ANALISI DEI DATI PER LE IMPRESE INDUSTRIALI (4)

Apprendimento non supervisionato

Nell'ultimo articolo (pubblicato su Filo Diretto dicembre 2021) abbiamo proseguito ad approfondire la preparazione dei dati, affrontando fenomeni come l'underfitting e l'overfitting, che possono penalizzare fortemente le prestazioni del nostro sistema.
Approfondiamo ora l'analisi degli algoritmi stessi. Nel numero di settembre 2021 abbiamo già visto due tipi di approccio, quello supervisionato (oggetto del prossimo articolo) e quello non supervisionato.
Gli algoritmi di apprendimento supervisionato utilizzano dati che hanno già un risultato atteso, che utilizzano per istruire il sistema; il modello creato è poi usato **per predire il valore delle nuove osservazioni**.
Gli algoritmi di apprendimento non supervisionato cercano di definire gruppi con simili caratteristiche nei dati. Non si ha un risultato atteso, ma si cerca di raggruppare i dati per similitudine**.**

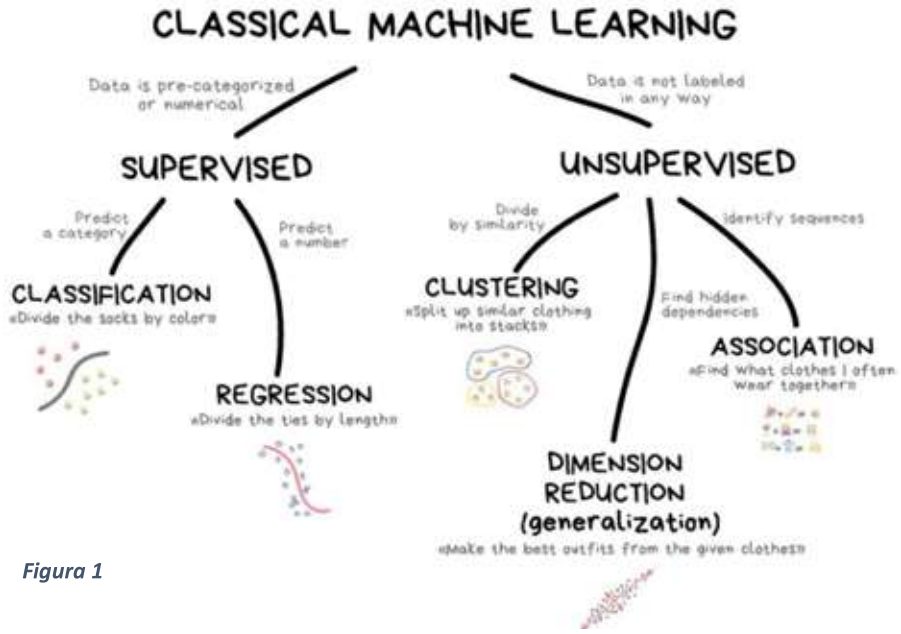

Figura 1

Con l'aiuto della figura 1 ricordiamo che l'oggetto del **Machine Learning in generale è quello di fornire descrizioni, previsioni e prescrizioni**, in ordine di complessità crescente, quindi **descrivere ciò che è successo, anticipare ciò che accadrà e fornire raccomandazioni per il futuro**.
Non ci occupiamo in questi articoli di algoritmi per le Reti Neurali, che richiedono una trattazione dedicata.

CLUSTERING

Un **algoritmo non supervisionato esplora i dati senza ricevere nessuna indicazione di come identificare i modelli che deve ricercare**. Si usa quando non si è certi di come classificare i dati e si desidera quindi che l'algoritmo classifichi i dati autonomamente.
L'algoritmo riceve i dati (ad esempio, tipi di mele, luoghi di produzione, data di raccolta, sapore) e analizza se ci sono relazioni tra i dati. Ovviamente c'è una relazione tra il tipo e il sapore. Come può essere il rapporto tra il luogo di produzione e il sapore? Si può valutare la data migliore di raccolta, se esiste la relazione con altri dati.
Ci sono vari metodi per raggruppare i dati (clustering), di seguito parliamo dei più comuni.

