**Machine Learning**

Varie aree Statistica, I.A, Data Science, Data Mining e Neuro Scienze contribuiscono alla realizzazione di algoritmi per l’apprendimento automatico (machine learning).

Data la grande mole di dati di chi oggi disponiamo, utilizzando algoritmi di autoapprendimento nel campo del machine learning, siamo in grado di trasformare questi dati in conoscenza. Si possono utilizzare questi algoritmi per individuare schemi nei dati ed eseguire previsioni sugli eventi futuri.

Nella seconda metà del ventesimo secolo, il ML si è evoluto come branca dell’Intelligenza Artificiale. Piuttosto che richiedere lka presenza una per individuare manualmente delle regole e costruire dei modelli per l’analisi di grandi quantità di dati, il ML offre un’alternativa più efficiente per catturare la conoscenza insita nei dati, modelli previsionali e poi prendere decisioni guidate dai dati.

Big Data: Enormi quantità di dati che non possono essere analizzati con metodologie classiche

Hardware: Gli algoritmi di ML sono degli algoritmi che eseguono calcoli matematici più o meno complessi e che quindi richiedono una potenza computazionale minima

Abbiamo tre metodi di apprendimento, tre tipi diversi di Machine Learning:

* **Apprendimento Supervisionato**

Alberi decisionali, Reti Neurali, SVM

Si basa sul concetto di addestramento.

All’algoritmo viene dato in input il set di dati da analizzare sia i risultati che ci aspettiamo ci restituisca. In questo modo sarà l’algoritmo ad addestrarsi. In questo modo passando dei nuovi dati l’algoritmo sarà in grado di creare le regole interne per determinare un risultato.

I dati in input ed output non so come sono correlati.

Consiste nel trarre un modello a partire da dati di addestramento etichettati, i quali ci consentono di effettuare previsioni relative a dati non disponibili o futuri. Nell’insieme di campioni i segnali di output desiderati (le etichette) sono già noti.

Un compito di apprendimento supervisionato, sulla base di etichette di classi discrete, è chiamato anche compito di classificazione. Un’altra sottocategoria di apprendimento con supervisione e la regressione dove il segnale risultante è un valore continuo.

Si possono risolvere due tipi di problemi:

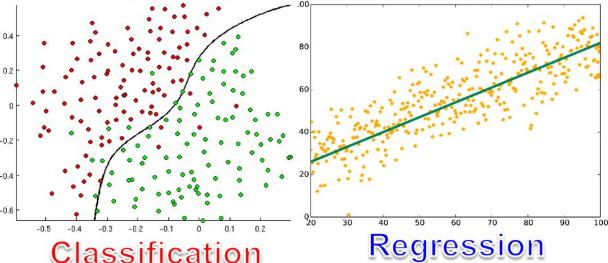
**Regressione**: Ho un Input continuo ed output continuo.

La previsione di risultati continui, chiamata anche analisi di regressione. Nell’analisi di regressione, abbiamo un certo numero di variabili predittive (descrittive) ed una variabile target continua (risultato): cerchiamo di trovare una relazione fra queste variabili, tale che ci consenta di prevedere un risultato.

Nell’immagine è illustrata la regressione lineare, date delle variabili predittive (asse x) e variabili risposta (asse y), tracciamo una linea retta attraverso questi dati in modo da minimizzare la distanza (distanza quadratica media) fra i punti del campione e la linea. Possiamo utilizzare il punto d’intersezione e la pendenza che abbiamo appreso da questi dati per prevedere la variabile target per nuovi dati.

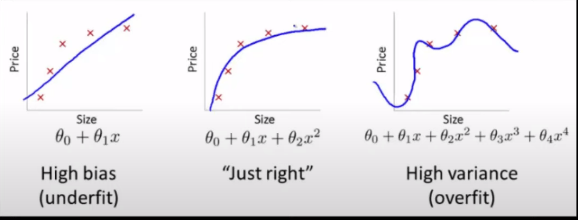
**Classificazione**: Ho un input discreto output discreto

La classificazione è una sottocategoria dell’apprendimento con supervisione, dove l’obiettivo è quello di prevedere le etichette di categoria delle classi per le nuove istanze, sulla base delle osservazioni compiute nel passato. Queste etichette sono valori discreti, non ordinati, che possono essere considerati come appartenenti ad un gruppo delle istanze.

*Classificazione binaria*: l’algoritmo d’apprendimento impara un insieme di regole con lo scopo di distinguere fra due possibili classi.

*Classificazione multiclasse*: un dataset d’apprendimento che è costituito da più esempi di scrittura a mano di ciascuna lettera dell’alfabeto, fornendo un nuovo carattere scritto a amano, il nostro modello predittivo sarà in grado di prevedere con una certa precisione la lettera corretta dell’alfabeto. Il compito di classificazione consiste nell’assegnare alle istanze etichette di categoria non ordinate.

Per valutare se il modello si utilizza un “funzione di costo”. Riesco a capire quando il modello ha sbagliato, il costo è proprio l’errore.

