabilità a posteriori F(K; x) che la particella sia una K data la misura x fatta è data da:

$$F(K;x) = \frac{L(K;x) \cdot P_A(K)}{L(\pi;x) \cdot P_A(\pi) + L(K;x) \cdot P_A(K)} = \frac{L(K;x)}{L(\pi;x) \cdot 7 + L(K;x) \cdot 1}$$

  dove L(K; x) e L(π; x) sono le likelihood per le due ipotesi K e π, data la misura x effettuata. La F(H; x) è detta anche probabilita' condizionale o anche relativa.

- Questa probabilità a posteriori può essere utilizzata per calcolare la purezza aspettata da una certa selezione fatta.

Statistiche di Test

- Indichiamo con x = $x_1, x_2, .. , x_n$ il vettore delle n variabili discriminanti in ogni evento che vogliamo utilizzare per distinguere tra due ipotesi semplici ed alternative H_0 e H_1. Vedremo poi come scegliere le n variabili discriminanti

- Come statistica di test **t(x)** possiamo usare il lemma di Neyman-Pearson che mi assicura il taglio più potente per una desiderata efficienza:

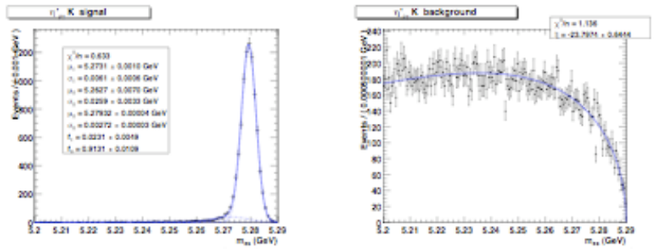
$$t(\mathbf{x}) = \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)}$$

- Per fare il rapporto delle likelihood, devo conoscere le p.d.f. per tutte e due le ipotesi H_0 e H_1. Questo lo potrei fare utilizzando eventi simulati MC

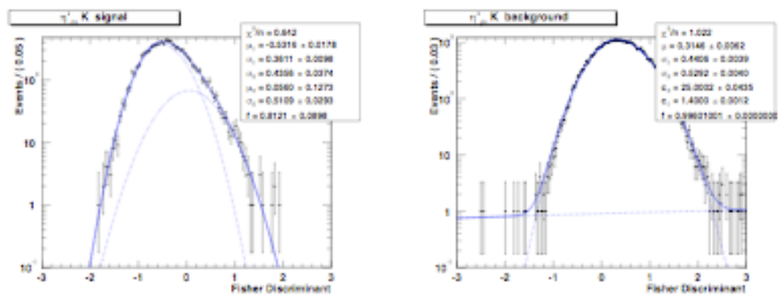
- Si noti però che le p.d.f. nelle due ipotesi sono istogrammi ad n dimensioni. Se prendo M bin in ogni istogramma dovrei determinare M^n parametri con i dati MC. Per grandi n questo è poco o del tutto non pratico

Scelta delle Variabili Discriminanti

- Le variabili discriminanti tra due ipotesi possono essere diverse e tutte in generale hanno un diverso potere di separazione. In figura è mostrato un esempio di variabile discriminante con elevato potere di separazione.



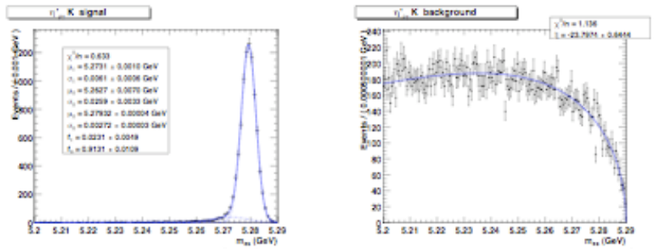
- Spesso vi sono variabili che hanno scarso potere di separazione. Quello che si osserva però è che combinando assieme in modo opportuno diverse di queste deboli variabili discriminanti, il loro potere discriminante aumenta (e talvolta di molto)



Questo Fisher è fatto con 5 variabili debolmente discriminanti

Scelta delle Variabili Discriminanti

- La selezione delle variabili discriminanti appartiene alla prima fase dell'analisi statistica di dati sperimentali. In talune situazioni l'importanza nella selezione di alcune variabili discriminanti è nota a priori o da analisi precedenti o da considerazioni di carattere cinematico (o dinamico) in Fisica. Vediamo ad esempio la distribuzione della massa del mesone B come ricostruita in eventi di segnale (a sinistra) e in eventi di fondo combinatorio (a destra) :



- In generale però è necessario fare uno studio mirato per determinare l'ordine di importanza delle variabili discriminanti.

- Si hanno diversi classificatori (discriminante di Fisher, reti neurali, boosted decision tree , random forest, ecc) che vedremo in seguito. La scelta della variabili discriminanti in questi classificatori dipende anche dal classificatore usato.

Forward Stepwise Addition

- Vi sono diversi metodi usati per valutare l'importanza (discriminatoria) relativa delle osservabili.

- Un metodo ben noto e molto usato è il Forward Stepwise Addition (FSA).

