Alma Mater Studiorum · Università di Bologna

**Scuola di Scienze**
**Dipartimento di Fisica e Astronomia**
**Tesi triennale in Fisica**

# APPLICAZIONI DEL MACHINE LEARNING ALLA FISICA DELLE ALTE ENERGIE

**Relatore:**
**Prof. Alberto Cervelli**

**Co-relatore:**
**Dr. Roberto Morelli**

**Presentata da:**
**Schiazza Filippo Antonio**

Anno accademico 2019/2020

**Sommario**

da scrivere...

# Indice

# 1 Introduzione

Fin dalla prima metà degli anni '60 era chiaro ai fisici che l'idea di una materia formato esclusivamente da elettroni, protoni e neutroni era limitante e non in grado di spiegare la moltitudine di particelle che erano ormai già state osservate. Per questo motivo nel 1964 Gell-Mann e Zweig proposero la Teoria dei Quark, che è stata arricchita negli anni successivi ed è oggi nota come teoria del Modello Standard. Il Modello Standard [...] è la teoria che ha permesso l'unificazione di tre delle quattro interazioni fondamentali (forte, debole ed elettromagnetica) e, presumibilmente, ha raggiunto il suo massimo con la scoperta del *Bosone di Higgs*[ Collaboration, 2012] nel 2012; tuttavia, mancando una formulazione dell'interazione gravitazionale, non può essere considerata una teoria del tutto.
Ci sono poi alcuni problemi che non possono essere spiegati con il MS, come lo *Hierarchy Problem* [...] o la presenza della *Black Mattrn* (BM, materia oscura) [...]. Per quanto riguarda la materia oscura risulta che nessuna delle particelle fondamentali del MS è una buona candidata a farne parte, per esempio se la BM fosse costituita da particelle cariche si sarebbe dovuta rilevare una qualche radiazione elettromagnetica proveniente dalle zone di universo nelle quali si stima esserci BM, ma così non è stato.
Da queste considerazioni sembrerebbe essere ormai arrivati ad un punto di stallo per quanto riguarda il MS, tuttavia è evidente da ciò che è stato accennato precedentemente che rimangono aperte molte domande; una ipotesi è che il MS rappresenti il limite a basse energie di una teoria più complessa, quindi una serie di fenomeni o non avvengono alle attuali energie raggiungibili al *Large Hadron Collider* oppure sono estremamente rari.
Per esempio, la teoria della *Supersimmetria* (SUSY) [...], che è una estensione del MS, risolve il Problema della Gerarchia introducendo un nuovo fermione/bosone per ogni fermione/bosone del MS; inoltre tali particelle sarebbero stabili e poco interagenti e quindi costituirebbero delle ottime candidate per la spiegazione della materia oscura.
Tutte queste considerazioni inducono a pensare che esista una fisica *"beyond the Standard Model"* (BSM) e la grande sfida dei prossimi anni è quella di capire in che modo si possa indagarla.
Il run3 del Large Hadron Collider è previsto per Maggio 2021, tuttavia non vi è stato un miglioramento notevole da un punto di vista energetico. Allo stesso tempo si stima che la produzione di dati sarà fino a dieci volte maggiore rispetto al run precedente, quindi la domanda è se sia possibile trattare in maniera innovativa questa enorme mole di dati per cercare di estrarre segnale di nuova fisica.
Nello specifico la domanda è se sia possibile utilizzare delle metodologie di Machine Learning per la separazione del segnale dal background in modo da riuscire ad osservare eventuali segnali rari.
Con il termine machine learning si intende una serie di metodologie di natura statistico-computazionale che permettono di estrarre informazione utile da enormi moli di dati, altrimenti difficilmente processabili dall'uomo. I dati, per la loro stessa natura, sono disomogenei e caotici, quindi risulta particolarmente complesso analizzarli per ottenerne dei risultati. Qui entra in gioco il machine learning, ovvero l'apprendimento automatico della "macchina", perché permette di trovare relazioni nascoste fra i dati autonomamente, ovvero senza la continua supervisione dell'essere umano.
In particolare verrà affrontato un metodo di ML, il *Variational Autoencoders* (VAEs), che si basa essenzialmente su un processo di diminuzione della dimensionalità dei dati ed una successiva fase di ricostruzione; tale algoritmo viene addestrato sui dati di background in modo che sia capace di riconoscere eventuali segnali di nuova fisica come delle anomalie.

# 2 Analisi multivariata e Machine Learning

Quando si parla di analisi dati ci si può essenzialmente ricondurre a tre macro-categorie di operazioni:

1. CLASSIFICAZIONE
   Questa tipologia è probabilmente la principale quando si ha a che fare con la fisica delle alte energie e consiste nell'associare un evento/oggetto ad una categoria. Nel caso specifico di questa trattazione l'obiettivo sarà esattamente quello di classificare gli eventi/oggetti nelle due categorie di background e segnale.

2. STIMA DI PARAMETRI
   In questa tipologia ricadono tutti quei processi attraverso i quali si estraggono dei parametri (ad esempio la massa di una tipologia di particelle) attraverso un fitting del modello teorico con i dati sperimentali;

3. STIMA DI FUNZIONI
   Si ricava una funzione continua di una o più variabili a partire dai dati sperimentali.

Questa trattazione si concentrerà essenzialmente sulle possibili metodologie di classificazione, proprio perché l'obiettivo è il discernimento fra segnale e background alla ricerca di nuova fisica.
Nelle prime righe di questo capitolo è stato introdotto il termine *evento* o *oggetto*, senza meglio specificare come questo fosse collegato ai dati. Un evento può essere pensato come una collezione di dati e quindi lo si può rappresentare come un vettore in uno spazio n-dimensionale:

$$\mathbf{x} = (x_1, ..., x_n) \tag{1}$$

In realtà l'utilizzo del termine vettore è improprio ogni qual volta si abbia a che fare con componenti (i dati) disomogenee tra loro, tuttavia lo si continuerà ad utilizzare per una questione di comodità tenendo a mente questa specifica. Nelle pagine successive con i termini evento, oggetto, vettore di input e pattern ci si riferirà sempre alla stessa entità appena introdotta.
Una volta presentate queste notazioni è possibile focalizzarsi sui diversi approcci che possono essere seguiti per tentare di separare il segnale dal background quando si ha a disposizione una grande mole di dati. Nelle righe che seguono verranno presentati tre diverse metodologie, dalla più semplice ed inefficiente a quella più complessa e raffinata:

1. Sistema di tagli sulle variabili, ovvero il metodo più semplice ma anche meno efficiente. Non può essere considerato un metodo multi-variato per le ragioni che si vedranno a breve;

2. Analisi multi-variata e, nello specifico, l'analisi discriminante lineare;

3. Machine Learning, dove entra in gioco il concetto di apprendimento automatico e quindi si dispone di tecniche che permettono all'algoritmo di imparare in maniera autonoma direttamente dai dati che gli vengono forniti.

## 2.1 Sistema di tagli

Un primissimo semplice approccio prevede di adoperare dei tagli sulle varie componenti dei pattern, in modo da ricavare un ipercubo nello spazio n-dimensionale dei pattern stessi. Per capire meglio questa metodologia si consideri un caso semplificato nel quale lo spazio in questione è bi-dimensionale e quindi i vettori di input sono del tipo $\mathbf{x} = (x_1, x_2)$; per raggiungere l'obiettivo di separazione bisognerà dunque applicare due tagli, uno sulla variabile $x_1$ ed uno su $x_2$, stabilendo un punto $\mathbf{P} = (P_1, P_2)$ come riportato in figura 1.

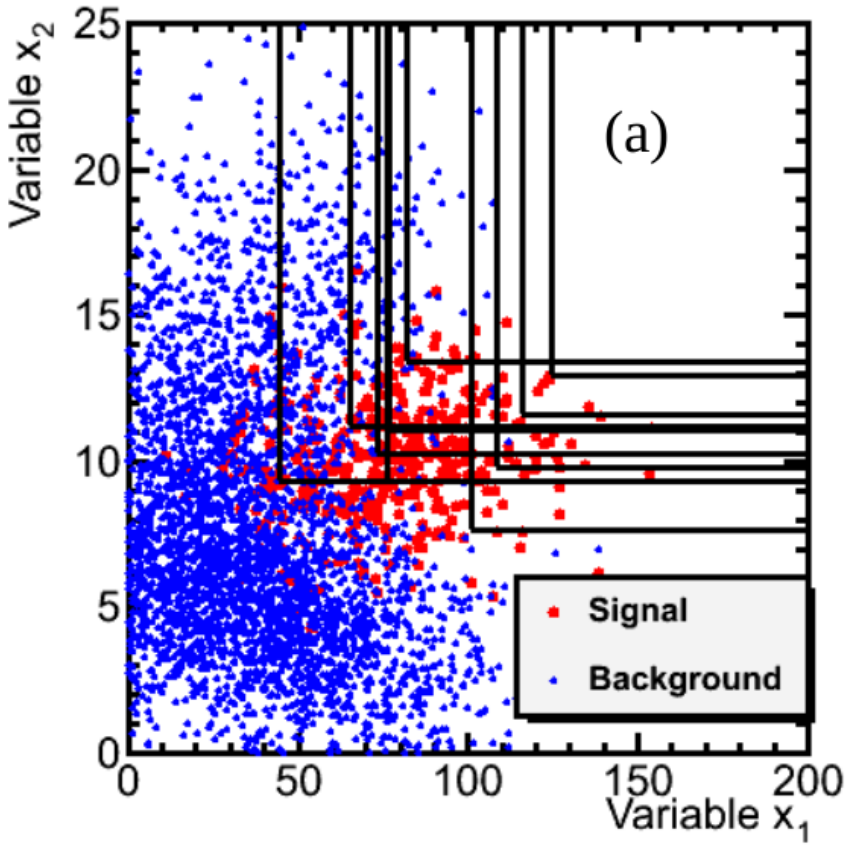


Figura 1: risultato grafico di un processo di taglio sulle due variabili $x_1$ e $x_2$ per la separazione del segnale dal background (Bhat, 2011).

Chiaramente la scelta del punto $\mathbf{P}$ non è casuale e deve essere ottimizzata; un metodo per raggiungere tale obiettivo è il Random Grid Search (RGS) che verrà presentato nella sezione 3.5.
In questo caso, nonostante si abbia a che fare con pattern multi-dimensionali, non è possibile parlare di approccio multivariato, perché i tagli sulle singole dimensioni sono indipendenti fra loro; ecco che si capisce che il prezzo da pagare per la grande semplicità di questo metodo è la sua scarsa efficienza, infatti i tagli che possono essere ottenuti sono paralleli agli assi e ciò è estremamente limitante.

## 2.2 Processi multivariati e analisi discriminante lineare

Per poter fare un passo in avanti nella trattazione bisogna considerare la possibilità che le varie componenti dei vettori evento siano tra loro correlate o, più in generale, bisogna considerare tali pattern nella loro totalità, ovvero senza focalizzarsi sulle singole componenti in maniera separata (come è stato fatto con il sistema di tagli). Per la ragione appena presentata è necessario considerare i processi multi-variati. Dato che le componenti dei pattern possono essere tra loro correlate è possibile ridurre la dimensionalità dello spazio n-dimensionale da n a d (con d < n). Verrà posto un focus particolare sul problema della dimensionalità degli input della sezione 3.8, che servirà da trampolino di lancio per affrontare un metodo particolare del ML, il Variational Autoencoders.
Come detto nella sezione precedente, i sistemi di tagli possono essere utilizzati per la separazione del segnale dal fondo ma hanno dei limiti piuttosto considerevoli che sono già stati illustrati. L'analisi discriminante è un metodo che permette di raggiungere lo stesso obiettivo di separazione, ma in modo più efficiente.
L'analisi discriminante si definisce lineare quando la funzione classificatrice è, appunto, lineare.
Si immagini di avere a disposizione un determinato set di eventi in input $\mathbf{x_i}$, ciascuno caratterizzato da un numero n di variabili (spazio n-dimensionale) e di volerli ripartire fra segnale e background.
Si definisce la funzione discriminante lineare nel seguente modo:

$$D(x_1, x_2, ..., x_n) = c_0 + c_1 x_1 + ... + c_n x_n = c_0 + \sum_{i=0}^{n} c_i x_i \tag{2}$$

quindi come una combinazione lineare delle componenti del vettore che rappresenta l'evento; il valore assunto dalla funzione per ogni singolo evento ne permette la separazione nelle due classi (nel presente caso segnale e background), utilizzando un valore di riferimento $D_0$.
A questo punto l'obiettivo è quello di massimizzare la distanza fra le due classi, ovvero rendere massima la differenza dei valori assunti dalla funzione $D(\mathbf{x})$ fra gli eventi appartenenti al background e quelli relativi al segnale.
Un esempio di questo approccio è il metodo proposto da Fisher: si consideri un campione di eventi appartenenti al segnale e se ne definisca la media $\boldsymbol{\mu_s}$ e la deviazione standard $\sigma_s$ ed un campione appartenente al background, definendo anche qui la media $\boldsymbol{\mu_f}$ e la deviazione standard $\sigma_f$. A questo punto la migliore configurazione dei parametri è quella che massimizza la seguente funzione:

$$F(\mathbf{c}) = \frac{(\boldsymbol{\mu_s} - \boldsymbol{\mu_f})^2}{\sigma_s^2 + \sigma_f^2} \tag{3}$$

Il pregio di un'analisi di questo tipo è quello di non doversi necessariamente limitare a dei tagli paralleli agli assi.

