Come funziona un Modello di Machine Learning basato su Alberi di Decisione

Prof. Fedeli Massimo

Introduzione

Un **albero di decisione** è un tipo di modello di *Machine Learning*, cioè un sistema che impara dai dati per fare previsioni o prendere decisioni.

Per capire cos'è, pensiamo ad un gioco di domande a cui si può rispondere solo con **sì** o **no**. Ad esempio: *"L'animale ha le ali?"* → sì/no *"Vive in acqua?"* → sì/no

Dopo alcune domande si arriva alla risposta finale. Un albero di decisione funziona in modo molto simile: fa una serie di domande sui dati finché non arriva a una conclusione.

La struttura dell'albero

Si chiama "albero" perché assomiglia a un albero rovesciato.

- **Radice**: è la prima domanda che il modello fa
- **Rami**: sono le possibili risposte (ad esempio sì/no)
- **Nodi**: sono le altre domande che seguono
- **Foglie**: sono le decisioni finali

Ogni percorso dalla radice a una foglia è come una **regola**. Per esempio: *Se età ¿ 18 e reddito ¿ 1000 → cliente affidabile*

Come impara l'albero

L'albero non nasce già pronto: deve **imparare** osservando molti esempi già risolti.

Immaginiamo di voler insegnare a un sistema a distinguere email normali da email di spam. Gli mostri tante email già etichettate: "spam" o "non spam".

Il modello prova a fare domande del tipo:

- L'email contiene tante parole in maiuscolo?
- Ci sono molti punti esclamativi?
- Contiene parole come "gratis" o "offerta"?

Per ogni domanda possibile, il modello valuta quanto quella domanda riesce a separare bene i casi.

L'idea chiave: creare gruppi simili

Per capire cosa fa davvero un albero di decisione, dobbiamo prima chiarire cosa si intende per **gruppi**.

Quando parliamo di dati, non stiamo parlando di numeri isolati, ma di un insieme di **casi** o **situazioni**. Ogni caso è una riga di informazioni. Per esempio:

- uno studente con età, voti e assenze
- un cliente con età, reddito e storico acquisti

- una email con testo, lunghezza e presenza di certe parole

All'inizio dell'addestramento, **tutti i casi sono messi insieme in un unico grande gruppo**. Questo gruppo è molto "mischiato": contiene esempi di tipi diversi (per esempio promossi e bocciati, spam e non spam, clienti affidabili e non affidabili).

Cosa fa l'albero

L'albero prova a **dividere questo grande gruppo in sottogruppi più piccoli**. Ogni divisione avviene facendo una domanda su una caratteristica dei dati.

Esempio con studenti:

"Il numero di assenze è maggiore di 10?"

Dopo la domanda otteniamo due gruppi:

- Gruppo 1: studenti con poche assenze
- Gruppo 2: studenti con molte assenze

Questi sono i "gruppi" di cui parliamo: insiemi di casi che hanno qualcosa in comune secondo una certa regola.

Cosa significa "simili"

Dire che un gruppo è **simile** o **omogeneo** significa che al suo interno gli elementi si assomigliano rispetto al risultato che vogliamo prevedere.

Se stiamo cercando di prevedere se uno studente sarà promosso o bocciato:

- Un gruppo con quasi tutti promossi è un gruppo molto omogeneo
- Un gruppo con metà promossi e metà bocciati è poco omogeneo

Quindi "simili" non vuol dire identici in tutto, ma simili **nel comportamento finale**.

Un'analogia semplice

Immagina di avere un cesto con frutta mista: mele, banane e arance tutte insieme.

All'inizio è tutto mescolato. Se separi le mele dalle altre, hai creato due gruppi:

- un gruppo con solo mele (molto omogeneo)
- un gruppo con banane e arance (ancora mescolato)

Poi potresti separare banane e arance, creando gruppi sempre più "puri".

L'albero di decisione fa la stessa cosa, ma invece di frutta separa dati in base a caratteristiche (età, reddito, parole in un testo, voti, ecc.).

Perché è così importante

Un gruppo molto mescolato rende difficile prendere una decisione corretta. Un gruppo molto omogeneo rende la decisione più facile e affidabile.

Per questo l'albero continua a dividere i dati in gruppi sempre più simili, finché ogni gruppo contiene casi che portano quasi tutti alla stessa risposta.

Questo è il cuore del funzionamento interno di un albero di decisione.

Cos'è il criterio di Gini (spiegato semplice)

L'indice di Gini misura il **disordine** dentro un gruppo.

Possiamo pensarlo così:

- Gini = 0 → gruppo perfettamente ordinato (tutti uguali)
- Gini alto → gruppo disordinato (tutti mescolati)

Quindi l'albero cerca sempre di fare domande che rendano i gruppi **più ordinati possibile**.

La formula del Gini

La formula matematica è:

$$Gini = 1 - \sum_{i=1}^{k} p_i^2$$

dove:

- k è il numero di categorie possibili
- p_i è la percentuale (probabilità) della categoria i nel gruppo

Non è necessario memorizzarla: serve solo a trasformare in un numero quanto un gruppo è mescolato.

Esempio molto intuitivo

Immagina un sacchetto con palline:

- 10 palline rosse → gruppo ordinato → Gini = 0
- 5 rosse e 5 blu → gruppo molto mescolato → Gini alto

L'albero cerca di dividere il sacchetto in modo che in ogni nuovo sacchetto le palline abbiano quasi tutte lo stesso colore.

Come sceglie la domanda migliore

Per ogni possibile domanda, l'albero:

1. Calcola quanto è disordinato il gruppo prima della divisione
2. Divide i dati in due gruppi
3. Calcola quanto sono disordinati i nuovi gruppi

Viene scelta la domanda che riduce di più il disordine totale.
Questo procedimento si ripete tante volte, creando nuovi rami, finché:

- i gruppi sono quasi puri, oppure
- si decide di fermarsi per non creare un albero troppo grande

Cosa succede quando deve fare una previsione

Quando arriva un nuovo dato (ad esempio una nuova email), il modello:

- parte dalla radice
- risponde alle domande una dopo l'altra
- segue il ramo corrispondente
- arriva a una foglia con la risposta finale

È come seguire un percorso in un quiz a risposta sì/no.

Perché è un modello facile da capire

Gli alberi di decisione sono tra i modelli più semplici da interpretare perché:

- assomigliano a un diagramma di flusso
- ogni decisione è una regola chiara
- si può spiegare il risultato seguendo il percorso fatto nell'albero

Per questo sono molto usati quando è importante non solo avere una risposta, ma anche **capire il motivo**.