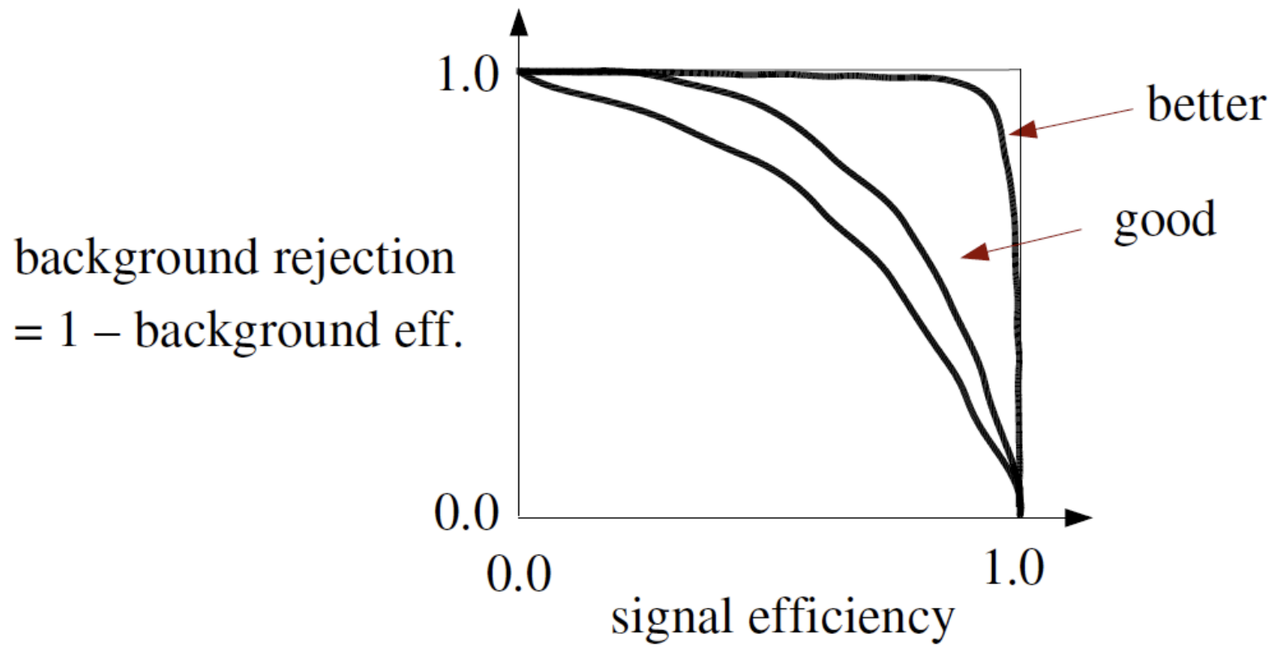
- Si individua un classificatore (per esempio una rete neurale) e si definisce una figura di merito (FOM) in base alla quale si valuta il potere discriminatorio di una variabile.

- Esistono molti tipi di FOM (tipo significanza statistica $S/\sqrt{(S+B)}$, rapporto segnale/fondo, ecc) ognuno dei quali ottimale in particolari tipi di analisi.

- Una FOM molto usata in Statistica è la curva Receiver Operating Characteristics (ROC)

Forward Stepwise Addition

- ROC è l'efficienza di reiezione del fondo (asse y) in funzione della efficienza del segnale (asse x). Più grande è l'area sotto la ROC , migliore la performance del classificatore.



Forward Stepwise Addition

- Scelto il classificatore le variabili vengono aggiunte una alla volta. Si calcola la FOM e si sceglie la variabile con il più grande aumento della FOM

- L'addizione di nuove variabili si arresta quando non è pù possibile aumentare la FOM

- Questa tecnica può essere migliorata. Si può decidere che ad ogni passo si aggiungono n variabili e se ne tolgono r. Viene tenuto sempre il sottoinsieme con la minor perdita sul test

Discriminante di Fisher

- Si può per semplicità selezionare come statistiche di test particolari funzioni lineari o non lineari delle misure sperimentali.

- Consideriamo ad esempio un campione di eventi costituito da due diversi tipi (o classi) di eventi. Un tipo lo chiamiamo segnale (questo è il tipo di eventi a cui siamo interessati) e l'altro lo chiamiamo fondo.

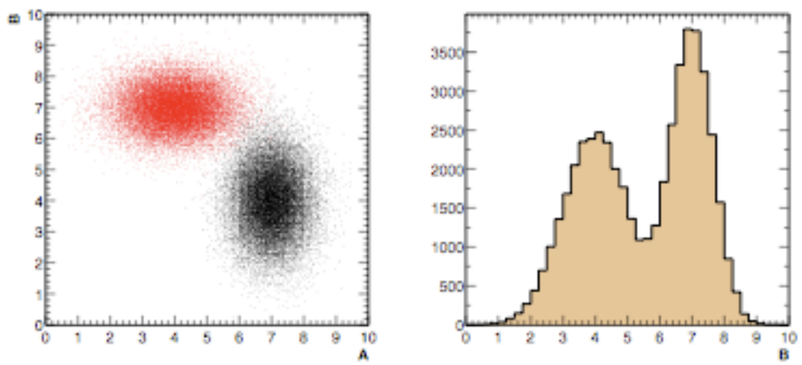
- Noi vogliamo cercare una statistica di test che mi permetta di separare al meglio questo campione nelle due classi segnale e fondo.

