Machine learning

L'apprendimento automatico (noto anche come machine learning) è una parte dell'intelligenza artificiale che raggruppa un insieme di metodi sotto diversi nomi quali:

- statistica computazionale,
- riconoscimento di pattern,
- reti neurali artificiali,
- filtraggio adattivo,
- teoria dei sistemi dinamici,
- elaborazione delle immagini,
- data mining,
- algoritmi adattivi,
- ecc.

Utilizza metodi statistici per migliorare progressivamente la prestazione di un algoritmo nell'identificare pattern (relazioni) nei dati.
L'apprendimento automatico è strettamente legato al riconoscimento di pattern e alla teoria computazionale dell'apprendimento ed esplora lo studio e la costruzione di algoritmi che possano apprendere da un insieme di dati e fare delle predizioni su questi, costruendo in modo induttivo un modello basato su dei campioni. L'apprendimento automatico viene impiegato in quei campi dell'informatica nei quali progettare e programmare algoritmi espliciti è impraticabile; tra le possibili applicazioni citiamo il filtraggio delle email per evitare spam, l'individuazione di intrusioni in una rete o di intrusi che cercano di violare dati, il riconoscimento ottico dei caratteri, i motori di ricerca e la visione artificiale.
È strettamente collegato, e spesso si sovrappone con la statistica computazionale, che si occupa dell'elaborazione di predizioni tramite l'uso di computer. L'apprendimento automatico è anche fortemente legato all'ottimizzazione matematica, che fornisce metodi, teorie e domini di applicazione a questo campo. Per usi commerciali, l'apprendimento automatico è conosciuto come analisi predittiva.

Il Machine Learning in sé è un insieme di tecniche e di metodologie che provengono da aree diverse, quindi abbiamo la Statistica, il Data Science, il Data Mining, l'intelligenza artificiale, si intrecciano tutte queste metodologie e vanno a creare quindi quelli che sono gli algoritmi di Machine Learning.

Data Mining

Il Data Mining è il processo di scoperta di relazioni, pattern, ed informazioni precedentemente sconosciute e potenzialmente utili, all'interno di grandi basi di dati. Un pattern indica una struttura, un modello, o, in generale una rappresentazione sintetica dei dati.
Le tecniche e gli algoritmi di data mining hanno lo scopo di analizzare vasti campioni di dati, allo scopo di identificare interessanti regolarità dette pattern. Un concetto correlato al "data mining" è quello di machine learning (apprendimento automatico); infatti, l'identificazione di pattern può paragonarsi all'apprendimento, da parte del sistema di data mining, di una relazione causale precedentemente ignota, cosa che trova applicazione in ambiti come quello degli algoritmi euristici e della intelligenza artificiale. Tuttavia, occorre notare che il processo di data mining è sempre sottoposto al rischio di "rivelare" relazioni causali inesistenti.

Le tecniche e le strategie applicate alle operazioni di data mining sono per larga parte automatizzate, consistendo in specifici software e algoritmi adatti al singolo scopo. Ad oggi, in particolare, si utilizzano reti neurali, alberi decisionali, clustering e analisi delle associazioni.
Le finalità del data mining sono applicabili ai più svariati campi: economico, scientifico, operativo, etc.

Una tecnica molto diffusa per il data mining è l'apprendimento mediante classificazione. Questo schema di apprendimento parte da un insieme ben definito di esempi di classificazione per casi noti, dai quali ci si aspetta di dedurre un modo per classificare esempi non noti. Tale approccio viene anche detto con supervisione (supervised), nel senso che lo schema di apprendimento opera sotto la supervisione fornita implicitamente dagli esempi di classificazione per i casi noti; tali esempi, per questo motivo, vengono anche detti training examples, ovvero esempi per l'addestramento. La conoscenza acquisita per apprendimento mediante classificazione può essere rappresentata con alberi di decisione.

Esempi data mining

- L'analisi cluster permette di identificare all'interno di un archivio un determinato gruppo di utenti secondo caratteristiche comuni. Queste caratteristiche possono essere l'età, la provenienza geografica, il titolo di studio e così via. Si tratta di una tecnica di data mining che nel marketing risulta utile per segmentare il database e inviare, ad esempio, una certa promozione al target giusto per quel prodotto o servizio (giovani, mamme, pensionati, ecc). Le combinazioni di variabili sono infinite e rendono l'analisi cluster più o meno selettiva a seconda delle esigenze di ricerca.
- Come classificare una email di risposta di un cliente? E come individuare possibili correlazioni tra potenziali acquirenti dei tuoi prodotti prima e dopo l'attuazione di una campagna di advertising? La risposta è una sola: analisi classificatoria, la tecnica di data mining che consente di riconoscere i cosiddetti pattern (schemi ricorrenti) all'interno di un database. Una soluzione efficace per rendere più performante la tua strategia di marketing, eliminare il superfluo e creare sotto-archivi ottimizzati.
- Marketing e sicurezza sono due aspetti che sembrano non avere alcuna relazione e che invece vanno (o dovrebbero andare) di pari passo. Per ovviare l'impiego di archivi infetti da intrusi (singoli valori aggiunti da cracker o veri e propri virus che duplicano i dati), è sufficiente procedere con la ricerca degli intrusi, una tecnica di data mining che bonifica il database e garantisce una maggiore sicurezza dell'intero sistema.

