

machine learning

Alberto Ferrari – Analisi dei Dati

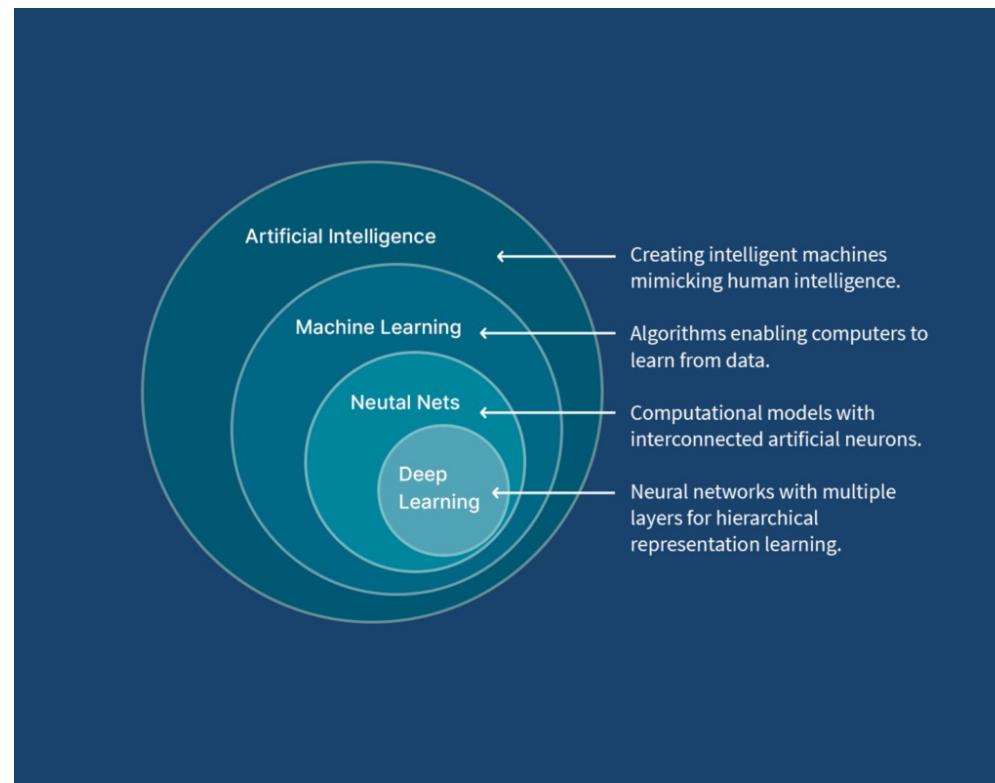
machine learning – apprendimento automatico

- dare ai computer la capacità di ***apprendere dai dati*** senza ricevere regole esplicite da un essere umano (programmatore)
 - *imparare dai dati senza essere programmati*
 - *computer addestrato non programmato*
- ***algoritmi classici***
 - è l'uomo a ***specificare*** il modo in cui individuare la ***soluzione migliore*** (algoritmo risolutivo)
 - il computer (programma) raggiunge la soluzione (esegue l'algoritmo-programma) in modo più veloce ed efficiente di un essere umano

machine learning - algoritmi

- capacità di un *algoritmo*
 - di **prendere decisioni** sulla base di una base di conoscenza (*knowledge-base*)
 - di **apprendere nuove informazioni** sulla base dell'esperienza (decisioni prese precedentemente)
 - al modello **non** viene fornita la **soluzione** migliore
 - riceve vari **esempi** del problema e gli viene chiesto di **decidere** qual è la soluzione migliore

machine learning – artificial intelligence



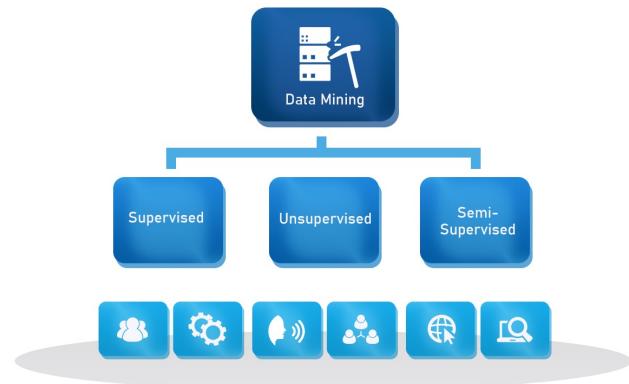
Alberto Ferrari – Analisi dei Dati

ciclo di vita di un progetto ML

- raccolta dati
- pulizia
- training
- test
- valutazione
- produzione

machine learning - tipologie

- tre tipologie fondamentali
 - ***supervised learning*** (apprendimento supervisionato)
 - ***unsupervised learning*** (apprendimento non supervisionato)
 - ***semi-supervised learning*** (apprendimento semi-supervisionato)