Non dobbiamo modellare i dati, ma creare un modello che impari a generalizzare, se gli si da un nuovo dato l’output sarà corretto.

La soluzione a questo problema e dividere il set di dati in due set TrainSet e TestSet. Dati in input i dati del TrainSet l’algoritmo apprenderà su questi dati, ma quello che ci interessa è che abbia imparato anche sul TestSet ovvero se l’algoritmo ha generalizzato dev’essere in grado di funzionare su dati su chi non è stato mai allenato (capire se funziona). La potenza sta appunto nel poter apprendere da molti parametri.

* **Apprendimento Non** **Supervisionato**

Clustering, Regole di associazione

Abbiamo una grande mole di dati l’algoritmo troverà dei pattern, delle relazioni…

Nell’apprendimento con supervisione, conosciamo in anticipo la risposta corretta quando descriviamo il nostro modello, mentre nell’apprendimento di rinforzamento definiamo una misura, o ricompensa, per le specifiche azioni messe in atto dall’agente. Nell’apprendimento senza supervisione, al contrario, abbiamo a che fare con dati non etichettati o dati dalla struttura ignota. Utilizzando tecniche di apprendimento senza supervisione siamo in grado di osservare la struttura dei nostri dati, per estrarre da essi informazioni cariche di significato senza però poter contare sulla guida né di una variabile nota relativa al risultato, né una funzione di ricompensa.

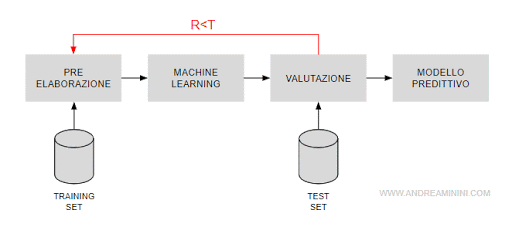
**Clustering:** Ricerca di sottogruppi tramite il clustering. Il clustering è una tecnica esplorativa di analisi dei dati che ci consente di organizzare una serie di informazioni all’interno di gruppi significativi detti cluster senza avere alcuna precedente conoscenza delle appartenenze a tali gruppi. Un gruppo di oggetti che condividono un certo grado di similarità, ma che sono più dissimilari rispetto agli oggetti presenti negli altri cluster.

Il clustering è un’ottima tecnica per la strutturazione dell’informazione e per individuare relazioni significative nei dati.

* **Apprendimento con** **rinforzo**

L’algoritmo impara dagli errori. Non partiamo dai dati ma (es. algoritmo della scacchiera impara dagli errori non dalle giocate degli altri). L’obiettivo è quello di sviluppare un sistema (agente) che migliori le proprie prestazioni sulla base delle interazioni con l’ambiente, le informazioni relative allo stato corrente dell’ambiente includono anche un cosiddetto segnale di ricompensa (reward), questo feedback non p l’etichetta o il valore corretto di verità, ma una misura della qualità con cui l’azione è stata misurata da una funzione di ricompensa.

Tramite l’interazione con l’ambiente, un agente può quindi utilizzare l’apprendimento di rafforzamento per imparare una serie di azioni che massimizzano questa ricompensa tramite un approccio esplorativo di tipo trial-and-error.

****

**Pre – elaborazione: date una “forma” ai dati**

La pre-elaborazione dei dati è uno dei passi cruciali di qualsiasi applicazione di apprendimento automatico. Molti algoritmi d’apprendimento automatico richiedono anche che, per ottenere le massime prestazioni, le caratteristiche scelte adottino la stessa scala, il che spesso viene ottenuto trasformano le caratteristiche in un intervallo [0,1] oppure in una distribuzione normale standard con media 0 e varianza 1.

Per determinare se il nostro algoritmo d’apprendimento non solo si comporta bene su un set di addestramento, ma segue generalizzazioni corrette sui nuovi dati, lo suddividiamo in modo casuale in due set distinti: set di addestramento / training ed il set di test.

Utilizziamo il set di addestramento per informare ed ottimizzare il modello di apprendimento automatico, mentre teniamo da parte fino all’ultimo il set di test, per valutare il modello finale.

**Addestramento e selezione di un modello predittivo**

È essenziale confrontare almeno un certo gruppo differente di algoritmi, in modo da addestrarli e selezionare poi il modello che offre le migliori prestazioni. Ma prima di poter confrontare modelli differenti, dobbiamo decidere le metriche da impiegare per misurare le prestazioni.

Una metrica comunemente utilizzata è l’accuratezza della classificazione, che è definita come la proporzione tra le istanze classificate correttamente.

Dobbiamo dunque capire quale modello si comporta meglio sul dataset di test finale e sui dati reali non utilizzando però questo dataset di test per la scelta del modello ma per la valutazione finale del modello stesso. Quindi il modello deve comportarsi al meglio sul dataset di test ma senza però essere addestrato sul dataset di test bensì su quello di training.

Possono essere utilizzate varie tecniche di convalida incrociata nelle quali il dataset di addestramento viene ulteriormente suddiviso in sottoinsiemi di addestramento di convalida, in modo da stimare le prestazioni di generalizzazione del modello.