## 2.3 Machine Learning

Proseguendo nel percorso intrapreso per l'ottimizzazione nella separazione del segnale dal background, si è giunti a trattare uno degli argomenti centrali di questo lavoro, il Machine Learning (ML).
Perché è utile il ML per l'obiettivo che è stato prefissato? La risposta sta nel fato che un algoritmo di ML è in grado di apprendere in maniera semi-autonoma a partire dai dati che gli vengono presentati e quindi sarà l'algoritmo stesso a stabilire quale sia la migliore forma con la quale delimitare lo spazio n-dimensionale dei vettori di input nelle due zone relative al segnale ed al background.
Tuttavia, prima di presentare il metodo di ML che verrà utilizzato per raggiungere l'obiettivo di classificazione segnale-background, verrà svolta un'ampia panoramica sul ML e sulle metodologie più comuni.
L'approccio classico all'analisi dei dati prevede la disponibilità di un modello matematico, che dipende da una serie di parametri incogniti. Questi parametri vengono ricavati a partire dai dati sperimentali attraverso processi che possono essere sia analitici che numerici. Quando si parla di machine learning la prospettiva viene ribaltata, perché il modello matematico non è noto a priori.
Bisogna distinguere tre macro-tipologie di approccio all'analisi dati nel machine learning:

- APPRENDIMENTO SUPERVISIONATO
  In questa tipologia di apprendimento vengono presentati al computer degli input di esempio ed i relativi output desiderati, con lo scopo di apprendere una relazione generale che lega gli input con gli output; in questo caso si utilizza il così detto "training data set", mentre per testare il modello ottenuto si considera il "test data set" dove non vengono forniti al computer gli output. L'apprendimento supervisionato verrà trattato in maniera più approfondita nel prossimo paragrafo.

- APPRENDIMENTO NON SUPERVISIONATO
  In questo caso non vengono forniti al computer gli output attesi fin dalle prime fasi di apprendimento del modello e quindi lo scopo è quello di scoprire una qualche struttura fra i dati di input.Come si vedrà, questo modello viene addestrato per identificare le caratteristiche peculiari degli eventi fisici di background in modo da contrapporli successivamente a quelle relative ad eventi di segnale.
  Verrà posta attenzione su un particolare metodo di apprendimento non supervisionato, il Variational Autoencoder (VAEs), del quale verrà svolta una trattazione teorica approfondita ed una applicazione al campo della fisica delle particelle

- APPRENDIMENTO PER RINFORZO
  Il Reinforcement Learning è basato sul concetto di ricompensa, ovvero si permette all'algoritmo di esplorare un così detto ambiente e, in base all'azione compiuta, gli si fornisce un feedback positivo, negativo o indifferente. Un esempio classico prevede di voler addestrare un algoritmo per un particolare gioco: si farà in modo di fargli compiere una serie di partite in maniera iterativa e gli si assegnerà una ricompensa in caso di vittoria o un malus in caso di sconfitta.

Una ulteriore distinzione che è necessario fare è fra algoritmi di classificazione, regressione e clustering:

- CLASSIFICAZIONE
  Gli algoritmi di classificazione sono caratterizzati da un output discreto, cioè una serie di classi alle quali l'input può appartenere.Questa tipologia di meccanismo viene in genere portata avanti tramite metodi di apprendimento supervisionato. Un esempio di algoritmo di classificazione è quello che permette di distinguere se un particolare oggetto è presente o meno in un'immagine.

- REGRESSIONE
  La regressione è simile alla classificazione con la differenza che, in questo caso, l'output è continuo. Anche gli algoritmi di regressione sono adatti ad essere trattati con metodologie di apprendimento supervisionato.

- CLUSTERING
  Nel clustering l'obiettivo è sempre quello di dividere gli input in delle classi, tuttavia in questo caso tali classi non sono stabilite a priori. La natura di algoritmi di questo tipo li rende adatti ad essere trattati tramite metodi di apprendimento non supervisionato.

# 3 Machine Learning : metodi e caratteristiche

Lo schema logico che verrà seguito in questo capitolo prevede di approfondire inizialmente i due approcci principali al ML, ovvero l'apprendimento supervisionato( 3.1) e non supervisionato ( 3.3), per poi presentare due metodi di apprendimento supervisionato, ovvero le *Reti Neurali* ( 3.6) e gli *Alberi Decisionali* ( 3.7) ed un metodo di apprendimento non supervisionato, il *Variational Autoencoders* ( 3.10). Nel fare ciò verranno presentati due importanti concetti del ML, come quello di *Iperparametri* ( 3.5) ed il *Curse of dimensionality*.

## 3.1 Apprendimento supervisionato

In questa sezione viene portata avanti una descrizione più approfondita e formale dell'apprendimento supervisionato.
Come già accennato precedentemente, quando si parla di apprendimento supervisionato si hanno a disposizione sia gli input $\mathbf{x}$ che i corrispettivi target di output $\mathbf{y}$; esisterà quindi una funzione $\mathbf{y} = \mathrm{f}(\mathbf{x})$ che mette in relazione gli input con gli output. Tuttavia, come detto, tale funzione è incognita ed è quindi ciò che viene ricercato con l'algoritmo di apprendimento. Nella pratica si cerca di approssimare la funzione agendo su una serie di parametri $\boldsymbol{\theta}$, quindi si avrà un qualcosa del tipo: $\hat{\mathbf{y}} = f(\mathbf{x}, \boldsymbol{\theta})$.

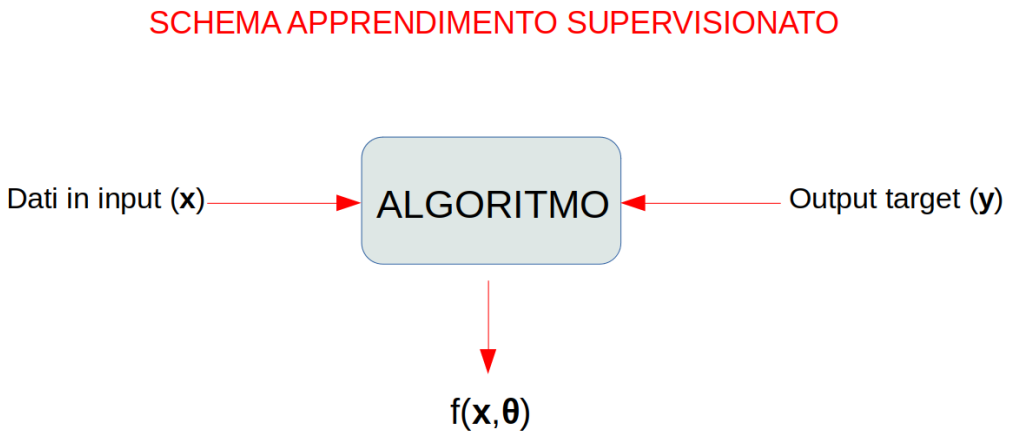


Figura 2: si riporta uno schema intuitivo del funzionamento di un algoritmo di apprendimento supervisionato

Per ogni vettore $\mathbf{x}$ del training data set è possibile definire una particolare funzione detta "Loss function" L($\mathbf{y}$,f($\mathbf{x}$,$\boldsymbol{\theta}$)); a questo punto è possibile fare una media della funzione di perdita sull'intero set di dati a disposizione, ottenendo la funzione di rischio:

$$R(\boldsymbol{\theta}) = \frac{1}{N} \sum_{k=1}^{N} L(\mathbf{y}, f(\mathbf{x}, \boldsymbol{\theta})) \tag{4}$$

dove N è il numero di eventi del training data set.
Un esempio di funzione di rischio molto diffusa è l'errore quadratico medio:

$$R(\boldsymbol{\theta}) = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{y}_k - f(\mathbf{x}_k, \boldsymbol{\theta}))^2 \tag{5}$$

Quando si addestra un modello si vuole inoltre evitare il così detto overfitting, che consiste in un eccessivo adattamento del modello ai dati di training, non raggiungendo la generalità richiesta. Un modo per verificare un eventuale overfitting è quello di verificare se il modello è nettamente migliore per il data set di allenamento rispetto al data set di test.
Per arginare questo problema è possibile modificare la funzione di rischio, aggiungendo una funzione $Q(\boldsymbol{\theta})$; in questo modo si ottiene la funzione di costo:

$$C(\boldsymbol{\theta}) = R(\boldsymbol{\theta}) + \lambda Q(\boldsymbol{\theta}) \tag{6}$$

con $\lambda$ parametro che esprime la rigidità del vincolo.
A questo punto l'obiettivo è quello di minimizzare la funzione di rischio (o di costo in caso di overfitting) e per fare ciò esistono diversi metodi, fra i quali il più comune è il metodo di discesa del gradiente.

## 3.2 Discesa del gradiente

La discesa del gradiente è una tecnica di ottimizzazione utilizzata per minimizzare l'errore che si introduce stimando la $\hat{\mathbf{y}} = f(\mathbf{x}, \boldsymbol{\theta})$ rispetto alla funzione "vera" $\mathbf{y} = f(\mathbf{x})$; quindi si avranno una Loss function $L(\mathbf{y}, f(\mathbf{x}, \boldsymbol{\theta}))$ ed un vettore dei parametri $\boldsymbol{\theta}$.
Esistono tre varianti del metodo di discesa del gradiente:

- *Batch Gradient Descent.*
  L'aggiornamento del vettore dei pesi $\boldsymbol{\theta}$ avviene solo dopo che sono stati presentati tutti i pattern all'algoritmo. Si calcola

$$\mathbf{G} = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{\nabla}_{\theta} L(\mathbf{y}_k, f(\mathbf{x}_k, \boldsymbol{\theta})) \tag{7}$$

  e con tale risultato viene aggiornato il vettore dei parametri nel seguente modo:

$$\boldsymbol{\theta} - \epsilon \mathbf{G} \rightarrow \boldsymbol{\theta} \tag{8}$$

  Qui $\epsilon$ prende il nome di *learning rate* e regola l'aggiornamento del vettore dei pesi nella direzione opposta a quella del gradiente $\mathbf{G}$.
  Quindi con questa tecnica si calcola la discesa del gradiente una sola volta, tuttavia si impiega molto tempo per arrivare ad una convergenza ed è quindi poco adatta quando si hanno grandi moli di dati a disposizione.

- *Stochastic Gradient Descent.*
  Viene calcolata la discesa del gradiente per ogni pattern fornito all'algoritmo:

$$\mathbf{G}_i = \boldsymbol{\nabla}_{\theta} L(\mathbf{y}_i, f(\mathbf{x}_i, \boldsymbol{\theta})) \tag{9}$$

  e quindi anche l'aggiornamento dei pesi avviene tante volte quanti sono i pattern iniziali:

$$\boldsymbol{\theta} - \epsilon \mathbf{G}_i \rightarrow \boldsymbol{\theta} \tag{10}$$

  Questa tecnica è, all'opposto della precedente, utile quando il numero di pattern di input è molto elevato.

- *Mini Batch Gradient Descent.*
  Si tratta di una via di mezzo fra i due metodi appena presentati perché l'aggiornamento dei pesi avviene più volte dopo che sono stati presentati dei sottogruppi dell'intero data set di addestramento.

## 3.3 Apprendimento non supervisionato

Nella sezione precedente è stata mostrata l'utilità degli algoritmi di apprendimento supervisionato, osservando che, nel caso in cui si abbiano a disposizione sia i vettori di input che i corrispettivi output target, si può ottenere un'approssimazione della relazione esistente input-output. Tuttavia non è sempre possibile avere a disposizione gli output target e bisogna capire se è comunque possibile ottenere informazioni utili dai dati.
Come già accennato nelle prime pagine di questa trattazione quando non si hanno a disposizione gli output target si possono applicare tecniche di apprendimento non supervisionato, dove l'obiettivo è quello di trovare eventuali partizioni degli input (Clustering). Si consideri la figura 3 dove sono riportate tre diverse configurazioni possibili nel caso di input bidimensionali: è evidente che nel caso a) sia possibile la separazione in due sotto gruppo e nel caso b) in un unico sotto gruppo, mentre nel caso c) sembrerebbe non si possano stabilire graficamente eventuali separazioni.

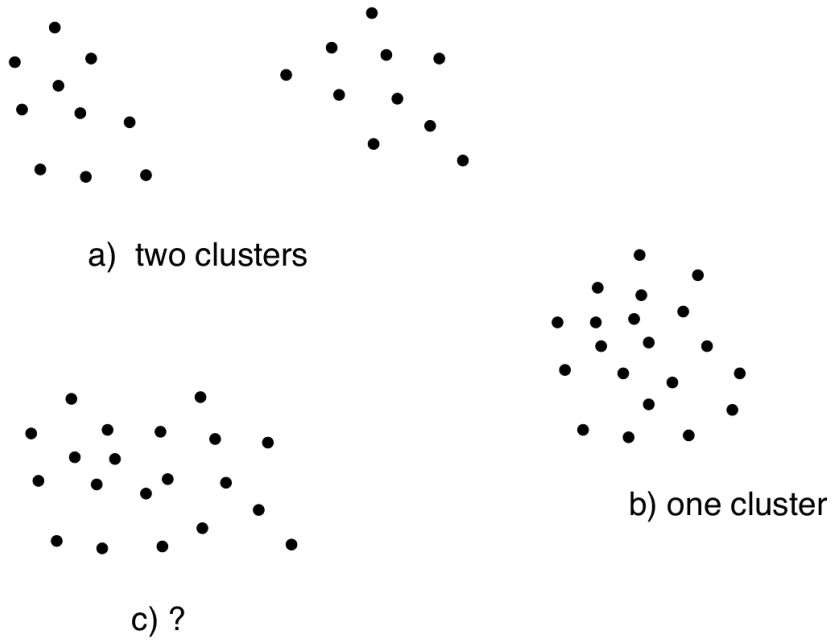


Figura 3: vettori di input in uno spazio bidimensionale in tre situazioni differenti. L'immagine è presa da Nilsson, 1998

Quindi un algoritmo di clustering si occupa della suddivisione del set di input $\Sigma$ in un numero N di sottogruppi $\Sigma_1$,...,$\Sigma_n$, detti appunto cluster; si noti che lo stesso numero N non viene stabilito a priori e fornito all'algoritmo, ma viene anch'esso ricavato a partire dai dati. Una volta fatto sarà possibile implementare un classificatore per collegare nuovi vettori di input con i cluster precedentemente individuati.
Inoltre, aumentando il livello di complessità, è possibile trovare eventuali gerarchie di partizionamento, ovvero cluster di cluster.

