

# AI & Machine Learning

## Introduzione

Gianluca Moro  
DISI – Università di Bologna  
Cesena Campus  
nome.cognome@unibo.it

## CONTENUTI

*Data Driven Learning*  
*Data Driven Decisions*

- Introduzione
- Alcuni progetti di Machine Learning
- Tipologie e qualità dei dati
- Apprendimento ***Supervisionato, Non Supervisionato e Auto-Supervisionato***
- Classificazione
- Clustering
- Scoperta di regole associative

# Introduzione al Machine Learning

## *Data Driven Decisions*

Introduzione

## Il Ruolo dei Dati nel Machine Learning

- '60 – prime collezioni di dati e prime basi di dati
- '70 – primi sistemi per la gestione di basi di dati (DBMS)
- '80 – maturità dei DBMS, nuovi tipi di dato, nuovi paradigmi di accesso, primi tentativi di estrazione di pattern
- '90 – web, data warehousing, scoperta di conoscenza da basi di dati (KDD)
- 2000 – esplosione dei big data
- 2010 – sviluppo della visione artificiale con deep learning (DL)
- 2017 – oggi: **breakthrough** nel Natural Language Processing con (DL) chatGPT e numerosi nuovi Large Language Model

## Rivoluzione in Corso: Cos'hanno in Comune queste Applicazioni Straordinarie ?

- Primo **assistente virtuale** con capacità umana nel conversare al telefono per fare una prenotazione <https://bit.ly/2Kyjbgw>
- Negozio al dettaglio interamente automatizzato senza nemmeno cassieri e casse per pagare <https://bit.ly/2FuQgb>
- Diagnosi automatica di melanoma con precisione superumana
  - **Dermatologist-level classification of skin cancer with deep neural networks**, Andre Esteva et. al. (Nature)
- **Profilazione utenti**: capacità di prevedere da pochi post e like caratteristiche personali altamente sensibili <https://bit.ly/1mG7Go6>
  - age, gender & sexual orientation, ethnicity, religious and political views, use of illegal substances, happiness,intelligence,personality traits, parental separation
- **Robot che imparano da zero e da soli** a svolgere nuovi compiti: camminare, correre, evitare ostacoli <https://bit.ly/2sQmx5z>
- **Auto a guida autonoma**: Tesla, Google e tanti altri ...

Gianluca Moro – DISI, Università di Bologna

5

## Rivoluzione in Corso: Cos'hanno in Comune queste Applicazioni Straordinarie ?



6

# Ottenere Intelligenza dai Dati: Rivoluzione Con Impatto Equiparato alla Scoperta dell'Elettricità

- L'intelligenza di queste applicazioni **non è cablata** nel codice, ma **è appresa da insiemi di dati** con algoritmi standard **di dominio pubblico** e indipendenti dal tipo di problema
  - ["Artificial Intelligence is the New Electricity"](#) Prof. Andrew Ng (Stanford) tra i massimi esperti mondiali di machine learning
  - *l'intelligenza artificiale applicata cambierà radicalmente interi settori della vita economica e sociale dalla salute, all'industria, all'agroalimentare, ai trasporti. L'impatto sulla società è equiparato a quello dell'invenzione del motore a vapore o dell'elettricità ([Ministero dello Sviluppo Economico](#)).*
- E.g. applicazioni che sfruttano l'**intelligenza ricavata dai dati**
  - **prevedere l'andamento delle vendite** di prodotti/servizi
  - scoprire le **propensioni di acquisto** di ogni utente
  - **riconoscere condizioni di rischio** di pazienti da sensori indossabili (IoT)
  - prevedere **consumi energetici, andamento della borsa** e tanto altro ...

Gianluca Moro – DISI, Università di Bologna

7

# Ottenere Intelligenza dai Dati: Rivoluzione Con Impatto Equiparato alla Scoperta dell'Elettricità

- L'intelligenza di queste applicazioni **non è cablata** nel codice, ma **è appresa da insiemi di dati** con algoritmi standard **di dominio pubblico** e indipendenti dal tipo di problema
  - Repubblica: L'economia 4.0 cambia il lavoro: i più richiesti saranno i **DATA SCIENTIST**
  - SOLE 24 ORE: Data Scientist, il lavoro da 100mila euro l'anno
  - IBM Predicts Demand For **DATA SCIENTIST** Will Soar 28% By 2020
  - LinkedIn's Fastest-Growing Jobs Today Are In **DATA SCIENCE**
  - Gartner Group: **Applied Artificial Intelligence** a top strategic technology
- E.g. applicazioni che sfruttano l'**intelligenza ricavata dai dati**
  - **prevedere l'andamento delle vendite** di prodotti/servizi
  - scoprire le **propensioni di acquisto** di ogni utente
  - **riconoscere condizioni di rischio** di pazienti da sensori indossabili (IoT)
  - prevedere **consumi energetici, andamento della borsa** e tanto altro ...

## DATA SCIENCE ??

Principi, tecniche e algoritmi per ricavare  
intelligenza dai dati con metodo scientifico

Gianluca Moro – DISI, Università di Bologna

8

## Dalle operazioni quotidiane all'analisi

- operazioni quotidiane:
  - inventario, fatturazione, anagrafiche, ...
- utilizzati una sola volta
  - archiviazione
- dati → migliorare i nostri ***processi decisionali?***
- ***apprendere dai dati?***

Gianluca Moro - DISI, Università di Bologna

9

## A story about the hurricane Frances

Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons... predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could "start **predicting** what's going to happen, instead of waiting for it to happen", as she put it.

*New York Times, 2004*

Cited from "Provost, Fawcett – Data Science for Business"

## Traduzione - Una storia sull'uragano Frances

L'uragano Frances stava arrivando, sfrecciando attraverso i Caraibi, minacciando un colpo diretto sulla costa atlantica della Florida. I residenti si sono diretti verso una zona più elevata, ma lontano, a Bentonville, Ark., i dirigenti dei Wal-Mart Stores hanno deciso che la situazione offriva una grande opportunità per una delle loro più recenti armi basate sui dati: la tecnologia predittiva.

Una settimana prima dell'arrivo della tempesta, Linda M. Dillman, chief information officer di Wal-Mart, ha sollecitato il suo staff ad elaborare previsioni basate su ciò che era accaduto quando l'uragano Charley aveva colpito diverse settimane prima. Sostenuta dai trilioni di byte di cronologia degli acquirenti archiviati nel data warehouse di Wal-Mart, ha ritenuto che l'azienda potesse "iniziare a prevedere cosa accadrà, invece di aspettare che accada", come ha detto lei.

*New York Times, 2004*

Cited from "Provost, Fawcett – Data Science for Business"

11

Gianluca Moro - DISI, Università di Bologna

## Recommendation: Una Storia Vera (i)

- Minneapolis, Supermercati TARGET
  - Un padre un po' arrabbiato si presenta in una filiale della catena chiedendo di parlare con un manager
  - *"mia figlia riceve da settimane vostri coupon sconto su prodotti per la maternità come vestiti, culle e pannolini per neonati... ma sta facendo ancora la scuola superiore, la state incoraggiando a rimanere incinta?"*
  - Il manager rileva che il materiale indirizzato alla figlia dell'uomo conteneva coupon sconto e pubblicità per abbigliamento e articoli premaman, ma anche foto di neonati sorridenti ...
  - Il manager: *"ci scusi deve esserci stato un errore"* e alcuni giorni dopo richiama il padre per scusarsi nuovamente
  - Al telefono, però, il padre era piuttosto imbarazzato. *"Ho parlato con mia figlia... ci sono state alcune attività in casa mia di cui non ero consapevole. Lei partorirà in Agosto. Le devo io delle scuse"*



fonte: Forbes - Charles Duhigg of New York Times

## Recommendation: Una Storia Vera (ii)

- Come il supermercato ha previsto che la ragazza era incinta e quali prodotti sarebbero stati più appropriati per lei ?
- il supermercato associa ad ogni scontrino un codice cliente che dipende dalla carta di credito, dall'email o dall'indirizzo
  - perciò ha lo **storico degli acquisti** di ogni cliente
  - ed anche **dati demografici** acquisiti direttamente o acquistati
- Le spese per neonati e bambini sono un grande business
  - il supermercato avviò anni prima un progetto mirato ad “agganciare” i genitori imminenti, prima che diventino clienti della concorrenza
- Cos’è stato scoperto dall’analisi dei dati ?
  - le **donne incinta** (*stato desumibile anche a posteriori dall’età del figlio*) acquistano di **più lozioni inodore all’inizio del 2°trimestre di gravidanza**
  - nelle **prime 20 settimane anche integratori** di calcio, magnesio e zinco
  - l’aumento del consumo di **battufoli di cotone, disinfettanti per le mani e salviette** indicano che sono prossime al parto

Gianluca Moro - DISI, Università di Bologna

## Recommendation: Una Storia Vera (iii)

- **C’è di più ...**
  - sono stati **identificati circa 25 prodotti** che insieme permettono di **stimare la data di nascita** in un piccolo intervallo
  - ciò consente di proporre coupons diversificati in base alla fase della gravidanza
  - uno score predice lo stato di gravidanza in base agli acquisti:
    - *e.g. donna di 23 anni, a Marzo raddoppia il consumo di lozioni inodore, magnesio, zinco, acquista un tappeto blu brillante -> parto in Agosto 87%*
- il supermercato ha rilevato anche un **comportamento inatteso**
  - un alto numero di futuri genitori che ricevevano solo materiale pubblicitario per la maternità non diventavano clienti
  - distribuendo invece articoli di maternità in mezzo ad altri, ha procurato maggiori vendite
  - Perchè ? la conclusione fu che a **nessuno piace essere troppo “spiato”**
- Aumentato il fatturato del 50%

Gianluca Moro - DISI, Università di Bologna

# Predizione?

- Perché è utile?
- A quale scopo specifico?
- Cosa dovremmo misurare?
- Ricerca di variazioni inusuali
- Decisioni guidate dai dati
  - Basate sui *dati*, piuttosto che sull'*intuizione*

Gianluca Moro - DISI, Università di Bologna

15

## Perché "scavare" nei dati per apprendere? Punto di vista commerciale

- quantità enormi di dati
  - web data, e-commerce
  - acquisti
  - transazioni bancarie
- memorizzazione e calcolo potenti ed economici
- Elevata pressione competitiva
  - fornire ai clienti servizi migliori, personalizzati, tempestivi
    - ... Customer Relationship Management (CRM)

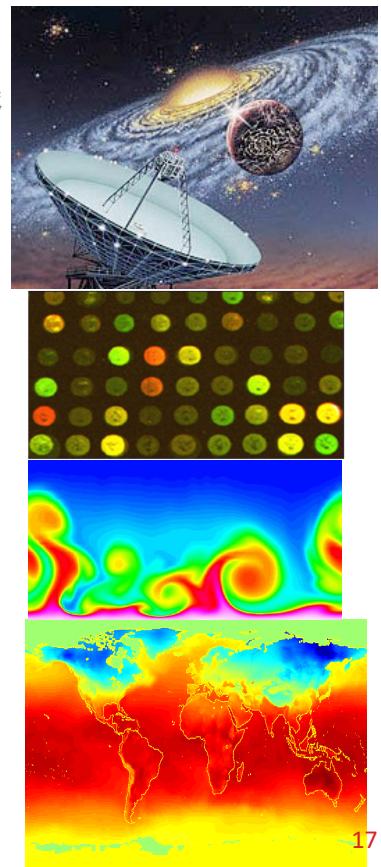
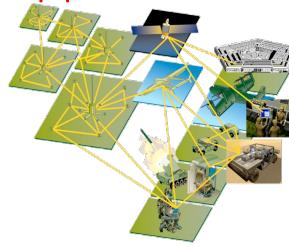


Gianluca Moro - DISI, Università di Bologna

16

## Perché "scavare" nei dati per apprendere? punto di vista scientifico

- I dati sono collezionati e memorizzati a velocità enormi (GB/hour)
  - sensori remoti su satelliti
  - telescopi
  - microarray elettronici che generano espressioni geniche
  - simulazioni scientifiche che generano terabyte di dati
- Le tradizionali tecniche di analisi non sono applicabili ai dati grezzi in tali quantità
- Data Science e Machine learning aiutano gli scienziati
  - classificazione e segmentazione dei dati
  - formulazione di ipotesi



Gianluca Moro - DISI, Università di Bologna

## MACHINE LEARNING - INTRODUZIONE

### Un primo successo del "machine learning"

- identificazione di malattie della soia
- erano disponibili 680 esempi di piante malate
- 35 attributi, ciascuno con un numero limitato di valori possibili
- ogni esempio di malattia era diagnosticato (etichettato) da un esperto di biologia vegetale
  - 19 categorie di malattia
- gli esperti sono soliti aiutarsi con un insieme di regole

## Malattie della soia

- Esempio di regole usate dall'esperto
  - se [la condizione della foglia è normale e la condizione dei germogli è anormale e l'ulcera è sotto la linea del terreno e l'ulcera è di colore marrone]  
allora la diagnosi è rhizoctonia root rot
  - se [è assente malformazione della foglia e la condizione dei germogli è anormale e l'ulcera è sotto la linea del terreno e l'ulcera è di colore marrone]  
allora la diagnosi è rhizoctonia root rot
- problema: regole non indipendenti
- prestazione delle regole se applicate alla lettera, senza l'aiuto di un esperto: 72%