supervised learning

apprendimento supervisionato

Alberto Ferrari – Analisi dei Dati

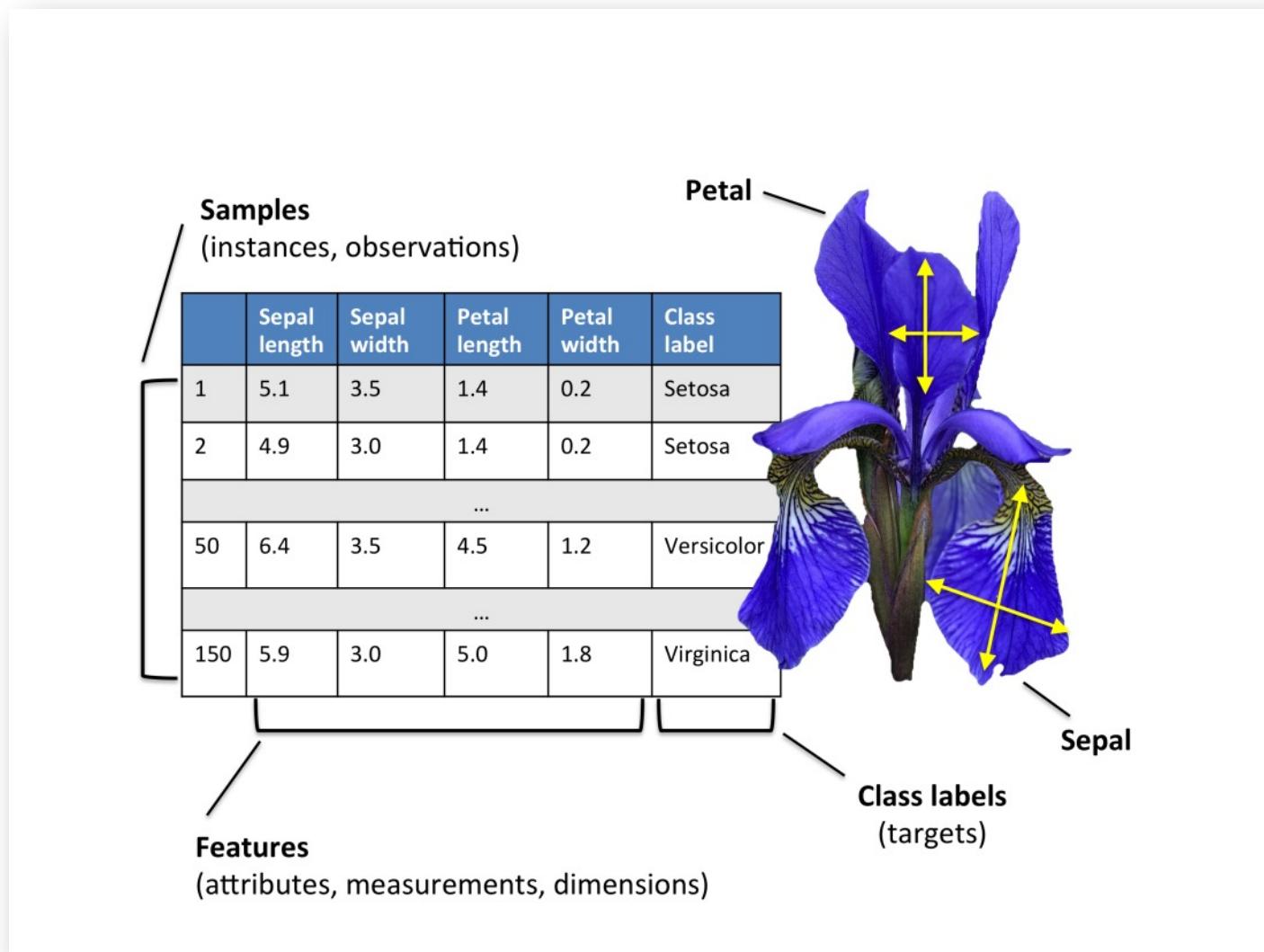
supervised learning

- apprendimento ***supervisionato***
- tecnica di apprendimento automatico che ha l'obiettivo di ***istruire*** un sistema
- per elaborare automaticamente ***previsioni*** sui valori di ***uscita***
- rispetto ad un ***input***
- sulla base di una serie di ***esempi*** ideali
- esempi ***forniti*** inizialmente e costituiti da ***coppie*** di input e di output

classico esempio: iris dataset

- il dataset *Iris* è un dataset **multivariato** (*più features per ogni occorrenza*) introdotto da Ronald Fisher nel 1936
- consiste in **150 istanze** di Iris
- **classificate** secondo tre specie: Iris **setosa**, Iris **virginica** e Iris **versicolor**
- le variabili (features) sono
 - *lunghezza del sepalo*
 - *larghezza del sepalo*
 - *lunghezza del petalo*
 - *larghezza del petalo*

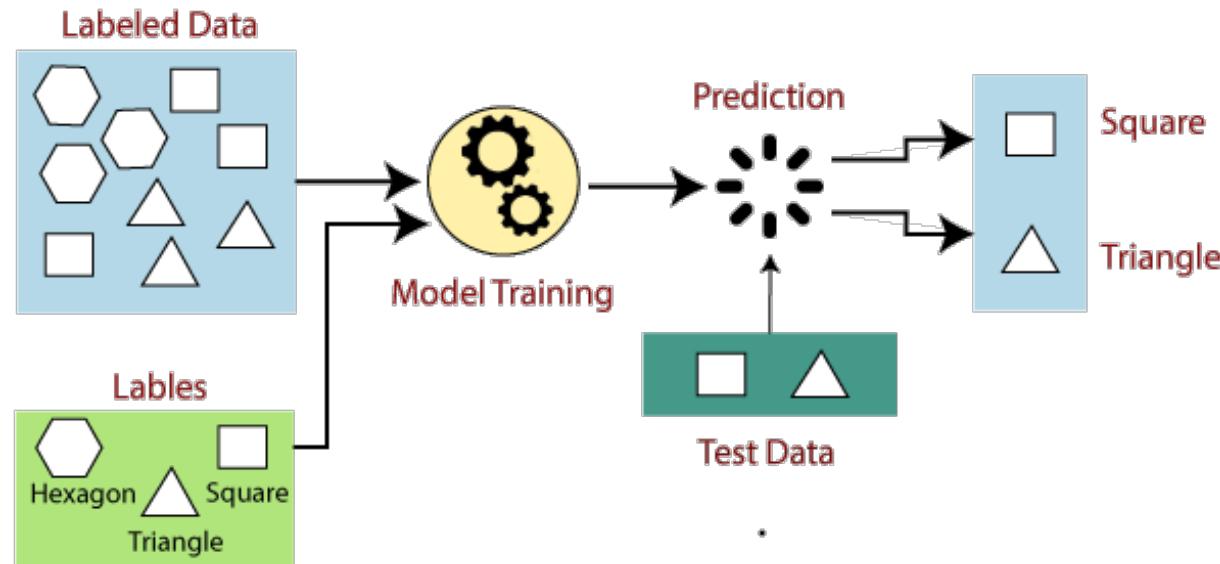




supervised learning - training set – test set

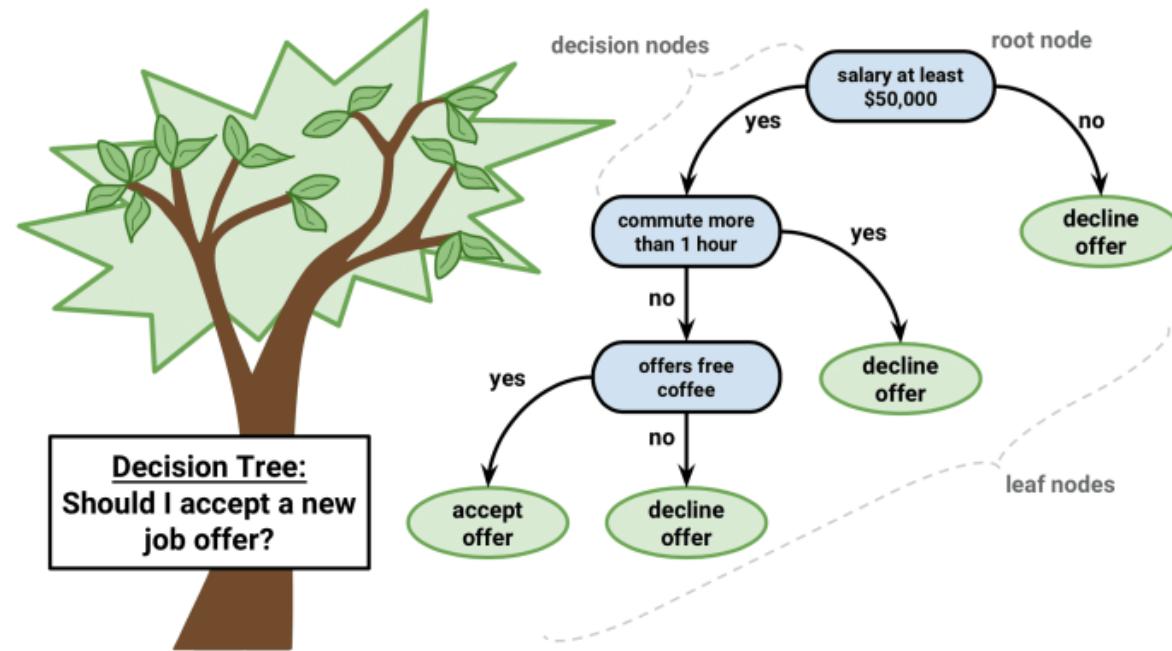
- ***training set***
 - insieme di dati di ***addestramento***
 - contiene informazioni etichettate (***labeled***) per permettere all'algoritmo di trovare il modo migliore per indovinare il maggior numero di casi
- ***labeling***
 - ***annotazione e classi***
 - permette all'algoritmo di imparare a discernere un esempio dagli altri
- ***test set***
 - insieme di dati di ***confronto***
 - contiene informazioni del tutto simili al training set
 - serve per ***verificare*** l'accuratezza dell'algoritmo addestrato