## 3.4 Metodo di clustering basato sulla distanza euclidea

Gli algoritmi di apprendimento non supervisionato sfruttano una qualche misura di similarità per separare i pattern (gli input) nei vari cluster. Una possibilità è quella di utilizzare la semplice distanza euclidea per poter separare lo spazio n-dimensionale dei pattern in delle sotto-aree, che sono appunto i cluster.
Per fare ciò viene implementato un metodo iterativo, basato sulla definizione di alcuni punti particolari nello spazio dei pattern, detti "cluster seekers" (letteralmente "cercatori di cluster").
Si definiscono M punti nello spazio n-dimensionale $\mathbf{C_1}, ..., \mathbf{C_M}$ e l'obiettivo è quello di fare in modo che ogni punto si muova verso il centro di ogni singolo cluster, in modo che ogni cluster abbia al suo centro uno di questi cluster seekers.
Come è già stato spiegato precedentemente, l'algoritmo non conosce a prescindere il numero di cluster ma riesce a ricavarlo dai pattern stessi; per questa ragione il numero di cluster seekers M è inizialmente casuale ed esiste un procedura per ottimizzarlo, che verrà illustrata in seguito.
I pattern del training data set $\Sigma$ vengono presentati all'algoritmo uno alla volta: per ognuno di essi ($\mathbf{x_i}$) si cerca il cluster seekers più vicino ($\mathbf{C_k}$) e lo si sposta verso $\mathbf{x_i}$ nel seguente modo:

$$\mathbf{C_k} + \alpha_k(\mathbf{x_i} - \mathbf{C_k}) \rightarrow \mathbf{C_k} \tag{11}$$

dove $\alpha_k$ è un parametro di apprendimento che determina di quanto il cluster seeker k-esimo si muove verso il punto $\mathbf{x_i}$.
A questo punto è utile fare in modo che più il cluster seeker è soggetto a spostamenti minore diventa l'entità dello spostamento. Per fare ciò si definisce una massa $m_k$ e le si assegna un valore pari al numero di volte in cui $\mathbf{C_k}$ è stato soggetto a spostamenti (quindi anche il valore della massa verrà aggiornato di volta in volta); dopodiché si assegna ad $\alpha_k$ il seguente valore

$$\alpha_k = \frac{1}{1 + m_k} \tag{12}$$

e, dato che ad ogni iterazione che coinvolge $\mathbf{C_k}$ il valore di $m_k$ aumenta di una unità, il parametro di apprendimento $\alpha_k$ diminuisce di volta in volta.
Il risultato di questo aggiustamento è che il cluster seeker si trova sempre nel punto che rappresenta la media dei punti del cluster.
Una volta che sono stati presentati tutti i pattern del training data set all'algoritmo, i vari cluster seeker andranno a convergere ai "centri di massa" dei cluster e la classificazione (cioè la delimitazione dei cluster nello spazio n-dimensionale) può essere fatta con una partizione dello spazio di Voronoi, di cui si riporta la seguente definizione:

> *In ogni insieme (topologicamente) discreto S di punti in uno spazio euclideo e per quasi ogni punto x, c'è un punto in S che è il più vicino a x. Il "quasi" è una precisazione necessaria dato che alcuni punti x possono essere equidistanti da 2 o più punti di S. Se S contiene solo due punti, a e b, allora il luogo geometrico dei punti equidistanti da a e b è un iperpiano, ovvero un sottospazio affine di codimensione 1. Tale iperpiano sarà il confine tra l'insieme di tutti punti più vicini ad a che a b e l'insieme di tutti i punti più vicini a b che ad a. È l'asse del segmento ab. In generale, l'insieme dei punti più vicini a un punto c ∈ S che ad ogni altro punto di S è la parte interna di un politopo (eventualmente privo di bordi) detto dominio di Dirichlet o cella di Voronoi di c. L'insieme di tali politopi è una tassellatura dell'intero spazio e viene detta tassellatura di Voronoi corrispondente all'insieme*

*S. Se la dimensione dello spazio è solo 2, è facile rappresentare graficamente le tassellazioni di Voronoi; è a questo caso che si riferisce solitamente l'accezione diagramma di Voronoi.*
*(Wikipedia, Diagramma di Voronoi.)*

Un esempio didattico del risultato di questa partizione è riportato in figura 4.

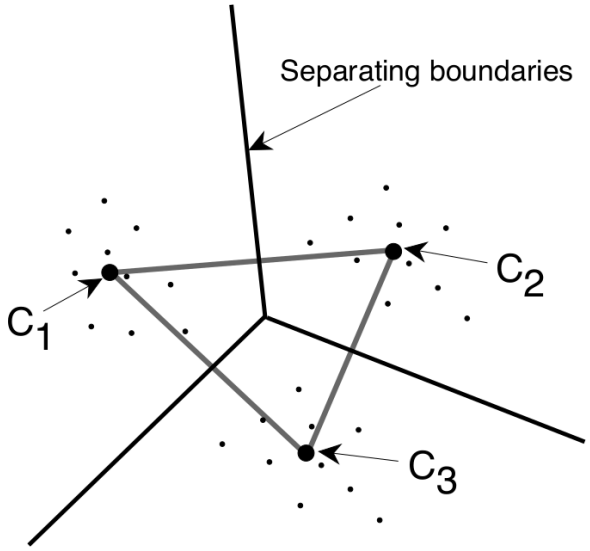


Figura 4: Si riporta un esempio di partizione dello spazio (bi-dimensionale) di Voronoi in tre sotto regioni . L'immagine è presa da Nilsson, 1998

Si è accennato qualche riga fa che il numero di cluster seeker è inizialmente scelto in maniera casuale, per poi essere ottimizzato; per il processo di ottimizzazione si utilizza la varianza dei pattern $\{\mathbf{x_i}\}$ per ogni cluster:

$$\sigma^2 = \frac{1}{L}\sum_{i=1}^{L}(\mathbf{x_i} - \boldsymbol{\mu})^2 \tag{13}$$

dove L è il numero di pattern nel cluster e $\boldsymbol{\mu}$ ne è la media:

$$\boldsymbol{\mu} = \frac{1}{L}\sum_{i=1}^{L}\mathbf{x_i} \tag{14}$$

A questo punto, se la distanza $d_{ij}$ fra due cluster seeker $\mathbf{C_i}$ e $\mathbf{C_j}$ è minore di un determinato valore $\epsilon$, allora si sostituiscono i due cluster seeker con uno nuovo posto nel loro centro di massa (tenendo conto delle due masse $m_i$ e $m_j$); dall'altro lato, se vi è un cluster per il quale la varianza $\sigma^2$ è più grande di un valore $\delta$, si aggiunge un nuovo cluster seeker vicino a quello già esistente e si eguagliano entrambe le loro masse a zero.
Come osservazione finale bisogna dire che nei metodi che si basano sul concetto di distanza è importante ri-scalare i valori delle componenti dei pattern (in linea di principio si possono avere componenti diverse con ordini di grandezza di molto differenti) in modo da evitare che alcune componenti pesino più di altre.

## 3.5 Iperparametri e Grid Search

Prima di parlare del Grid Search è necessario introdurre il concetto di **iperparametro**. Come detto nelle sezioni precedenti, un modello di apprendimento è caratterizzato da una serie di parametri che vengono modificati in maniera iterativa in modo da minimizzare la Loss function e, come noto, tale processo avviene attraverso un continuo confronto con il training data set. Quando si parla di iperparametri si intende invece una serie di parametri che caratterizzano il modello implementato che non sono modificati nel processo di addestramento con il training data set ma vengono prestabiliti dall'utente.
Chiaramente al variare degli iperparametri cambia anche la qualità del processo di apprendimento del modello e quindi anch'essi devono essere sottoposti ad un processo di ottimizzazione. A questo punto entra in gioco il metodo del Grid Search che è appunto un metodo di ottimizzazione degli iperparametri.
Il Grid Search è piuttosto semplice sia da comprendere concettualmente sia da implementare nella pratica; fa parte dei così detti "Brute-Force Search", cioè di quei metodi che si basano sulla sistematica verifica di tutte le possibili soluzioni ad un problema per poi considerare la migliore. Per esempio si consideri il problema di dover cercare i divisori di un numero n: un approccio "Brute-Force" prevedrebbe di considerare tutti i numeri minori di n e verificare quelli per i quali la divisione non dà resto. Questo esempio permette anche di mettere in evidenza il limite principale di tale tipologia di approccio: il numero di possibilità da esplorare può aumentare molto velocemente, soprattutto se si considera un processo multivariato.
Tornando ora nello specifico al Grid Search, si consideri un modello caratterizzato da un numero k di iperparametri. Si può definire, in analogia a ciò che è stato fatto con i parametri, un vettore le cui componenti sono appunto gli iperparametri:

$$\boldsymbol{\mu} = (\mu_1, ..., \mu_k) \tag{15}$$

Tale vettore apparterrà ovviamente ad uno spazio k-dimensionale, sul quale può essere costruita una griglia i cui nodi corrispondono a particolari combinazioni degli iperparametri.
A questo punto si può avviare l'apprendimento del modello per ogni particolare configurazione degli iperparametri ed ottenere un valore per la Loss function. Si arriva allora ad avere una valore della Loss per ogni nodo della griglia e quindi basta considerare quello per il quale la Loss è minore, ottenendo la miglior configurazione degli iperparametri.
In Figura 5 nella pagina seguente è riportato per chiarezza un esempio visivo dell'esito di un processo di ottimizzazione degli iperparametri attraverso il metodo Grid Search.

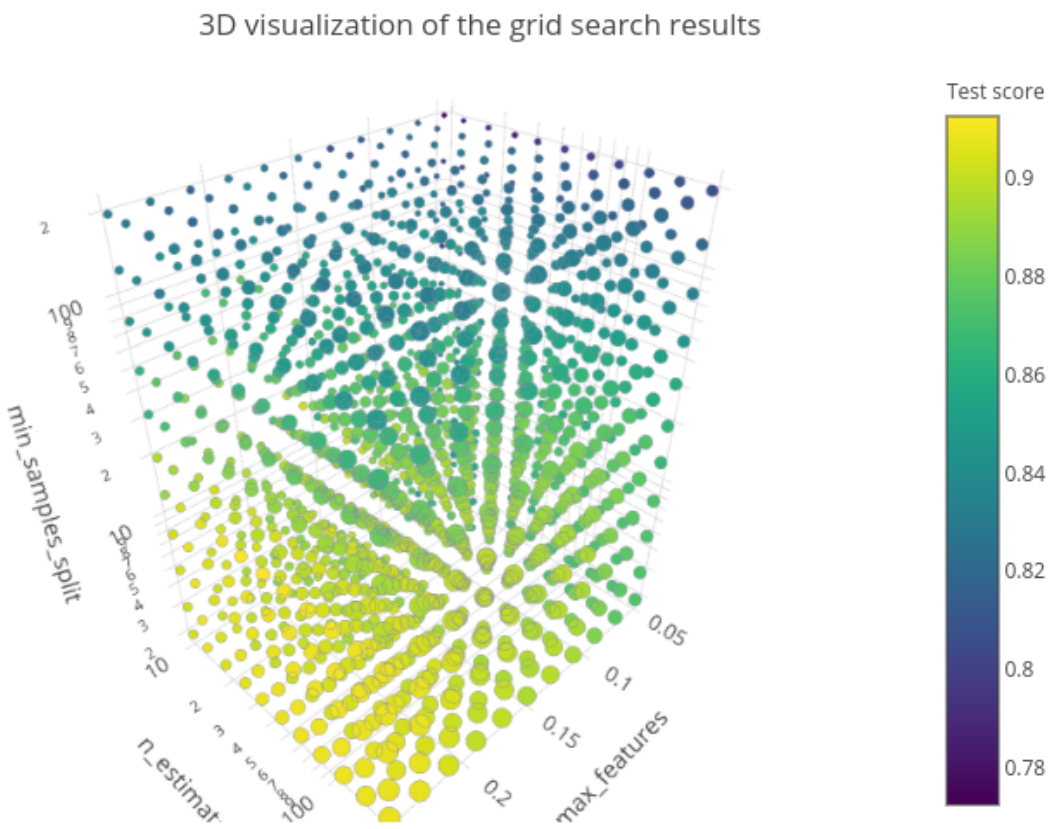


Figura 5: la figura illustra visivamente l'esito di un processo di ottimizzazione degli iperparametri attraverso il metodo Grid Search ( *Knuth: Computers and Typesetting*)

Come accennato precedentemente, man mano che aumenta la complessità del modello è molto probabile che aumenti il numero degli iperparametri e quindi la dimensionalità dello spazio introdotto precedentemente; ciò implica l'aumento considerevole del numero di configurazioni degli iperparametri da esplorare attraverso il Grid Search e quindi il tempo necessario per concludere l'ottimizzazione.
E' possibile ovviare parzialmente a questo problema attraverso il Random Grid Search (RGS), dove non sono considerati tutti i nodi della griglia, ma solo una loro parte selezionata in maniera casuale secondo una particolare distribuzione (ciò permette anche di tener conto di conoscenze pregresse).
Un ulteriore accorgimento può essere quello di arrestare le configurazioni meno promettenti prima di portarle a termine, risparmiando tempo e risorse computazionali. Per far ciò, basta impelmentare degli *scheduler* che tengano conto dell'andamento dei diversi training relativi alle diverse configurazioni. Un altra possibilità riguarda l'esplorazione di nuove varianti a partire dalle configurazioni iniziali. In quest'ultimo caso non è nemmeno necessario definire in modo rigido lo spazio degli iperparamentri da esplorare dato che le nuove configurazioni vengono ricercate a partire dall'andamento delle precedenti (*popultion based training*).

## 3.6 Reti Neurali

Le reti neurali sono probabilmente il metodo di apprendimento supervisionato più conosciuto ed utilizzato nel campo dell'analisi dati.
La struttura di una rete neurale prevede la presenza di unità fondamentali, dette neuroni, che sono organizzate in strati e legate fra di loro mediante delle connessioni (sinapsi), ciascuna delle quali è caratterizzata da un peso. Sono proprio questi pesi a giocare un ruolo fondamentale nel processo di apprendimento della rete perché sono loro i parametri soggetti a modifica.
Il nome rete neurale (artificiale) deriva dal fatto che la loro struttura è inspirata dalle corrispondenti strutture biologiche (seppur di molto semplificata).
In una rete neurale è sempre presente uno strato di input ed uno di output, mentre il numero di livelli nascosti può variare a seconda della complessità della rete; In figura 6 è riportato un esempio di rete neurale con un singolo strato interno nascosto.