## Generazione automatica di regole

- inizio anni '70
- i dati grezzi degli esempi sono stati trattati con un algoritmo di apprendimento
- le regole generate hanno dato luogo a una classificazione con accuratezza del 97.5%
- l'accuratezza è risultata migliore di quella di botanici non molto esperti

## Un altro tra i primi progetti di successo ...

- il dipartimento dell'agricoltura degli Stati Uniti ogni anno eroga indennizzi per danni da maltempo a centinaia di migliaia di agricoltori
- una frazione delle richieste di indennizzo è fraudolenta
- un'analisi a campione delle richieste per verificarne l'autenticità ha un costo molto elevato rispetto alla resa
- un progetto di data science volto a individuare le frodi ha reso oltre venti volte il suo costo

## Definizione di *knowledge discovery*

*The nontrivial extraction of implicit, previously unknown and potentially useful information from data*

W. Frawley, G. Piatetsky-Shapiro, and C. Matheus:  
 “Knowledge Discovery in Databases: An Overview”.  
 AI Magazine, Fall 1992, pgs 213-228

## Knowledge Discovery

- scoprire (e presentare) “conoscenza” in una forma facilmente comprensibile, e utilizzabile a scopi gestionali/decisionali
  - tecniche statistiche
  - di visualizzazione
  - di machine learning
- scalabilità
  - efficienza computazionale su DB di notevoli dimensioni (Giga-Tera bytes)

## Knowledge Discovery (ii)

- non solo algoritmi
  - processo complesso di manipolazione dei dati
- **data integration**
  - formato eterogeneo (ad es: rappresentano gli stessi dati con schemi differenti)
  - riconciliazione delle varie fonti
- **data cleaning**
  - dati affetti da “rumore” (errori, dati non interessanti, ecc.)
  - preprocessing per “pulire” i dati

## Estrazione di Conoscenza dai Dati - Data Mining

- i dati sono memorizzati digitalmente
- le ricerche sono assistite da computer
- la quantità di dati memorizzati nel mondo raddoppia ogni 20 mesi (stima)
- ⇒ scoprire informazioni nascoste nei database con processo (semi-)automatico
  - processo di estrazione/selezione/trasformazione
  - individuazione di *pattern strutturali*

## Learning

- apprendere (learn)
  - ottenere conoscenza tramite studio, esperienza, trasmissione
  - divenire consapevole tramite informazione o osservazione
  - basarsi sulla memoria
  - essere informato
  - ricevere istruzioni
- allenare (train)
  - apprendimento più passivo (animali, ...)
- definizione operazionale
  - *saper modificare il proprio comportamento in modo da migliorare le proprie prestazioni su problemi noti*
  - *saper risolvere problemi nuovi che prima dell'apprendimento non potevano essere risolti*

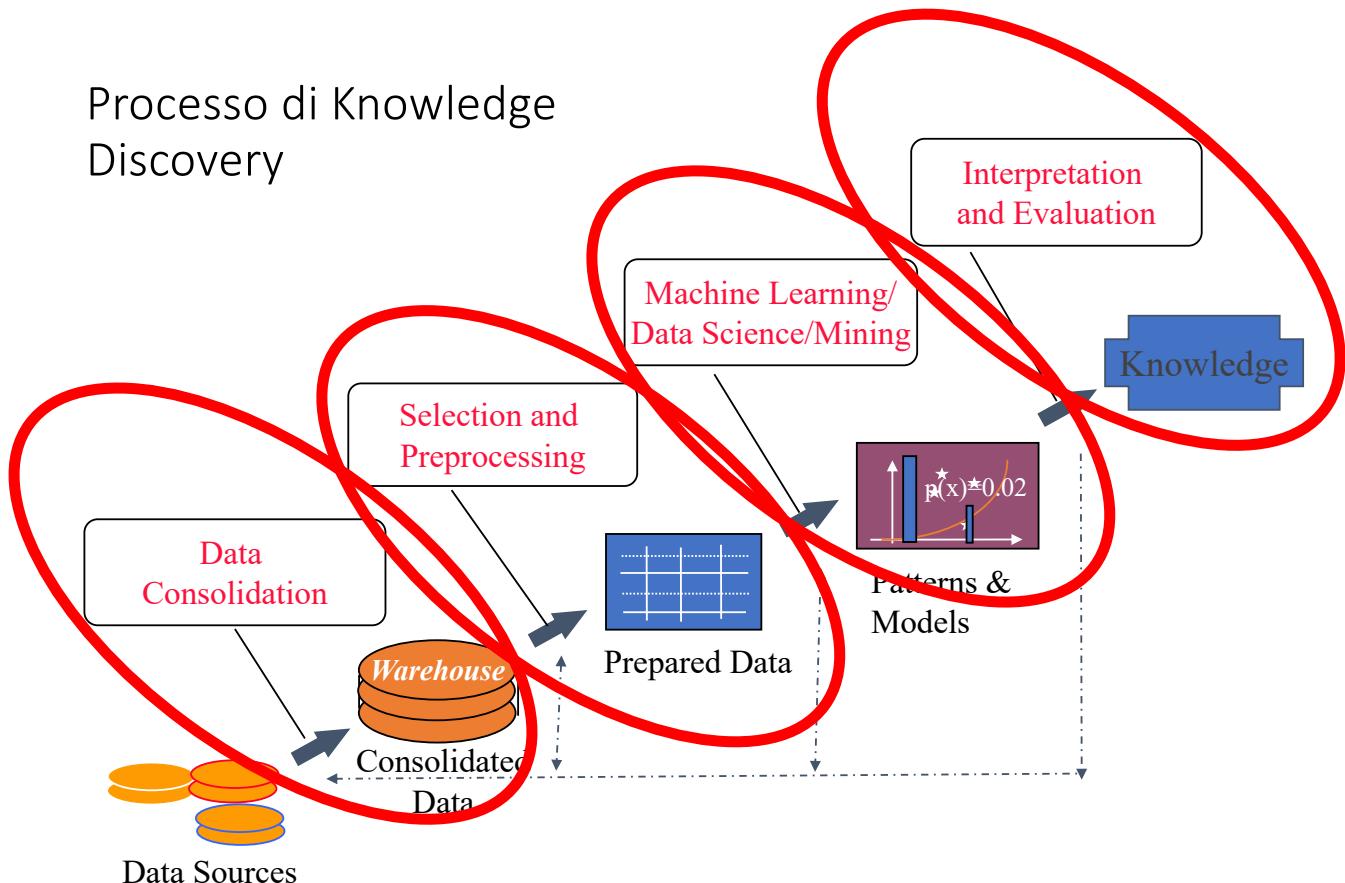
## Hanno detto ...

- Learning is constructing or modifying representations of what is being experienced (Michalski 1986)
- [Machine] Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time (Simon 1983)
- It's all very well to speak of adaptation over time, but how can this be quantified? (Stoutamire)
- We do believe that there is a process that explains the data we observe (Alpaydin 2004)

## Business Intelligence

- un insieme di [processi aziendali](#) per raccogliere ed analizzare informazioni strategiche
- la tecnologia utilizzata per realizzare questi processi,
- le informazioni ottenute come risultato di questi processi
- Questo termine è stato coniato da [Howard Dresner](#), analista del gruppo [Gartner](#), nei primi anni '90.

## Processo di Knowledge Discovery



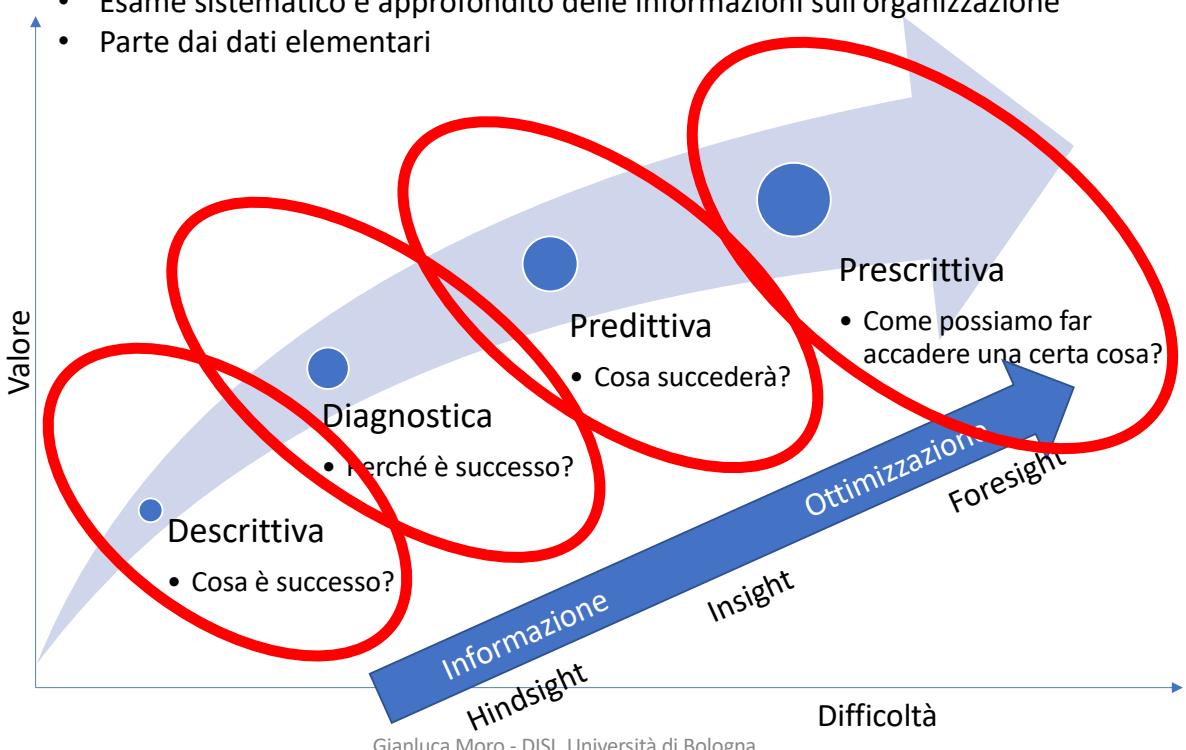
Gianluca Moro - DISI, Università di Bologna

29

## MACHINE LEARNING - INTRODUZIONE

### Analytics

- Esame sistematico e approfondito delle informazioni sull'organizzazione
- Parte dai dati elementari



Gianluca Moro - DISI, Università di Bologna

30

## Descriptive Analytics

- Aggregare i dati con tecniche DB
- Comprendere i dati
  - produrre report
  - produrre **statistiche descrittive**
  - individuare gruppi (clustering)
  - **profiling**

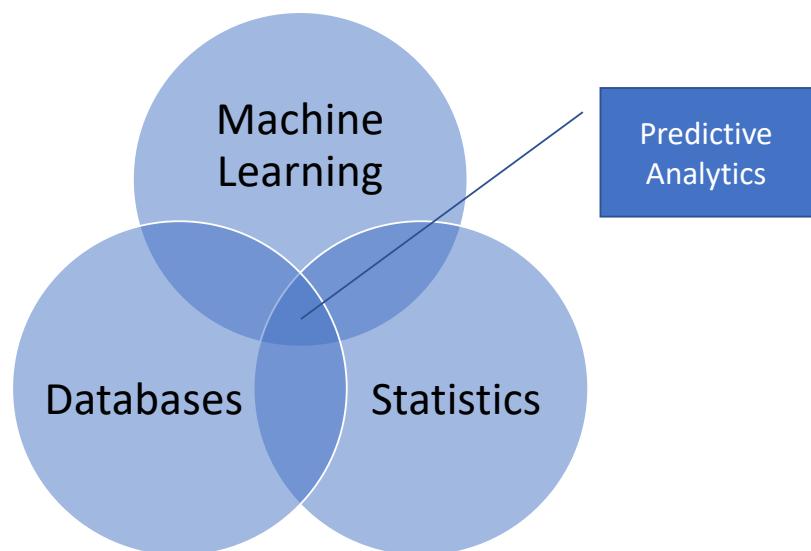
## Diagnostic Analytics

- Utilizzare insieme le **statistiche descrittive** e la **conoscenza di dominio**
- Comprendere le **cause** dei fenomeni
- Esempio:
  - I ricavi sono aumentati in Europa Orientale, la possibile causa è l'abbandono del nostro segmento di mercato da parte di un competitor

## Predictive Analytics

- Calcolare il valore di una determinata variabile in un momento **futuro**, data la disponibilità della storia di un determinato insieme di variabili
- Livello micro:
  - C'è il 60% di probabilità che il nostro più grande fornitore dell'Europa Orientale concluda un accordo con un nostro competitor il prossimo anno
- Livello macro:
  - I ricavi nell'Europa Orientale potrebbero aumentare fra il 6 e l'8% il prossimo anno

## Predictive Analytics



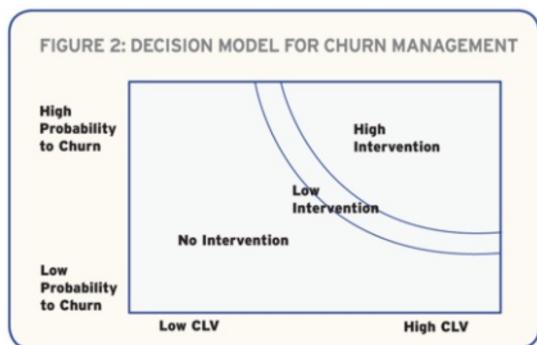
# Prescriptive Analytics

- Decidere le azioni da intraprendere per ottenere il comportamento futuro desiderato
- Scegliere tra varie opzioni e strategie
- Ottimizzazione

