Introduzione all'Apprendimento Non Supervisionato

Docente: Tommaso Muraca

Apprendimento Non Supervisionato

L'**apprendimento** non supervisionato viene **chiamato "non supervisionato"** perché l'analisi inizia con **dati senza etichetta (non c'è y).**

I **due compiti** di apprendimento non supervisionato che esploreremo sono il **Clustering dei dati in gruppi per somiglianza** e la **riduzione della dimensionalità** per comprimere i dati mantenendone la struttura e l'utilità.

Ecco 2 esempi pratici in cui i metodi di **apprendimento non supervisionato** potrebbero essere **utilizzati**:

- La **segmentazione** che potrebbe fare una **piattaforma pubblicitaria** della popolazione in gruppi più piccoli con dati demografici e abitudini di acquisto simili in modo che gli inserzionisti possano raggiungere il loro mercato target con annunci pertinenti.

- La **riduzione di un set di dati di grandi dimensioni** che potrebbe fare un team di **data science** per semplificare la modellazione e ridurre le dimensioni del file.

Una **grande differenza** che si ha con **l'apprendimento supervisionato** è che non è sempre facile fornire **parametri relativi al rendimento** di un algoritmo di apprendimento non supervisionato, la **"prestazione"** è infatti **spesso soggettiva** e specifica del dominio.

Clustering

Un **esempio** interessante di **clustering nel mondo reale** è il sistema di **clustering delle fasi di vita del fornitore di dati di marketing Acxiom Personicx.**

Questo servizio **segmenta le famiglie statunitensi in 70 cluster** distinti all'interno di **21 gruppi di fasi della vita** utilizzati dagli inserzionisti quando scelgono come target annunci Facebook, annunci display, campagne di direct mailing, ecc.

Il loro **white paper** rivela che hanno **utilizzato il centroid clustering e l'analisi delle componenti principali** , entrambe tecniche che tratteremo.

Potete immaginare come **l'accesso a questi cluster** sia **estremamente utile** per gli inserzionisti che desiderano:

- **Comprendere** la propria base di clienti esistente;

- **Utilizzare** la propria spesa pubblicitaria in modo efficace indirizzando potenziali nuovi clienti con dati demografici, interessi e stili di vita pertinenti.

Esaminiamo ora un paio di **metodi di clustering** per sviluppare l'intuizione su come eseguire questa attività.

Clustering: K-means

L'**obiettivo** del **clustering** è **creare gruppi di punti** dati in modo tale che i **punti in cluster diversi siano diversi mentre i punti all'interno di un cluster siano simili.**

Con clustering **k-means**, vogliamo **raggruppare i nostri punti dati in k gruppi,** l'output dell'algoritmo sarebbe un insieme di "etichette" che assegnano ciascun punto dati a uno dei k gruppi.

Nel clustering k-mean, il modo in cui **questi gruppi vengono definiti** è **creando un centroide** per ciascun gruppo, i **centroidi sono come il cuore del cluste**r, "catturano" i punti a loro più vicini e li aggiungono al cluster.

Ecco i **passaggi per il clustering k-means:**

1. Definire i k centroidi . Inizializzali in modo casuale (esistono anche algoritmi più elaborati per inizializzare i centroidi che finiscono per convergere in modo più efficace).

2. Trovare il centroide più vicino e aggiornare le assegnazioni dei cluster. Assegnare ciascun punto dati a uno dei k cluster, ogni punto dati viene assegnato al cluster del centroide più vicino. Qui, la misura della "vicinanza" è un iperparametro – spesso una distanza euclidea.

3. Spostare i centroidi al centro dei rispettivi cluster. La nuova posizione di ciascun baricentro viene calcolata come la posizione media di tutti i punti nel suo cluster.

Continuare a ripetere i passaggi 2 e 3 finché il **baricentro non smette di muoversi** molto ad ogni iterazione (ovvero finché l'algoritmo non converge).

Clustering: K-means

In questo **grafico a destra** si vede in breve **come funziona il clustering k-means:**

- ogni punto del piano è colorato in base al baricentro a cui è più vicino in ogni momento.