supervised learning



esempio di classificazione

Decision Tree Classifier

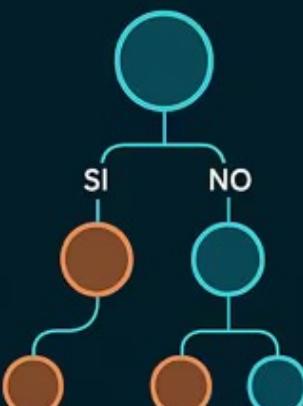


ALGORITMI DI CLASSIFICAZIONE SUPERVISIONATA

SUPPORT VECTOR MACHINE



ALBERO DECISIONALE

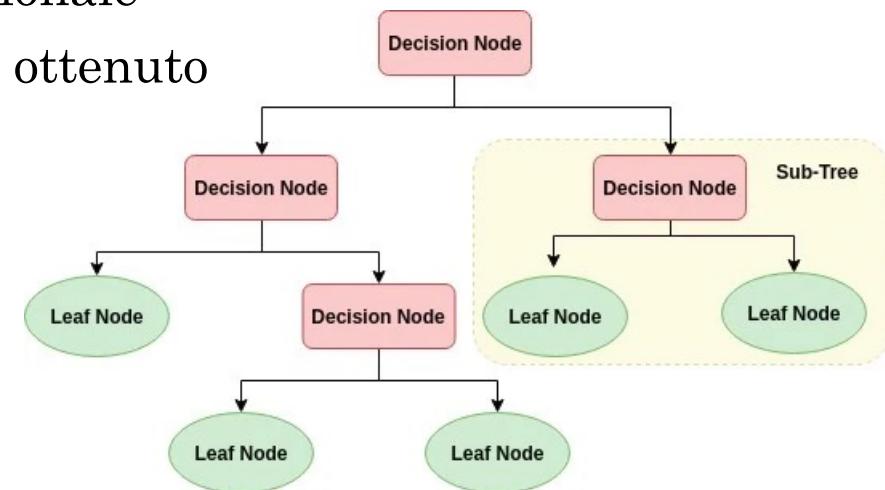


k-NN



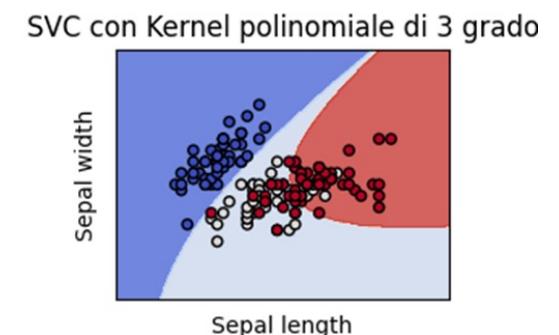
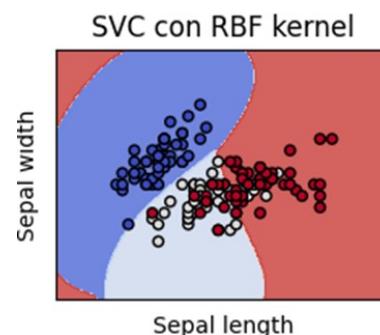
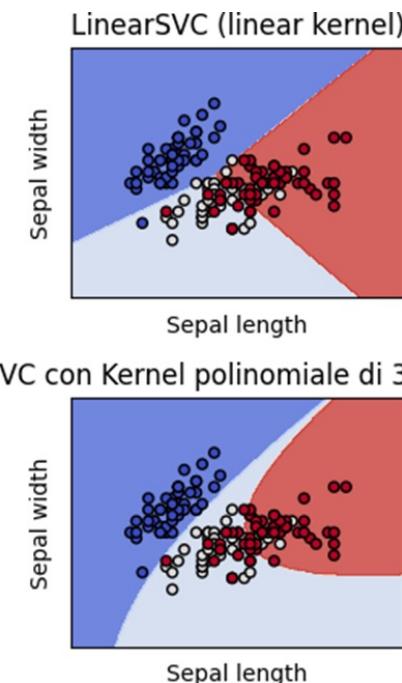
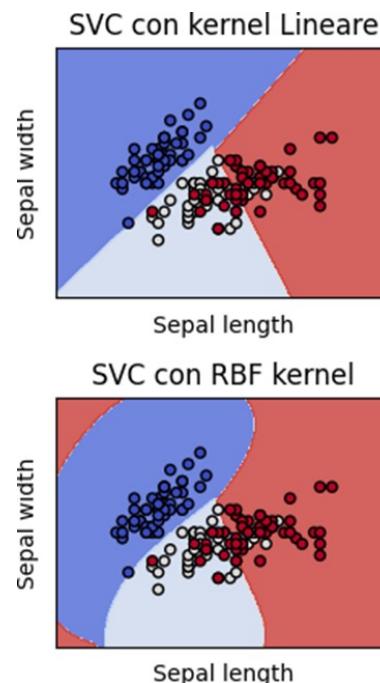
decision tree – albero decisionale

- è una struttura ad albero simile a un diagramma di flusso
- ogni un nodo interno rappresenta una feature (attributo)
- ogni ramo rappresenta una regola decisionale
- ogni nodo foglia rappresenta il risultato ottenuto



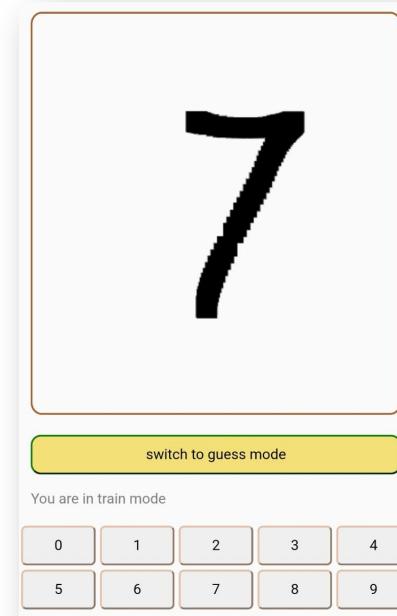
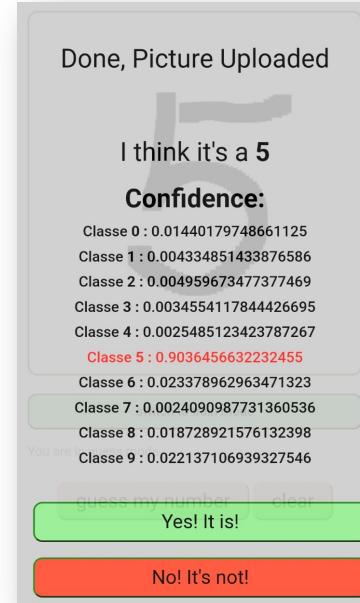
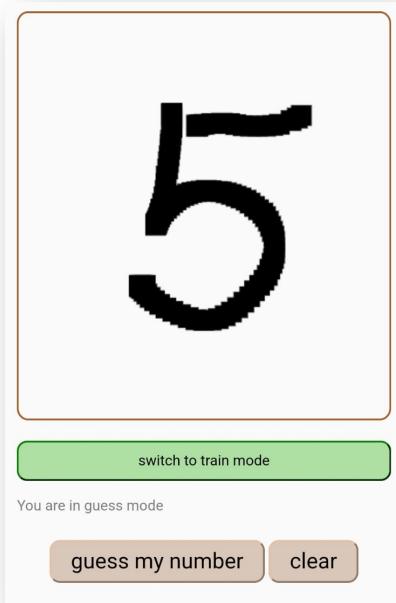
SVM (support-vector machines)

- l’obiettivo dell’algoritmo è di individuare un iperpiano che separi nel miglior modo possibile i punti che rappresentano di dati di una classe da quelli di un’altra classe.
- l’iperpiano “migliore” è quello che ha il margine maggiore tra le due classi