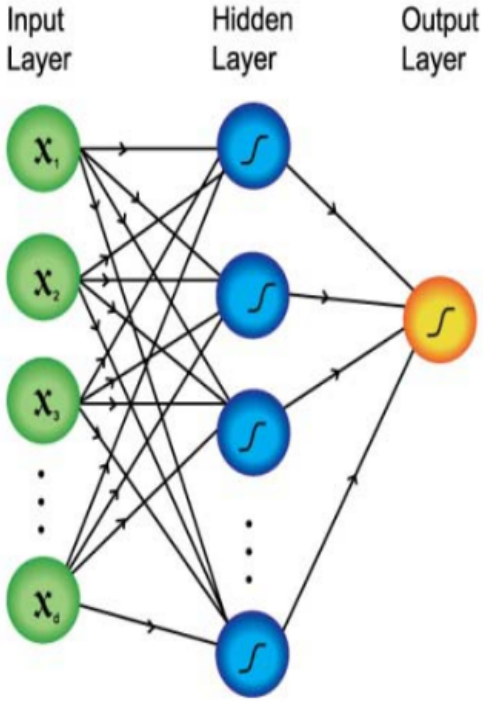


Figura 6: si riporta un esempio grafico di rete neurale formata da un unico strato nascosto. L'immagine è presa da Bhat, 2011.

Si può passare ora a presentare il modello del singolo neurone per capire com'è strutturato e quale compito svolge. Gli elementi che caratterizzano il singolo neurone sono:

1. Una serie di connessioni in ingresso (ciascuna caratterizzata da un proprio peso);
2. Un sommatore che ha il compito di svolgere la somma pesata degli input, utilizzando i pesi caratteristici delle connessioni;
3. Un output e la relativa funzione di attivazione, che viene usata per limitarne l'ampiezza (tipicamente ad intervalli [0,1] o [-1,1]);

4. Un valore di soglia che viene usato per aumentare o diminuire il valore ottenuto dalla somma pesata.

Si riporta in figura 7 lo schema grafico di un singolo neurone (k), dove $\mathbf{x} = (x_1, ..., x_m)$ è il vettore degli input input, $\mathbf{w}_k = (w_{k1}, ..., w_{km})$ è il vettore dei pesi, $\phi$(x) è la funzione di attivazione, $b_k$ è il valore di soglia e $y_k$ è l'output.

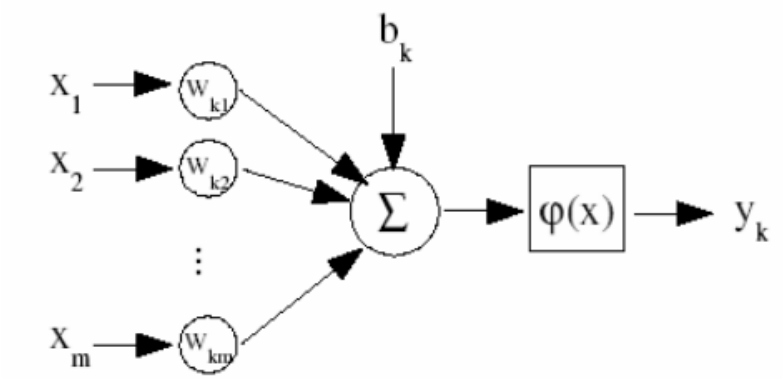


Figura 7: Illustrazione della struttura di un neurone ( *Appunti di reti neurali*).

Quindi il neurone opera la seguente somma pesata:

$$s_k = \mathbf{x} \bullet \mathbf{w_k} = \sum_{i=1}^{m} x_i w_{ki} \tag{16}$$

e si ottiene l'output attraverso la funzione di attivazione:

$$y_k = \phi(s_k + b_k) \tag{17}$$

Risulta utile spendere qualche parola in più sul tipo di funzione di attivazione più utilizzata, ovvero la funzione sigmoide:

$$sig(x) = \frac{1}{1 + e^{-\alpha x}} \tag{18}$$

dove $\alpha$ è un parametro che permette di regolare la pendenza della curva, come si evince dalla figura 8

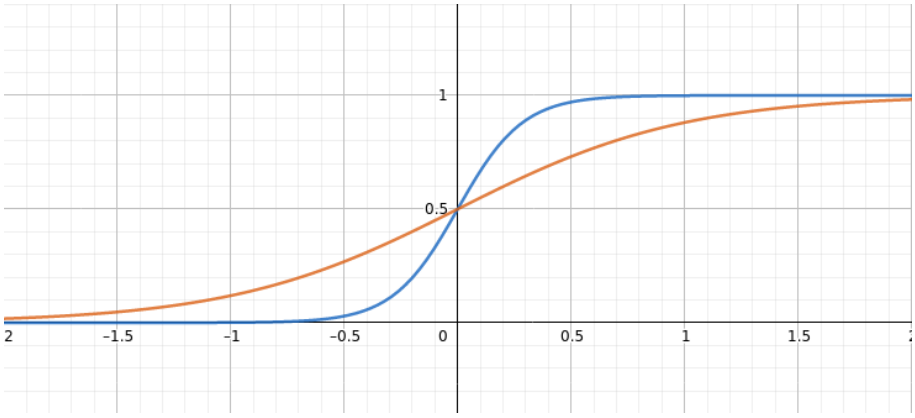


Figura 8: Si riportano due sigmoidi, dove per quella in rosso si ha $\alpha$=2 e per quella in blu $\alpha$=7.

Un singolo neurone è l'ingrediente fondamentale di una rete neurale e nella gran parte dei casi non è in grado di svolgere da solo nessun compito interessante ma deve sempre essere inserito in una rete. Tuttavia esiste una caso particolare nel quale un singolo neurone può portare a termine un compito di classificazione; affinché ciò sia possibile è necessario che i vettori evento siano riconducibili a due sole categorie e che il loro spazio possa essere separato (in relazione alle due categorie) da un singolo iper-piano. In questo caso esiste un teorema di convergenza che garantisce appunto la convergenza dei pesi nel processo di addestramento.

A questo punto è possibile passare ad un livello di complessità superiore, osservando in che modo possono essere organizzati i neuroni per formare la rete neurale; si distinguono due tipologie di reti:

1. Reti feedforward con uno o più strati: in questo caso il segnale si propaga dai nodi di input verso quelli di output, senza connessioni fra i neuroni di uno stesso strato;
2. Reti feedback: sono reti cicliche dove il segnale si propaga anche fra i neuroni di uno stesso strato.

Bisogna chiedersi ora in che modo apprende il singolo neurone. Una delle possibilità (nel caso in cui sia noto l'output target) è l'apprendimento con correzione di errore, che viene presentato velocemente nel seguito. Si consideri un singolo neurone che ha in ingresso una serie di input $(x_1, ..., x_n)$ e quindi produce un valore di output $y$ attraverso la somma pesata già introdotta precedentemente; tale valore può quindi essere confrontata con il risultato atteso $R$, ottenendo così un errore $err = R - y$; si può quindi definire la funzione di costo

$$E = \frac{1}{2} err^2 \tag{19}$$

sulla quale si applicherà il metodo di discesa del gradiente già discusso nel paragrafo 2.1 per ottimizzare i parametri.
Una volta capito in che modo avviene l'addestramento di un singolo neurone, è possibile trattare le diverse funzioni che può svolgere una rete neurale:

1. L'associazione, a sua volta divisibile in autoassociazione ed eteroassociazione. Nel primo caso vengono presentati alla rete neurale una serie di vettori evento nella

fase di training per poi verificare se uno di questi vettori, ripresentato parzialmente, viene nuovamente riconosciuto dalla rete e completato (tale funzione ben si presta ad essere ottenuta a seguito di un processo di apprendimento non supervisionato); nel secondo caso si utilizza un vettore non ancora noto alla rete neurale come richiamo di una già processato;

2. Il riconoscimento consiste nell'associazione da parte della rete di un vettore evento ad una delle varie categorie possibili. Tale obiettivo può essere ottenuto a seguito di una fase di addestramento dove vengono forniti alla rete sia i vettori in input che le categorie alle quali questi appartengono (si tratta chiaramente di un processo di apprendimento supervisionato). Si ipotizzi di avere a disposizione dei vettori evento con un numero n di componenti (i dati) e, chiaramente, possono essere pensati come dei punti in uno spazio n-dimensionale; questo spazio potrà essere allora diviso in delle regioni che corrispondono alle varie categorie di cui si è parlato precedentemente ed i confini di queste zone si ottengono a seguito del processo di addestramento;

3. L'approssimazione di funzioni, dove si hanno a disposizione gli input $\mathbf{x_i}$ ed i corrispettivi output $\mathbf{y_i}$. Quello che si cerca di fare è approssimare al meglio la funzione $\mathbf{y} = f(\mathbf{x})$ vera con una $g(\mathbf{x})$, tale per cui la distanza euclidea è inferiore ad un valore prefissato positivo (piccolo) $\epsilon$:

$$\|g(\mathbf{x}) - f(\mathbf{x})\| < \epsilon \tag{20}$$

Arrivati a questo punto è possibile esporre la trattazione su come una rete neurale viene addestrata. Precedentemente si è introdotta la struttura di una rete neurale, specificando le differenze fra lo strato di input, quello di output e gli strati nascosti. Ogni singolo neurone nel suo processo di addestramento deve aggiornare i suoi pesi, in modo che l'output della rete neurale sia simile a quello atteso.
Uno dei metodi migliori per addestrare la rete neurale è l'algoritmo di back-propagation. Una rete neurale è caratterizzata da due tipologie di segnale: da un lato vi è un segnale di funzione che si propaga dallo strato di input verso quello di output e, dall'altro, vi è un segnale di errore che ha origine nello strato di output e si propaga verso quello di input. E' il segnale di errore a giocare un ruolo fondamentale nel processo di apprendimento tramite ottimizzazione dei pesi che caratterizzano la rete neurale.
Addentrandosi nell'algoritmo di back-propagation bisogna fare una distinzione fra il modo in cui esso viene applicato allo strato di output ed il modo in cui viene applicato agli strati nascosti:

1. Neurone nello strato di output.
   Si consideri uno strato di output con un numero n di neuroni e ci si focalizzi sul k-esimo. In un certo momento del processo di apprendimento, alla rete neurale si starà presentando il j-esimo elemento del training data set, quindi per il neurone k si otterrà il seguente segnale di errore:

$$err_k^{(j)} = R_k^{(j)} - y_k^{(j)} \tag{21}$$

   dove con la lettera y si intende il valore ottenuto in output dal neurone e con R il valore atteso.
   L'errore totale dello strato di output per il vettore evento j-esimo viene definito nel

seguente modo:

$$E^{(j)} = \frac{1}{2}\sum_{k=1}^{n}(err_k^{(j)})^2 \tag{22}$$

Se poi N è il numero totale di elementi del training data set, allora la funzione di costo può essere definita nel seguente modo:

$$E_{tot} = \frac{1}{N}\sum_{j=1}^{N} E^{(j)} \tag{23}$$

e l'obiettivo è quello di minimizzare tale funzione di costo. Per fare ciò si procede aggiustando i pesi a seguito della presentazione di ogni singolo vettore evento. Si utilizza il metodo di discesa del gradiente, procedendo nel seguente modo:
il gradiente è dato da

$$\frac{\partial E^{(j)}}{\partial w_{ki}^{(j)}} \tag{24}$$

e gli aggiornamenti del peso vengono applicati nel verso opposto del gradiente, ovvero

$$\Delta w_{ki}^{(j)} = -\mu\frac{\partial E^{(j)}}{\partial w_{ki}^{(j)}} \tag{25}$$

con $\mu$ fattore di apprendimento, definito nella sezione precedente come *learning rate*. Manca a questo punto il calcolo esplicito del gradiente, che può essere eseguito con la regola della catena

$$\frac{\partial E^{(j)}}{\partial w_{ki}^{(j)}} = \frac{\partial E^{(j)}}{\partial err_k^{(j)}}\frac{\partial err_k^{(j)}}{\partial y_k^{(j)}}\frac{\partial y_k^{(j)}}{\partial S_k^{(j)}}\frac{\partial S_k^{(j)}}{\partial w_{ki}^{(j)}} \tag{26}$$

dove $S_k^{(j)} = s_k^{(j)} + b_k^{(j)}$ (si faccia riferimento all'equazione (16)).
Una volta calcolate le quattro derivate si ottiene:

$$\frac{\partial E^{(j)}}{\partial w_{ki}^{(j)}} = -err_k^{(j)}\phi'(S_k^{(j)})y_i^{(j)} \tag{27}$$

e quindi:

$$\Delta w_{ki}^{(j)} = err_k^{(j)}\phi'(S_k^{(j)})y_i^{(j)}\mu \tag{28}$$

2. Neurone in uno strato nascosto
   In questo caso l'output del neurone non ha un diretto valore con il quale può essere confrontato, quindi il segnale di errore deve essere determinato a partire dai segnali di errore di tutti i neuroni dello strato successivo, da cui il nome di back-propagation proprio perché il segnale di errore prosegue all'indietro dall'output verso l'input.

Come ultima considerazione sulle reti neurali bisogna sottolineare che il coefficiente di apprendimento deve essere scelto in maniera accurata, infatti se fosse troppo piccolo si avrebbe una convergenza estremamente lenta e, viceversa, un valore troppo grande porterebbe ad una instabilità con comportamento oscillatorio. Per gestire meglio questo aspetto, nella pratica viene definito un learning rate variabile. Generalmente si fa in modo che questo fattore parta da un certo valore per decrescere mano mano che si ci avvicina ad una soluzione ottimale. In prima istanza infatti è utile avere un valore abbastanza grande da poter seguire in modo efficace la discesa del gradiente, rendendo più rapida la ricerca della soluzione ottimale. Successivamente, però, risulta utile diminuire questo stesso parametro per evitare oscillazioni attorno al punto di minimo.