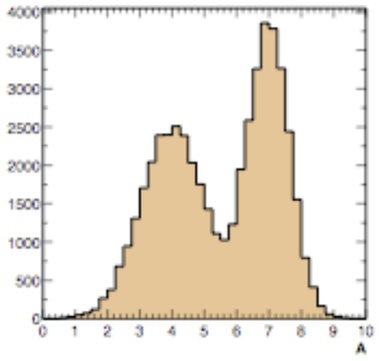
- Consideriamo in ogni evento n variabili discriminanti che possano in qualche misura avere p.d.f. diverse per gli eventi di segnale e per quelli di fondo

- Per avere un'idea di quello che vogliamo fare consideriamo il caso di due sole variabili discriminanti A e B

Idea di Base del Discriminante di Fisher

Scatter plot delle due variabili discriminanti A e B

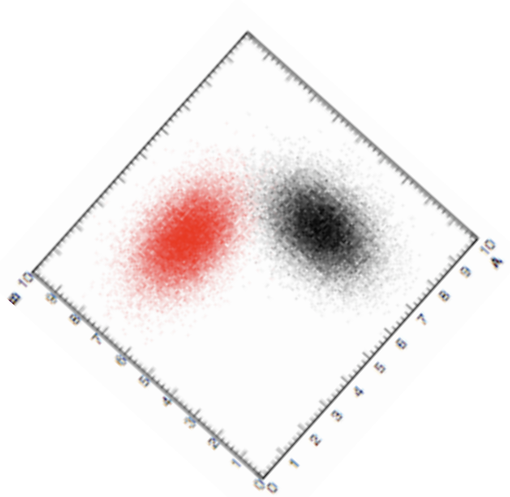


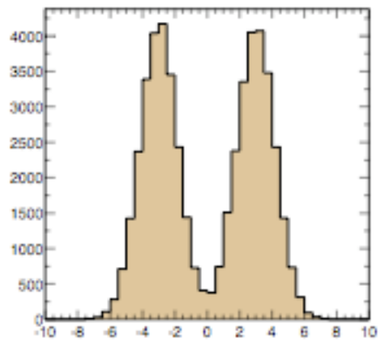


Scegliamo due variabili discriminanti A e B per ogni evento e con queste cerchiamo di separare il campione di misure nelle due classi (eventi in rosso e nero). Per separare le due classi potrei fare le proiezioni sugli assi e fare un taglio sulle variabili A e B

Da queste proiezioni (alto a destra e in basso a sinistra) osservo che la la separazione non è ottimale. Cosa potrei fare per migliorare la separazione dei due tipi di evento?

Idea di Base del Discriminante di Fisher





Immaginiamo di ruotare le variabili A e B

Come si vede dalla proiezione in basso a destra, ora la separazione tra le due classi è molto migliorata.

Per fare questo devo ruotare il sistema di riferimento passando dal riferimento iniziale a quello ruotato. Le nuove coordinate si ottengono mediante una combinazione lineare delle coordinate iniziali (in questo caso si ha una matrice di rotazione 2x2).

Naturalmente a seconda della rotazione effettuata il livello di separazione varia: quindi i coefficienti della combinazione lineare devono essere ottimizzati (per avere la massima separazione possibile)

Discriminante di Fisher

- Scelte in ogni evento le n variabili discriminanti linearmente indipendenti $x_1, x_2, .. , x_n$, la statistica di test

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

è detta **discriminante (lineare) di Fisher**. a^T è il vettore trasposto del vettore **a** dei coefficienti $a_1, a_2, .., a_n$

- Devo ottimizzare i coefficienti in modo da massimizzare la distanza (separazione) tra la pdf di una classe e la pdf dell'altra classe. Questo può essere fatto in diversi modi. Qui seguiamo l'approccio di Fisher.

- Consideriamo i valori medi e matrice di covarianza per le due ipotesi H_0 e H_1 (k=0 e k=1)

$$(\mu_k)_i = \int x_i f(\mathbf{x}|H_k) dx_1 \cdots dx_n$$

$$(V_k)_{ij} = \int (x - \mu_k)_i (x - \mu_k)_j f(\mathbf{x}|H_k) dx_1 \cdots dx_n$$

Discriminante di Fisher

- Analogamente consideriamo valori medi e varianze per il discriminante di Fisher per le due ipotesi H_0 e H_1

$$\tau_k = \int t \cdot g(t|H_k)dt = \mathbf{a}^T \mu_k$$

$$\Sigma_k^2 = \int (t - \tau_k)^2 g(t|H_k)dt = \mathbf{a}^T V_k \mathbf{a}$$

- Per aumentare la separazione tra i due tipi posso aumentare nello spazio ad n dimensioni la distanza $|\tau_0 - \tau_1|$.