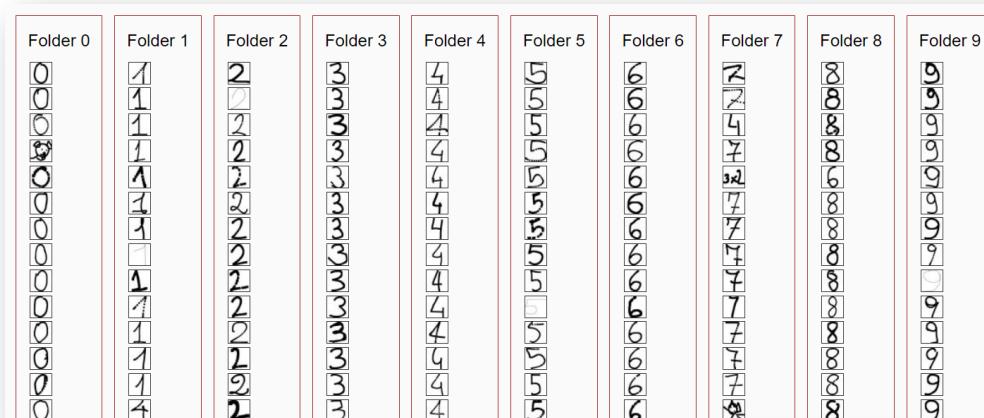
riconoscimento cifre numeriche

- <https://ml.webweb.cloud/digit/draw>



Machine Learning Tatami

- apprendimento supervisionato
- l'algoritmo utilizzato per il modello di apprendimento è SVM (support-vector machines)
- le features (caratteristiche) sono ottenute dal colore bianco o nero dei pixel delle immagini che rappresentano le cifre



Alberto Ferrari – Analisi dei Dati

unsupervised learning

apprendimento non supervisionato

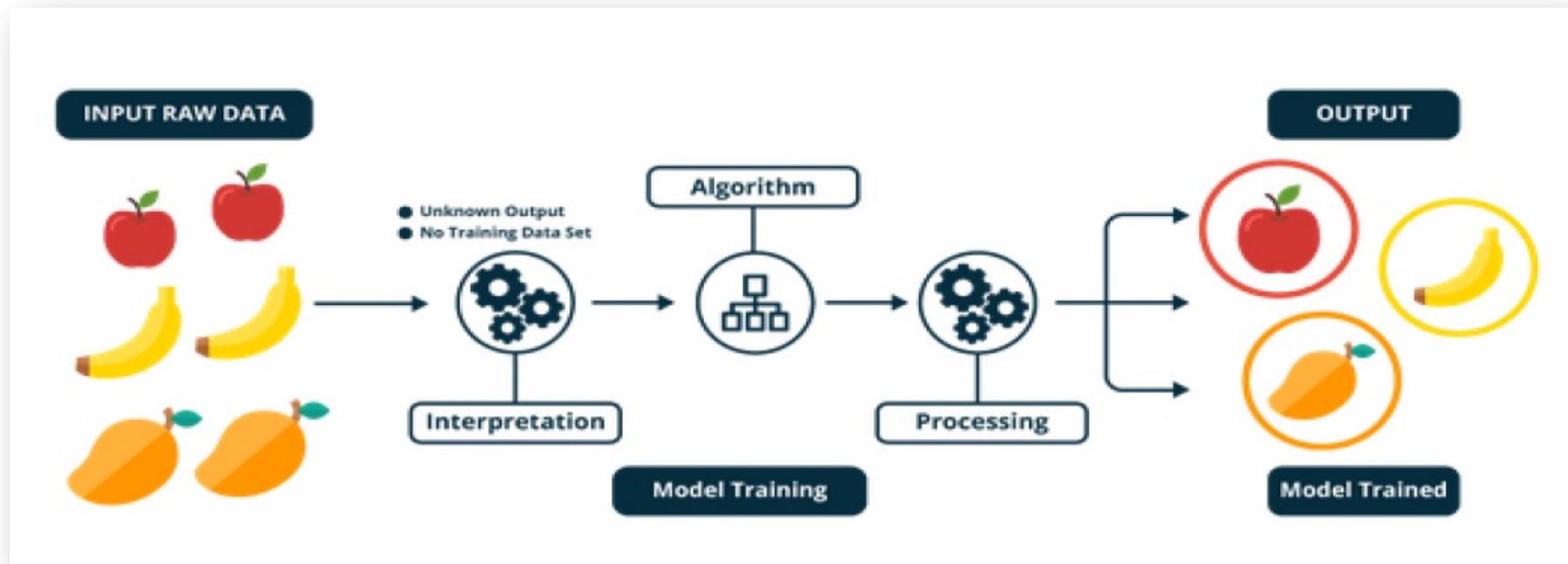
Alberto Ferrari – Analisi dei Dati

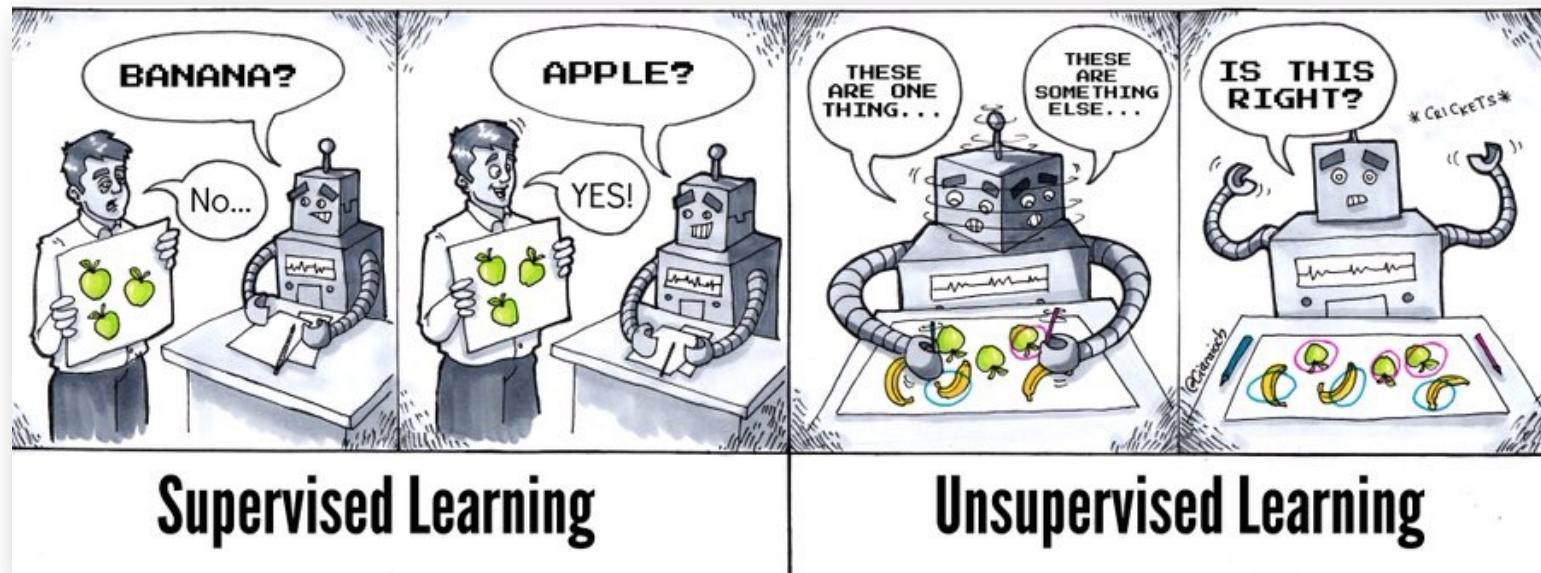
unsupervised learning

- apprendimento ***non supervisionato***
- consiste nel fornire al sistema informatico una serie di ***input*** (esperienza del sistema)
- il sistema ***classifica*** i dati sulla base di caratteristiche comuni
- lo scopo è cercare di effettuare ragionamenti e previsioni sugli input successivi
- durante la fase di apprendimento vengono forniti ***solo esempi non annotati***
- le ***classi non sono note a priori*** ma devono essere apprese automaticamente