Clustering K-means

Raccoglie i dati in un certo numero di gruppi (K), che hanno caratteristiche determinate dall'algoritmo e sono posizionati intorno ad alcuni "centri". Il numero dei gruppi viene definito dall'operatore prima di iniziare, poi in caso viene corretto se necessario. Ogni punto sulla figura 2 rappresenta ad esempio un cliente.
È molto utilizzato nel marketing per segmentare i clienti in base a caratteristiche distinte, quindi per focalizzare meglio le campagne di promozione o prevenirne l'abbandono. **L'algoritmo lavora su più dimensioni immaginarie** quindi la figura mostrata è una rappresen-

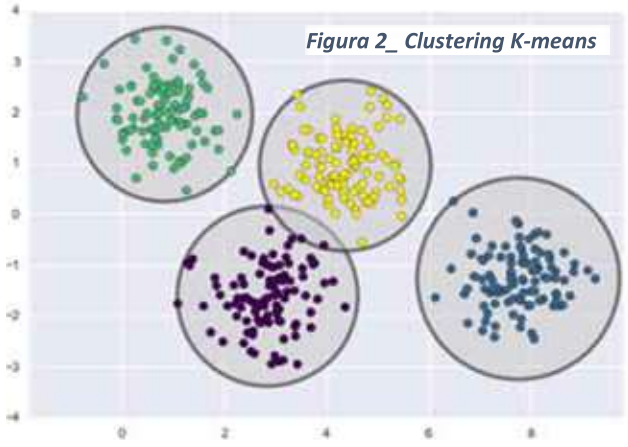

Figura 2_ Clustering K-means

tazione semplificata per fornire un esempio su due dimensioni (cerchi su un piano). **Questo metodo ha alcune limitazioni** perché riconosce meglio i dati con rappresentazioni regolari, es. circolari o sferiche, a seconda delle dimensioni geometriche in cui si lavora. Viene quindi definito come "hard clustering".

Gaussian Mixture

Per i dati con forma irregolare, si può usare il Gaussian Mixture (Fig. 3)**.**
Il metodo deriva dalla distribuzione normale statistica. Si ipotizza che i dati siano distribuiti in maniera gaussiana e che ciascuna di queste distribuzioni rappresenti un cluster. **Il numero di cluster viene determinato in automatico dal sistema.**
Questo metodo permette di segmentare i clienti in maniera più accurata nel marketing, usando caratteristiche meno evidenti, oppure permette di classificare i dipendenti in base alla probabilità che cerchino un'altra occupazione e che lascino l'azienda. È considerato come "soft clustering".

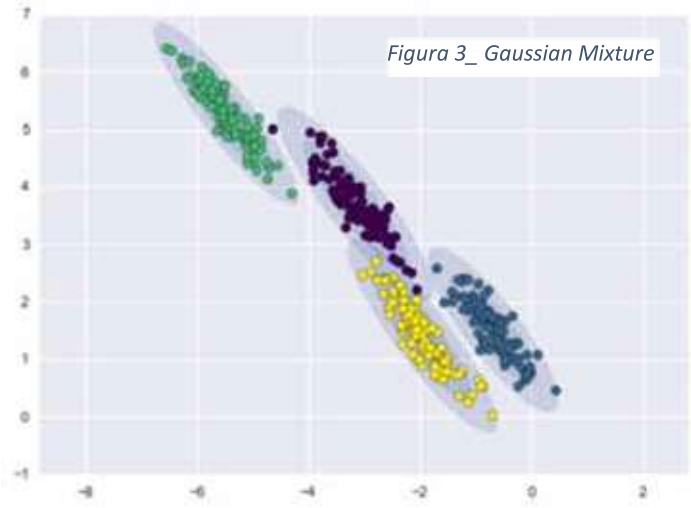

Figura 3_ Gaussian Mixture

Clustering gerarchico

Divide o aggrega i cluster, definiti con i metodi precedenti, **creando un albero gerarchico** (Fig. 4) per permettere una classificazione.
Ad esempio, **raggruppa i clienti dotati di tessere fedeltà** in gruppi progressivamente più caratterizzati per creare un'offerta dedicata, oppure **si usa per proporre nuovi prodotti**, raggruppando i clienti in base a vari parametri come, ad esempio, parole chiave nei social media.

Riduzione della dimensione

Esistono inoltre algoritmi non supervisionati **per la riduzione dimensionale dei dati**, evitando i problemi che si creano se i dati sono troppi e non portano informazioni utili al sistema.
Pensiamo, come concetto, all'indice BMI (Body Mass Index, che ci segnala il sovrappeso…) che da altezza e peso (due dimensioni) ci fornisce un risultato ad una sola dimensione.

Sistemi di associazione e raccomandazione

Facendo riferimento alla Figura 1 nella pagina precedente, parliamo ora dei sistemi di associazione e raccomandazione. Spesso si usa la **previsione del comportamento derivata dai cluster** per identificare i parametri importanti per fare una raccomandazione.
Possiamo pensare a **Netflix che ci raccomanda quali film potremmo vedere** in base alle preferenze di altri clienti con attributi simili.
Amazon utilizza moltissimo questi sistemi. Basta ordinare un libro che ci vengono subito raccomandati articoli che potremmo voler leggere.
L'immagine è ripresa da Amazon che raccomanda i "Recommender Systems" …

Immagini tratte da https://towardsdatascience.com, Amazon, https://vas3k.com

Nel prossimo numero di Filo Diretto proseguiremo l'analisi.

Figura 4 Clustering gerarchico