BookSales UniBG: iterazione 2

Del Prete Giovanni, Ghilardi Nicola, Polver Marco

May 23, 2020

Contents

0.1 Algoritmo: il clustering

Il clustering è un algoritmo di apprendimento non supervisionato (machine learning) che ha l'obiettivo di ricercare gruppi di oggetti tali che gli oggetti appartenenti a un gruppo siano "simili" tra loro e differenti dagli oggetti negli altri gruppi. Dunque, in statistica, è una tecnica di analisi multivariata dei dati volta al raggruppamento di elementi omogenei in un insieme di dati.
Oggi le tecniche di clustering sono diffuse in diversi campi applicativi, tra i quali: il marketing, l'analisi del territorio, le assicurazioni gli studi sismici e l'image processing. Uno degli usi più comuni che viene fatto dei cluster, in ambito economico, è proprio la segmentazione di mercato, che può essere riferita a consumatori o a categorie di prodotti con lo scopo di valutare le caratteristiche e i comportamenti dei consumatori, personalizzando l'offerta e incrementando dunque la quantità di prodotti venduti.

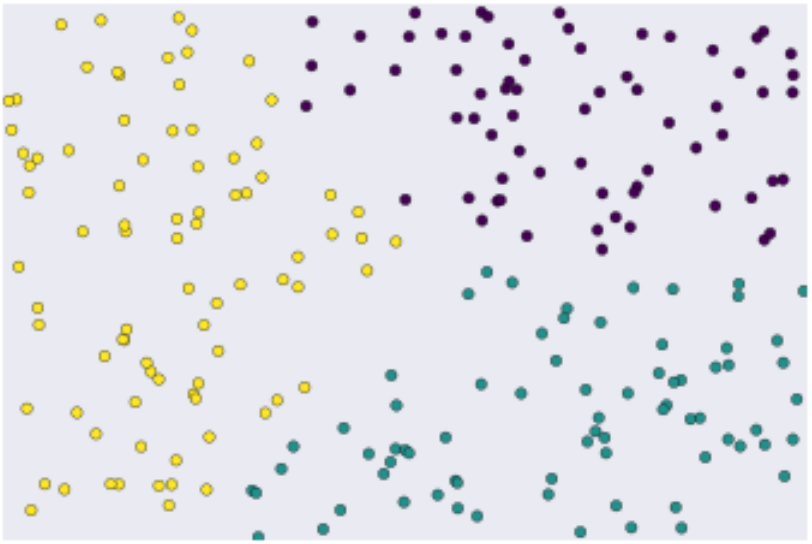

Figure 1: Cluster

0.1.1 K-means clustering

Il k-means clustering è l'algoritmo di clustering più utilizzato sia in ambito accademico che in ambito industriale e funziona nel seguente modo.

1. Scegliere il numero di k clusters in cui raggruppare il dataset.
2. Selezionare casualmente k centroidi iniziali (con centroide si intende il centro geometrico di ogni cluster).
3. Calcolare la distanza tra ogni centroide e tutte le osservazioni.
4. Assegnare ognuna delle osservazione al cluster rappresentato dal centroide più vicino.
5. Ricalcolare la posizione dei centroidi come la posizione media delle osservazioni appartenenti al cluster che il centroide rappresenta.
6. Ripetere dal punto 3, finchè nessuna osservazione cambierà più il proprio cluster di appartenenza.

Problema: come scegliere a priori il numero di clusters?
La soluzione a questo problema sta nel testare in modo iterativo più valori di k e confrontarli. Per valutare il risultato viene analizzata la funzione di costo:

$$J(k) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n}|p_{\mathrm{j}} - c_{\mathrm{i}}|$$

Tale funzione di costo è la media delle distanze medie delle osservazioni dal centroide del proprio cluster. E' facile notare che aumentando il numero di cluster diminuirà, di conseguenza, la media delle distanze delle osservazioni dai centroidi; dunque la funzione di costo è decrescente.
L'*elbow method* è un metodo "grafico" che consente di determinare il numero di cluster ideale valutando la pendenza della funzione di costo. Come si evince dalla figura 4, il valore ottimo di k è prossimo al punto dove la funzione di costo inizia a decrescere più lentamente.