unsupervised learning

- dati raw, ***no labels*** => il modello può raggruppare gli items per ***similitudine***





unsupervised learning - esempio

- esempio classico in ambito **marketing**
- i **consumatori** con caratteristiche simili vengono raggruppati (**classificati**) in base ai loro acquisti in categorie
- in base all'appartenenza alle varie categorie vengono attivate **campagne di marketing** specifiche



semi-supervised learning

apprendimento semi-supervisionato

Alberto Ferrari – Analisi dei Dati

semi-supervised learning

- apprendimento semi-supervisionato
- combina una **grande** quantità di dati **non etichettati** con una **piccola** quantità di dati **etichettati**
- l'apprendimento non supervisionato insieme a quello supervisionato permette all'algoritmo di **suddividere in cluster** gli esempi e poi di assegnare a tutti gli elementi di un certo gruppo la **label** di quelli etichettati presenti nel gruppo
- **vantaggi**
 - permette il **labeling automatico** di grandi quantità di dati altrimenti non etichettabili
- **svantaggi**
 - i dati al confine fra due gruppi potrebbero avere etichette di entrambi introduce un po' di **bias** nel training

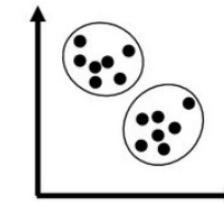
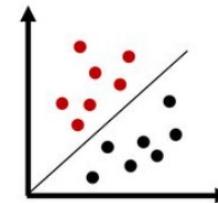
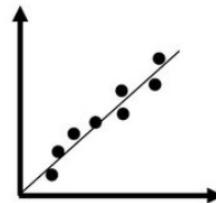
tecniche di apprendimento automatico

Alberto Ferrari – Analisi dei Dati

tipi di apprendimento automatico

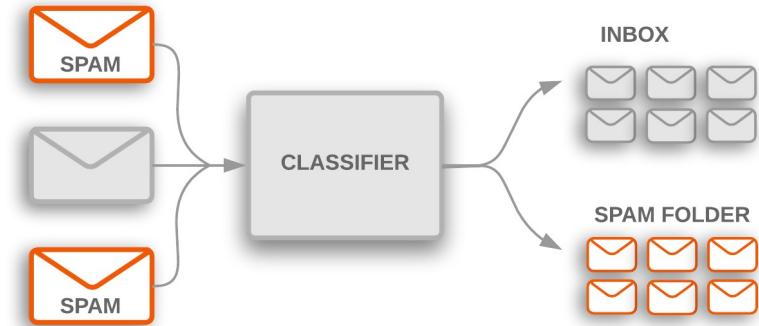
- *classificazione*
- *regressione*
- *clustering*

Regression	Classification	Clustering
<ul style="list-style-type: none">• Predict continuous valued output• Supervised	<ul style="list-style-type: none">• Predict discrete valued output• Supervised	<ul style="list-style-type: none">• Predict discrete valued output• Unsupervised



classificazione

- permette di **categorizzare** un insieme di dati in classi
- problemi affrontabili con algoritmi di classificazione sono
 - riconoscimento vocale
 - riconoscimento facciale
 - interpretazione della scrittura a mano
 - classificazione dei documenti
 - riconoscimento di immagini
 - ...



matrice di confusione (confusion matrix)

- utilizzata per ***valutare*** un modello di machine learning
- è una matrice in cui
 - le ***previsioni*** sono rappresentate nelle righe
 - lo stato ***effettivo*** è rappresentato nelle colonne

		True Class		
		Apple	Orange	Mango
Predicted Class	Apple	7	8	9
	Orange	1	2	3
	Mango	3	2	1

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

matrice di confusione – esempio con due classi (mail)

- ***true positive*** (TP veri positivi)
 - il modello ha classificato la mail come spam e lo è realmente
- ***true negative*** (TN veri negativi)
 - il modello ha classificato la mail come non spam e non lo è realmente
- ***false positive*** (FP falsi positivi)
 - il modello ha classificato la mail come spam ma in realtà non lo è
 - definito errore di primo tipo
- ***false negative*** (FN falsi negativi)
 - il modello ha classificato la mail come non spam ma in realtà si tratta di uno spam
 - definito errore di secondo tipo

metriche di valutazione

- **tasso di errore** (error rate) ERR

- numero di tutti i pronostici errati diviso per il numero totale del set di dati
- il miglior tasso di errore è 0, il peggiore è 1

$$ERR = \frac{FP + FN}{TN + FP + FN + TP}$$

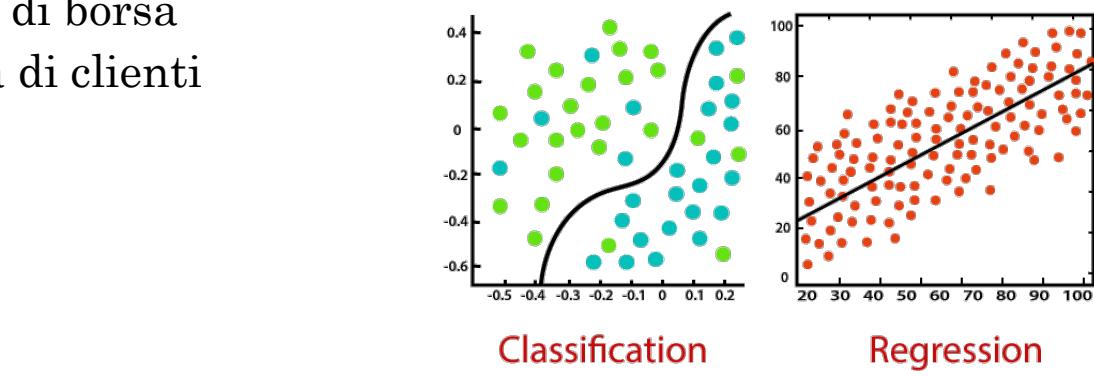
- **accuratezza** (accuracy)

- la migliore accuratezza è 1, la peggiore è 0

$$Accuracy = \frac{TP + TN}{TN + FP + FN + TP}$$

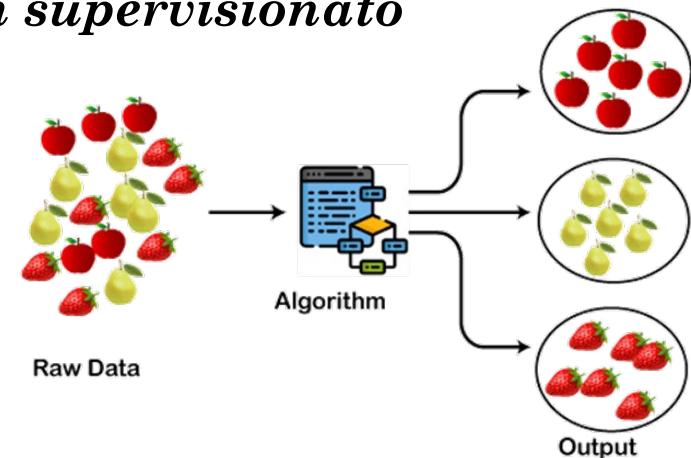
regressione

- modello di calcolo statistico che, a differenza della classificazione, non assegna una classe ad ogni item esaminato ma assegna un ***valore reale stimato***
- il calcolo statistico è il risultato di un algoritmo di ***minimizzazione di errore***
- la regressione fa sempre riferimento all'apprendimento ***supervisionato***
- problemi affrontabili con algoritmi di classificazione
 - previsioni temperature meteo
 - previsioni andamento azioni di borsa
 - stima della capacità di spesa di clienti
 - ...