Gianluca Moro - DISI, Università di Bologna

35

# Prescriptive Analytics



Fonte: McKnight Consulting Group

FIGURE 3: DETERMINING MULTIPLE ACTIONS BASED ON CUSTOMER SEGMENTS

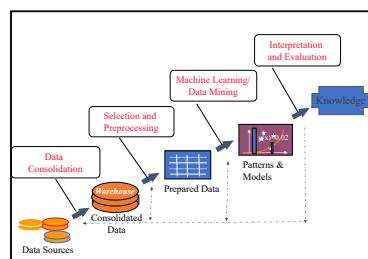
SEGMENT	ACTION
Greater than 50% probability to purchase and top 3 decile customer	Email and call
Greater than 25% probability to purchase and top 5 decile customer	Email
Unlikely to purchase and/or low decile customer	No promotion

Gianluca Moro - DISI, Università di Bologna

36

## Il ciclo virtuoso

Problema →



Conoscenza ↓

Identificare problema o opportunità

Agire in base alla conoscenza

Strategia ↑

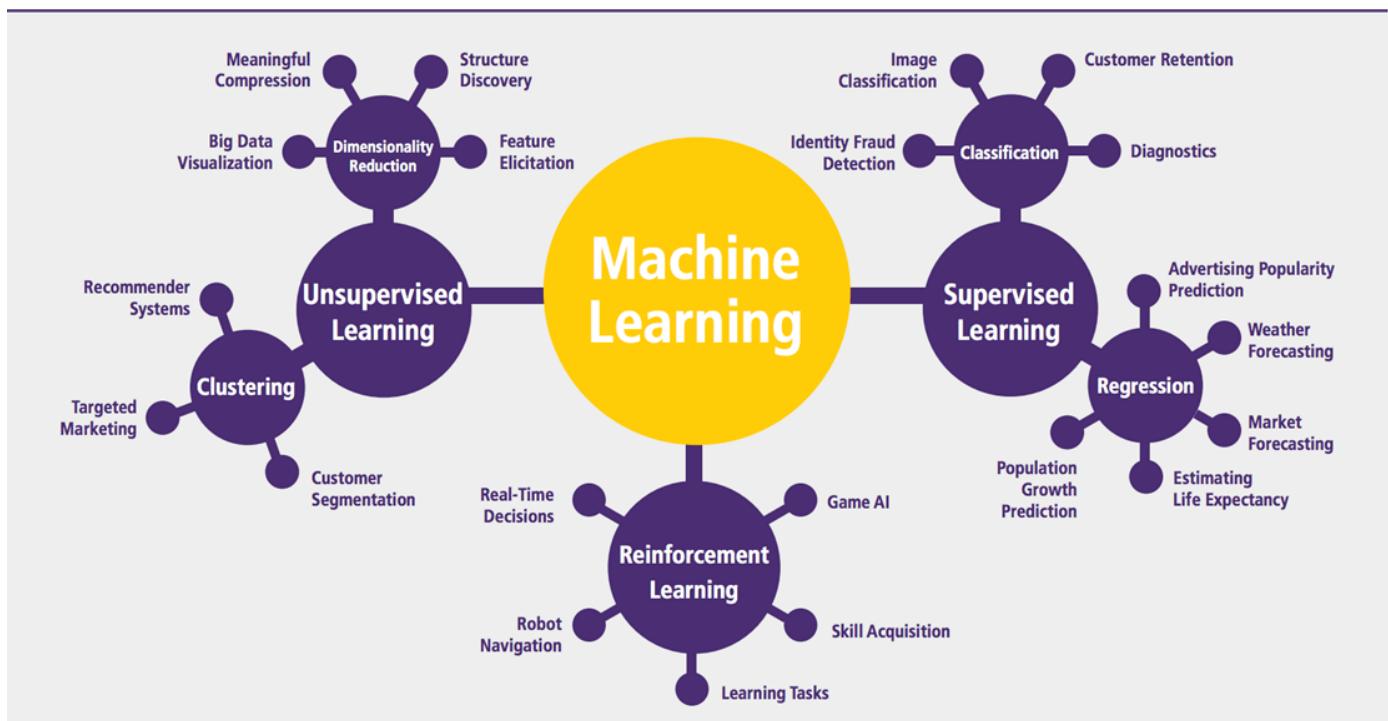
Misurare l'effetto dell'azione

Risultati ↓

Gianluca Moro - DISI, Università di Bologna

37

## Machine Learning a “colpo d’occhio”



Fonte: <https://medium.com/analytics-vidhya/which-machine-learning-algorithm-should-you-use-by-problem-type-a53967326566>

Gianluca Moro - DISI, Università di Bologna

## Tre tipi di learning + 1

- Supervisionato

- apprendere dall'esperienza di altri



Gianluca Moro - DISI, Università di Bologna

39

## Tre tipi di learning + 1

- Non Supervisionato

- scoprire *pattern*

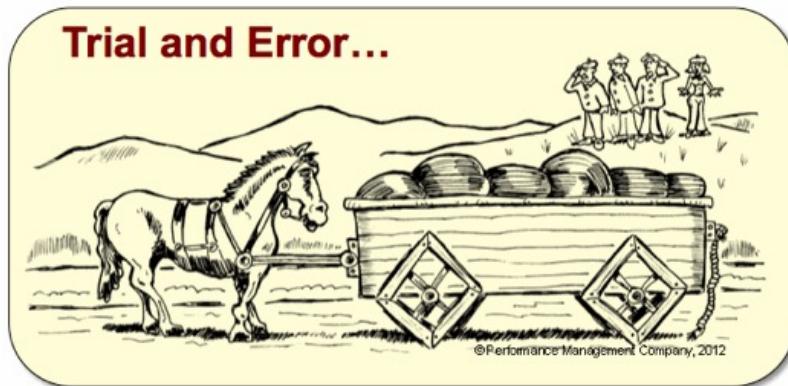


Gianluca Moro - DISI, Università di Bologna

40

## Tre tipi di learning + 1

- Con rinforzo (reinforcement)
  - prova e sbaglia



Gianluca Moro - DISI, Università di Bologna

41

## Tre tipi di learning + 1

- +1 = Self-Supervised
  - Algoritmi e metodi supervisionati applicati a problemi non supervisionati
  - *E.g. Apprendere il significato di parole e frasi con algoritmi di classificazione che predicono le parole volontariamente eliminate in modo casuale dal testo fornito in input*
  - *Quali sono i dati disponibili ? Qualsiasi documento, pagina web, post etc. ossia Terabyte di dati disponibili*
  - *Esempi: Language Model per l'elaborazione del linguaggio naturale in centinaia di lingue*
    - *BERT (Bidirectional Encoder Representation from Transformers) 2500 milioni di parole di training*
  - Modelli alla base di quasi tutti i task moderni di Natural language processing
    - Text classification by topic, sentiment analysis e opinion mining, information extraction, named entity recognition, semantic parsing, text summarization, question answering, chatbot ...

Gianluca Moro - DISI, Università di Bologna

42

## Compiti del learning

- Metodi predittivi
  - uso di alcune variabili per predire valori sconosciuti o futuri di altre variabili
- Metodi descrittivi
  - trovare schemi (pattern) che descrivono i dati in forma facilmente interpretabile dagli utenti

Gianluca Moro - DISI, Università di Bologna

43

## Metodi...

### Supervisionati

- c'è un *target* specifico
- occorre ricavare dai dati un meccanismo che permetta di derivare informazioni sul target a partire dalle altre informazioni disponibili

### Non Supervisionati

- non c'è un *target* specifico
- occorre fare emergere *spontaneamente* un **modello dai dati**

# Task non supervisionati

## Metodi

- Clustering
- Scoperta di regole associative
- Riduzione di dimensionalità non supervisionata
  - Esempio: analisi componenti principali (PCA)

## Task

- Profiling
- ....

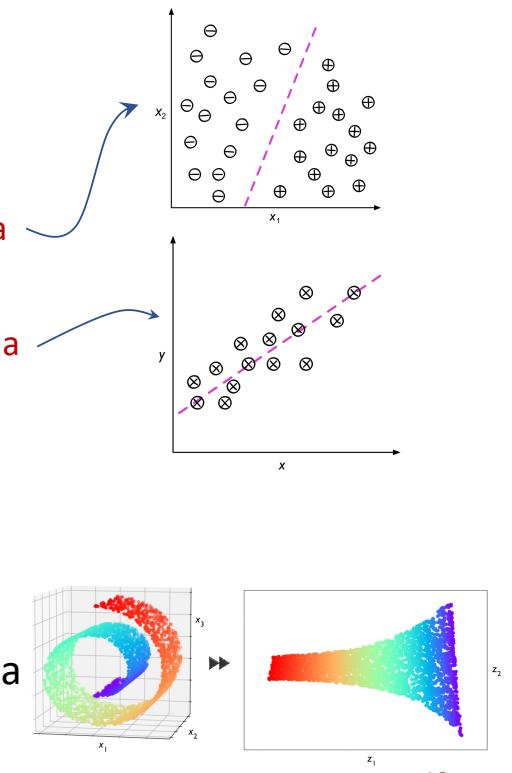
# Task Supervisionati

## Metodi

- Classificazione
  - prevedere il valore di una variabile **discreta**
- Regressione
  - prevedere il valore di una variabile **continua**

## Task

- Individuazione di anomalie
- Similarity matching
- Predizione di link
- Riduzione dimensionalità supervisionata
  - esempio: Linear Discriminant Analysis



# Metodi Supervisionati / Non Supervisionati

- tecniche sostanzialmente differenti
- caratteristica del problema e/o dei dati → non è un elemento di progetto
- informazione supervisionata ← aggiunta di attributo a ciascun individuo

# Metodi Supervisionati / Non Supervisionati

Come ottenere informazione supervisionata?

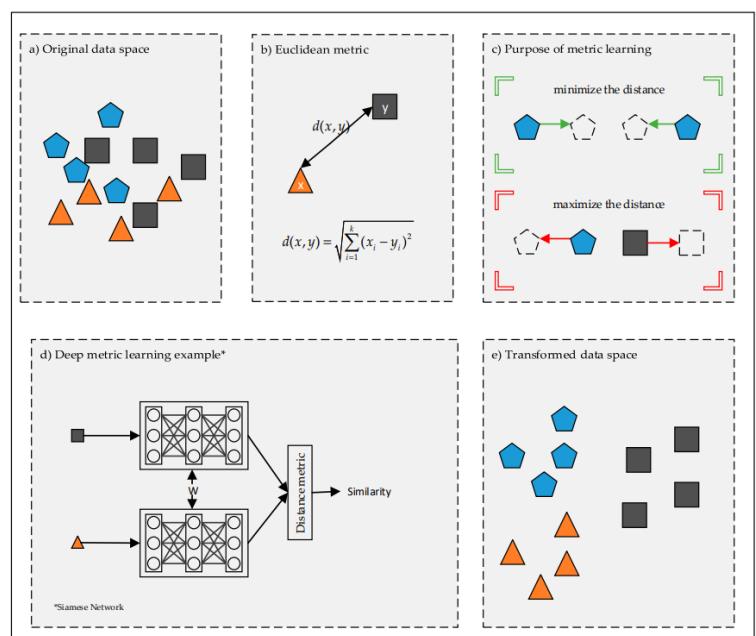
- etichettatura di individui da parte di *esperti*
- storia
  - l'informazione non è disponibile quando devo decidere cosa fare, ma col tempo vedrò quello che succede
  - dalla storia passata apprendiamo come prevedere l'attributo supervisionato che al momento non è disponibile
- I dati supervisionati contengono relazioni nascoste tra gli attributi che usiamo come input e l'attributo target da predire
- Un dataset è non supervisionato se non contiene l'attributo che interessa predire, ma i dati contengono comunque altre relazioni
  - addestrare ad apprendere queste relazioni spesso ci permette comunque di risolvere il problema

## Esempio di Problema non Supervisionato Risolvibile con Metodi Supervisionati – Self-Supervision

- Un'azienda intende realizzare un sistema di information retrieval, i.e. un motore di ricerca, capace di rispondere correttamente alle query degli utenti
  - ad esempio per la ricerca di prodotti da una breve descrizione
- Il problema sarebbe supervisionato se l'azienda avesse un elenco di query degli utenti e per ciascuna avesse associata la lista di k prodotti ordinata in modo decrescente di rilevanza
- L'azienda non ha un dataset del genere, per cui il problema è non supervisionato e non possiamo usare algoritmi supervisionati
- ma nei dati ci sono relazioni, quali ?
  - ogni prodotto ha un nome associato ad una descrizione e ad una o più immagini
  - che succede se addestriamo un modello di machine learning a predire queste relazioni ?
  - **DEEP METRIC LEARNING** (prossima slide)

## Deep Metric Learning

- Deep Metric Learning (DML) originates from the introduction of *deep learning* in *metric learning*
- It uses deep architectures for creating the embeddings of the samples
- The three main parts of a DML model are:
  - Input sampling
  - Structure of the network model
  - Metric loss function