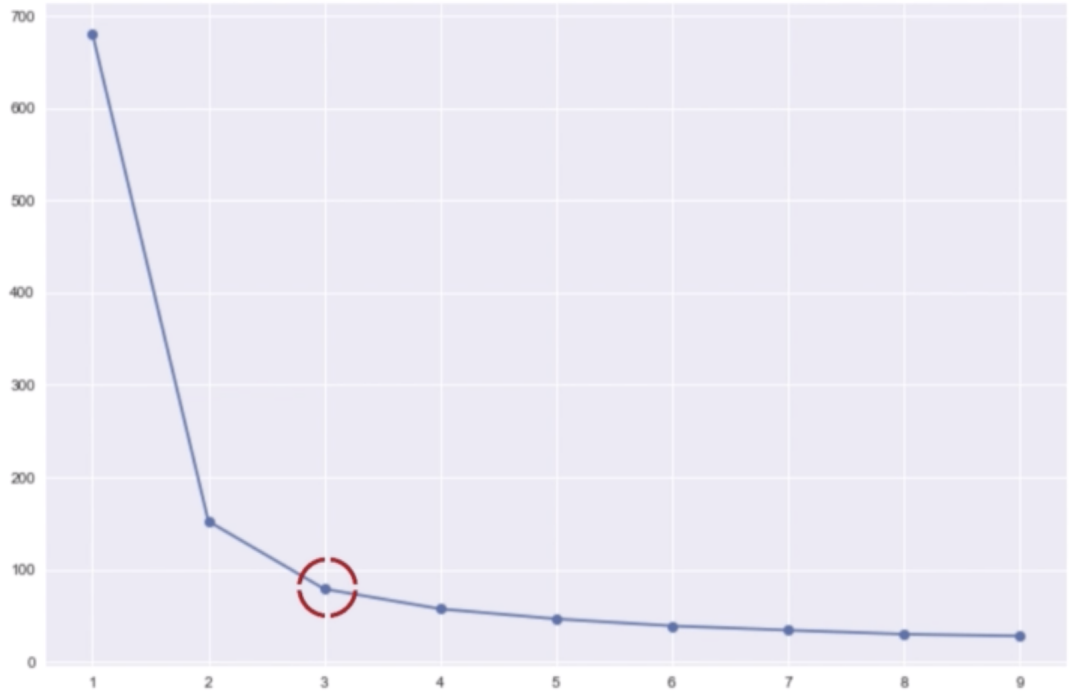

Figure 2: Elbow method

Di seguito si può osservare in pseudocodice la funzione che assegna ad ogni osservazione un cluster di appartenenza.

Data: Matrice delle osservazioni, k cluster
Result: Cluster di appartenenza di ogni osservazione
creazione k centers casuali;
while *centers non cambia* **do**
    distanza ← centers - osservazione;
    for *ogni cluster* **do**
        if *distanza minima* **then**
            clusters ← osservazione;
        end
    end
    centers ← media osservazioni clusters
end
return clusters

La funzione di calcolo dei cluster è chiamata da una funzione che calcola, tramite l'elbow method, il numero ottimale di cluster in cui dividere le osservazioni. Essendo l'elbow method un metodo di valutazione grafico, è stato necessario impostare sperimentalmente e compatibilmente con il problema, un valore target di derivata da raggiungere per ottenere il numero adeguato di cluster. E' stato necessario impostare altri due parametri per limitare la complessità temporale del processo di clustering: una valore massimo di k_max da testare e un contatore per limitare le iterazioni.

Data: media delle distanze medie delle osservazioni dal centroide per ogni cluster
Result: numero k di cluster ottimale
k = 2;
fisso un k_max di cluster;
while $|derivata| < 0.1$ *e* $k < k\_max$ *e* $count < 100$ **do**
    calcolo cluster e centroidi;
    cluster_cost ← media delle distanze dal proprio centro di un cluster;
    total_cost ← media di cluster_cost;
    derivata ← $total\_cost_{k-1} - total\_cost_k$;
    k ← k + 1;
    count ← count + 1;
end
return k

0.1.2 Il clustering in BookSales Unibg

L'algoritmo di k-means clustering è stato utilizzato in BookSales Unibg con l'obiettivo di suddividere in cluster gli utenti sulle basi dei loro interessi. L'algoritmo assegna ad ogni studente un cluster di appartenenza basandosi sui titoli presenti nella sua Wish List e Interesting Title. Ogni libro presente in BookSales Unibg appartiene ad una categoria: le features di ogni studente sono il numero di libri che interessano allo studente appartenenti ad ogni categoria. Le categorie in cui sono stati suddivisi i libri sono:

- Fisica
- Matematica
- Informatica
- Meccanica
- Elettronica
- Economia

- Automazione

L'algoritmo viene eseguito dall'applicazione automaticamente una volta al giorno. Dopo i risultati del clustering, ogni studente vedrà nella pagina Suggested Adds i titoli in linea con i propri interessi. Dato il numero elevato di features e di cluster non è possibile visualizzare graficamente i risultati di un'esecuzione dell'algoritmo.

0.2 Funzionalità aggiunte

All'applicazione sono state aggiunte funzionalità per renderla più user friendly. La prima è quella di fornire un'anteprima della copertina dei libri in copertina. Le immagini vengono caricate nel database e in fase di ricerca vengono lette e caricate nella schermata.
Ogni utente inoltre ha la possibilità di aggiungere un libro nella propria wishlist. Per permettere tale operazione si è aggiunto un tasto *Add to wishlist*.