Data Science

La data science è un settore interdisciplinare che utilizza metodi scientifici, processi, algoritmi e sistemi per estrarre valore dai dati. I data scientist combinano le competenze in varie discipline, tra cui statistica, informatica ed economia aziendale, per analizzare i dati raccolti dal Web, dagli smartphone, dai clienti, dai sensori e da altre fonti.
La data science mostra i trend e produce insight che le aziende possono utilizzare per prendere decisioni più mirate e creare prodotti e servizi più innovativi. I dati costituiscono la base dell'innovazione, ma il loro valore deriva dalle informazioni che i data scientist possono ottenere e in base alle quali agire.

Chi è un data scientist

il Data Scientist non può non avere forti competenze interdisciplinari, deve padroneggiare gli Advanced Analytics e i Big Data e dunque deve dimostrare solide competenze informatiche, ma deve saper leggere oltre il dato, individuare i pattern con competenze a livello di statistica e di matematica, deve poi saper dialogare con le aree di business e deve avere un quadro dei modelli di business attuali e potenziali che si "appoggiano" sui dati.

Strumenti per i data scientist
Tra gli strumenti più comunemente utilizzati dai data scientist, ci sono i notebook open source, ovvero applicazioni Web che consentono di scrivere ed eseguire codice, visualizzare dati e vedere i risultati in un unico ambiente. Alcuni dei notebook più diffusi includono Jupyter, RStudio e Zepplin. I notebook sono molto utili per eseguire analisi, ma presentano delle limitazioni quando devono essere utilizzati dai data scientist per lavorare in team. Per risolvere questo problema, sono state concepite le piattaforme di data science.

Impieghi della Data Science
Le organizzazioni utilizzano i team di data science per trasformare i dati in un vantaggio competitivo ridefinendo i prodotti e i servizi. Ad esempio, le aziende analizzano i dati raccolti dai call center per identificare i clienti propensi all'abbandono e utilizzano strategie di marketing per tentare di fidelizzarli. Le aziende di logistica analizzano i modelli di traffico, le condizioni meteorologiche e altri fattori per migliorare la velocità di consegna e ridurre i costi. Le aziende farmaceutiche analizzano i dati degli esami clinici e i sintomi segnalati per aiutare i medici a diagnosticare le malattie in anticipo e trattarle in modo più efficace.

La base dei Big Data e Data Science
Per comprendere lo sviluppo dei Big Data occorre anche saper individuare i modelli di utilizzo degli Analytics nelle imprese e ancora una

volta è necessaria una distinzione duale nelle tipologie di dati:
- dati strutturati
- dati destrutturati

Nel caso dei dati destrutturati si tratta poi tipicamente di
- testo
- immagini
- video
- audio
- elementi di calcolo

Dato destrutturato vuol dire dato eterogeneo, che significa, banalizzando un po', dati che rispecchiano la "eterogeneità" della realtà.

Il percorso Data Driven

Dietro a questi "segni particolari" ci sta la funzione centrale dei Big Data che è quella di fornire la miglior rappresentazione possibile della realtà attraverso i dati. Ma per rappresentare in modo verosimile prima e veritiero poi la realtà con i dati è necessario sviluppare metodiche e logiche di rappresentazione con processi di verifica e di controllo. Con questo approccio si va a collocare l'impresa all'interno di uno scenario di tipo Data Driven costituito da 4 grandi tipologie di Data Analysis.

1 – Descriptive Analytics

Si parte dall'Analisi Descrittiva che è costituita da tutti i tool che permettono di rappresentare e descrivere anche in modo grafico la realtà di determinate situazioni o processi. Nel caso delle imprese parliamo ad esempio della rappresentazione di processi aziendali. La Descriptive Analytics permette la visualizzazione grafica dei livelli di performance.

2 – Predictive Analytics

Si passa poi all'Analisi Predittiva basata su soluzioni che permettono di effettuare l'analisi dei dati al fine di disegnare scenari di sviluppo nel futuro. Le Predictive Analytics si basano su modelli e tecniche matematiche come appunto i Modelli Predittivi, il Forecasting e altri.