- La separazione migliora anche quanto più strette sono le distribuzioni attorno a τ_0 e τ_1 e quindi quanto più piccole sono le varianze Σ_0^2 e Σ_1^2

- La quantità che scelgo per ottimizzare la separazione è:

$$J(\mathbf{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}$$

Discriminante di Fisher

- Riscriviamo numeratore in termini delle misure

$$(\tau_o - \tau_1)^2 = \sum_{i,j=1}^{n} a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j = \sum_{i,j=1}^{n} a_i a_j B_{ij} = \mathbf{a}^T B \mathbf{a}$$

  con la matrice B definita da: $B_{ij} = (\mu_o - \mu_1)_i (\mu_0 - \mu_1)_j$

- Per il denominatore si ha: $\Sigma_0^2 + \Sigma_1^2 = \sum_{i,j=1}^{n} a_i a_j (V_0 + V_1)_{ij} = \mathbf{a}^T W \mathbf{a}$

  con $W_{ij} = (V_0 + V_1)_{ij}$

- Sostituendo si ha: $J(\mathbf{a}) = \dfrac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T W \mathbf{a}}$

- Per massimizzare questa quantità, pongo uguali a zero le derivate rispetto ai coefficienti e ottengo i valori ottimizzati dei parametri

Discriminante di Fisher

$$\mathbf{a} \propto W^{-1}(\mu_0 - \mu_1)$$

- Come si vede i coefficienti sono determinati a meno di un fattore di scala
  La definizione del discriminante può essere generalizzata nel moto seguente

$$t(\mathbf{a}) = a_0 + \sum_{i=1}^{n} a_i x_i$$

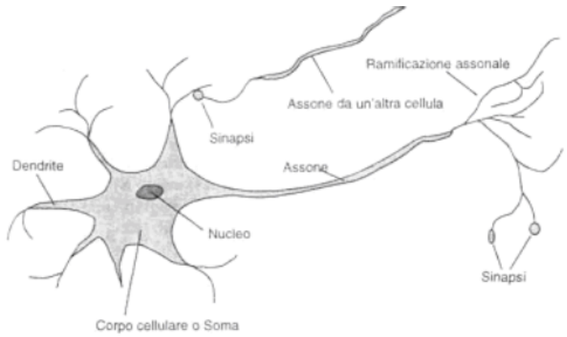
dove a_0 (offset) e il fattore di scala sono scelti in modo da fissare i valori di τ_0 e τ_1 a qualunque valore desiderato

- La matrice W ed i valori di aspettazione μ_0 e μ_1 sono determinati utilizzando dati di training generalmente generati con tecniche MC.
  Si simulano eventi MC per il segnale e per il fondo.
  Uso questi eventi per ottimizzare il discriminante di Fisher, calcolandone i coefficienti

- Quindi uso il discriminante di Fisher (con i coefficienti già ottimizzati) sui dati per discriminare il segnale dal fondo

Reti neurali Artificiali

- Le reti neurali artificiali (o semplicemente reti neurali) imitano le reti neurali biologiche come il nostro cervello.



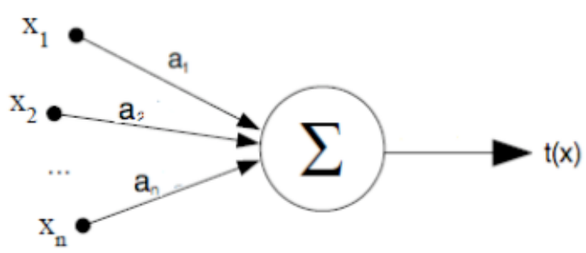
- il neurone è una speciale cellula in grado di ricevere impulsi da altri neuroni tramite le ramificazioni (dette dendriti). Le informazione ricevute vengono elaborate dal corpo centrale del neurone e trasmesse ad un altro neurone (denominato neurone post-sinaptico) o verso altre cellule tramite una lunga estensione denominata assone.

- Il neurone ha quindi porte di ingresso da cui riceve informazioni (stimoli) . In base alla intensità di questi stimoli si attiva (si eccita) oppure no.

- Il neurone ha una porta di uscita (l'assone) da cui (se attivato) trasmette informazione al neurone post-sinaptico.

Percettrone

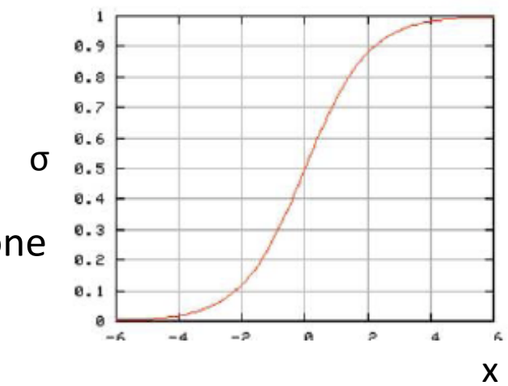
- Il percettrone è la rete neurale più semplice . È costituito da un solo neurone (detto nodo) che ha un certo numero n di ingressi (i valori delle variabili discriminanti x_1, x_2, ….. X_n)



- Nel nodo le informazioni entranti vengono opportunamente pesate con i pesi a_1, a_2, …, a_n e sommate in modo da calcolare un potenziale di attivazione.

- La funzione di attivazione può avere forme diverse (dare il segno della funzione, o essere funzione a scalino (0,1) oppure dare in uscita una distribuzione continua mediante la funzione sigmoidea:

$$\sigma(x) = \frac{1}{1 - e^{-x}}$$



Rete Neurale Multistrato

- La formula di uscita della rete è data da:

$$t(x) = \sigma(a_0 + \sum_{i=1}^{n} a_i x_i)$$

dove il termine a_0 è un termine di offset denominato bias.