# Predizione

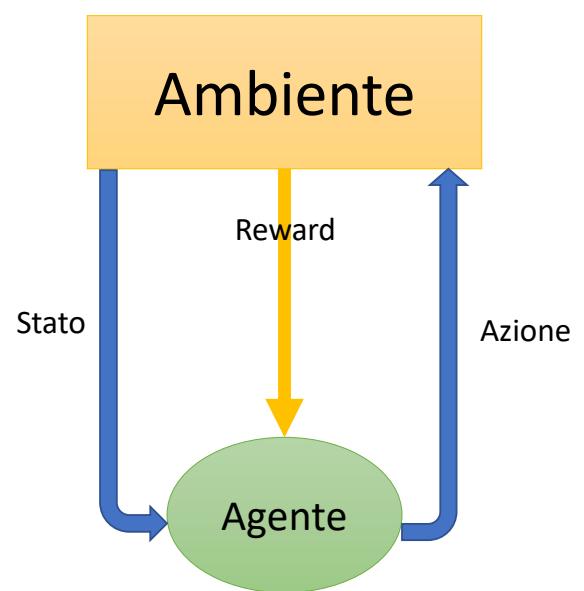
- Più dati → migliori predizioni
  - andare indietro nel tempo
- Devo convivere con l'impossibilità di raggiungere il 100% di accuratezza
- Fare “un po' meglio” può generare un grande valore di business
  - Las Vegas è stata costruita sul 51%



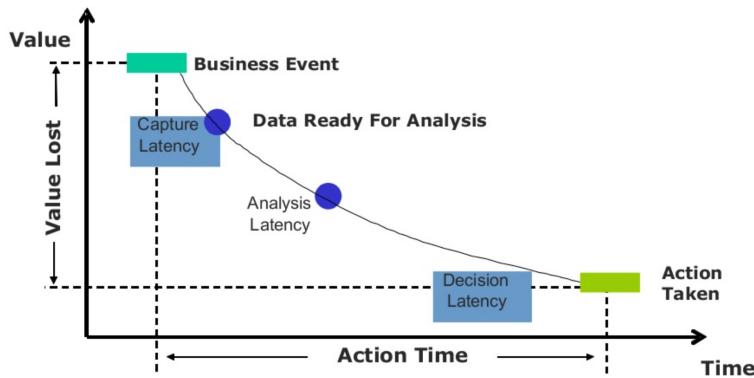
# Reinforcement Learning

Un altro *approccio* di ML

- Agente preposto a *prendere decisioni*
- Riceve premi o penalità come conseguenza delle sue azioni
- Modifica le sue strategie in base a premi/penalità
- Apprende la sequenza di azioni che massimizza il totale dei premi



# Real Time Analytics



Richard Hackathorn, Bolder Technology, Inc.

Real-time analytics aim at shortening the period of time between the occurrence of a business event that requires an appropriate action by the organization and the time the action is finally carried out. According to Hackathorn (2002), the additional business value of an action decreases, the more time elapses from the occurrence of the event to taking action. The elapsed time is called action time and can be seen as the latency of an action. Action time comprises three components:

- Data latency is the time from the occurrence of the business event until the data is stored and ready for analysis.
- The time from the point when data is available for analysis to the time when information is generated out of it is called analysis latency.
- Decision latency is the time it takes from the delivery of the information to taking the appropriate action. This type of latency mostly depends on the time the decision makers need to decide and implement their decisions. Value of reducing the action time In order to reduce action time, latency has to be reduced in all of the three components.

53

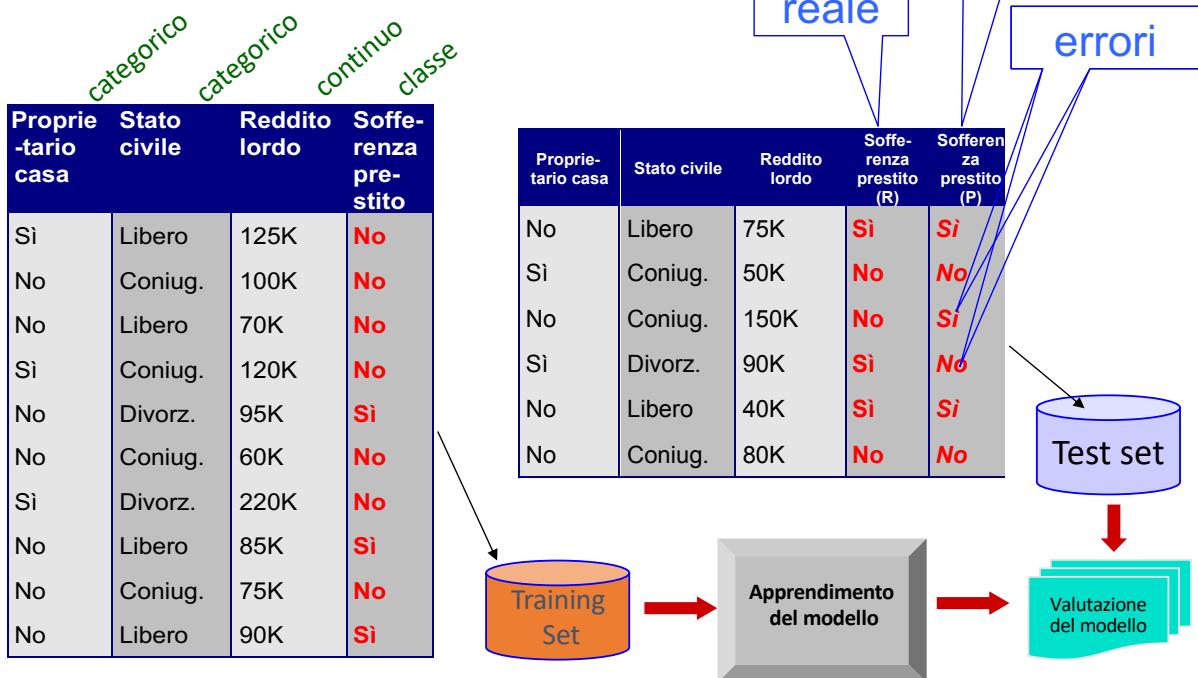
Gianluca Moro - DISI, Università di Bologna

Classificazione

## Classificazione: Definizione

- Data una collezione di record (*training set*)
  - Ogni record contiene valori per un insieme di **attributi**, uno di questi attributi è la **classe** (valori discreti)
- Trovare un **modello** per l'attributo classe come funzione dei valori degli altri attributi (**preditori**)
- Obiettivo: assegnare con la massima accuratezza la classe (ignota) di record **non visti in precedenza**
- Obiettivo 2: convalidare il modello stimandone l'accuratezza
  - Si utilizza un **test set** di record la cui classe è nota e viene confrontata con le predizioni del modello

# Esempio di generazione di un classificatore



Gianluca Moro - DISI, Università di Bologna

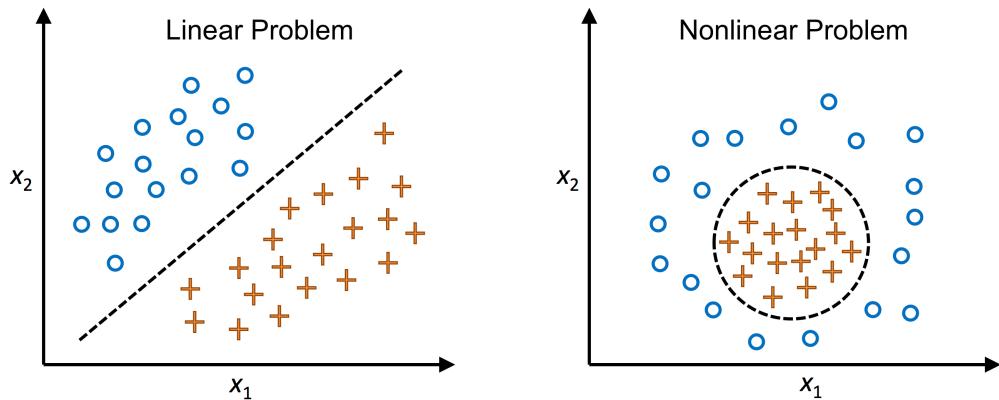
55

## Classificazione

### Tipi di classificazione

- **Binaria**
  - la classe ammette soltanto due valori: convenzionalmente *vero* e *falso*
- **Multiclasse**
  - la classe ammette più di due valori, ogni individuo ha un solo valore di classe
- **Multi-etichetta (i.e. multi-label)**
  - ogni individuo può avere contemporaneamente più valori di classe,
  - esempio: gli argomenti di un testo
- **Crisp**
  - predice un valore di classe
- **Probabilistica**
  - assegna a ogni classe una probabilità
  - può essere convertita in *crisp* scegliendo un *valore soglia*

# Tipi di classificazione – lineare e non lineare



## Esempio con due classi

- Nel primo caso il modello è un iperpiano  $y = w_1 \cdot x_1 + w_2 \cdot x_2 + b$
- Nel secondo è  $y^2 = w_1 \cdot x_1^2 + w_2 \cdot x_2^2 + b$

Gianluca Moro - DISI, Università di Bologna

57

## Classificazione - Valutazione

# Valutazione di un modello di classificazione

- **Accuratezza:** percentuale di predizioni corrette
- Per classificazione binaria
  - **precisione:** percentuale di veri positivi rispetto al numero di predizioni positive
    - $TP/(TP+FP)$
  - **recupero (recall):** percentuale di veri positivi rispetto al numero di positivi presenti nel dataset
    - $TP/(TP+FN)$

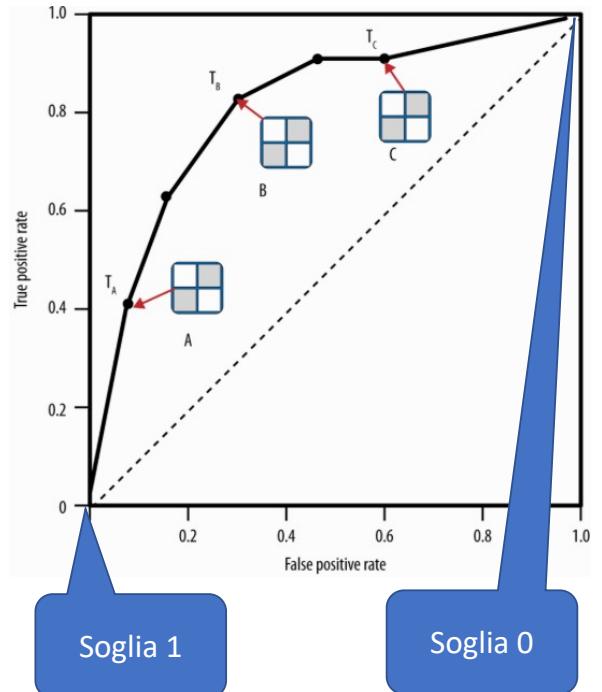
matrice di confusione

	<b>p</b>	<b>n</b>
<b>Y</b>	True positives	False positives
<b>N</b>	False negatives	True negatives

Y = predetto positivo  
N = predetto negativo

# Valutazione di un modello di classificazione

- Se si dispone di un classificatore probabilistico, variando il valore di soglia
  - si modifica la matrice di confusione
  - si modificano precisione e recupero
    - abbassando la soglia aumenta il recupero (TP) ma anche il tasso di falsi positivi (FP, quindi diminuisce la precisione)
- Curva ROC
  - ogni valore di soglia disegna un punto della curva



Gianluca Moro - DISI, Università di Bologna

Classificazione

## Applicazione 1

- Marketing diretto
  - Obiettivo: ridurre il costo postale delle promozioni selezionando un insieme di clienti a maggiore probabilità di acquisto di un nuovo telefono cellulare
  - Approccio:
    - Usare i dati relativi a un prodotto similare trattato in precedenza
    - Sappiamo quali clienti hanno acquistato o no
      - {acquisto, non acquisto} è l'attributo di classe
    - Raccogliamo informazioni sui clienti
      - demografia, reddito, stili di vita, abitudini di acquisto
    - Usiamo queste informazioni come attributo di input per costruire il modello di classificazione

Gianluca Moro - DISI, Università di Bologna

## Applicazione 2

- Individuazione di frodi
  - Obiettivo: predire casi fraudolenti nelle transazioni delle carte di credito
  - Approccio:
    - usare le transazioni passate e le informazioni sui clienti
      - quando acquista, dove acquista, regolarità nei pagamenti, ...
    - etichettare le transazioni passate come lecite o fraudolente
    - apprendere un modello di classificazione
    - usare il modello come preditore delle nuove transazioni
- Problema: sbilanciamento dei dati supervisionati
  - Le frodi rappresentano una piccola frazione dei dati

Gianluca Moro - DISI, Università di Bologna

## Applicazione 3

- Customer Attrition/Churn:
  - Obiettivo: prevedere se un cliente potrebbe essere perso a favore della concorrenza
  - Approccio
    - usare le registrazioni delle interazioni del cliente con l'azienda per trovare attributi adatti
      - con che frequenza chiama il servizio clienti, a che ora e in quale giorno chiama, stato finanziario, stato civile, ...
    - etichettare il cliente come leale/non leale
    - costruire un modello di "lealtà"