3 – Prescriptive Analytics

Con le Analisi Prescrittive si entra nell'ambito di strumenti che associano l'analisi dei dati alla capacità di assumere e gestire processi decisionali. Le Prescriptive Analytics sono tool che mettono a disposizione delle indicazioni strategiche o delle soluzioni operative basate sia sull'Analisi Descrittiva sia sulle Analisi Predittive.

4 – Automated Analytics

Il quarto punto scenario è rappresentato dalle Automated Analytics che permettono di entrare nell'ambito dell'automazione con soluzioni di Analytics. A fronte dei risultati delle analisi descrittive e predittive le Automated Analytics sono nella condizione di attivare delle azioni definite sulla base di regole. Regole che possono essere a loro volta il frutto di un processo di analisi, come ad esempio lo studio dei comportamenti di una determinata macchina a fronte di determinate condizioni oggetto di analisi.

Statistica

La statistica, il data mining ed il machine learning hanno un ruolo importante nel processo di apprendimento dei dati (estrapolazione di informazioni da essi), descrivendo quindi le caratteristiche dei dati e costruendo un modello. Principalmente il machine learning viene usato per la risoluzione dei problemi classici della statistica, usando il concetto del ***data mining****.*

Data Mining

Processo basato sui principi della statistica che consiste nell'esplorare un vasto bacino di dati per trovare dei pattern (regole) che li lega.

Gli algoritmi di data mining vengono usati per la risoluzione di vari problemi, come per esempio il riconoscimento di frodi, analisi market based.

Il punto forte del dm è che il suo scopo è quello di capire i dati e non di fare previsioni su essi, infatti essi vengono usati per trovare dei pattern sui quali verranno fatte delle previsioni.

Es. Un venditore potrebbe essere interessato ad avere una visione delle caratteristiche di chi risponde ad una promozione e di chi no, per poi (in base ai pattern trovati), cambiarla.

Quali tecniche della statistica possono essere utilizzati per il machine learning?

- Problem Framing
- Data Understanding
- Data Cleaning
- Data Selection
- Data Preparation
- Model Configuration
- Model Selection
- Model Presentation

Problem Framing

A volte il principale problema delle leve finanziarie è la strutturazione del problema.
Per esempio metodi come la regressione e la classificazione permettono di dividere gli input con gli output in classi ben precise.
La strutturazione del problema non è sempre ovvia.
Per i nuovi nel campo, potrebbe richiedere una esplorazione significativa, per l'osservazione del dominio, mentre gli esperti potrebbero bloccarsi nell'osservazione in modo convenzionale.
L'indagine statistica nell'esplorazione dei dati include due parti importantissime:
Exploratory Data Analysis, riepilogo e visualizzazione per esplorare ad hoc i dati; Data Mining, che serve per la scoperta automatica di strutture e relazioni e pattern di dati.

Data Understanding

Il data understanding consiste nell'avere una profonda comprensione sia della distribuzione delle variabili che delle relazioni tra di esse.
Per questo tipo di metodo statistico vengono usati due metodi statistici principali:

- summary statistics: metodo usato per riassumere la distribuzione e le relazioni tra le variabili usando quantità statistiche;
- data visualization: metodo usato per riassumere la distribuzione tra le variabili usando i grafici

Data Cleaning

Di solito I dati che vengono osservati non sono sempre incontaminati.
Anche se i dati sono digitali potrebbero essere soggetti a danneggiamenti, o da valori incoerenti tra di loro, per evitare ciò si usano alcuni metodi per "pulire", come ad esempio:

- Outlier detection: metodo che consiste nell'individuare i dati lontani dai valori aspettati dalla distribuzione;

- Imputation: metodo usato per la correzione dei dati e per il loro riempimento nel caso in cui mancassero.

Data Selection

Non tutti i dati osservati o tutte le variabili correlate potrebbero essere utili nella produzione del modello: il processo che si occupa della riduzione del bacino dei dati, tenendo solo quelli più importanti si chiama data selection.

Data Preparation

I dati spesso non possono essere usati direttamente per la creazione del modello.
Alcuni processi, spesso, sono necessari per strutturarli in maniera da renderli usabili.

Model Configuration

Un algoritmo di machine learning spesso ha una suite di hyper parameters, che permettono al metodo di apprendimento di essere sfruttato per un problema specifico e non più generico.

La configurazione degli hyper parameters è spesso di natura empirica, invece che analitica, richiede una vasta disponibilità di esempi per valutare gli effetti dell'efficacia del modello.