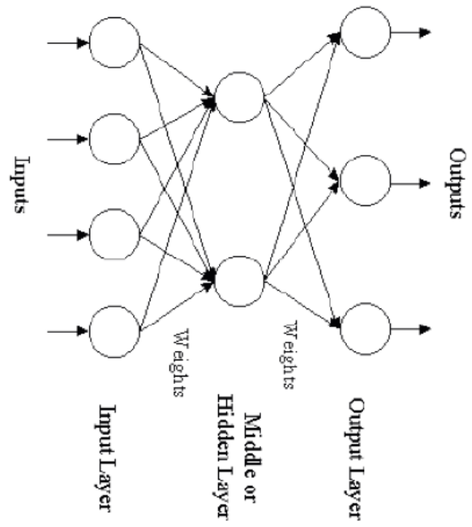
- Il bias può essere considerato il peso di un nodo fittizio e la formula vista può essere riscritta cosi:

$$t(x) = \sigma(\sum_{i=0}^{n} a_i x_i)$$

- L'architettura di una rete neurale può essere varia. Oltre allo strato in ingresso, si può avere uno strato in uscita con uno o più nodi e tra lo strato in ingresso e quello in uscita si può avere uno o più strati intermedi detti anche strati nascosti. Tipicamente vi è un solo strato nascosto.

Reti Neurali Multistrato

- In queste reti multistrato si può fare in modo che i valori in input in un certo strato derivino solo da nodi dello strato precedente (come nella rete in figura).



- Questo tipo di rete neurale è detta "feed-forward".

- Una volta definita l'architettura della rete, questa deve essere istruita (fase di addestramento)
  Volendo usare la rete per esempio per separare due classi di eventi (tipo H_0 e tipo H_1) dobbiamo insegnare alla rete come fare queta separazione.

- Usiamo un campione di eventi di tipo H_0 (chiamiamoli segnali) e un campione di eventi di tipo H_1 (chiamiamoli fondo). Questi campioni (training set) possono essere o simulati oppure campioni di dati di controllo.

Apprendimento e Overtraining

- Si danno in pasto alla rete (in modo casuale) eventi di segnale ed eventi di fondo. La rete conosce il tipo di evento in ingresso.

- Per ogni ciclo la rete riaggiusta i parametri (pesi) delle varie variabili in modo da ridurre l'errore tra il valore in uscita generato nel nodo ed il valore vero (che la rete conosce). Cosi facendo la rete impara a distinguere un evento di un tipo (segnale) da un evento di altro tipo (fondo).

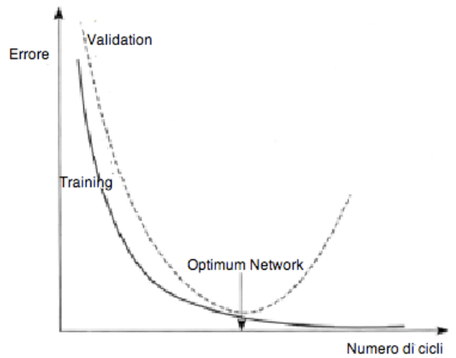
- Questo tipo di apprendimento è detto supervisionato

- Come faccio a controllare che non ci siano bias nell'addestramento? Una possibilità è di suddividere il training set in K sottocampioni. Addestro la rete in un sottocampione e la verifico sull'insieme dei K-1 sottocampioni (aggregati). Itero K volte e prendo la media dei risultati (K-fold cross-validation).

- L'apprendimento da parte della rete ha però un problema detto overtraining. Aumentando il numero di cicli nella fase di training, l'errore della rete nella separazione segnale-fondo tende a zero. Questo perché la rete si adatta sempre più alle caratteristiche del training set.

Validazione e Test

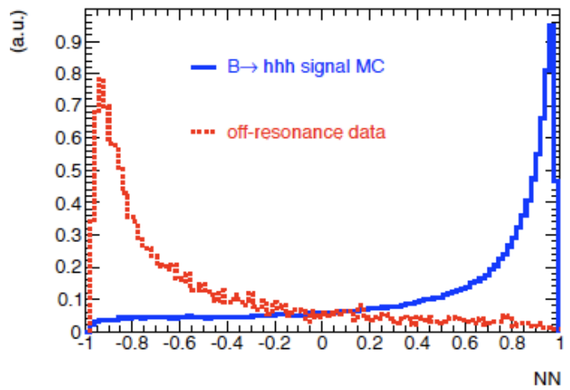
- È necessario perciò usare la rete già istruita con un altro campione di dati (validation set), indipendente dal training set. In questo caso al crescere del numero di cicli di addestramento, verifico la qualità dell'addestramento sul validation set. Quando noto che l'errore di identificazione sul validation set comincia ad aumentare, arresto il training.



- Quando la rete è stata validata, si utilizza un altro campione di test indipendente (test set) per valutare l'accuratezza finale della rete.

- Una volta addestrata, la rete ricevendo in ingresso un evento (di tipo non noto) è in grado di identificare (con una certa probabilità) il tipo di evento

- Fasi di addestramento e problema dell'overtraining sono comuni a tutti i classificatori multivariati.



Significanza (Statistica) di un Segnale

- Abbiamo visto un livello di significanza nel confronto tra due ipotesi

  ed un livello di significanza (detta anche consistenza) che mi dice quanto la misura che ho fatto è consistente con una certa ipotesi. Lo stesso termine è usato per indicare due cose **completamente** diverse
- Nel primo modo si tratta di un test a due ipotesi dove la regione di accettanza va definita **prima** che si faccia l'esperimento o che si utilizzino i dati sperimentali.
- Nel secondo metodo la significanza dipende solo dalle **misure fatte** e dalla **p.d.f.** della ipotesi assunta vera. Molto spesso si quota per quantificare quanto una misura sperimentale è inconsistente con una certa ipotesi.