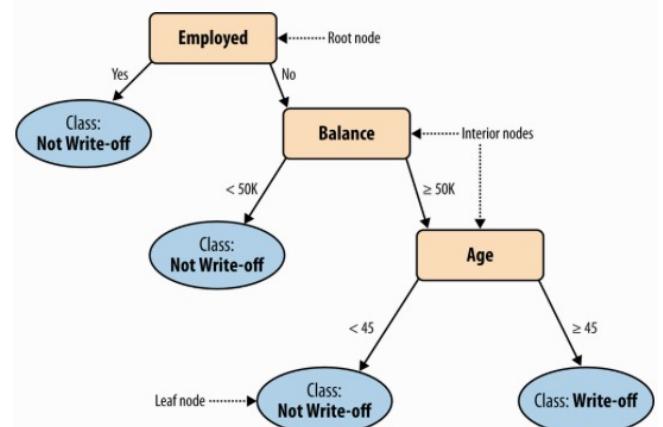
## Applicazione 4

- Catalogazione di osservazioni celesti
  - obiettivo: predire la classe di oggetti celesti (stella o galassia) in base a immagini telescopiche (osservatorio Palomar)
    - 3000 immagini per oggetto con 23,040 x 23,040 pixels per immagine.
  - Approccio:
    - segmentare l'immagine
    - misurare attributi dell'immagine (features – 40 per oggetto)
    - costruire un modello della classe
    - Storia di successo:
    - Individuati 16 nuovi quasar, lontanissimi e molto difficili da individuare

Gianluca Moro - DISI, Università di Bologna

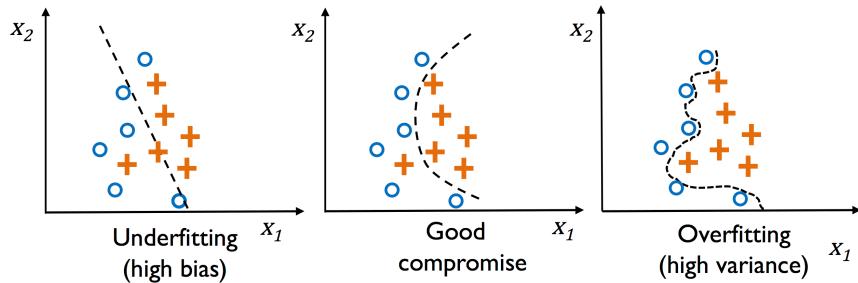
## Albero di decisione

- Classificazione mediante albero
- i test (yes/no oppure  $\geq \leq$ ) sono individuati cercando gli attributi maggiormente informativi/divisivi rispetto alla classe
- Non solo classificazione: Regression tree per fare regressione (slide dopo)



Gianluca Moro - DISI, Università di Bologna

# Overfitting



- Eccessivo adattamento del modello ai dati di test
- Quando si verifica overfitting, il classificatore avrà un comportamento nel "*mondo reale*" peggiore di quanto era stato previsto in fase di test
- Vari metodi per ridurre/eliminare l'overfitting
  - Semplificare il modello mediante tecniche di "regolarizzazione"
  - Ridurre le dimensioni del training set a favore del test set
  - Ridurre il numero di variabili di input

Gianluca Moro - DISI, Università di Bologna

# Regessione: Definizione

- Data una collezione di record (***training set***)
  - Ogni record contiene valori per un insieme di attributi (***predittori***), un altro attributo (***predetto***) non è noto al momento di prendere una decisione
  - Il suo valore è **continuo**
- Trovare un **modello** per il ***predetto*** come funzione dei ***predittori***
- Obiettivo: assegnare con la massima accuratezza il valore dell'attributo predetto (**ignoto**) per record **non visti in precedenza**
- Obiettivo 2: convalidare il modello stimandone l'accuratezza
  - Si utilizza un ***test set*** di record il cui predetto è noto e viene confrontato con le predizioni del modello

## Caratteristiche ed esempi

- Modelli di dipendenza lineari e non lineari
- Ampiamente studiata in statistica
- Esempi
  - Predire l'ammontare delle vendite di un nuovo prodotto in base alle spese in pubblicità
  - Predire la velocità del vento come funzione di temperatura, pressione, umidità, ...
  - Predire l'andamento nel tempo di indici di borsa

Gianluca Moro - DISI, Università di Bologna

## Clustering: Definizione

- Dato un insieme di punti (dati), ciascuno con un insieme di valori negli attributi, e una *misura di similarità* fra i punti, trovare i raggruppamenti tali che
  - i punti all'interno dello stesso cluster sono "più simili" gli uni agli altri
  - i punti in cluster diversi sono "meno simili"
- Misure di similarità
  - Distanza euclidea in spazi vettoriali
  - Altre misure specifiche per singoli problemi

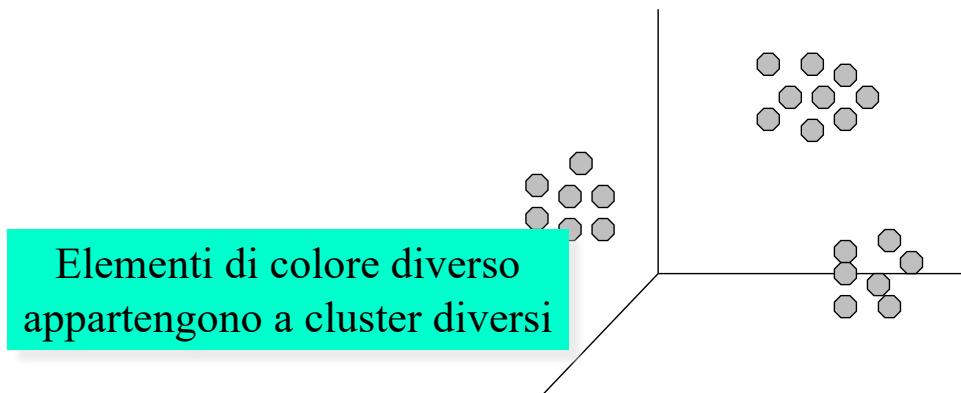
Gianluca Moro - DISI, Università di Bologna

## Esempi

I Clustering in spazio 3d sulla base della distanza euclidea

minimizzazione delle  
distanze intra-cluster

massimizzazione delle  
distanze inter-cluster



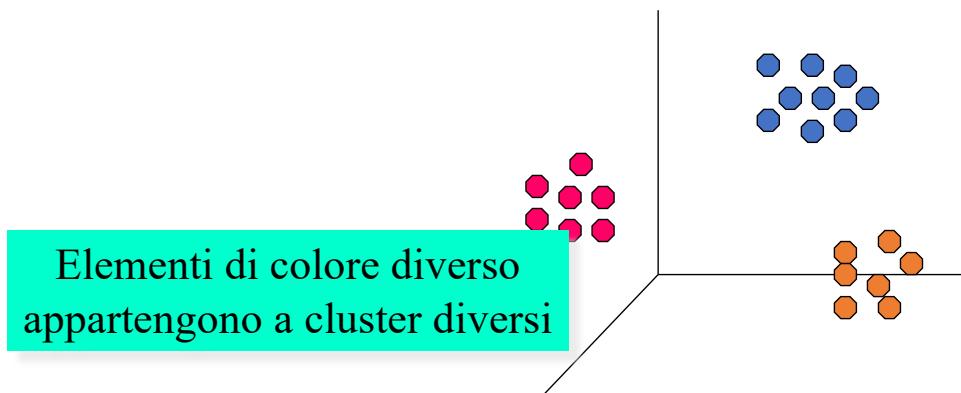
Gianluca Moro - DISI, Università di Bologna

## Esempi

I Clustering in spazio 3d sulla base della distanza euclidea

minimizzazione delle  
distanze intra-cluster

massimizzazione delle  
distanze inter-cluster



Gianluca Moro - DISI, Università di Bologna

# Valutazione

- Misurare la **coesione**
  - quanto gli oggetti all'interno di uno stesso cluster sono simili l'uno all'altro
- Misurare la **separazione**
  - quanto i cluster sono separati gli uni dagli altri

Gianluca Moro - DISI, Università di Bologna

# Applicazione 1

- Segmentazione di mercato:
  - Obiettivo: suddividere un insieme di clienti in sottoinsiemi distinti ciascuno dei quali potrebbe essere destinatario di una specifica politica di marketing
  - Approccio
    - collezionare gli attributi dei clienti (anagrafe, stili di vita)
    - trovare cluster di clienti simili
    - misurare la qualità dei cluster rispetto agli obiettivi, confrontando i pattern di acquisto di clienti nello stesso cluster e in cluster diversi

Gianluca Moro - DISI, Università di Bologna

## Applicazione 2

- Clustering di documenti:
  - Obiettivo: trovare gruppi di documenti simili basandosi sulle parole che contengono
  - Approccio: identificare le parole che compaiono con maggiore frequenza (ignorando però le **stopword**: parole ovvie, come congiunzioni, articoli, verbi generici), calcolare una misura di similarità in base alla frequenza dei termini, usare la similarità per il clustering
  - Risultato: le tecniche di "information retrieval" possono usare i cluster per collegare un nuovo documento

Gianluca Moro - DISI, Università di Bologna

## Esempio di Document Clustering

- Elementi: 3204 Articles del Los Angeles Times.
- Misura di similarità: quante parole sono in comune (dopo il filtraggio delle stopword)

<b>Category</b>	<b>Total Articles</b>	<b>Correctly Placed</b>
<b>Financial</b>	555	364
<b>Foreign</b>	341	260
<b>National</b>	273	36
<b>Metropolitan</b>	943	746
<b>Sports</b>	738	573
<b>Entertainment</b>	354	278

Gianluca Moro - DISI, Università di Bologna

# Clustering di S&P 500 Stock Data

- Osservare gli scambi di borsa ogni giorno
- Punti di clustering: area-{UP/DOWN}
- Misura di Similarità: due punti sono più simili se gli eventi che descrivono sono frequentemente insieme nello stesso giorno
  - quantificabile con le regole associative

	<i>Discovered Clusters</i>	<i>Industry Group</i>
<b>1</b>	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Telabs-Inc-Down, Natl-Semiconduct-DOWN,Oracle-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
<b>2</b>	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOW N,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
<b>3</b>	Fannie-Mae-DOWN,Fed-Ho me-Loan-DOWN, MBNA-Corp -DOWN,Morgan-Stanley-DOWN	Financial-DOWN
<b>4</b>	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Gianluca Moro - DISI, Università di Bologna

MACHINE LEARNING - INTRODUZIONE

## Scoperta di regole associative: Definizione

- Dato un insieme di record ciascuno dei quali contiene degli oggetti presi da una collezione data
  - produrre regole di dipendenza che predicono la presenza di un oggetto in base alla presenza di altri

<i>TID</i>	<i>Items</i>
<b>1</b>	<b>Bread, Coke, Milk</b>
<b>2</b>	<b>Beer, Bread</b>
<b>3</b>	<b>Beer, Coke, Diaper, Milk</b>
<b>4</b>	<b>Beer, Bread, Diaper, Milk</b>
<b>5</b>	<b>Coke, Diaper, Milk</b>

Regole scoperte:  
 $\{\text{Milk}\} \rightarrow \{\text{Coke}\}$   
 $\{\text{Diaper}, \text{Milk}\} \rightarrow \{\text{Beer}\}$

## Applicazione 1

- Marketing e vendite promozionali
  - supponiamo di aver scoperto la regola  
 $\{Bagels, \dots\} \rightarrow \{Potato\ Chips\}$
  - Potato Chips come conseguente=> Può essere usato per determinare cosa dovrebbe essere fatto per incrementarne le vendite
  - Bagels nell'antecedente=> può essere usato per vedere quali prodotti potrebbero essere influenzati se il negozio cessa la vendita di Bagels
  - Bagels nell'antecedente e Potato chips nel conseguente=> Può essere usato per vedere quali prodotti dovrebbero essere venduti con i Bagels per promuovere le vendite di Potato chips!

Gianluca Moro - DISI, Università di Bologna

## Applicazione 2

- Gestione degli scaffali dei supermercati
  - Obiettivo: Identificare gli oggetti acquistati insieme da un numero sufficientemente elevato di clienti
  - Approccio: Trattare i dati del punto vendita raccolti alle casse per trovare dipendenze tra gli oggetti
  - Una regola classica --
    - Se un cliente acquista pannolini per bambini e latte è probabile che acquisti anche birra (!?!)
      - (tipico del mercato americano...)

Gianluca Moro - DISI, Università di Bologna

## Applicazione 3

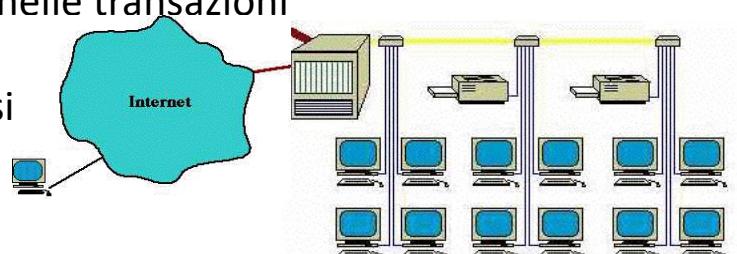
- Gestione dell'assortimento
  - Obiettivo: una società di riparazione attrezzi vuole prevedere la natura delle riparazioni che saranno richieste dai clienti e mantenere i veicoli di servizio equipaggiati con i giusti pezzi di ricambio, per ridurre il numero di visite necessarie per risolvere una chiamata
  - Approccio: trattare i dati su strumenti e parti impiegati in precedenti riparazioni e scoprire gli schemi di "co-occorrenza"

Gianluca Moro - DISI, Università di Bologna

Altre funzioni

## Deviazione/Individuazione di anomalie

- Trovare variazioni significative rispetto al comportamento normale
- Applicazioni:
  - Individuazione di frodi nelle transazioni di carte di credito
  - Individuazione di intrusi



Gianluca Moro - DISI, Università di Bologna

## Similarity matching

- Trovare individui simili ad altri individui
- Esempi
  - Un'azienda potrebbe voler trovare potenziali clienti che abbiano caratteristiche simili ai suoi attuali clienti migliori
  - Recommendation
    - Amazon
    - ...

