Introduzione

Prefazione

Questi appunti si rifanno alle lezioni 2023/2024 del corso Introduction to Machine Learning tenuto dalla docente Elisa Ricci, al libro 'Deep Learning' di Ian Goodfellow e Yoshua Bengio; ed infine al libro 'Hands on machine learning' di Aurélien Géron pubblicato da O'Reilly.

Gli appunti sono scritti con typst, senza una panoramica sui diversi argomenti, ma affrontandoli uno ad uno a seconda della necessità. All'interno di questa introduzione troverete solo i concetti basilari, utili alla comprensione dei successivi argomenti.

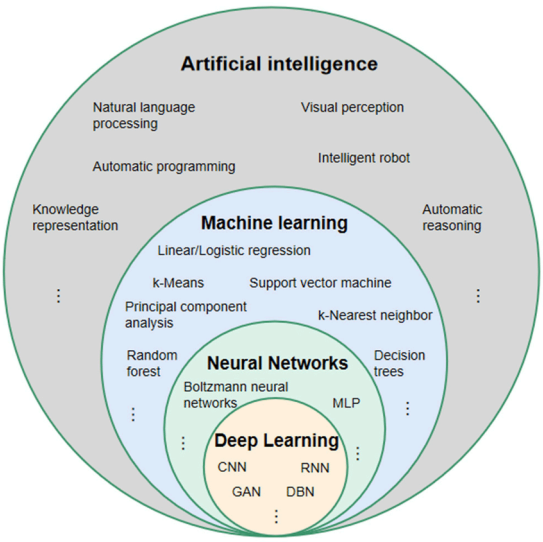

Figure 1: La relazione tra intelligenza artificiale, machine learning e deep learning.

Dataset

Il dataset è l'insieme dei dati disponibili per l'analisi. Su questo dataset si effettuano le operazioni di training e testing.

Il training set è il sottoinsieme del dataset utilizzato per addestrare il modello; mentre il test set è il sottoinsieme utilizzato per testare il modello. Il validation set è un sottoinsieme del training set utilizzato per regolare gli iperparametri del modello, prima della fase di testing.

Per generare questi sottinsiemi è necessario fare due assunzioni sui dati (*i.i.d. assumption*), ovvero che siano:

- **indipendenti** (non ci sia correlazione tra i dati del training set e del test set)
- **identicamente distribuiti** (prelevati dalla stessa distribuzione di probabilità p_{data})

Modello

L'obbiettivo, nel Machine Learning, è che il nostro modello performi bene su dati che non ha mai visto prima; questa abilità è detta *generalizzazione*. Ogni modello ha le sue peculiarità, e la scelta del modello giusto dipende dal problema (*task*) che si vuole risolvere. I modelli possono essere divisi in categorie, anche se con eccezioni e sfumature, a seconda del tipo di apprendimento:

- Supervised Learning: il modello apprende da un training set etichettato precedentemente.
- Unsupervised Learning: il modello apprende pattern o strutture dai dati senza etichette.
- Reinforcement Learning: il modello apprende attraverso il feedback di un ambiente.
- Semi-Supervised Learning: il modello apprende sia da dati etichettati che non etichettati. Viene utilizzato in sostituzione al supervised learning nei casi in cui etichettare i dati risulti troppo costoso o, richieda troppo tempo.

Task

Le principali task per cui viene adottato il Machine Learning sono:

- Classification: classificare un input in una delle classi predefinite.
- Regression: predire un valore numerico (continuo), dato un input.

- Transcription: convertire un input in testo. L'input può essere un'immagine, un audio, ecc.
- Machine Translation: tradurre un testo in un'altra lingua.
- Anomaly Detection: identificare pattern anomali nei dati.
- Synthesis: generare nuovi dati che seguano la stessa distribuzione dei dati originali. (*e.g.* textures, speech, ecc.)
- Denoising: in questo task il modello, ha come input un dato corrotto $\tilde{\boldsymbol{x}}$ e deve predire il dato originale bold(x); o meglio la distribuzione di probabilità $p(\boldsymbol{x}|\tilde{\boldsymbol{x}})$.
- Density Estimation:
- assd