## 3.7 Alberi Decisionali

Gli alberi decisionali sono, al pari delle reti neurali, un metodo ML di apprendimento supervisionato e la loro caratteristica fondamentale è il presentarsi in maniera particolarmente intuitiva perché è possibile avere una semplice rappresentazione grafica del meccanismo del loro funzionamento.
Gli alberi decisionali rappresentano un mezzo estremamente interessante per le operazioni di classificazione (sia per output continui che discreti) ed operano attraverso una serie di test sugli attributi degli input (con il termine attributo o *features* si intende una componente del vettore di input, come già specificato nella sezione introduttiva di questa tesi [ 2] ).
Come primo passo è utile discutere come è strutturato un albero decisionale, introducendo alcune notazioni (come ausilio alla trattazione si riporta in figura 9 un esempio di albero decisionale molto semplice).

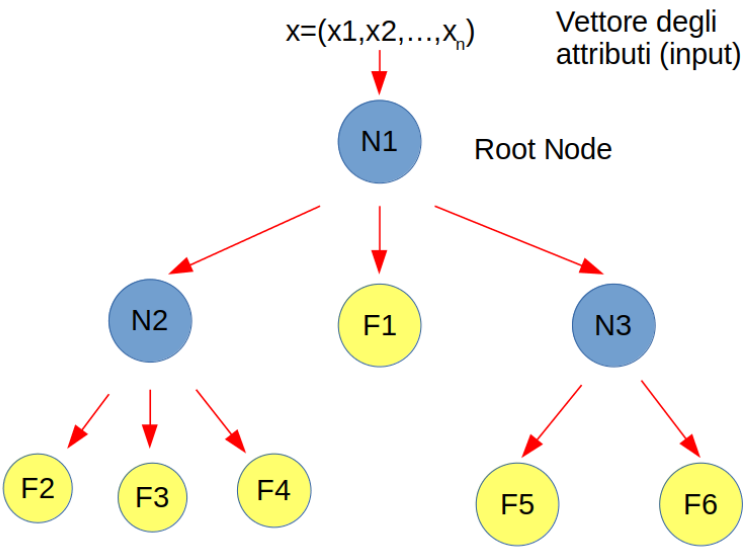


Figura 9: esempio di come è strutturato un albero decisionale

Gli elementi che caratterizzano un albero decisionale sono:

- Nodo.
  I nodi sono riportati in figura 9 come dei cerchi colorati in azzurro e contrassegnati dalla lettera N. Ogni nodo si occupa di eseguire un test su di un singolo attributo (il nodo iniziale dove avviene il primo test e quindi la prima differenziazione degli input è detto ”root node”);

- Ramo.
  I rami (detti anche archi) determinano le regole di ”splitting”, ovvero le regole attraverso le quali vengono separati gli esempi nelle rispettive categorie a seconda dei loro attributi; tali rami determinano quindi i percorsi all'interno degli alberi decisionali e quindi, in ultima analisi, la classificazione finale;

- Foglie.
  Le foglie sono una classe particolare di nodi, ovvero i nodi finali che, in quanto tali,

non generano nuove diramazioni ma rappresentano il risultato finale del processo di classificazione.

Bisogna tenere a mente che si sta parlando di una metodologia di ML, che è quindi volta ad estrarre dai dati di input (training data set) e dai corrispettivi output target l'informazione generale, per poter poi applicare l'algoritmo a casi per i quali non si è in possesso degli output di riferimento.
Detto ciò è necessario capire in che modo possa essere costruito un albero decisionale e l'idea di base è quella di stabilire, di volta in volta, il criterio sugli attributi che più discrimina gli input.
Nella costruzione degli alberi decisionali bisogna distinguere due fasi successive:

- "Building", ovvero costruzione.
  In questo prima fase l'obiettivo è quello di far crescere l'albero in dimensione, quindi in termini di rami e nodi per avere un numero adeguato di regole di splitting così da ottenere una classificazione in classi omogenee nella fase finale. In questa prima fase si ottiene un albero particolarmente folto e quindi probabilmente soggetto all'overfitting (come primo argine a ciò è possibile introdurre un criterio d'arresto capace di fermare la crescita dell'albero al realizzarsi di particolari condizioni);

- "Pruning", ovvero potatura.
  Questa fase è quella che permette di ridurre l'overfitting perché vengono eliminati i rami che non contribuiscono in maniera significativa al processo di classificazione.

Come già detto la prima fase è quella di "Building", durante la quale viene costruito l'albero decisionale aggiungendo nodi e rami, con il fine di ottenere una classificazione finale il più possibile omogenea; è altresì noto che ad ogni nodo corrisponde un attributo sul quale è effettuato un test, quindi è evidente che, dato un particolare numero di attributi, il numero di alberi possibili è molto elevato. Bisogna trovare un modo per disporre nella maniera più efficace i nodi all'interno dell'albero: l'idea è quella di scegliere per primo l'attributo attraverso il quale si ha una maggiore discriminazione dei dati in input. Per fare ciò si possono percorrere due strade distinte, utilizzando:

- Coefficiente di impurità di Gini.

- Guadagno informativo.

E' chiaro che entrambi questi indici vengono utilizzati con la stessa finalità, tuttavia ogni algoritmo ha una differente logica di costruzione e quindi adotterà uno solo dei due indici, con la possibilità di ottenere risultati differenti.
A questo punto l'albero decisionale è stato costruito e quindi si deve passare alla fase di "pruning", con l'obiettivo di ridurre le dimensioni dell'albero per evitare l'ormai noto overfitting. Per fare ciò le strade sono due, infatti da un lato si può seguire un approccio "top-down", partendo dalla radice e suddividendo l'intera struttura in sotto alberi e dall'altro un approccio "bottom-up", partendo dalle foglie ed analizzando l'impatto di ogni singola potatura; è altresì possibile introdurre nella fase di costruzione stessa un criterio di *early-stopping* o *pre-pruning* richiedendo un valore minimo di miglioramento dell'algoritmo fra un'iterazione e l'altra: ad ogni passaggio di separazione dell'albero decisionale viene

fatto un controllo sulla Loss function e, se tale errore non diminuisce significativamente tra un passaggio e l'altro, si interrompe il processo di costruzione dell'albero. Il problema dell'*early stopping* è che potrebbe portare ad una classificazione non ottimale, infatti non è detto che nell passaggio successivo di separazione non ci sarebbe potuta essere una riduzione significativa dell'errore; per questa ragione in genere si utilizzano entrambi i metodi di *pruning* e di *early stopping* parallelamente, per poi confrontare i risultati.

In conclusione bisogna sottolineare che gli alberi decisionali sono particolarmente utilizzati per i seguenti motivi:

- semplicità nell'interpretazione e nella visualizzazione;
- tolleranza ad eventuali attributi mancanti per alcuni input nel training data set o nel test data set;
- insensibilità ad eventuali attributi irrilevanti nella classificazione;
- invarianza per trasformazioni monotone effettuate sugli attributi, che rende la fase di pre- processamento dei dati non necessaria.

Gli alberi decisionali hanno tuttavia una serie di limiti elencati nelle righe che seguono:

- instabilità rispetto al variare del training data set, cioè data set di allenamento di poco differenti fra loro producono risultati molto diversi;
- frequente problema dell'overfitting.

Tuttavia è possibile combinare vari alberi decisionali insieme per ottenere migliori prestazioni predittive rispetto all'utilizzo di un solo albero; infatti, come è stato appena illustrato, gli alberi decisionali singoli hanno alcuni problemi non di poco conto, che possono però essere parzialmente arginati considerando più alberi e prendendo come decisione finale una media delle decisioni di ciascun albero.
Arrivati a questo punto ci si chiede in che modo possano lavorare insieme vari alberi decisionali per ottenere una decisione finale migliore rispetto a quella del singolo albero; una nota tecnica prende il nome di *Bagging*, nella quale si generano in maniera casuale dei sottogruppi del training data set ed ognuno di questi viene utilizzato per l'addestramento di un albero decisionale. Il risultato sarà una collezione di alberi decisionali e, come decisione finale, si utilizzerà la media delle decisioni dei singoli alberi.
Esiste un'estensione di questo metodo conosciuta con il nome di *Random Forest*; in questo caso viene aggiunto un passaggio ulteriore al processo appena illustrato perché viene scelto casualmente anche un sottogruppo degli attributi dei pattern, ottenendo così un metodo capace di agire anche su data set ad alta dimensionalità e che permette di ridurre sensibilmente il problema dell'overfitting.

## 3.8 Curse of dimensionality e riduzione della dimensionalità

A questo punto si è giunti finalmente al cuore di questa trattazione, dove vengono illustrate le basi teoriche di un metodo di apprendimento non supervisionato, il Variational Autoencoders (VAEs), del quale si studierà nel prossimo capitolo un'applicazione al campo della fisica delle alte energie.
Gli argomenti trattati nelle prossime pagine per presentare le basi teoriche del VAEs seguono la seguente struttura logica:

- Presentazione del problema della dimensionalità;
- Una delle possibili soluzioni: Autoencoders;
- Evoluzione dell'autoencoders: il Variational Autoencoders.

Come è stato più volte detto in questa trattazione, quando si parla di input (o pattern) ci si riferisce a dei vettori, le cui componenti sono i dati veri e propri; questi vettori, in quanto tali, possono essere pensati all'interno di un opportuno spazio n-dimensionale (con n=numero di componenti del vettore).
Quando si parla di "Curse of dimensionality" (letteralmente "la maledizione della dimensionalità") ci si riferisce ad una serie di problemi che ci si trova ad affrontare quando bisogna trattare spazi con un'alta dimensionalità, che altrimenti non comparirebbero in spazi a bassa dimensionalità.
Dato che all'aumentare della dimensionalità i volumi nello spazio aumentano in maniera significativa, ci si troverà nella situazione per cui i pattern risultano sparsi nello spazio e questo è chiaramente un problema per ogni analisi che ne si vuole fare basata sulla statistica; infatti, per ottenere dei risultati significativi a livello statistico, la quantità di dati necessari aumenta in maniera esponenziale e questo risulta essere un problema a livello pratico.
Quando si parla di riduzione della dimensionalità ci si riferisce ad una serie di tecniche, attraverso le quali viene ridotto il numero delle variabili che caratterizzano i vettori di input; l'obiettivo di base è quello di proiettare gli elementi dello spazio n-dimensionale (i vettori di input) su di uno spazio a dimensione inferiore, cogliendo l'essenza stessa dei dati.
Ciò che è necessario notare è che avere degli input con più bassa dimensionalità permette di avere anche meno parametri (gradi di libertà) e quindi una struttura più semplice del modello. Il prediligere la semplicità alla complessità, oltre che per gli ovvi motivi, deriva dal fatto che la seconda è molto soggetta al fenomeno dell'overfitting.
Il processo di riduzione della dimensionalità è una metodologia di preparazione dei dati, per poi essere presentati all'algoritmo di apprendimento, che si troverà di fronte delle informazioni più compatte e quindi più facilmente processabili.
Inoltre bisogna notare che, se il processo di riduzione della dimensionalità viene svolto sul training data set, allora deve essere attuato anche sul test data set, per garantire un processo di verifica valido.

Il processo di riduzione della dimensionalità può essere portato avanti attraverso due metodologie differenti:

- Selezione, dove solo alcune componenti dei vettori di input vengono conservate;
- Estrazione, dove viene creato un numero ridotto di nuove componenti a partire da quelle originali.

A prescindere da questa distinzione, bisogna sottolineare che tutti i processi di riduzione della dimensionalità hanno una struttura comune, ovvero sono caratterizzati da una fase di *encoding* (che rappresenta il vero e proprio processo di riduzione della dimensionalità) e da una fase di *decoding*, nella quale si verifica quanta informazione è stata persa nel processo.
Si riporta in figura 10 l'illustrazione grafica del processo *encoding-decoding*

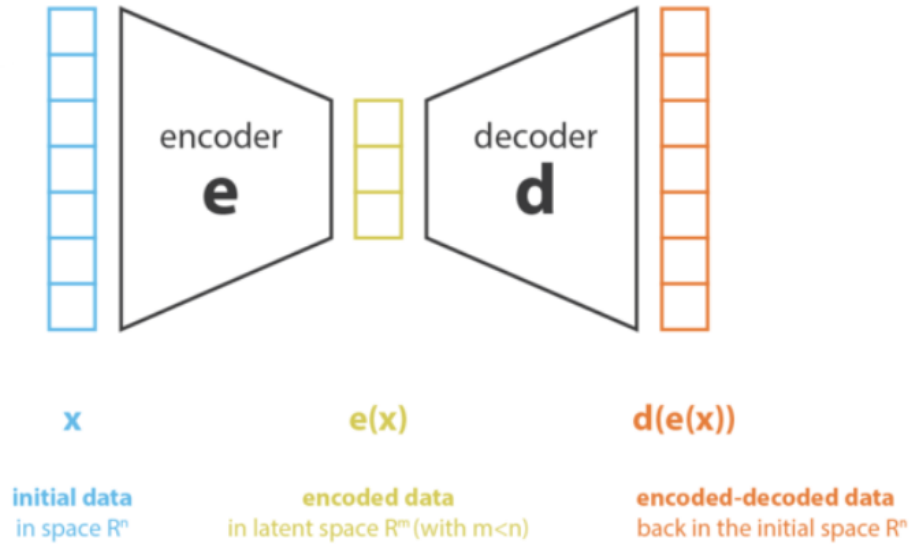


Figura 10: Strutture generale di un processo di riduzione della dimensionalità . L'immagine è presa da Rocca, 2019

Il vettore di input $\mathbf{x}$ (n-dimensionale) viene compresso dall'encoder $\mathbf{e}$ in un vettore $\mathbf{e}(\mathbf{x})$ di uno spazio m-dimensionale (con m<n), detto *spazio latente*; l'encoder, come detto, può agire per selezione o per estrazione.
Il decoder $\mathbf{d}$ svolge la funzione opposta, ovvero decomprime il vettore $\mathbf{e}(\mathbf{x})$ in $\mathbf{d}(\mathbf{e}(\mathbf{x}))$ per tornare allo spazio originario n-dimensionale.
Nel caso in cui $\mathbf{x} = \mathbf{d}(\mathbf{e}(\mathbf{x}))$ (caso ideale) si dice che il processo è un *lossless encoding*, ovvero non c'è stata perdita di informazioni nella riduzione della dimensionalità; viceversa, se $\mathbf{x} \neq \mathbf{d}(\mathbf{e}(\mathbf{x}))$, si parla di un *lossy encoding*, cioè un processo nel quale parte dell'informazione viene persa e non può essere recuperata con la fase di decoding.