Gianluca Moro - DISI, Università di Bologna

## Profiling

- Descrizione del comportamento
- Comportamento tipico di un individuo o di un gruppo di individui
- Esempio
  - Qual'è l'utilizzo tipico di smartphone di un certo segmento di clientela?
- Utilizzato anche per l'individuazione (per contrasto) delle anomalie
  - Es: frodi nelle carte di credito

Gianluca Moro - DISI, Università di Bologna

## Predizione di link

- Principale applicazione nelle reti sociali
  - Poiché Andrea e Marta condividono 10 amici potrebbero essere a loro volta amici
- Scoprire nuovi principi attivi per lo sviluppo di nuovi farmaci mediante previsione di nuove molecole rappresentate con grafi (nodi e link)
- Esistono grafi che modellano parte delle conoscenze umane per diverse discipline
  - E.g. Unified Medical Language System (UMLS)
  - Incremento di UMLS mediante estrazione automatica di conoscenza riportata negli articoli scientifici pubblicati in PUBMED (27 milioni di articoli con tasso di incremento esponenziale negli ultimi vent'anni)

Gianluca Moro - DISI, Università di Bologna

## Riduzione di dimensioni

- Utile, spesso necessario in dataset molto grandi
- Individuare un sottoinsieme degli attributi che garantisce le stesse prestazioni dell'intero insieme
- Obiettivi:
  - Facilità di elaborazione
  - Maggiore comprensibilità
- Compromesso tra semplificazione e perdita di informazione

Gianluca Moro - DISI, Università di Bologna

## ML e Industry 4.0

- Sensori
- Connessione
- Conoscenza approfondita dello *stato* dei sistemi produttivi
- Condizioni ideali per applicare tecniche di ML

Manutenzione predittiva

Ottimizzazione

Gianluca Moro - DISI, Università di Bologna

MACHINE LEARNING

## Rischi del Machine Learning

Valutazione della **qualità** dei modelli

- Previsione dei **costi** di utilizzo
  - Quanto valore mi daranno le predizioni corrette?
  - Quanto mi costeranno le predizioni errate?

Gianluca Moro - DISI, Università di Bologna

## Rischi (i)

### Rappresentatività dei dati

- I dati su cui fondo l'apprendimento sono rappresentativi dei dati che rileverò in condizioni di utilizzo?
  - Es: modello di rischio finanziario usato in aree in cui la situazione economica generale è diversa da quella dei dati usati per generare il modello stesso
  - Es: situazione che cambia nel tempo rispetto ai dati usati per generare il modello

Gianluca Moro - DISI, Università di Bologna

## Rischi (ii)

### Analisi errate

- Indispensabile curare la fase di valutazione

### Privacy

- Quando i dati coinvolgono le persone la privacy deve essere garantita
- Questo può generare costi/conflitti

Gianluca Moro - DISI, Università di Bologna

## Rischi (iii)

### Costi e *commitment*

- I costi, in particolare quelli di acquisizione/preparazione dei dati, non devono essere sottostimati
- È necessario *commitment* da parte degli *stakeholders* per garantire il completamento del progetto e il corretto inserimento nel contesto di destinazione

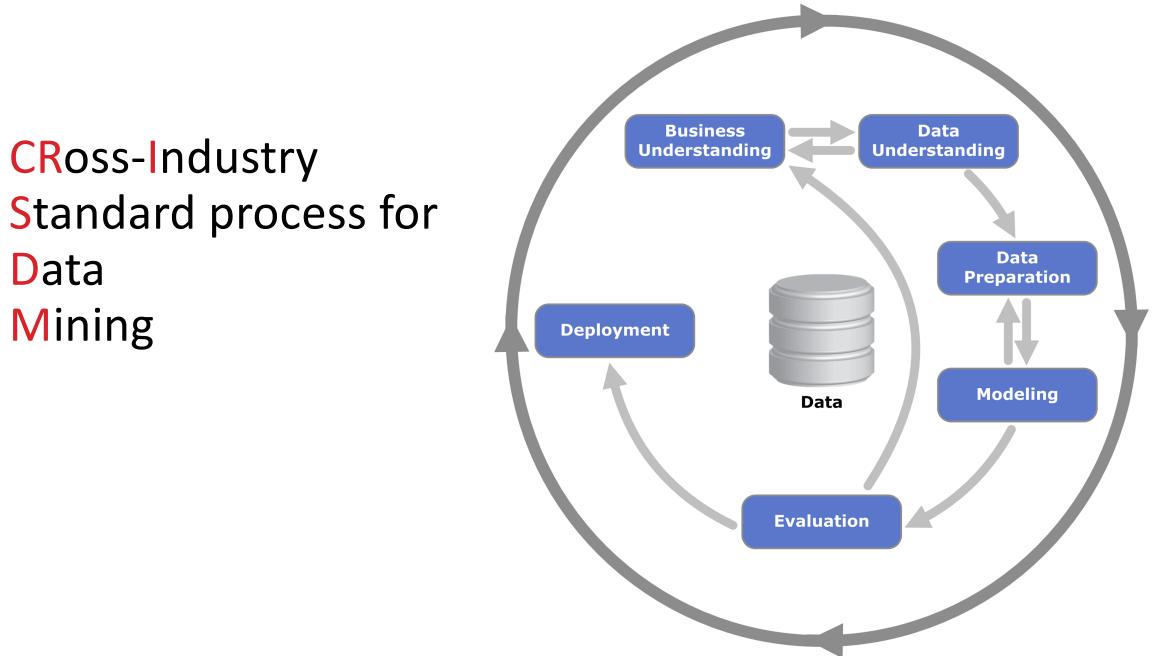
Gianluca Moro - DISI, Università di Bologna

# Metodologia di Sviluppo di Progetti di Data Science

Breve introduzione

## CRISP-DM

[https://en.wikipedia.org/wiki/Cross\\_Industry\\_Standard\\_Process\\_for\\_Data\\_Mining](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)



91

Gianluca Moro - DISI, Università di Bologna

## Perché è necessaria una metodologia?

- Linee guida per condurre il processo
- Standardizzazione dei passi da seguire
- Documentazione del processo

## Business Understanding

- Riformulare il problema in molti modi, secondo necessità
- Pensare allo scenario applicativo
- Raffinamento iterativo della formulazione del problema e dello scenario

## Data Understanding

- Quali dati grezzi sono disponibili?
  - *raramente corrispondono esattamente alle esigenze*
  - *generalmente sono collezionati per scopi diversi*
    - *es: db clienti, db transazioni, db marketing hanno generalmente un'intersezione, ma non sono coincidenti*
      - *possono avere anche gradi diversi di affidabilità*
- Costo di accesso ai dati?
  - Campagna di raccolta ad hoc per integrare i dati?
- Possibili orientamenti diversi del progetto a seconda dei dati trovati

## Data Preparation

- Tecniche di analisi applicabili soltanto a dati di un determinato tipo
- Alcune trasformazioni possono produrre migliore qualità dei risultati
- Dati mancanti o incompleti
- Generalmente questa è la fase più costosa

## Modellazione

- Trovare i pattern nascosti nei dati
  - usare una o più delle funzioni del machine learning
- Mettere a punto e validare il modello di analisi

## Valutazione

- Stimare la qualità del risultato
- Confrontare diverse scelte di modellazione sia dal punto di vista *qualitativo* che *quantitativo*
  - usare anche tecniche statistiche di stima
- Stimare l'impatto sul business
  - Quante decisioni sbagliate posso aspettarmi?
  - Di quale tipo?
  - Con quale impatto sul business?

## Deployment

- Inserire i modelli prodotti nei sistemi software che assistono il processo aziendale su cui si vuole operare
  - es: un modello di predizione degli abbandoni può essere inserito in un sistema CRM per fare offerte allettanti ai clienti

# Strumenti software

## Open source

- Python – Scikit-Learn
- R
- Tensorflow
- Keras
- Pytorch
- Apache Spark
- WEKA

## Commercial

- Rapid Miner
- *SQL*
- *Excel*
- Tableau
- Microsoft ML server
- IBM Watson
- IBM SPSS Modeler
- SAS Enterprise Miner
- KNIME
- MATLAB

Gianluca Moro - DISI, Università di Bologna

[Informazioni generali](#)

## Sorgenti dal web

- KD Nuggets: <https://www.kdnuggets.com>
- Kaggle: <https://www.kaggle.com>
- DataCamp: <https://www.datacamp.com>
- Data Science Central: <https://www.datasciencecentral.com>

Gianluca Moro - DISI, Università di Bologna

# Deep Learning

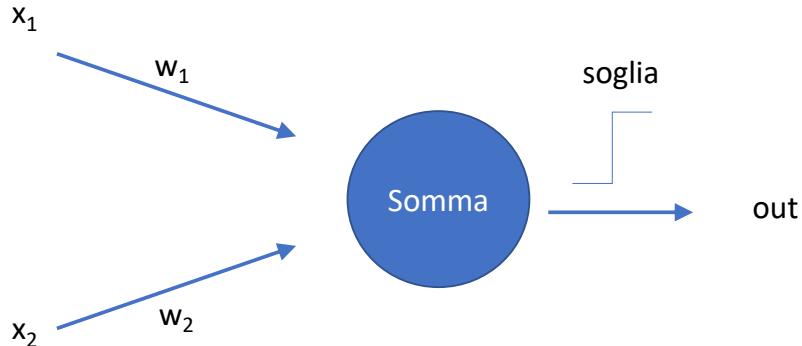
Gianluca Moro - DISI, Università di Bologna

Deep Learning

## Apprendimento con *reti profonde*

- Supervisionato o non supervisionato
- Richiede grandi risorse di calcolo
- Richiede elevate quantità di dati
- Richiede grande padronanza degli strumenti
  - e tempi di apprendimento anche molto lunghi
- Può dare eccellenti risultati

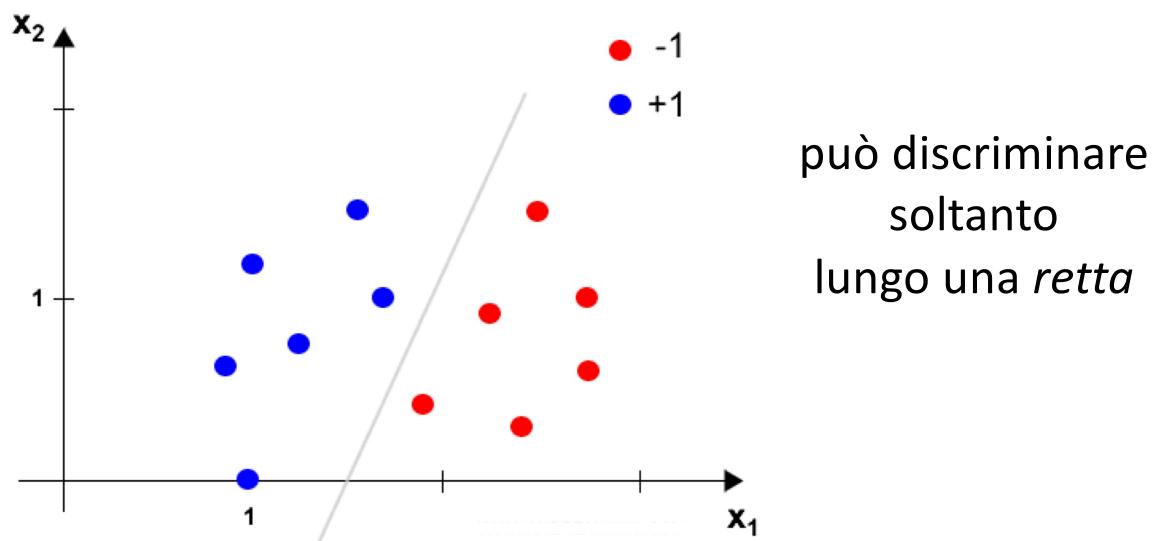
In principio era il *percettrone lineare*  
(Rosenblatt 1958)