  Di fatto questo non è altro che un p-value cioè la probabilità sotto l'ipotesi fatta di ottenere un risultato compatibile o meno compatibile di quello effettivamente osservato.

- Quando si cercano cose nuove o si trovano cose inaspettate è in questo secondo modo che usualmente è intesa la significanza (in HEP)

Significanza in un Esperimento di Conteggio

- In un esperimento di conteggio si contano in una zona detta di segnale il numero totale di eventi n accumulati e il numero di eventi di fondo n_b aspettati nella stessa regione.

- Il numero di eventi di segnale è $n_s = n - n_b$. Per ora supponiamo che n_b sia noto con errore nullo. Le tre variabili n, n_s e n_b sono variabili poissoniane con valori di aspettazione ν_s, ν_b e $\nu = \nu_s + \nu_b$

- La probabilità di osservare n candidati assumendo una distribuzione poissoniana è:

$$f(n;\nu_s,\nu_b) = \frac{(\nu_s+\nu_b)^n}{n!}\exp\left[-(\nu_s+\nu_b)\right]$$

- Gli eventi che considero come segnale potrebbero essere effetto di una fluttuazione in alto del numero di eventi di fondo. Se osservo n_{oss} candidati io devo calcolare quanto è la probabilità che il fondo fluttui dando un numero di eventi uguale o maggiore ad n_{oss} supponendo che non ci siano segnali ($n_s = 0$)

Significanza in un Esperimento di Conteggio

- Questa probabilità (p-value) è data da:

$$P(n \geq n_o ss) = \sum_{n=n_{oss}}^{\infty} f(n;\nu_s = 0,\nu_b) = 1 - \sum_{n=0}^{n_{oss}-1} f(n;\nu_s = 0,\nu_b)$$

$$= 1 - \sum_{n=0}^{n_{oss}-1} \frac{\nu_b^n}{n!} \exp\left[-\nu_b\right]$$

- Per esempio ho osservato 5 eventi mentre mi aspetto v_b = 0.5. In questo caso la probabilità che i 5 eventi siano dovuti a fluttuazione del fondo è 1.7 10^{-4}. Questo in termini frequentisti significa che se accettassi l'ipotesi che sia fluttuazione del fondo a questo p-value farei una cosa giusta una su 5882 volte. Quindi questa ipotesi viene rigettata

- Noi stiamo cercando una fluttuazione in alto dal valore medio. Si può esprimere il p-value riportando in una gaussiana standard l'area a destra da $+\infty$ sino al punto tale che l'area racchiusa sia pari al p-value. Questo punto indica a quante sigma sono dal'ipotesi rigettata. Nel caso precedente l'ipotesi di flutttuazione del fondo è esclusa con una significanza di 3.6 σ

Significanza in un Esperimento di Conteggio

- Se il numero di eventi di fondo aspettati è noto con un certo errore si determina un intervallo di possibili valori di ν_b e per ognuno di questi conseguentemente si determina un intervallo di possibili valori di p-value.

- In questo esperimento abbiamo cercato se c'è un eccesso di eventi sopra il fondo aspettato in una zona ben precisa (e nota a priori) che abbiamo chiamato regione del segnale.

- **Da quanto detto è chiaro che il p-value permette di rigettare una ipotesi con una certa significanza ma NON permette mai di avvalorare un'ipotesi.**

Test del χ^2 di Pearson

- Supponiamo di aver misurato una variabile che distribuiamo in un istogramma di N bin. Supponiamo che la statistica di misure permetta di avere almeno 5 eventi per ogni bin. In una regione **dove mi aspetto un segnale** trovo effettivamente un eccesso di eventi sul fondo.

- Faccio un fit sui dati sovrapponendo una curva che mi descrive il fondo ad una curva che mi descrive il segnale. Dal fit trovo che nella regione del segnale trovo un numero di eventi di segnale n_s su un fondo di n_b eventi.

- Come posso convincermi che sto osservando veramente un segnale e non una fluttuazione del fondo?

- Faccio l'ipotesi che ci sia solo fondo e con questa ipotesi fitto i dati sperimentali. Calcolo quindi il χ^2 del fit:

Test del χ^2 di Pearson

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

con n_i numero di eventi trovati nel bin i-esimo e v_i il numero di eventi aspettati nell'ipotesi di solo fondo.

- Il p-value lo trovo integrando la distribuzione del χ^2 , con n_d gradi di libertà, dal valore di χ^2 osservato all'infinito

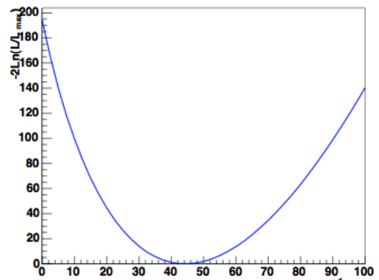
$$P = \int_{\chi^2}^{\infty} f(z; n_d) dz$$

- Da questo calcolo posso determinare con quale significanza posso eventualmente rigettare l'ipotesi che l'eccesso trovato nella regione del segnale sia dovuto a fluttuazione statistica del fondo