Come conseguenza di ciò che è stato appena illustrato, l'obiettivo di un processo di riduzione della dimensionalità è quello di trovare la coppia encoder-decoder (e,d) fra una famiglia di encoder E e di decoder D, che minimizzi l'informazione persa:

$$(e, d) = \min_{E \times D} \epsilon(\mathbf{x}, \mathbf{d}(\mathbf{e}(\mathbf{x}))) \tag{29}$$

dove $\epsilon(\mathbf{x}, \mathbf{d}(\mathbf{e}(\mathbf{x})))$ è la grandezza attraverso la quale viene quantificata la quantità di informazione persa nel processo di riduzione.

A questo punto è possibile illustrare le varie metodologie di riduzione della dimensionalità, secondo la distinzione già incontrata fra selezione ed estrazione.
I metodi di selezione ("*Feature Selection Methods* (FSM)") sono metodi attraverso i quali vengono selezionate le componenti dei vettori di input da tenere e quelle da eliminare perché irrilevanti per le analisi successive. I FSM si includo i *wrapper methods* ed i *filter methods*: i primi valutano il modello con varie combinazioni di subset delle variabili originali e selezionano quella con la più alta efficienza, mentre i secondi utilizzano un metodo basato su dei punteggi per valutare eventuali correlazioni fra le variabili di partenza.
I metodi di estrazione, invece, si basano fortemente sull'algebra lineare; in particolare vengono utilizzati spesso per la riduzione della dimensionalità i metodi di fattorizzazione delle matrici per cogliere la parte più importante dei dati.
Il più comune di questi metodi prende il nome di *Principal Component Analysis* (PCA), del quale verrà presentata brevemente l'idea di base, evitando di addentrarsi troppo nella trattazione matematica che è essenzialmente riconducibile ad un calcolo di autovalori ed autovettori.
L'idea del PCA è quella di costruire un numero $n_e$ di nuove variabili indipendenti che siano combinazione lineare delle $n$ variabili di partenza; tale costruzione viene fatta in modo tale che la proiezione delle vecchie variabili sul nuovo sottospazio generato da quelle nuove sia il più possibile vicina ai dati iniziali, dove la vicinanza è da intendere in termini della distanza euclidea. In altre parole con il PCA si ricerca il sottospazio dello spazio dei pattern di partenza per il quale l'errore che viene compiuto nell'approssimazione dei dati tramite proiezioni sia il più piccolo possibile. Si riporta in figura 11 un'illustrazione di ciò che è stato appena detto.

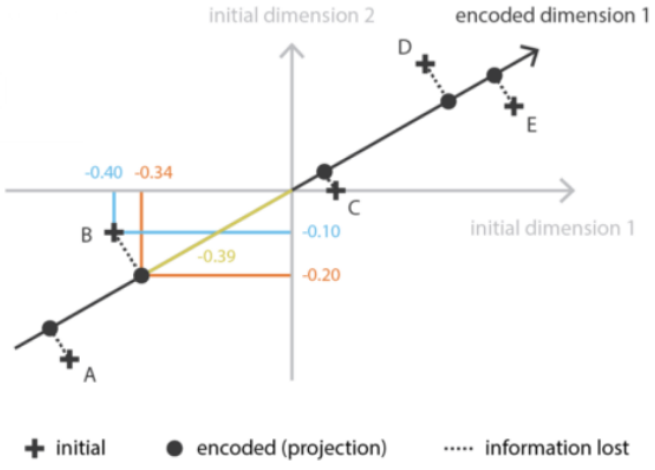


Figura 11: Illustrazione del processo di PCA nel caso di uno spazio dei pattern iniziali bi-dimensionale (Rocca, 2019).

## 3.9 Autoencoders

Gli autoencoders, come ogni altro metodo di riduzione della dimensionalità, sono costituiti da un encoder e da un decoder; tuttavia in questo caso la peculiarità è che sia l'encoder che il decoder sono delle reti neurali, come è possibile vedere in figura 12.

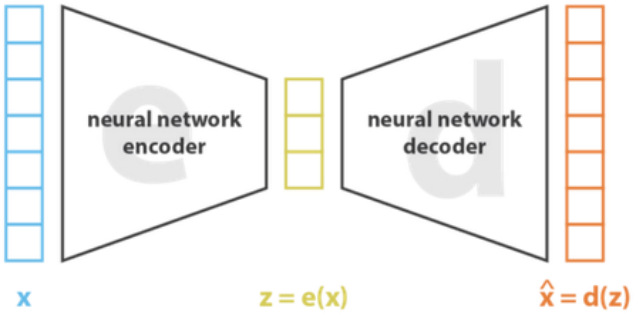


Figura 12: Struttura di un generico autoencoder (Rocca, 2019).

L'obiettivo è chiaramente quello di individuare la coppia encoder-decoder che ottimizza il processo di ricostruzione degli input e ciò viene fatto attraverso il seguente processo iterativo: si presentano all'encoder i pattern di partenza uno alla volta, subiscono un processo di riduzione della dimensionalità e poi vengono ricostruiti (tornando alla dimensionalità di partenza), viene calcolato l'errore dal confronto fra l'input iniziale e quello ricostruito ed avviene l'aggiornamento dei pesi della rete neurale mediante il meccanismo di back-propagation, già incontrato nella sezione 3.6.
Intuitivamente l'autoencoder può essere pensato come un collo di bottiglia, attraverso il quale solo una parte dell'informazione riesce a passare oltre e a formare i vettori dello spazio latente. Facendo riferimento alla figura 12 si osserva che, a partire dai pattern in input $\mathbf{x}$, come primo passo si costruisce lo spazio latente degli $\mathbf{z} = e(\mathbf{x})$ per poi procedere alla fase di decodifica nella quale si ottengono i pattern ricostruiti $\hat{\mathbf{x}} = d(\mathbf{z})$; si procede successivamente al calcolo degli errori nel seguente modo:

$$L = \|\mathbf{x} - \hat{\mathbf{x}}\| \tag{30}$$

dove L è l'errore di ricostruzione.
Una considerazione necessaria circa l'errore, che potrebbe sembrare in contraddizione con quanto detto fino ad ora sul concetto di ottimizzazione, è che si vuole di norma evitare che $\mathbf{x} = \hat{\mathbf{x}}$, perché questo vuol dire che l'autoencoder ha imparato la funzione identità e, come conseguenza, la struttura dello spazio latente, che è quella interessante per il processo di riduzione della dimensionalità, non porta alcuna informazione interessante; ciò è dovuto al fatto che l'encoder non impara se vi siano variabili più o meno importanti di altre o se esse possano essere compattate in nuove variabili di dimensionalità minore.
Per fornire un esempio pratico di ciò che è stato appena affermato, si consideri un insieme di vettori di input N dimensionali; una possibilità è quella di prendere una per una le componenti dei pattern e disporle lungo una retta (spazio latente 1-dimensionale) nella fase di encoding, per poi procedere in maniera inversa nella fase di decodifica. L'errore con questo procedimento sarà nullo ma non si può essere soddisfati essenzialmente per due motivi, ovvero perché lo spazio latente non è interpretabile e sfruttabile e perché in

un processo di riduzione della dimensionalità si vuole fare in modo che i dati continuino a conservare una qualche struttura.
Una possibilità per evitare il risultato appena illustrato, che è in fin dei conti una sfaccettatura del concetto di overfitting, è di aggiungere alla funzione L un fattore di regolarizzazione che penalizza i risultati per i quali $\mathbf{x} = \hat{\mathbf{x}}$.
Quindi bisogna sempre porre particolare attenzione alla scelta della profondità dell'encoder, ovvero alla sua capacità di riduzione della dimensionalità.
Per completezza nella trattazione si osserva che gli autoencoder possono essere sia lineari che non; il primo caso si ottiene quando non si inserisce una funzione di attivazione non lineare e si utilizzano solo due strati, quindi le trasformazioni possono essere rappresentate come matrici e si ottiene un risultato simile a quello del PCA ( 3.8).
Il caso di autoencoder non lineari (*deep autoencoder*) può essere pensato come un passo successivo per quanto riguarda la riduzione della dimensionalità. Infatti, come è stato già detto, il PCA ricerca il miglior iperpiano nello spazio dei pattern originali sul quale questi possano essere proiettati in modo da ridurre la perdita di informazione; dall'altro lato gli autoencoder non lineari non si limitano alla ricerca di iperpiani, ma possono esplorare anche superfici più complesse, come si evince chiaramente dalla figura 13.

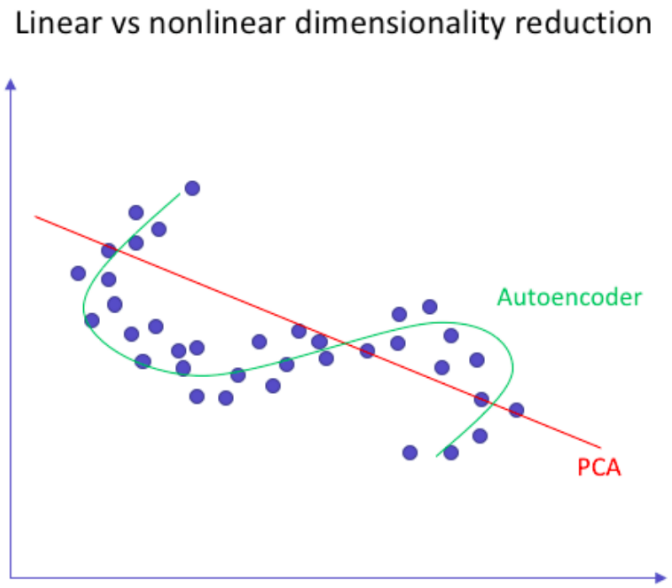


Figura 13: Differenza fra i metodi lineari (PCA) e gli autoencoder non lineari (Jordan, 2018).

## 3.10 Variational Autoencoders (VAEs)

Nella sezione precedente è stato presentato l'autoencoder come un metodo di riduzione della dimensionalità; tuttavia, apportando una sostanziale modifica, tale metodo può essere utilizzato per la generazione di pattern ed in questo caso si parla di Variational Autoencoders (VAEs).
Bisogna domandarsi a questo punto in che modo si possa passare dalla riduzione della dimensionalità al generare nuovi pattern e perché il semplice autoencoder, così come è stato incontrato, non permette di raggiungere tale obiettivo e debba essere modificato opportunamente.
Come primo passo è necessario capire perché gli autoencoders non siano adatti al processo di generazione e la risposta può essere intuita già da ciò che è stato detto nella sezione precedente ma si cercherà di argomentarla meglio. L'autoencoder, come noto, agisce su di un pattern originale $\mathbf{x}$ trasformandolo in un $\mathbf{z} = e(\mathbf{x})$ nello spazio latente (fase di encoding) e poi trasforma nuovamente $\mathbf{z}$ per tornare allo spazio originale; si potrebbe pensare che, per far svolgere la funzione di generazione, si potrebbe scegliere un punto a caso dello spazio latente (che è stato costruito nella fase di encoding) e darlo in pasto al decoder, ottenendo così un nuovo pattern per nulla collegato a quelli originali. Il problema nel ragionamento appena esplicato sta nel non tenere in conto la regolarità dello spazio latente. Come detto nella sezione precedente, ottenere uno spazio latente regolare non è affatto semplice e quindi la regolarità non può essere assunta a priori; questo non deve affatto stupire perché l'autoencoder è pensato per codificare e decodificare informazioni cercando di ridurre al minimo la perdita delle stesse e quindi tenderà naturalmente all'overfitting. Il risultato sarà uno spazio latente non regolarizzato e quindi, campionando da esso un punto a caso, si otterrebbe un pattern ricostruito assolutamente privo di senso, come si può osservare, in maniera didattica, dalla figura 14.

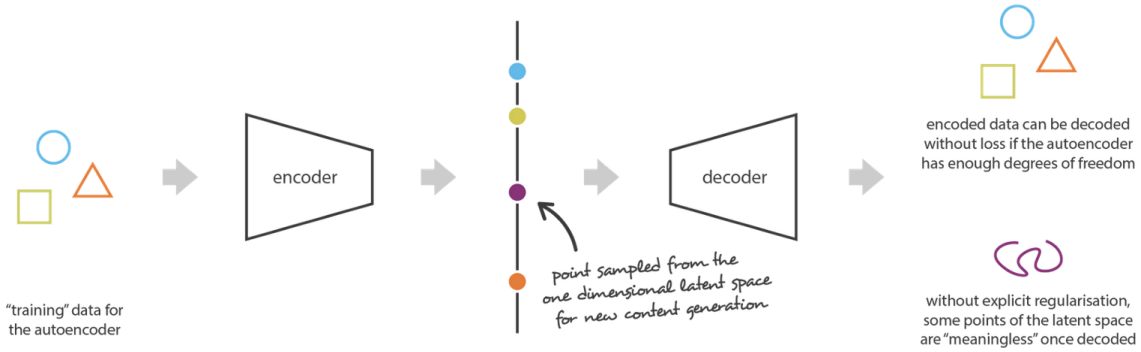


Figura 14: Illustrazione grafica e semplificata dei limiti nell'applicazione degli autoencoder per fini di generazione di nuovi pattern (Rocca, 2019).

Quindi, per raggiungere l'obiettivo prefissato, si attua un processo di regolarizzazione durante la fase di addestramento e si può iniziare a parlare di Variational Autoencoders (VAEs). Si può quindi definire il VAEs come un autoencoder per il quale si attua una regolarizzazione dello spazio latente durante il processo di addestramento ed è quindi adatto alla generazione di nuovi pattern.
I VAEs, al pari degli autoencoders, sono costituiti da una coppia di encoder e decoder con

una sostanziale differenza, nel senso che per gli autoencoders il pattern di partenza viene codificato come un punto nello spazio latente ($\mathbf{z}$), mentre per i VAEs la codifica avviene tramite una distribuzione nello spazio latente ($p(\mathbf{z}|\mathbf{x})$). Il processo seguito nella fase di addestramento è il seguente:

1. Il pattern iniziale viene codificato nello spazio latente come una distribuzione di probabilità;
2. Si campiona un punto dello spazio latente a partire dalla distribuzione del punto precedente;
3. Tale punto viene decodificato dal decoder, ottenendo il pattern ricostruito;
4. Avviene il confronto fra il pattern iniziale e quello ricostruito da cui si calcola l'errore, che viene poi propagato mediante il meccanismo della backpropagation.