clustering

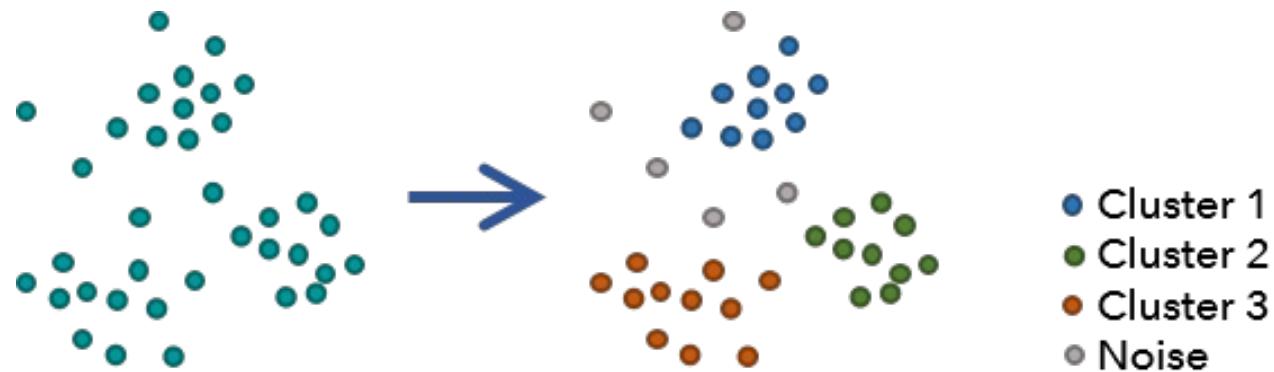
- lo scopo è di *raggruppare* gli items analizzati in gruppi con *caratteristiche simili*
- il *calcolo* effettuato per determinare le similitudini fra items è spesso la *distanza* in qualche spazio n-dimensionale
- fa riferimento ad analisi di apprendimento *non supervisionato*
- è di *supporto* in algoritmi semi-supervised



clustering

insieme di tecniche volte alla selezione e raggruppamento di elementi omogenei in un insieme di dati

le tecniche di clustering si basano su misure relative alla somiglianza tra gli elementi



bias

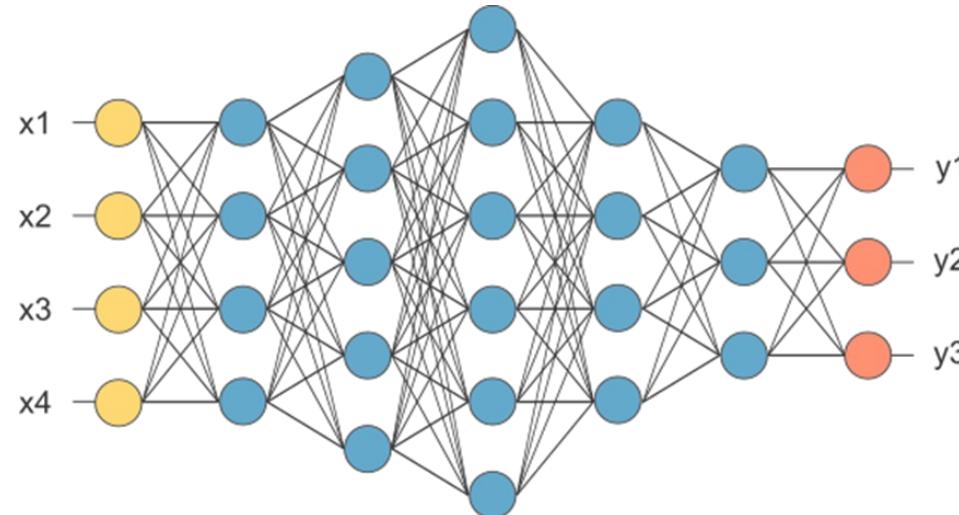
insieme di assunzioni che il classificatore usa per predire l'output dati gli input che esso non ha ancora incontrato

(Mitchell, 1980)



giudizi (o pregiudizi) che non corrispondono necessariamente alla realtà, sviluppati sulla base dell'interpretazione delle informazioni in possesso che portano a un errore di valutazione o mancanza di oggettività di giudizio

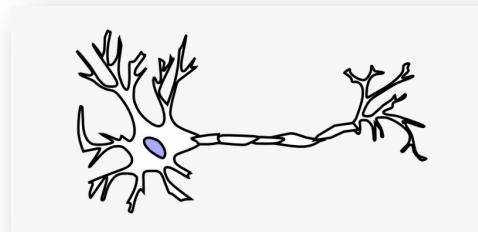
reti neurali artificiali



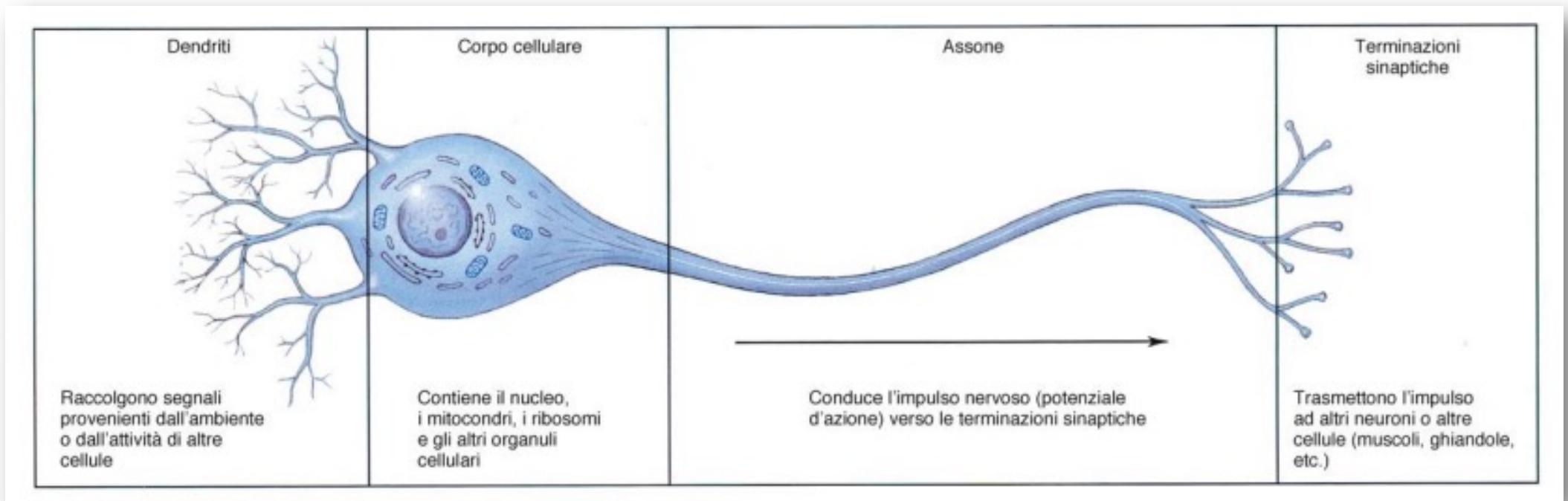
Alberto Ferrari – Analisi dei Dati

reti neurali artificiali

- **rete neurale** artificiale
 - “... un sistema di elaborazione costituito da una serie di elementi di elaborazione semplici e altamente interconnessi, che elaborano le informazioni mediante la loro risposta di stato dinamica agli input esterni.” (Robert Hecht-Nielsen)
- una rete neurale artificiale rappresenta un software che cerca di imitare come funziona il **cervello** umano
- gli elementi di elaborazione semplici e altamente interconnessi sono i **neuroni** (nodi)