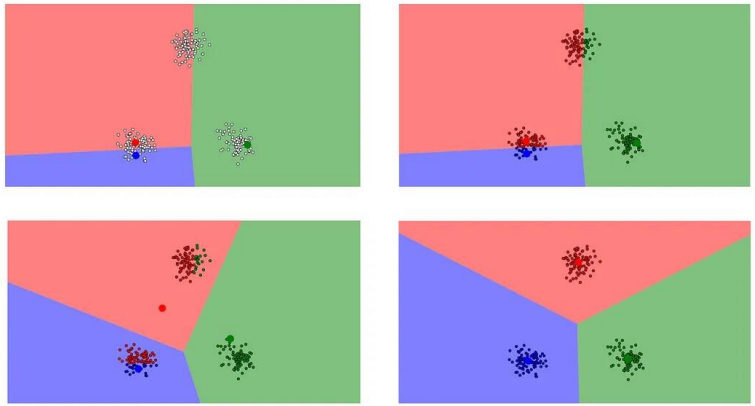
Noterete che **i centroidi** (i cerchi blu, rossi e verdi più grandi) **iniziano in modo casuale** e poi **si adattano rapidamente per catturare i rispettivi gruppi.**

Un'altra applicazione nella vita reale del clustering k-means è la **classificazione delle cifre scritte a mano.**

Partiamo sempre dal **solito dataset con immagini delle cifre** come un lungo vettore di luminosità dei pixel.

Diciamo che le **immagini sono in bianco e nero** e misurano **64x64 pixel**, ogni pixel rappresenta una dimensione.

Quindi il mondo in cui vivono queste immagini ha **64x64=4.096 dimensioni**, in questo mondo a 4.096 dimensioni, il clustering **k-means** ci consente di **raggruppare le immagini vicine tra loro e presupporre che rappresentino la stessa cifra**, il che può ottenere risultati piuttosto buoni per il riconoscimento delle cifre.



Clustering gerarchico

Il clustering gerarchico è **simile al clustering normale**, tranne per il fatto che il **suo obiettivo è creare una gerarchia di cluster,** ciò può essere **utile** quando desideriamo **flessibilità nel numero di cluster**.

Ad **esempio**, immaginate di **raggruppare articoli su un mercato online** come **Etsy o Amazon,** nella **home** page vorremmo alcune **ampie categorie di articoli per una navigazione semplice**, ma **man mano** che entriamo in **categorie di acquisto più specifiche vorremmo livelli crescenti di granularità**, cioè gruppi di articoli più distinti.

In **termini di output dell'algoritmo**, oltre alle assegnazioni dei **cluster costruiamo anche un bell'albero** che ci informa sulle **gerarchie tra i cluster,** possiamo quindi **scegliere il numero di cluster che desideriamo da questo albero.**

Ecco i **passaggi per il clustering gerarchico:**

1. **Iniziamo con N cluster**, uno per ciascun punto dati.

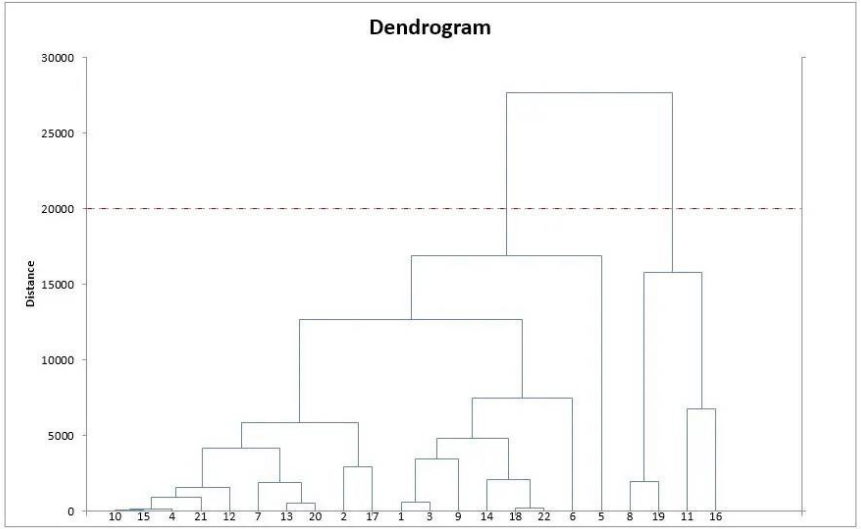
2. **Uniamo i due cluster più vicini tra loro,** ora abbiamo i cluster N-1.

Clustering gerarchico

3. Ricalcoliamo le distanze tra i cluster. Esistono diversi modi per farlo, uno di questi (chiamato clustering con collegamento medio) consiste nel considerare la distanza tra due cluster come la distanza media tra tutti i loro rispettivi membri.

4. Ripetere i passaggi 2 e 3 finché non si ottiene un cluster di N punti dati. Ottieniamo quindi un albero (noto anche come **dendrogramma**) come quello qui a destra.

5. Scegliere un numero di cluster e tracciare una linea orizzontale nel dendrogramma. Ad esempio, se desideriamo k=2 cluster, dovremmo tracciare una linea orizzontale attorno a "distanza=20000". Otterremo un cluster con i punti dati 8, 9, 11, 16 e un cluster con il resto dei punti dati. In generale, il numero di cluster che ottieniamo è il numero di punti di intersezione della nostra linea orizzontale con le linee verticali nel dendrogramma.



Andiamo ora alla pratica