- Se **non** si conosce la regione del segnale, bisogna tener conto del fatto che la fluttuazione del fondo osservata potrebbe essere in uno qualunque dei bin e questo abbassa la significanza nell'osservazione di un eventuale segnale (look elsewhere effect)

Significanza di un Segnale col ML

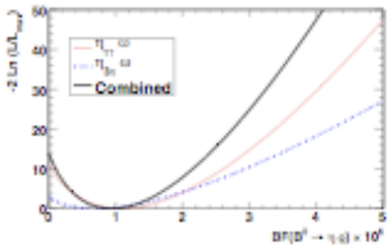
- Vediamo come calcolare la significanza statistica di un segnale in una analisi di ML. Faccio lo scan della likelihood (un esempio in figura) dove e riportato -2log(L/Lmax). Questa per grandi campioni di dati ha un andamento di tipo parabolico (la likelihood ha forma gaussiana) .

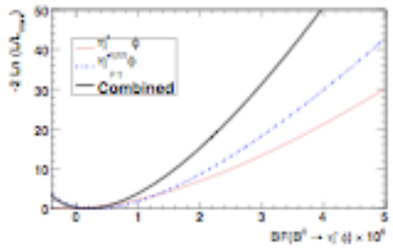


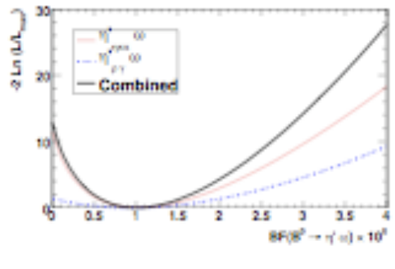
- In questa ipotesi -2log(L/Lmax) ha un andamento del χ^2 con un numero di dof pari alla differenza tra il numero di parametri liberi al massimo della L e il numero di parametri liberi con zero segnale. Se siamo nel caso che congeliamo un solo parametro libero ponendo n_s =0, allora la significanza statistica S è data in unita di σ dalla radice quadrata del valore del χ^2 a zero segnale (intercetta della L sull'asse y) :

$$S = \sqrt{\chi 2}\ (n_s = 0)\ \sigma$$

Significanza di un Segnale col ML







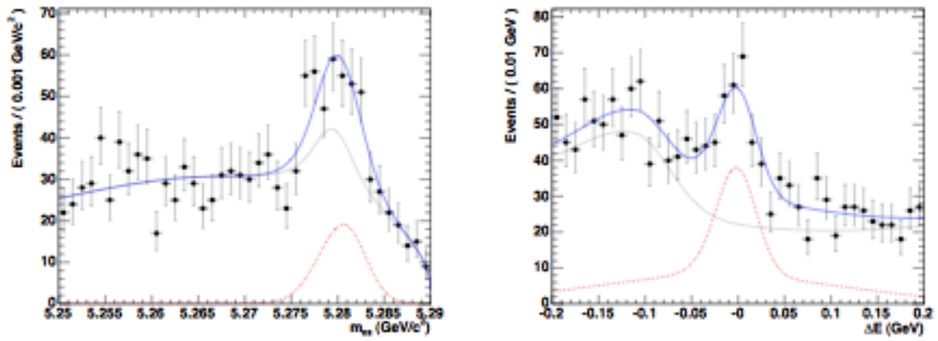
In questi altri tipi di decadimento del mesone B il numero di segnali è minore, la logL non è parabolica perché la L non è gaussiana.

- -2log(L/L_{max}) non va piu' come il χ^2 ma calcolo la significanza S ancora come la radice quadrata del χ^2 nell'ipotesi di zero segnale. Qui il calcolo della significanza è generoso!

- Nella prassi delle alte energie con S ≥ 5 σ si ha una osservazione ; con 3 σ ≤ S < 5 si ha una evidenza; con S < 3σ si dà un UL (spesso al 90%)

- Nel calcolo finale della significanza dovrò tener conto delle incertezze sistematiche.