Alberto Ferrari – Analisi dei Dati

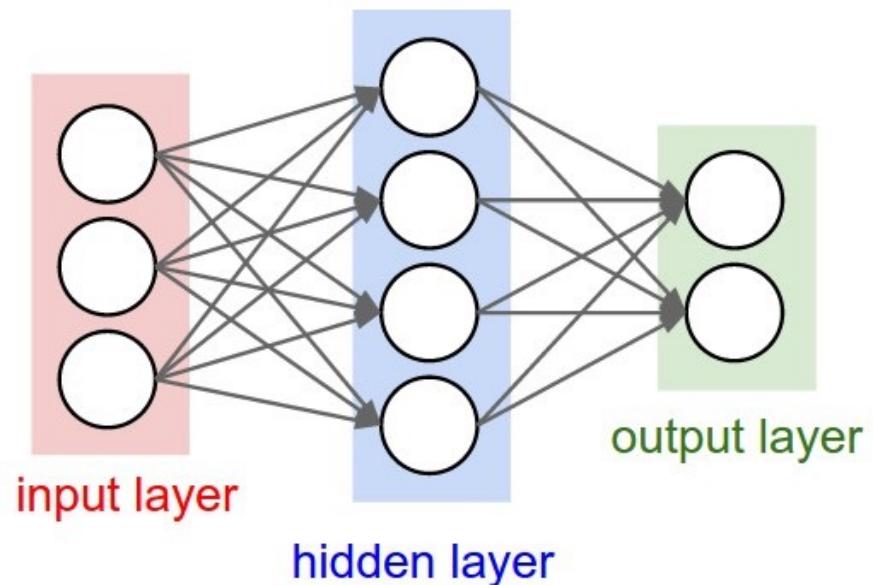


cervello umano

- il cervello è costituito da circa ***85 miliardi*** di neuroni
- i ***neuroni*** sono in grado di trasmettere informazioni sotto forma di impulsi elettrici ad alta velocità
- un neurone è costituito da due parti principali
 - il **soma**
 - i **neuriti** che si dividono in
 - **dendriti** (destinati alla ricezione dei messaggi che arrivano dagli altri neuroni)
 - **assone** (deputato alla trasmissione dei messaggi)
- la **sinapsi** è l'unione dell'assone di un neurone precedente con i dendriti di neuroni successivi e permette di trasmettere il messaggio elettrico da una cellula all'altra

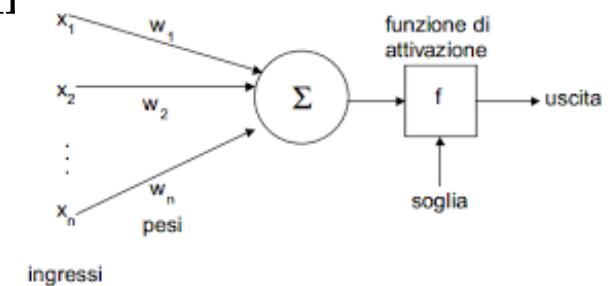
tipi di nodi - livelli

- ***input layer*** (livello di ingresso)
 - riceve le informazioni provenienti dall'esterno
- ***hidden layer*** (livello nascosto)
 - collega il livello di ingresso con quello di uscita e aiuta la rete neurale ad imparare le relazioni complesse
 - spesso i livelli nascosti sono più di uno
- ***output layer*** (livello di uscita)
 - mostra il risultato



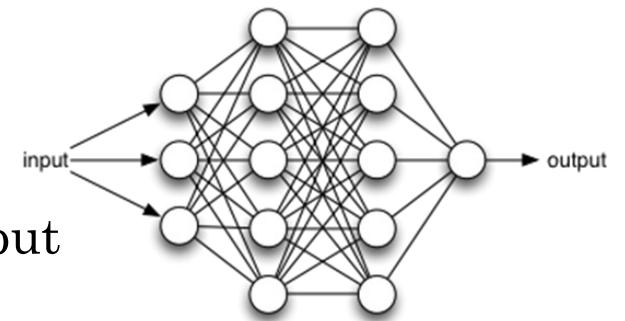
funzionamento

- ogni **livello** è formato da centinaia **neuroni** artificiali
- ogni neurone artificiale è connesso con neuroni di **altri livelli**
- ad ogni **connessione** è associato un **peso**
- i pesi iniziali sono impostati **casualmente**
- ad ogni neurone è associata una **funzione di attivazione** che dipende dai pesi delle connessioni in entrata
- la funzione **determina** l'**attivazione** o meno delle connessioni in uscita
- ogni serie di dati in ingresso attraversa tutti gli strati della rete e restituisce un output attraverso il livello di **uscita**



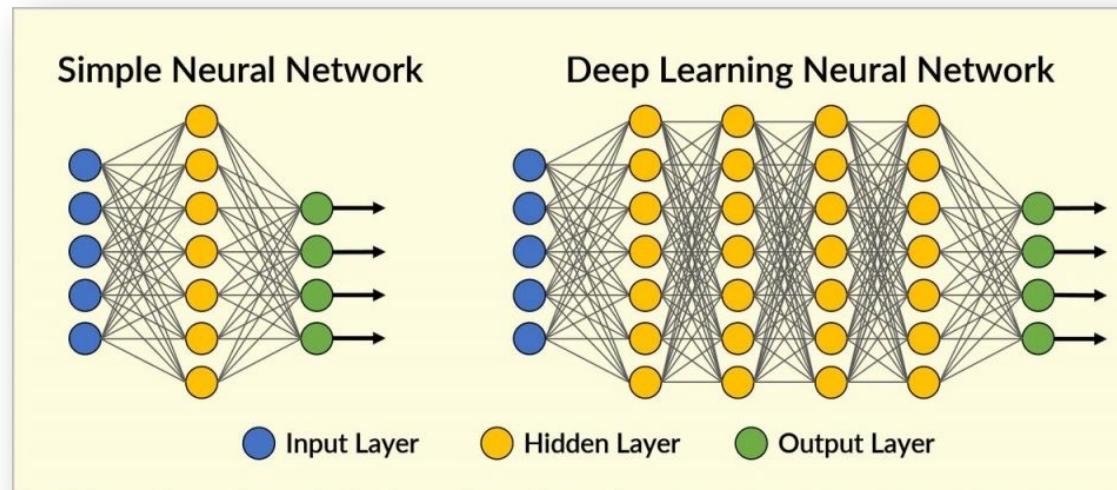
apprendimento

- ***feedback***
 - ***verifica*** della risposta in output in base ai dati di input
- algoritmo di ***backpropagation***
 - ***confronta*** il risultato ottenuto da una rete con l'output che si vuole in realtà ottenere
 - la differenza tra i due risultati prevede di ***modificare i pesi*** delle connessioni tra i livelli della rete partendo dal livello output
 - procedendo a ***ritroso*** modifica i pesi dei livelli nascosti e quelli dei livelli di input



deep learning

- si parla di deep learning quando una rete neurale artificiale che è composta da almeno **2 *livelli nascosti***
- le applicazioni di deep learning contengono normalmente **molte** più livelli nascosti (10, 20 o più)