Nella pratica si cerca di fare in modo che la distribuzione ottenuta alla fine del processo di codifica sia il più possibile vicina ad una distribuzione Normale, perché come si vedrà in seguito si riesce a garantire una certa regolarità dello spazio latente; in particolare si farà in modo che l'encoder restituisca la media e la matrice di covarianza della distribuzione Normale.
Seguendo questa linea, si ottiene che il processo di addestramento è regolato da una funzione di perdita, composta da un termine relativo alla ricostruzione ed uno relativo alla regolarizzazione dello spazio latente; tale termine di regolarizzazione viene espresso mediante la *Divergenza di Kullback-Leibler*, che rappresenta una misura della differenza ta due distribuzioni di probabilità.
Nelle righe precedenti è stata richiesta la regolarità dello spazio latente ed è giusto chiarire cosa si intenda effettivamente. Lo spazio latente è regolare se:

- è continuo, ovvero due punti vicini portano ad un risultato simile una volta decodificati;
- è completo, nel senso che un qualunque punto porta ad un risultato sensato una volta decodificato.

Il solo fatto di codificare i pattern come delle distribuzioni non garantisce la continuità e la completezza, perché se non si inserisce il termine di regolarizzazione nella funzione di perdita il VAE continuerà a comportarsi come un semplice autoencoder, tendendo semplicemente a minimizzare l'errore di ricostruzione. Questo può avvenire in due modi, ovvero codificando i pattern come distribuzioni o con varianze molto piccole (quasi come singoli punti) o con medie molto diverse fra loro (punti molto lontani nello spazio latente); nel primo caso non viene garantita la continuità e nel secondo la completezza.
Per ottenere la regolarità dello spazio latente si richiede allora che le distribuzioni con cui vengono codificati i pattern siano il più possibile vicine a distribuzioni Normali con media zero e matrice di covarianza uguale all'identità; le medie saranno allora vicine con conseguente sovrapposizione delle distribuzioni, anche perché la matrice di covarianza così fatta impedisce la codifica come punti nello spazio latente. Il prezzo da pagare sarà chiaramente un più alto errore nella fase di ricostruzione.
Prima di passare alla formulazione matematica dei VAEs, bisogna notare che la regolarità

dello spazio latente implica la presenza di un gradiente, il quale permette di mischiare le caratteristiche dei pattern in input e quindi di dare un significato al campionamento nello spazio latente.

### 3.10.1 Formulazione matematica dei VAEs

Per formulare in maniera rigorosa i VAEs è necessario utilizzare l' inferenza variazionale e verranno dati alcuni concetti fondamentali della teoria dell'informazione.
Per prima cosa bisogna introdurre una grandezza, che è in grado di quantificare la quantità di informazione di una proposizione ed è appunto detta *Informazione*:

$$I = \log p(x) \tag{31}$$

dove x è l'evento.
Questo concetto di informazione coincide con quello che si possiede intuitivamente, ovvero che ad eventi certi o molto probabili corrisponde una quantità di informazione nulla o molto bassa, mentre ad eventi poco probabili corrisponde una quantità di informazione più alta.
Un'altra quantità fondamentale nella teoria dell'informazione è l'*entropia*, ovvero l'informazione media, ed è definita nel seguente modo:

$$H = \sum p(x) \log p(x) \tag{32}$$

A questo punto si introduce la *KL divergency* (già accennata precedentemente) che è di fondamentale importanza nel processo di addestramento dei VAEs; si tratta di una misura della dissimilarità di due distribuzioni (p(x) e q(x)):

$$KL(p(x)||q(x)) = -\sum p(x) \log q(x) + \sum p(x) \log p(x) = -\sum p(x) \log \frac{q(x)}{p(x)} \tag{33}$$

si nota come essa sia molto simile alla differenza delle entropie delle due distribuzioni e ciò è in linea con l'intuizione perché due distribuzioni che hanno un'informazione media uguale saranno pressoché uguali. Le due proprietà fondamentale della *KL divergency* sono le seguenti:

1. $$KL(p(x)||q(x)) \geq 0 \tag{34}$$

2. $$KL(p(x)||q(x)) \neq KL(q(x)||p(x)) \tag{35}$$

Ora è possibile ricollegarsi al discorso sui VAEs e si definisca con $\mathbf{x}$ la grandezza osservabile e con $\mathbf{z}$ quella nascosta dello spazio latente. Il teorema di Bayes [ Del-Prete, 2010] permette di scrivere:

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \tag{36}$$

dove

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})dz \tag{37}$$

tuttavia tale integrale è difficile da calcolare ed è in questo punto che entra in gioco l'inferenza variazionale per approssimare $p(\mathbf{z}|\mathbf{x})$ con una qualche funzione $q(\mathbf{z}|\mathbf{x})$. Si assume inoltre che quest'ultima debba essere scelta dalla famiglia delle distribuzioni Normali, andando a variare i parametri in modo che risulti il più possibile simile a $p(\mathbf{z}|\mathbf{x})$ e per quantificare ciò si utilizza la *KL divergency*:

$$KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) = -\sum q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|\mathbf{x})} \tag{38}$$

e sostituendo $p(z|x)$ con la 36 si ottiene:

$$\begin{aligned} KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) &= -\sum q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})q(\mathbf{z}|\mathbf{x})} \\ &= -\sum q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} + \sum q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}) \end{aligned}$$

ma in entrambi i termini della somma la sommatoria è estesa sulle z, quindi:

$$\sum q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}) = \log(\mathbf{x}) \sum q(\mathbf{z}|\mathbf{x}) = \log p(\mathbf{x})$$

perché $\sum q(\mathbf{z}|\mathbf{x}) = 1$.
Si arriva quindi al seguente risultato:

$$\log p(\mathbf{x}) = KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) + \sum q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} \tag{39}$$

La sommatoria nell'equazione 39 prende il nome di *Variational lower bound* e viene indicata con la lettera $\mathcal{L}$.
A questo punto si deve osservare che $\mathbf{x}$ è fissato e quindi il termine sinistro dell'equazione 39 è una costante; di conseguenza minimizzare la *KL divergency* equivale a massimizzare la $\mathcal{L}$, che può essere riscritta in maniera semplificata:

$$\begin{aligned} \mathcal{L} &= \sum q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} \\ &= \sum q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}|\mathbf{z}) + \sum q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} \\ &= E_q \log p(\mathbf{x}|\mathbf{z}) - KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \end{aligned}$$

Quindi, ricapitolando, l'obiettivo iniziale è quello di trovare la distribuzione $p(\mathbf{z}|\mathbf{x})$ che però è molto complessa per essere calcolata; allora si cerca di approssimarla con una $q(\mathbf{z}|\mathbf{x})$ scelta tra un'opportuna famiglia e, per scegliere quella più vicina, si deve minimizzare $KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))$, che come visto equivale a massimizzare la $\mathcal{L}$.
In figura 15 è possibile osservare in che modo si passa da $\mathbf{x}$ a $\mathbf{z}$ e viceversa.

Il significato del termine di *KL divergency* che compare in $\mathcal{L}$ suggerisce che la distribuzione $q(\mathbf{z}|\mathbf{x})$ debba essere il più possibile ad una distribuzione $p(\mathbf{z})$ che può essere scelta e quindi

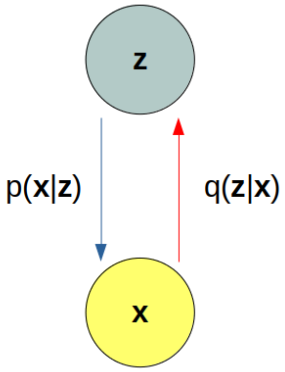


Figura 15: Illustrazione grafica della modalità attraverso la quale si ottiene il passaggio dalla variabile **x** osservabile alla **z** nello spazio latente e viceversa.

si assume una distribuzione Normale.
L'altro termine in $\mathcal{L}$ è riconducibile ad un errore di ricostruzione, infatti il processo di decodifica, una volta campionato **z** è deterministico e quindi si ottiene che:

$$p(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}|\hat{\mathbf{x}}) \tag{40}$$

dove $\hat{\mathbf{x}}$ è il pattern ricostruito. Inoltre se si considerano distribuzioni gaussiane si troverà che:

$$p(\mathbf{x}|\hat{\mathbf{x}}) \propto e^{-|\mathbf{x}-\hat{\mathbf{x}}|^2} \tag{41}$$

e quindi

$$\log p(\mathbf{x}|\hat{\mathbf{x}}) \propto -|\mathbf{x}-\hat{\mathbf{x}}|^2 \tag{42}$$

Quindi si osserva che l'autoencoder tende a minimizzare semplicemente $|\mathbf{x}-\hat{\mathbf{x}}|^2$, mentre il VAE tende a minimizzare la seguente quantità:

$$|\mathbf{x}-\hat{\mathbf{x}}|^2 + KL(q(\mathbf{z}|\mathbf{x})||N(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \tag{43}$$

Nella pratica si costruisce la rete neurale che si occupa della fase di codifica in modo che restituisca i parametri della distribuzione Normale e quindi la media e la matrice di covarianza, che si impone essere diagonale per semplicità; da qui viene campionato un punto dello spazio latente a partire da tale distribuzione e avviato verso il decoder per ottenere il pattern ricostruito da confrontare con quello iniziale (nella fase di addestramento). Lo schema di questo processo è riportato in figura 16.

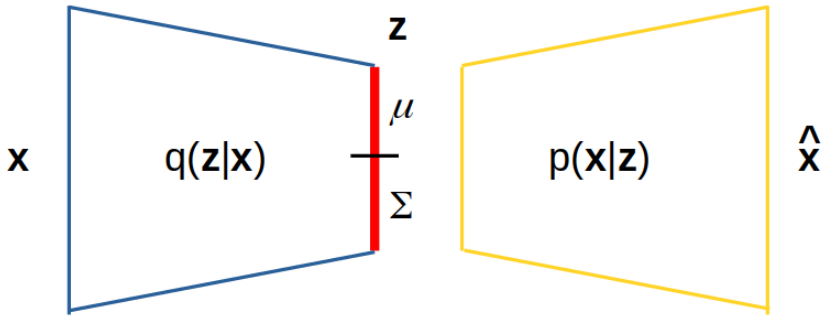


Figura 16: Schema di funzionamento del VAE, dove viene messo in evidenza che l'encoder è forzato a restituire come output i parametri della distribuzione Normale, ovvero la media e la matrice di covarianza.

Infine, per la fase di generazione di nuovi pattern, sarà sufficiente eliminare l'encoder e campionare dalla distribuzione che si è ottenuta alla fine dell'addestramento e fornire tale punto al decoder per costruire un nuovo pattern.

# 4 Ricerca di fisica Behind Standard Model con il VAEs

In quest'ultimo capitolo verrà presentata una possibile applicazione dei Variational Autoencoders nel campo della fisica delle alte energie, con lo scopo di ricercare segnali di nuova fisica BSM, ovvero oltre il Modello Standard.
Come noto, gli esperimenti portati avanti al *Large Hidron Collider* hanno l'obiettivo di esplorare la fisica spingendosi sempre a più alte energie; attualmente, dopo la scoperta del *Bosone di Higgs*, la teoria del Modello Standard sembrerebbe essere completa, anche se rimango alcuni problemi aperti, come lo *Hierarchy Problem* e la spiegazione della *Dark Mattern*.
Nella ricerca di nuova fisica BSM sono attualmente presenti due problemi:

1. Solitamente ad LHC tale ricerca avviene nel cosiddetto modo *model dependent*, ovvero viene ricercata nuova fisica con un preciso modello in mente ed i risultati sono ottimi nel caso in cui il modello che si sta utilizzando è giusto, come ad esempio con la scoperta del Bosone di Higgs; il problema è che tutti i nuovi modelli testati fino ad ora non hanno prodotto risultati e quindi è molto probabile che eventuale fisica BSM non sia spiegabile con tali teorie ed è in quest'ottica che una ricerca Model Dependent perde di efficacia.

2. Il secondo problema è di natura pratica, infatti ad LHC vengono prodotte 40 milioni di collisioni di protoni al secondo e solo i risultati di mille di queste al secondo possono essere conservati dagli esperimenti ATLAS e CMS. La scelta di questi mille viene svolta da degli algoritmi di selezione ed è quindi molto probabile che eventuali segnali di nuova fisica vengano tralasciati.

La possibile soluzione potrebbe essere di utilizzare per il sistema di selezione (*trigger*) degli algoritmi di natura *model independent*, che vengono addestrati sulla fisica del SM e quindi sono in grado di rilevare eventuale fisica BSM come anomalia.
Nelle pagine seguenti verranno illustrati i risultati che si ottengono nel caso in cui si utilizzino i VAEs per lo scopo appena esposto.

# 5 Introduction

## 5.1 Model description

Like many other deep generative models, the variational auto encoders' (VAEs) aim is to learn the underlying input data distributions to generate new samples with features similar to the original ones. In order to achieve this goal, these algorithms are built basically in two steps: the first one - the *encoding stage* - where the VAE compresses the input data in a lower-dimensional space (*latent space*) and the second step where it tries to reconstruct the original input - *decoding phase* - distribution starting from this incomplete information.
Contrarily to the vanilla autoencoders, where a point wise encoding results in a less efficient and precise regeneration, the VAEs compress data in a continuous latent space. Indeed, each input is associated with a distribution within this space and the reconstruction step starts only after sampling from that distribution. In this way, the model is able to reconstruct similarly points close togheter in the latent space. This approach gives continuity and completeness to this space, making the regeneration step easier and more robust.