Gianluca Moro - DISI, Università di Bologna

103

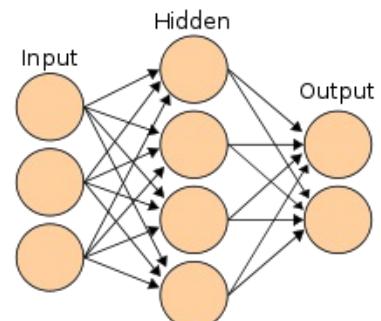
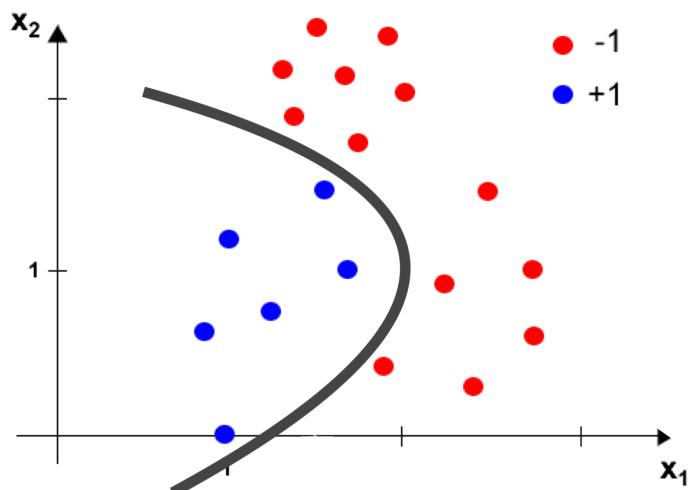
In principio era il *percettrone lineare*  
(Rosenblatt 1958)



Gianluca Moro - DISI, Università di Bologna

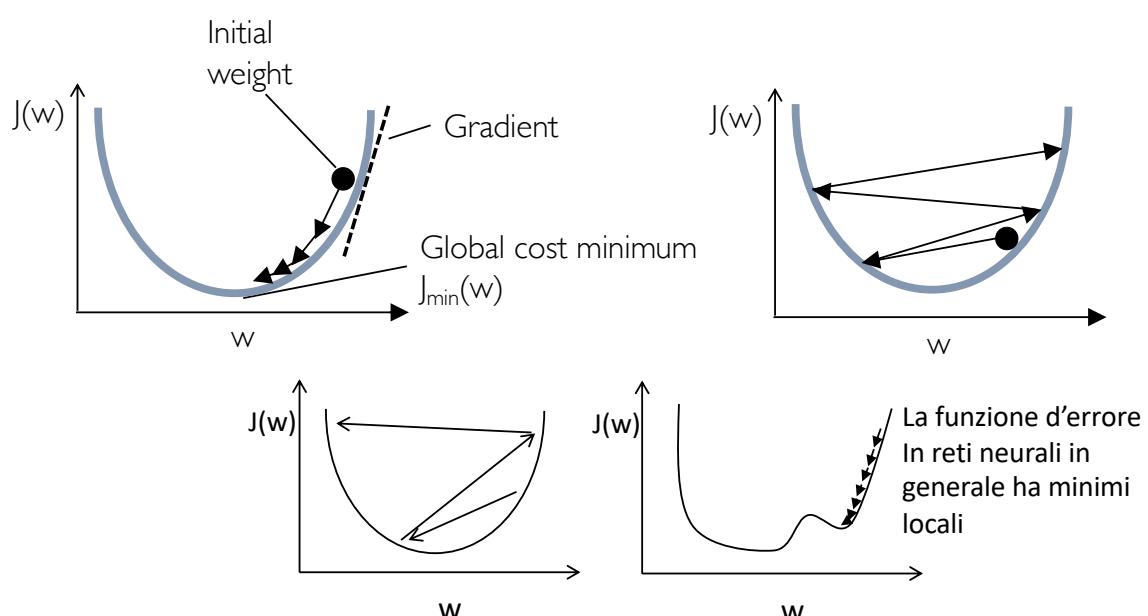
104

## Rete neurale connessione di più elementi «a strati»



permettono di discriminare lungo superfici complesse

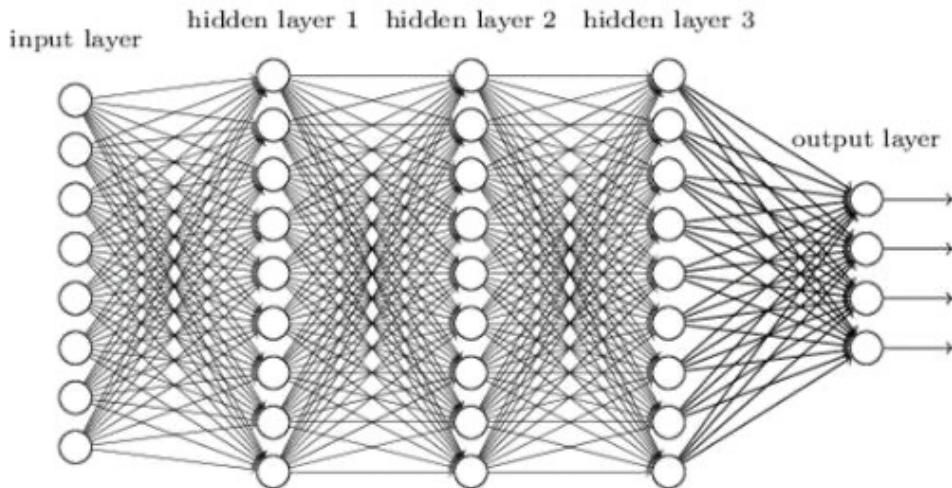
## Discesa del Gradiente: Minimizzazione della funzione d'errore



Learning Rate: troppo grande o troppo piccolo

# Deep Learning

## Rete neurale «profonda»



Gianluca Moro - DISI, Università di Bologna

107

# Deep Learning

## Rete neurale «profonda»

- Migliaia di connessioni → migliaia di parametri da *apprendere*
- Necessario hardware molto potente
- Molte diverse soluzioni architetturali
  - reti ricorsive
  - reti convoluzionali
  - autoencoder
  - macchine di Boltzmann
  - ...

Gianluca Moro - DISI, Università di Bologna

108

# Applicazioni

Self Driving Cars	News Aggregation and Fraud News Detection	Natural Language Processing	Virtual Assistants
Entertainment	Visual Recognition	Fraud Detection	Healthcare
Personalisations	Detecting Developmental Delay in Children	Colourisation of Black and White images	Adding sounds to silent movies
Automatic Machine Translation	Automatic Handwriting Generation	Automatic Game Playing	Language Translations
Pixel Restoration	Photo Descriptions	Demographic and Election Predictions	Deep Dreaming

Gianluca Moro - DISI, Università di Bologna

109

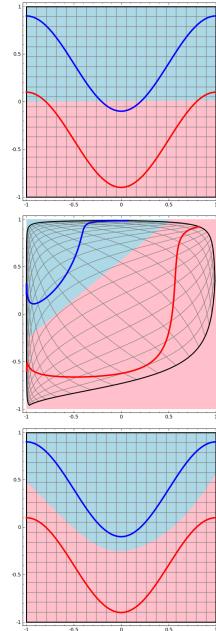
## Un esempio di Deep Dreaming



Gianluca Moro - DISI, Università di Bologna

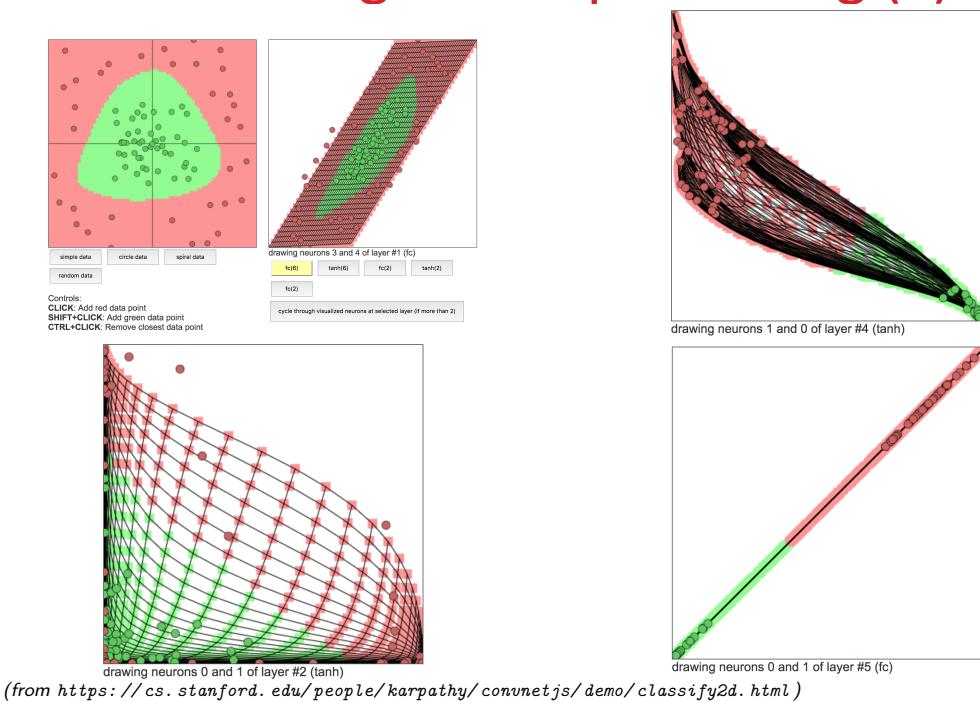
# Machine Learning Vs Deep Learning (i)

- 1st figure: a data set with two classes that overlap and are not well linearly separable
- 2nd figure: the minimisation of the Cost function (e.g. mean square error) corresponds geometrically to a deformation of the data space as an attempt to separate the two class data
- the transformed data space can be linearly separated by the last NN layer
- 3rd figure: the linear separation in the transformed space corresponds to a not linear separation in the original data space
- what is going to learn the NN ? a KERNEL FUNCTION



Gianluca Moro - DISI, Università di Bologna

# Machine Learning Vs Deep Learning (ii)



Gianluca Moro - DISI, Università di Bologna

# Informazioni Testuali

Gianluca Moro - DISI, Università di Bologna

Testi

## Informazioni testuali

- 99% delle pagine web sono testi non strutturati
- email
- documenti scientifici
- documenti legali
- interazioni con clientela

# Alcune Risultati di Successo

- **Stock Market Predictions via Twitter** – 86% accuracy to predict incr/ decr Dow Jones, J. of Computational Science, 2011, J. Bollen et al.
- **Wavii** – App for gathering, classifications and distribution of news – start-up acquired by Google in April 2013 for 30 Millions USD
- **Summly**, iOS App for organizing and summarizing news in 400 chars – start-up acquired by Yahoo! in March 2013 for a similar amount
- **Watson** – Born as a *question answering system* in english natural language without topic restrictions – it won the the TV Jeopardy quiz (USA) against the best world human competitors (IBM)
- **Social TV** – merging TV and social networks – several successful startups Yidio, Miso, Getglue, TVzap, Trendrr.tv, IntoNow, *Yahoo!*, *SKY TV*, *Nielsen* ...
- **Topsy** – tweet search engine acquired by Twitter for 200 mln USD
- **Publishing** – TwoReads: The right book for each reader, Italian startup
- **DeepMind** – Neural Turing Machine - funded by Oxford researchers, acquired by Google for 400 mln \$
- **Google** – Attention mechanism and Transformers – BERT and BERTology (today hundreds of Transformers, even to predict earthquake)
- **Facebook** – BlenderBot 2, first chatbot with memory and retrieval enhanced
- **OpenAI** – chatGPT, chatbot with reasoning capability based on a very large language model

Gianluca Moro - DISI, Università di Bologna

115

## Alcune applicazioni di successo

### Crime Prediction Systems

- **Criminal Reduction Utilising Statistical History (CRUSH)**
  - Large database of illegal events:
    - committed crimes, tens of features for each illegal event, information on known criminals and their behaviours, tip-off from informants, video surveillance data
    - weather data (if at night rain, more cars are stolen)
  - Goal: predicting crimes
  - Experimentation since 2006 at Memphis (USA)
    - the system offers to police the prediction of robberies, vandalism after a sport match, possibility that cars are stolen
    - Experimentation even if Florida and UK
  - 31% reduction of general crimes and 15% of violent crimes (Dept. of Criminology and Criminal Justice - University of Memphis)

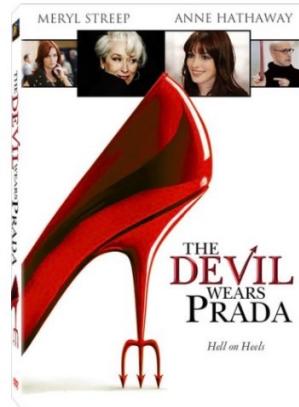
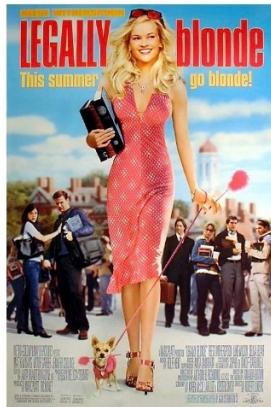
Gianluca Moro - DISI, Università di Bologna

116

## Alcune applicazioni di successo

### Quale film avrà successo?

- predire il successo di un film sulla base dello script
- Premio di \$1.000.000 alla migliore predizione



likes?  
→



117

Gianluca Moro - DISI, Università di Bologna

## Text Mining

- Text Classification, Clustering
- Classification, Clustering of documents, usually by topic
- Text Extraction & Summarization
- Extracting entities, such as persons, companies, brands, dates, events, places and generation of document abstracts
- Sentiment & Opinion mining
- Classifying reviews, posts, emails etc. by opinion orientation
- Question answering
  - Supplying answers to questions asked in natural languages
  - Chatbot
- Information Retrieval: finding relevant documents wrt searches