L'interpretazione e la comparazione dei risultati rispetto a diverse configurazioni degli hyper parameters viene fatta usando delle sottocategorie della statistica, chiamate rispettivamente:

Statistical Hypothesis Tests, metodo che quantifica quanto il risultato è vicino a quello aspettato, e l'estimation statistics per quantificare l'incertezza del risultato ottenuto.

Model Selection

Uno tra I vari algoritmi di ML potrebbe essere più appropriato di un altro per uno specifico tipo di modello predittivo: il processo per la selezione di uno dei metodi per la soluzione è chiamato model selection.

Model Presentation

Una volta che il modello finale è stato allenato, esso può essere presentato agli stakeholders, per essere usato per la sua applicazione nella predizione di dati reali.

Regressione

La regressione lineare è un algoritmo di apprendimento automatico supervisionato in cui l'uscita prevista è continua e ha una pendenza costante. Viene utilizzato per prevedere i valori all'interno di un intervallo continuo (ad esempio, vendite, prezzo) piuttosto che cercare di classificarli in categorie (ad es. Gatto, cane).

Regressione semplice

La regressione lineare semplice usa la forma tradizionale di intercettazione delle pendenze, dove m e **b** sono le variabili che il nostro algoritmo proverà a "imparare" per produrre le previsioni più accurate. **x**
rappresenta i nostri dati di input e **y** rappresenta la nostra previsione **y = mx + b**

Regressione multivariata

Un'equazione lineare più complessa e multi-variabile potrebbe apparire come questa, dove **w** rappresenta i coefficienti, o pesi, il nostro modello proverà ad imparare.
$\mathbf{f(x,y,z) = w_1x + w_2y + w_3z}$
Le variabili **x, y, z** rappresentano gli attributi, o pezzi distinti di informazioni, che abbiamo su ogni osservazione. Per le previsioni di vendita, questi attributi potrebbero includere la spesa pubblicitaria dell'azienda su radio, TV e giornali.

Regressione semplice

Diciamo che ci viene assegnato un set di dati con le seguenti colonne (caratteristiche): quanto spende un'azienda sulle pubblicità radiofoniche ogni anno e le sue vendite annuali in termini di unità vendute. Stiamo cercando di sviluppare un'equazione che ci consenta di prevedere le unità vendute in base a quanto spende un'azienda sulla pubblicità radiofonica.

Per ottimizzare la regressione utilizziamo la Discesa Graduale

MSE misura la differenza quadratica media tra i valori effettivi e quelli previsti di un'osservazione. L'output è un numero singolo che rappresenta il costo o il punteggio associato al nostro attuale insieme di pesi. Il nostro obiettivo è ridurre al minimo MSE per migliorare la precisione del nostro modello.

Matematica

Data la nostra semplice equazione lineare **y = mx+b**, possiamo calcolare **MSE** come:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Per ridurre al minimo MSE, utilizziamo la discesa per gradienti per calcolare il gradiente della nostra funzione.

Allenamento

La formazione di un modello è il processo di miglioramento iterativo dell'equazione di previsione eseguendo il loop del set di dati più volte.La formazione è completa quando raggiungiamo una soglia di errore accettabile o quando le successive iterazioni di addestramento non riducono i nostri risultati finali.

SITOGRAFIA

Cosa è il Machine Learning
http://www.intelligenzaartificiale.it/
https://it.wikipedia.org/wiki/Apprendimento_automatico
https://www.albertoolla.it/cose-machine-learning-reti-neurali/
https://www.ai4business.it/intelligenza-artificiale/machine-learning/machine-learning-cosa-e-applicazioni/

Data Mining:
https://www.math.unipd.it/~dulli/corso06/DMteoria.pdf
https://www.egon.com/it/blog/664-tecniche-esempi-data-mining-marketing

Data Science:
https://it.wikipedia.org/wiki/Scienza_dei_dati
https://www.youtube.com/watch?v=X3paOmcrTjQ

Alcuni esempi di machine learning:
https://medium.com/botsupply/il-machine-learning-%C3%A8-divertente-parte-1-97d4bce99a06

https://medium.com/botsupply/il-machine-learning-%C3%A8-divertente-parte-2-dec556e4855d

https://medium.com/botsupply/il-machine-learning-%C3%A8-divertente-parte-3-deep-learning-e-convolutional-neural-network-cnns-cc106559ffa9

Gradient Descent
https://hbfs.wordpress.com/2012/04/24/introduction-to-gradient-descent/

scikit-learn:
https://scikit-learn.org/stable/

Linear Regression:
https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html

Statistic:

https://machinelearningmastery.com/statistical-methods-in-an-applied-machine-learning-project/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6082636/pdf/nihms962164.pdf
Machine Learning For Dummies®, IBM Limited Edition

Training Machine Learning Images: https://cs.nyu.edu/~roweis/data.html