The variational autoencoder architecture could be thought of as the ensemble of two neural networks, one on top of the other, designed respectively with a contractive-path and an expansive one. The first of these provides a last layer/s with fewer neurons respect to the input layer, so to be able to compress the data in a lower-dimensional space. The second one, instead, starting from the latent representation, moves in the opposite direction and tries to reconstruct the compressed data in the higher-dimensional original space. Naturally, this separation is only speculative and joint optimization of all the network weights can be obtained minimizing an objective function.

The target *function loss* suitable for the final goal is constituted by two pieces. The first term is related to the model reconstruction performances, while the second one regards the Kulback-Lieber divergence between the latent space shape and its target distribution, usually a multivariate standard gaussian. So the general form of the final loss results from a weighted sum of these two terms:

$$Loss_{tot} = Loss_{reco} + \beta D_{KL} \tag{44}$$

where $\beta$ is a free parameter defining the relative weight of the divergence loss term in the total final score function and $D_{KL}$ is the Kulback-Lieber divergence, described in detail in Section 5.3.1. Under this loss definition, the model learns how to compress and successfully reconstruct the bulk of the variable distributions contained in the training samples. On the other hand, the model is prone to fail while reconstructing the more rare events. Due to this behavior, the latter kind of event is likely to end up in the right tail of the distribution loss (higher losses). This situation is even more evident when considering data samples characterized by different variable distributions with respect to the ones used for training, for example samples with non SM events which were not present in the SM-only background sample.
Given these considerations, in this study a model trained to reproduce a cocktail of Monte Carlo SM processes is presented, with the aim of testing the background modeling capabi-

lities. The signal sensitivity is tested including both background and signal events in the validation sample and expecting to find a region in the right tail of the loss distribution where maximize the purity of the signal.

## 5.2 Dataset

The Wh1Lbb analysis ntuples are used in this study. The same preselection of the original analysis is applied both on the background processes and on the signal events and it is reported in Table 1. In Figure 17 the distributions of the most discriminating variables are shown for the total background (sum of all the processes) and some signal example. The definition and the data/MC agreement of each variable considered for the training are explained and described in detail in Section 6 of. It is worth to remark that only the background events are used to train the model, while the signal events are only included in the evaluation step after the events selection.

Tabella 1: Preselection cuts applied both on signal and background samples.

| | Preselection |
|---|---|
| Exactly 1 signal lepton | True |
| met trigger fired | True |
| $2-3$ jets with $p_T > 30GeV$ | True |
| $b$-tagged jet | [1-3] |
| met | $>220$ GeV |
| mt | $>50$ GeV |

The model is trained in three different kinematics regions described in Table 2. The aim is to test the sensitivity to the signal model in different topological spaces. The trade-off between a model-independent analysis and better signal sensitivity is one of the starting points investigated in this study. In the first case, any or only loose selection requirements would be applied to avoid cuts tailored on a specific signal hypothesis. Nevertheless, training on more selected background events could allow the model to focus more on the relevant background distinctive features finally leading to a better signal selection, albeit reducing the generality of the selection, and possibly reducing the selection sensitivity to similar models (as pMSSM, or different signal models).

Tabella 2: Requirements for the three selected regions.

| | Preselection | mid. region | $2-3$b region |
|---|---|---|---|
| Exactly 1 signal lepton | True | True | True |
| met trigger fired | True | True | True |
| $2-3$ jets with $p_T > 30GeV$ | True | True | True |
| $b$-tagged jet | [1-3] | [1-3] | [2-3] |
| met | $>220$ GeV | $>220$ GeV | $>220$ GeV |
| mt | $>50$ GeV | $>50$ GeV | $>50$ GeV |
| mbb | | $[100-140]$ GeV | $[100-140]$ GeV |
| mct | | $>100$ GeV | $>100$ GeV |

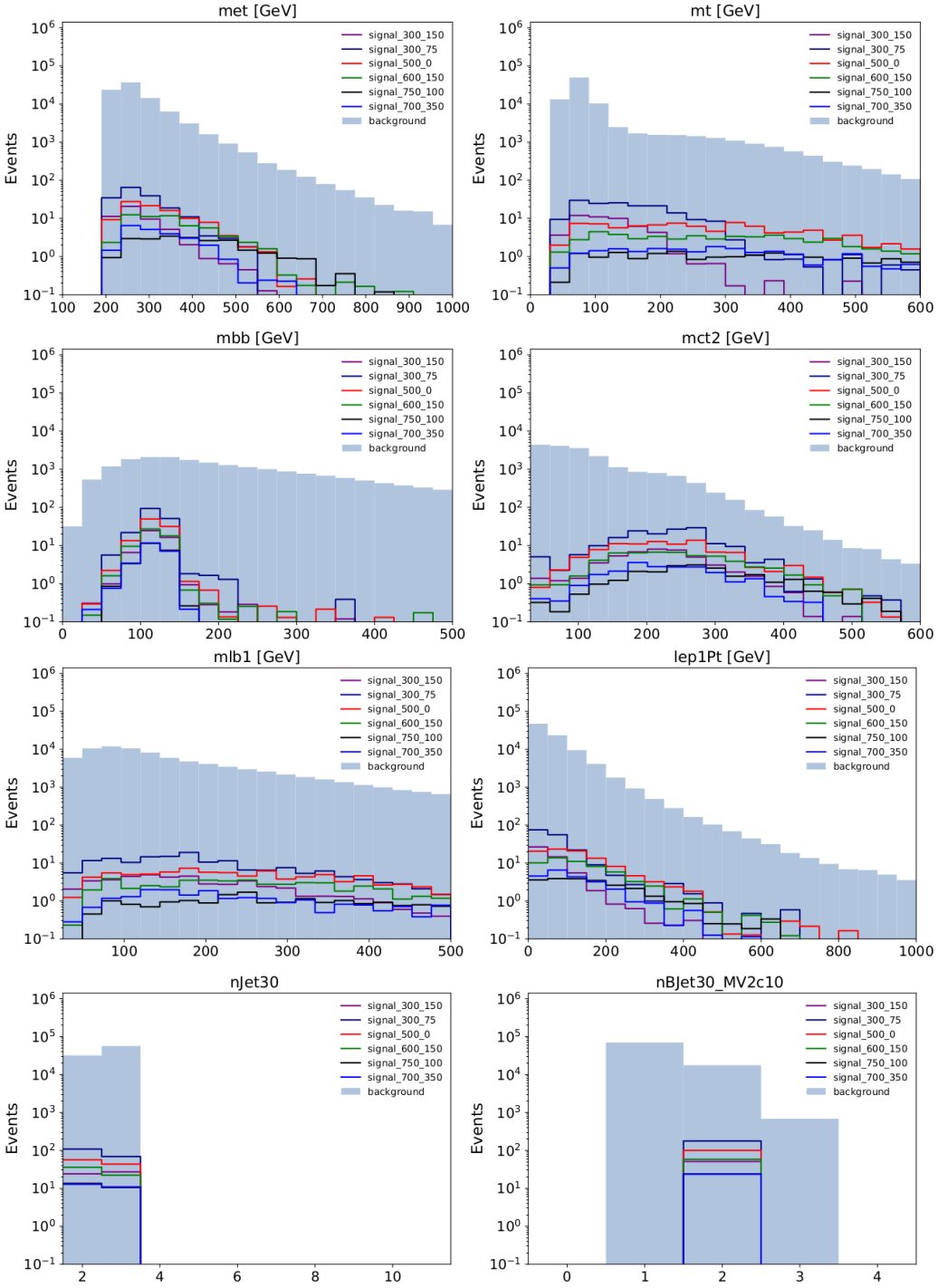


Figura 17: Distributions of the input variables for the sum of all the background processes and some signal points, as example.

## 5.3 Training

The model is trained in the three regions separately after a $60\% - 40\%$ training-validation split. Due to the different selection depth, the total training-validation event number varies region by region. In order not to reduce this number drastically, a threshold of $10^6$ events for the training-validation sample is fixed as a lower limit and the 2 - 3*b region* is taken as the region with the smallest selected number of events.

### 5.3.1 Loss function

The training loss function described above (eq. 44) consists of two terms. The goal of the training is to find the best trade-off between good reconstruction performances and a regular latent space. A closer description of these two terms is given in the following.

The Kullback-Lieber divergence term regards the encoding phase outputs and forces all the predicted distributions to stay as close as possible to the prior choice. For the sake of simplicity, usually a multivariate gaussian is selected as target distribution. In this study, the Kulback-Lieber divergence has the following closed analytical form:

$$D_{KL} = \frac{1}{2k} \sum_{i,j=1} \left(\sigma_p^j \sigma_z^{i,j}\right)^2 + \left(\frac{\mu_p^j \mu_z^{i,j}}{\sigma_p^j}\right)^2 + ln \frac{\sigma_p^j}{\sigma_z^{i,j}} - 1 \tag{45}$$

where,$k$ is the batch size selected for the training, $i$ runs over the samples and $j$ over the latent space dimensions. The parameters predicted from the model for each event are: (1) $\mu_z$ and $\sigma_z$ that define the distribution shape in the latent space; (2) $\mu_p$, and $\sigma_p$ that represents the prior shape parameters. It is important to observe that, following the work described in, also this latter couple of parameters are optimized during the training letting the model to select the optimal latent space target distribution.

The reconstruction loss term is instead represented by the average negative-log-likelihood of the inputs given the shape parameter values predicted by the model during the decoding phase:

$$\begin{aligned} Loss_{reco} &= -\frac{1}{k} \sum_{i,j=1} \ln \left[P\left(x|\alpha_1, \alpha_2, ...\alpha_n\right)\right] \\ &= -\frac{1}{k} \sum_{i,j=1} \ln \left[f_{i,j}\left(x_{i,j}|\alpha_1^{i,j}, \alpha_2^{i,j}, ...\alpha_n^{i,j}\right)\right] \end{aligned} \tag{46}$$

In the equation, $j$ runs over the input space dimensions, $f_{i,j}$ is the functional form chosen to describe the pdf of the *i-th* input variable and $\alpha_n$ are the parameters of this function and represent also the final output of the network. Two different functional forms are selected to describe the distribution of the variables defining each physical events inside the training dataset. Specifically:

- the clipped log-normal function is used for all the continuous variables: *met, mt, mbb, mct2, mlb1, lep1Pt*:

$$P\left(x|\alpha_1,\alpha_2,\alpha_3\right) = \begin{cases} \alpha_3\delta\left(x\right)\frac{1-\alpha_3}{x\alpha_2+\sqrt{2\pi}}e^{-\frac{(\ln x-\alpha)^2}{2\alpha_2^2}} & for x \geq 10^{-4} \\ 0 & \leq 10^{-4} \end{cases} \tag{47}$$

- A truncated discrete gaussian for for the discrete variables is: *njet30, nBjet30_MVc10*:

$$\Theta\left(x\right)\left[erf\left(\frac{n+0.5-\alpha_1}{\alpha_2+\sqrt{2}}\right)-erf\left(\frac{n-0.5-\alpha_1}{\alpha_2+\sqrt{2}}\right)\right] \tag{48}$$

  where the normalization factor N is set to:

$$N = \frac{1}{2}\left(\frac{-0.5-\alpha_1}{\alpha_2+\sqrt{2}}\right) \tag{49}$$

### 5.3.2 Model architecture

The network architecture is briefly described in the 5.1. It consists mainly of a contractive path followed by an expansive one. So, the feed-forward step goes first through a stack of fully-connected layers that progressively builds a representation of the inputs in the latent space and, then, towards a second fully-connected set of layers whose goal is to output the parameters shape (eq: 47, 48): the latter are used to describe the variables probability distributions for each event and represent the final output of the variational autoencoder.

The configuration of the network here trained is strongly inspired by the work **vae:Cerri_2019**.

Nevertheless, the model configuration is also affected by the choice of a set of hyperparameters: that is, all the parameters fixed a priori, whose value is not learned automatically during the training. The hyperparameters are opposed to the network weights, linking the neurons in the model architecture for which an optimization occurs through the back-propagation of the loss. The weights update happens batch by batch during the traininig, following the gradient descents related to the loss minimization. The training success also depends on the hyperparameters that, in some way, fix the boundary condition of the learning procedure. Usually, their choice proceeds by trial and error, but some hints to address the initial guesses come from the problem context. For example, increasing the number of hidden layers goes along the complexity of the model to be related to the dataset size. A more detailed list of hyperparameter examples is reported here:

- learning rate;
- number of neurons per layer;
- latent space dimension;
- batch size;
- beta weight in the total loss sum;

- kind of activation layer;
- penalization weights on one/more variable loss.

The list is quite long, and it is helpful to figure out how to handle the fine-tuning process in an effective way. For that reason, the *tune* python library has been exploited and modified accordingly to work for this study. One of the strengths of this library is the possibility to run a hyperparameters optimization at any scale. Deploying on a cluster of many GPUs, as in this case, allows for an extensive search among all the possible parameters configuration, reducing the time and the human effort. A training summary is stored during the training and all the model performances can be compared. A final rank based on the evaluation metric helps to focus on the more promising architectures and, finally, to select the best one.

# Riferimenti bibliografici

Bhat, Pushpa (2011). "Advanced Analysis Methods in Particle Physics". In: *Annual Review of Nuclear and Particle Science*.

Bulò, Samuel Rota. *Appunti di reti neurali*. URL: https://www.dsi.unive.it/~srotabul/files/AppuntiRetiNeurali.pdf.

Collaboration, ATLAS (2012). "Observation of a new particle in the search for the Standard Model Higgs bosonwith the ATLAS detector at the LHC". In: *Physics Letters B 716 (2012) 1–29*.

Del-Prete, T. (2010). *Methods of Statistical Data Analysis in High Energy Physics*. Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, Italia.

Jordan, Jeremy (2018). *Introduction to autoencoders*. URL: https://www.jeremyjordan.me/autoencoders/.

Knuth, Donald. *Knuth: Computers and Typesetting*. URL: https://towardsdatascience.com/using-3d-visualizations-to-tune-hyperparameters-of-ml-models-with-python-ba2885eab2e9.

Nilsson, Nils J. (1998). *Introduction to Machine Learning*. Department of Computer Science, Stanford University.

Rocca, Joseph (2019). *Understanding Variational Autoencoders (VAEs)*. URL: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73.