Controllo di Bontà del fit col ML

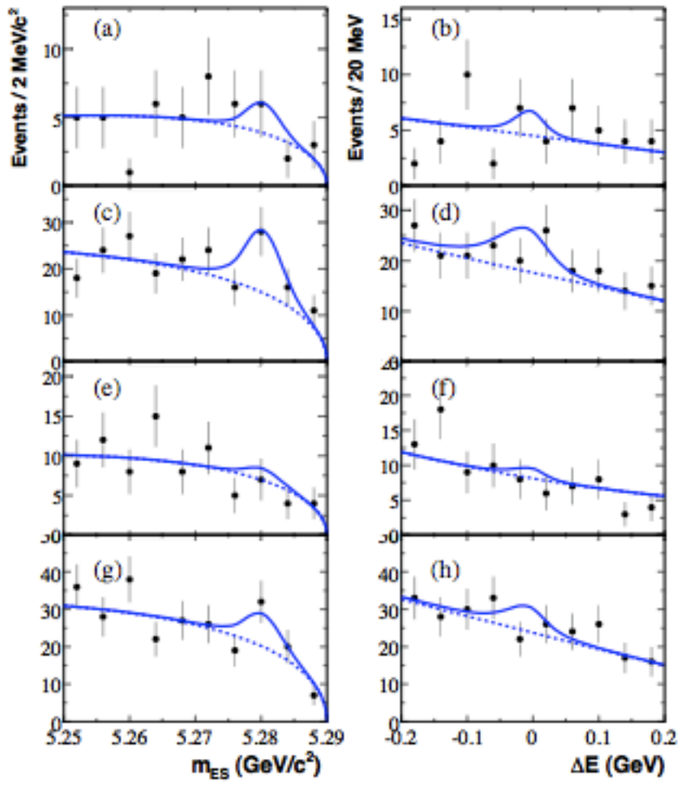


Un controllo della bontà del fit (e sulla significanza di un segnale) può essere fatto utilizzando le proiezioni degli eventi sulle variabili discriminanti. Sopra sono riportate le distribuzioni su due variabili discriminanti dove è ben visibile un fondo su sui c'è un segnale co massa intorno a 5.28 GeV/c^2 e ΔE attorno a zero (come aspettato)

Questo controllo può essere fatto ad esempio tagliando duro su tutte le variabili in modo da isolare un campione ricco di segnale (se sono veri). Si plottano le variabili discriminanti e si sovrappone il fit del Ml (scalato per l'effetto dei tagli).

Se il segnale è significativo (come nella figura riportata) allora ci sentiamo più sicuri nel dire che abbiamo osservato un segnale nuovo.

Controllo di Bontà del fit col ML



In questi decadimenti del mesone B invece il numero di segnali non è significativo e questo è confortato dalle proiezioni

Test di Kolmogorov-Smirnov

- Supponiamo di avere n misure della variabile casuale X

- Il test di Kolmogorov-Smirnov utilizza dati non istogrammati e permette di controllare quanto un campione di dati segue una certa p.d.f. f a parametri noti (cioè non estratti da fit sul campione !!).

- Possiamo calcolare la c.d.f. F della p.d.f. f e la c.d.f. $S_n(x)$, detta cumulativa empirica, costruita con i dati . Per calcolare $S_n(x)$:

- Ordino in modo crescente i dati del campione sommo via via i dati, ottenendo una curva a scalino dove ad ogni $x_{(i)}$ la funzione fa un salto di altezza 1/n:

$$S_n(x) = \begin{cases} 0 & \text{per } x < x_{(1)} \\ \frac{r}{n} & x_{(r)} \leq x \leq x_{(r+1)} \\ 1 & x_{(n)} \leq x \end{cases}$$

dove x(r) è la statistica di ordine r [$x_{(n/2)}$ è la mediana]

Test di Kolmogorov-Smirnov

- La c.d.f. F e quella empirica $S_n(x)$ dovrebbero avere gli stessi valori di aspettazione se i dati effettivamente seguono la p.d.f. f

- Posso vedere di quanto differiscono F e $S_n(x)$ e da questo stimare se effettivamente il campione di dati segue la p.d.f. f

- Nel test di Kolmogorov-Smirnov per questo confronto si usa la statistica

$$D_n \equiv Max|S_n(x) - F(x)|$$

- Moltiplicando D_n per la radice quadrata di n si ottiene : $d_n = D_n\sqrt{n}$

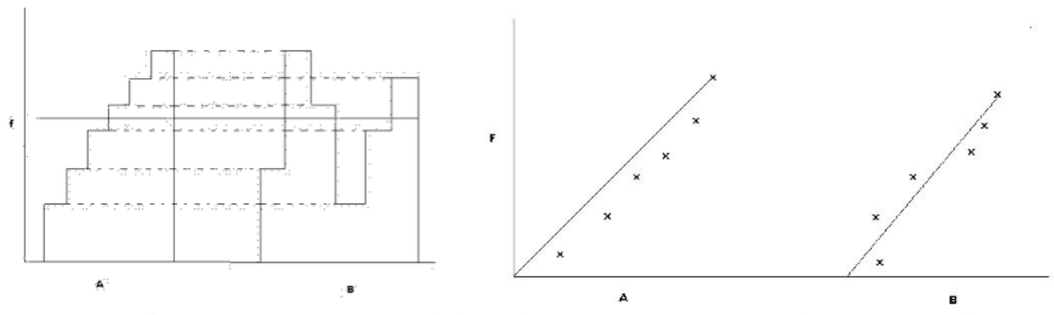
- Se l'accordo è buono, d_n dovrebbe essere piccolo. Queste funzioni sono tabulate ed i loro quantili si prendono da tavole statistiche o si calcolano.

- Questo test è molto usato quando si vuole controllare se due campioni di dati provengono dalla stessa popolazione :

$$D_{n_1-n_2} \equiv Max|S_{n_1} - S_{n_2}|$$

Test di Kolmogorov-Smirnov

- Il test di Kolmogorov-Smirnov è molto più sensibile del test del χ^2. Ci sono situazioni nelle quali il test del χ^2 può dare risultati che sono imprecisi. Il test di KS è anche un test non binnato (utilizzabile anche in piccoli campioni di dati)



- La funzione f costante potrebbe dare uno stesso buon risultato nei fit a sinistra per i due istogrammi. Questo perché nel χ^2 appaiono i quadrati delle differenze tra valore dell'istogramma e quello della funzione fittata. Questa situazione non si verifica per il test di Kolmogorov-Smirnov a destra.

- Per come è definito, il test di Kolmogorov-Smirnov è sensibile soprattutto nella parte centrale della distribuzione ma molto poco sensibile alle differenze (piccole) che si hanno nelle code