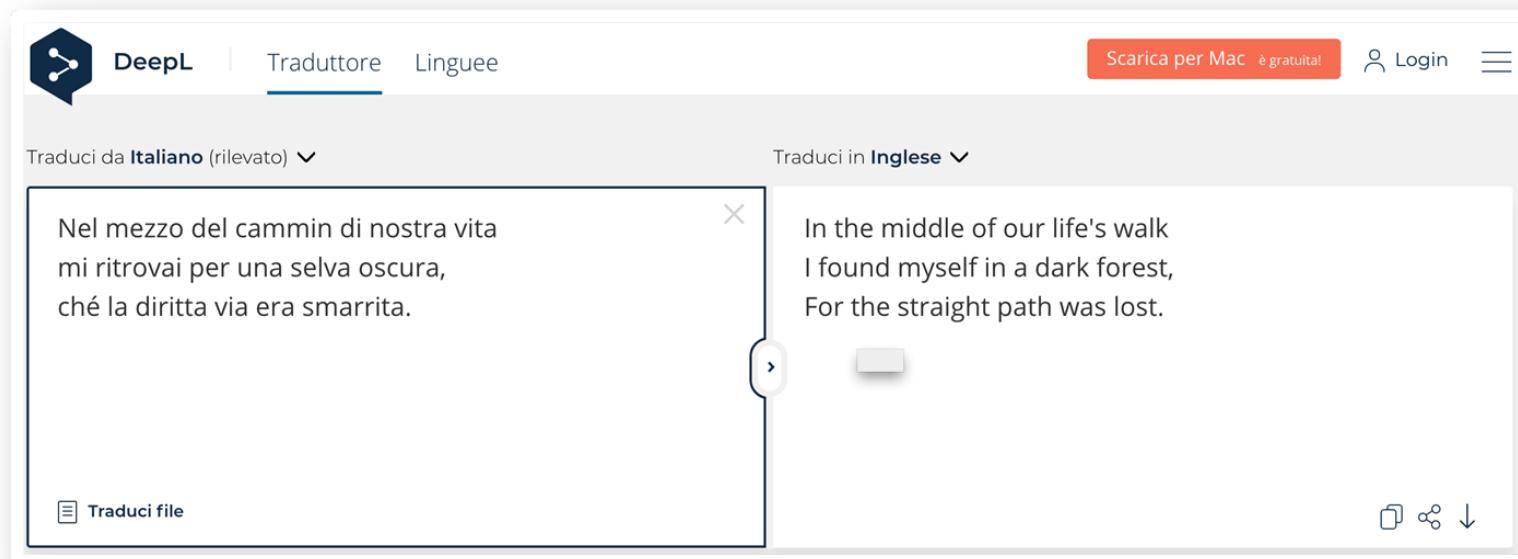


applicazioni deep learning

- ***traduzione automatica***
 - algoritmi di deep learning migliorano l'apprendimento delle relazioni tra parole e la loro mappatura in una nuova lingua
 - Google Neural Machine Translation
- ***classificazione*** di oggetti in immagini
 - algoritmi in grado di classificare gli oggetti di una immagine
- generazione automatica di ***linguaggio naturale***
 - applicazione che produce voce umana (es Wavenet)
- ***lettura delle labbra***
- ***colorazione*** automatica di immagini in bianco e nero

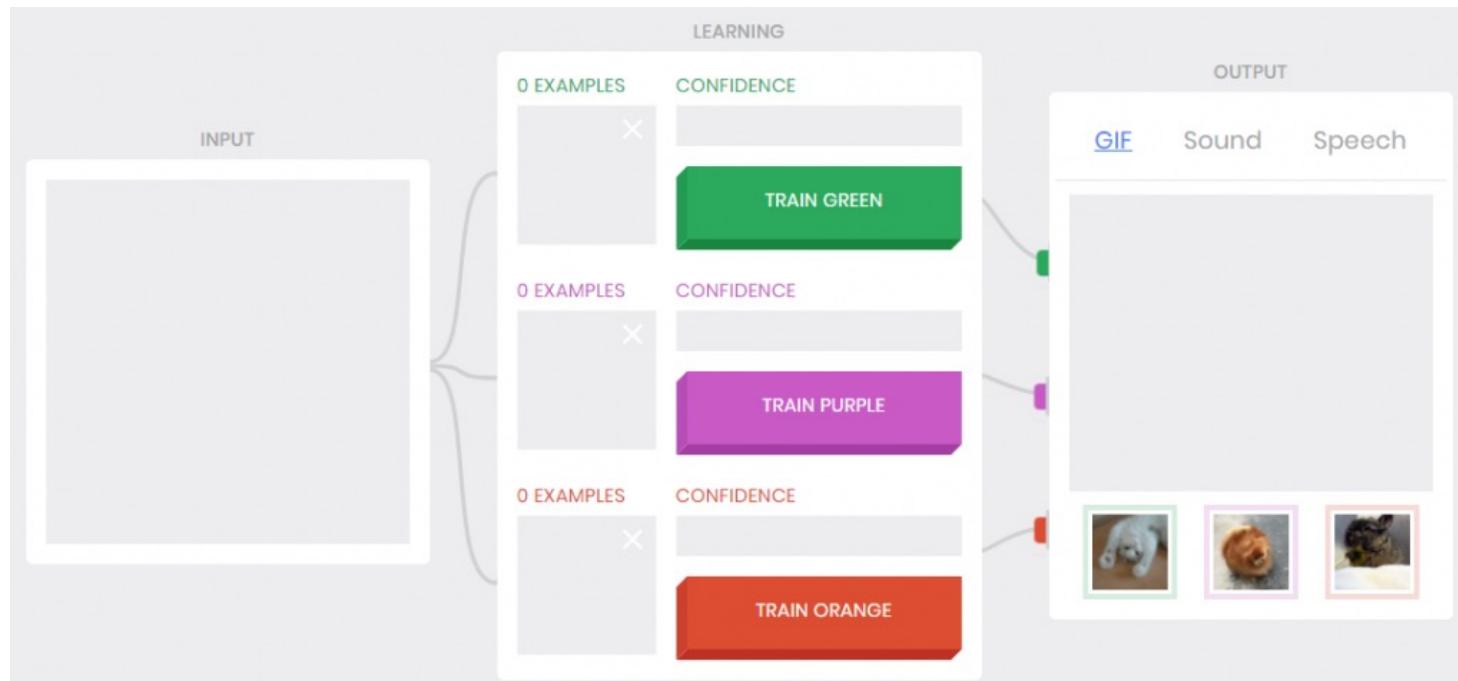
Traduzione automatica

<https://www.deepl.com/it/translator>



Alberto Ferrari – Analisi dei Dati

<https://www.youtube.com/watch?v=3BhkeY974Rg>



<https://teachablemachine.withgoogle.com/>

Alberto Ferrari – Analisi dei Dati

video

- *come funziona il machine learning*
 - https://www.youtube.com/watch?v=f_uwKZIAeM0
- *google immagini*
 - <https://www.youtube.com/watch?v=xkbBC9ZejI0&t=18s>
- *applicazioni*
 - <https://youtu.be/UwsrzCVZAb8>
- *test*
 - <https://teachablemachine.withgoogle.com>
- *apprendimento automatico spiegato in 5 minuti*
 - <https://www.youtube.com/watch?v=3bJ7RChxMWQ>

esempi

- esempio classificazione cifre numeriche
 - https://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html