118

Gianluca Moro - DISI, Università di Bologna

# Apprendimento per Rinforzo

Gianluca Moro - DISI, Università di Bologna

Reinforcement Learning

## Apprendimento per interazione

- Guidato dall'obiettivo
- Dalla situazione all'azione
- Non si dice esplicitamente *cosa fare*
- Trial-and-error
  - reward (+ o -)
- Unica soluzione quando il problema è non differenziabile
  - ossia sono richieste scelte discrete

## Diverso dall'apprendimento non supervisionato

- Impossibile avere una descrizione esaustiva degli output desiderati nei casi derivanti dalle interazioni
- non si cerca di estrarre una struttura nascosta nei dati
- si cerca di eseguire una sequenza di azioni che massimizza il *reward* date le condizioni dell'ambiente

## Esempi

- Un giocatore di scacchi decide una mossa (scelta discreta)
  - giudizio sulla vantaggiosità di particolari posizioni
  - previsione di mosse dell'avversario
- Un controllore adattativo regola i parametri di una raffineria di petrolio
  - compromesso resa/costo/qualità
- Un robot *pulitore* individua un'area di possibile intervento e decide se intervenire o andare alla stazione base per ricaricarsi di energia

## Elementi

- Politica
  - corrispondenza stati/azioni
- Segnale di *reward*
  - obiettivo dell'apprendimento
  - ciò che è utile *nell'immediato*
- Funzione di *valore*
  - ciò che è positivo *nel lungo termine*
  - potrebbe essere il cumulo dei premi che l'agente può aspettarsi nel futuro, a partire da un certo stato

123

Gianluca Moro - DISI, Università di Bologna

## Elementi

- Il *reward* giunge direttamente dall'ambiente
- Il *valore* è stimato in base all'esperienza passata
  - influenza le *scelte di azione*

124

Gianluca Moro - DISI, Università di Bologna

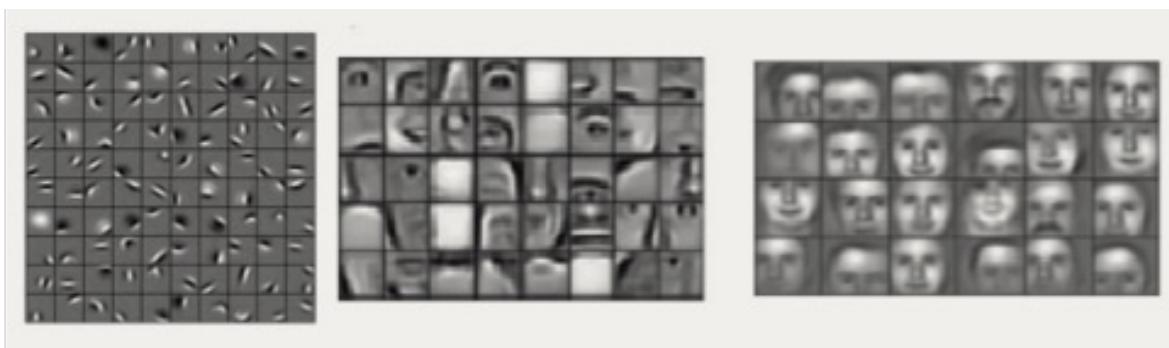
# Visione Artificiale

Gianluca Moro - DISI, Università di Bologna

Metodologia

**Processo complesso  
fortemente basato sul deep learning**

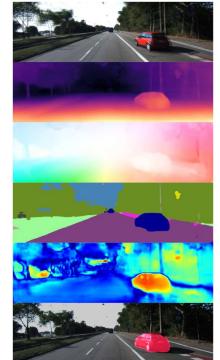
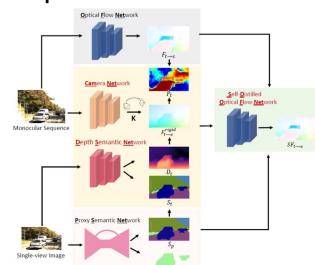
Strati successivi della rete riconoscono elementi via via più complessi



# Processo complesso fortemente basato sul deep learning

Esempio: Riconoscimento di oggetti e Comprensione di scene in tempo reale

Numerose tecnologie disponibili: OpenCV, SimpleCV, Yolo, Detectron2  
OpenPose, VGGNet, ResNet, Inception, Xception ...



Gianluca Moro - DISI, Università di Bologna

127

# Strumenti

# Strumenti più usati di Machine Learning

(Sorgente: [towardsdatascience.com](https://towardsdatascience.com/))

1. Knime <i>GUI</i>	Open-source machine learning tool that is based on GUI. The best thing about Knime is, it doesn't require any knowledge of programming. One can still avail of the facilities provided by Knime. It is generally used for data relevant purposes. For example, data manipulation, data mining, etc. Moreover, it processes data by creating different various workflows and then execute them. It comes with repositories that are full of different nodes. These nodes are then brought into the Knime portal. And finally, a workflow of nodes is created and executed.
2. Accord.net <i>Programming</i>	Computational machine learning framework. It comes with an image as well as audio packages. Such packages assist in training the models and in creating interactive applications. For example, audition, computer vision, etc. As .net is present in the name of the tool, the base library of this framework is C# language. Accord libraries are very much useful in testing as well as manipulating audio files. To upgrade your ML skills Explore — 70+ Machine Learning Datasets & Project Ideas!!
3. Scikit-Learn <i>Programming</i>	Open-source machine learning package. It is a unified platform as it is used for multiple purposes. It assists in regression, clustering, classification, dimensionality reduction, and preprocessing. Scikit-Learn is built on top of the three main Python libraries viz. NumPy, Matplotlib, and SciPy. Along with this, it will also help you with testing as well as training your models.
4. TensorFlow <i>Programming</i>	Open-source framework that comes in handy for large-scale as well as numerical ML. It is a blender of machine learning as well as neural network models. Moreover, it is also a good friend of Python. The most prominent feature of TensorFlow is, it runs on CPU and GPU as well. Natural language processing, Image classification are the ones who implement this tool.
5. Weka <i>GUI + Programming</i>	Open-source software. One can access it through a graphical user interface. The software is very user-friendly. The application of this tool is in research and teaching. Along with this, Weka lets you access other machine learning tools as well. For example, R, Scikit-learn, etc.

129

Gianluca Moro - DISI, Università di Bologna

# Strumenti più usati di Machine Learning

(Sorgente: [towardsdatascience.com](https://towardsdatascience.com/))

6. Pytorch <i>Programming</i>	Deep learning framework. It is very fast as well as flexible to use. This is because Pytorch has a good command over the GPU. It is one of the most important tools of machine learning because it is used in the most vital aspects of ML which includes building deep neural networks and tensor calculations. Pytorch is completely based on Python. Along with this, it is the best alternative to NumPy.
7. RapidMiner <i>GUI</i>	A piece of good news for the non-programmers. It is a data science platform and has a very amazing interface. RapidMiner is platform-independent as it works on cross-platform operating systems. With the help of this tool, one can use their own data as well as test their own models. Its interface is very user-friendly. You only drag and drop. This is the major reason why it is beneficial for non-programmers as well.
8. Google Cloud AutoML <i>Programming</i>	The objective of Google cloud AutoML is to make artificial intelligence accessible to everyone. What Google Cloud AutoML does is, it provides the models which are pre-trained to the users in order to create various services. For example, text recognition, speech recognition, etc. Google Cloud AutoML became very much popular among companies. As the companies want to apply artificial intelligence in every sector of the industry but they have been facing difficulties in doing so because there is a lack of skilled AI persons in the market.
9. Jupyter Notebook <i>Programming</i>	One of the most widely used machine learning tools among all. It is a very fast processing as well as an efficient platform. Moreover, it supports three languages viz. Julia, R, Python. Thus the name of Jupyter is formed by the combination of these three programming languages. Jupyter Notebook allows the user to store and share the live code in the form of notebooks. One can also access it through a GUI. For example, winpython navigator, anaconda navigator, etc.
10. Apache Mahout <i>Programming</i>	Mahout is launched by Apache which is an open-source platform based on Hadoop. It is generally used for machine learning and data mining. Techniques such as regression, classification, and clustering became possible with Mahout. Along with this, it also makes use of math-based functions such as vectors, etc. If you haven't yet started with ML you can start with these Machine Learning Tutorial for Beginners with Case Study!!

130

Gianluca Moro - DISI, Università di Bologna

## Strumenti più usati di Machine Learning

(Sorgente: [towardsdatascience.com](https://towardsdatascience.com))

11. Azure machine learning studio <i>Programming</i>	It is launched by Microsoft. Just like, Google's Cloud AutoML, this is Microsoft's product which provides machine learning services to the users. Azure machine learning studio is a very easy way to form connections of modules and datasets. Along with this, Azure also aims to provide AI facilities to the user. Just like TensorFlow, it also works on CPU and GPU.
12. MLLIB <i>Programming</i>	Like Mahout, MLLIB is also a product of Apache Spark. It is used for regression, feature extraction, classification, filtering, etc. It also often called Spark MLLIB. MLLIB comes with very good speed as well as efficiency.
13. Orange3 <i>Programming</i>	A data mining software which is the latest version of the Orange software. Orange3 assists in preprocessing, data visualization, and other data-related stuff. One can access Orange3 through the Anaconda Navigator. It is really very helpful in Python programming. Along with this, it can also be a great user interface.
14. IBM Watson <i>GUI</i>	A web interface that is given by IBM for using Watson. Watson is a human interaction Q and A system which is based on Natural Language processing. Watson is applied in various fields such as automated learning, information extraction, etc. IBM Watson is generally used for research and testing purposes. Its objective is to offer a human-like experience to the users.
15. Pylearn2 <i>Programming</i>	A machine learning library that is built on top of Theano. Therefore, there are many functions that are similar between them. Along with this, it can perform math calculations. Pylearn2 is also capable of running on the CPU and GPU as well. Before getting to Pylearn2, you must be familiar with Theano.

## Strumenti più usati di Deep Learning

(Sorgente: [upgrad.com](https://upgrad.com))

1. Neural Designer	Neural Designer is a professional application to discover unknown patterns, complex relationships, and predicting actual trends from data sets using neural networks. The Spain based startup company Artelnics developed Neural Designer, which has become one of the most popular desktop applications for data mining. Neural Designer uses neural networks as mathematical models mimicking the human brain function. It builds computational models that function as the central nervous system.
2. H2O.ai	H2O was developed from scratch using Java as the core technology and efficiently integrated with most other products like Spark and Apache Hadoop. This gives extreme flexibility to customers. With H2O, anyone can apply predictive analytics and machine learning easily to solve tough business problems. It uses an open-source framework with an easy-to-use web-based GUI, the most familiar interface. All common database and file types are supported using standard data-agnostic support. The tool is massively scalable and helps in real-time data scoring.
3. DeepLearningKit	Apple uses this deep learning framework in most of its products like iOS, OS X, tvOS, etc. Apple uses it to support pre-trained deep learning models on Apple's devices that have GPUs. DeepLearningKit uses Deep Convolutional Neural Networks like image recognition. It is currently trained with the Caffe Deep Learning framework, but the long-term goal is to support using other deep learning models like TensorFlow and Torch.
4. Microsoft Cognitive Toolkit	Microsoft Cognitive Toolkit is a commercially used toolkit that trains deep learning systems to learn precisely like human brain. It is free open-source and effortless to use. It provides exceptional scaling capabilities along with speed and accuracy and enterprise-level quality. It empowers users to harness the intelligence within massive datasets through deep learning. Microsoft Cognitive Toolkit describes neural networks as a sequence of computational steps through a directed graph. The leaf nodes of the directed graph represent input values or network parameters. The tools work exceptionally well with massive datasets. Microsoft products like Skype, Cortana, Bing, Xbox use the Microsoft Cognitive Toolkit to generate industry-level Artificial Intelligence.

## Strumenti più usati di Deep Learning

(Sorgente: upgrad.com)

5. Keras	Keras is a deep learning library that has minimal functionalities. It was developed with a focus on enabling fast experimentation and works with Theano and TensorFlow. The key benefit is that it can take you from idea to result in a swift speed. It is developed in Python and works as a high-level neural networks library capable of running on top of either TensorFlow or Theano. It allows for easy and fast prototyping using total modularity, extensibility, and minimalism. Keras supports convolutional networks, recurrent networks, a combo of both, and arbitrary connectivity schemes like multi-input and multi-output training.
6. ConvNetJS	ConvNetJS allows users to formulate and solve Neural Networks using JavaScript. It is an experimental reinforcement learning module based on Deep Q Learning. There is no need for other software, compilers, installations, or GPUs. Contributions from other communities have extended the library, and the complete code is available on GitHub under the MIT license. It can specify and train convolutional networks to process images.
7. Torch	The torch is a highly efficient open-source program. This scientific computing framework is supporting machine learning algorithms using GPU. It uses a dynamic LuaJIT scripting language and an underlying C/CUDA implementation. The torch has a powerful N-dimensional array feature, lots of routines for indexing, slicing, transposing, etc. It has excellent GPU support and is embeddable so that it can work with iOS, Android, etc.

## Riferimenti

- [Data Science con Python, Dai Fondamenti al Machine Learning.](#) Joel Grus, 2021. Gianluca Moro, curatore scientifico dell'edizione italiana. EGEA Casa Editrice dell'Università Bocconi. Edizione in inglese [Data Science from Scratch](#), O'Reilly Media ([versione precedente](#))