Raccolta di appunti di Data e Web Mining

Massimo, Vettori Stefano, Masiero
Thomas, Vego Scocco

December 20, 2022

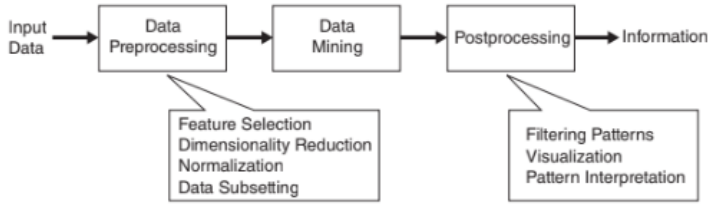
Contenuti

Chapter 1

Introduction

1.1 Cosa vuol dire fare Data Mining?

Il data mining è il processo di scoperta automatica di informazioni utili in grandi basi di dati. Le tecniche di data mining vengono implementate per setacciare grandi insiemi di dati al fine di trovare modelli nuovi e utili per prevedere l'esito di un'osservazione futura.



I dati passati in input ad un algoritmo di data mining possono essere memorizzati in una varietà di formati (file flat, fogli di calcolo, o tabelle relazionali) e possono risiedere in un repository di dati centralizzato o essere distribuiti su più siti.

1.2 Fasi di un algoritmo di Data Mining

Data Preprocessing

Durante questa fase i dati in input grezzi vengono trasformati, quindi vengono puliti dal rumore, vengono rimossi i duplicati e vengono selezionate le variabili/feature più significative. Questo processo è quello più laborioso e

dal quale dipende poi tutta la successiva fase di estrapolazione dei modelli, se viene fatto male quindi l'intero processo è compromesso.

Data Postprocessing

Durante questa fase solo i risultati validi e utili vengono tenuti per fare previsioni. Un esempio di post-elaborazione è la visualizzazione, che consente agli analisti di esplorare i dati e i risultati da diversi punti di vista. I metodi di verifica delle ipotesi possono essere applicati anche durante la post-elaborazione.

1.3 Quali sono i task del Data Mining?

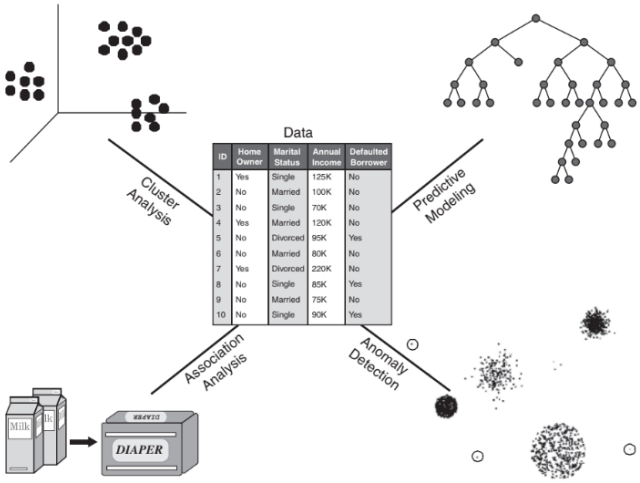
Le attività di data mining sono generalmente suddivise in due categorie principali:

1.3.1 Compiti predittivi

il cui obiettivo è quello di predire il valore di un particolare attributo basandosi sui valori di altri attributi. L'attributo che vogliamo predirre è comunemente chiamato(obbiettivo, variabile dipendente o etichetta da predirre), mentre gli attributi utilizzati per fare la previsione sono noti come (variabili esplicative, indipendenti o feature).

1.3.2 Compiti descrittivi

In questo caso l'obiettivo è derivare modelli (correlazioni, trend, cluster, traiettorie e anomalie) che riassumono le relazioni antecedenti dei dati.



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

1.3.3 La modellazione predittiva

Si riferisce al compito di costruire un modello per la variabile target in funzione delle variabili esplicative. Ci sono due tipi di attività di modellazione predittiva: la classificazione, utilizzata per classificare etichette discrete e la regressione, che viene invece utilizzata per le variabili target continue. Per esempio, prevedere se un utente Web effettuerà un acquisto online è un'attività di classificazione perché la variabile di destinazione è un valore binario. D'altra parte, la previsione del prezzo futuro di un'azione è un compito di regressione perché il prezzo è un valore continuo. L'obiettivo di entrambi i compiti è quello di apprendere un modello che riduca al minimo l'errore tra i valori previsti e quelli reali della variabile obiettivo.

1.3.4 L'analisi delle associazioni

Viene utilizzata per scoprire modelli che descrivono fortemente caratteristiche associate ai dati. I modelli scoperti sono tipicamente rappresentati sotto forma di regole di implicazione o sottoinsiemi di funzionalità. A causa dell'esponenziale dimensione del suo spazio di ricerca, l'obiettivo dell'analisi di associazione è quello di estrarre il massimo delle informazioni in modo efficiente. Utili applicazioni dell'analisi delle associazioni includono la ricerca di gruppi di geni che hanno funzionalità correlate, identificare le pagine Web

a cui si accede insieme o comprendere le relazioni tra i diversi elementi del sistema climatico terrestre.

1.3.5 L'analisi dei cluster

Cerca di trovare gruppi di osservazioni strettamente correlate in modo che le osservazioni che appartengono allo stesso cluster siano più simili tra loro rispetto alle osservazioni che appartengono ad altri cluster. Clustering è stato utilizzato per trovare gruppi di clienti correlati, trovare aree dell'oceano che hanno un significativo impatto sul clima terrestre e comprimere i dati.

1.3.6 Il rilevamento delle anomalie

è il compito di identificare le osservazioni le cui caratterestiche sono significativamente diverse dal resto dei dati. Tali osservazioni sono note come anomalie o valori anomali. L'obiettivo di un algoritmo di rilevamento delle anomalie è scoprire le vere anomalie ed evitare di etichettare erroneamente oggetti normali come anomali. In altre parole, un buon rilevatore di anomalie deve avere un alto tasso di rilevamento e un basso tasso di falsi allarmi. Applicazioni del rilevamento delle anomalie possono includere il rilevamento di frodi, intrusioni di rete, modelli insoliti di malattie, e disturbi dell'ecosistema, come siccità, inondazioni, incendi, uragani, ecc.

Chapter 2

Preprocessing

Vedere capitolo 2 del libro e riassumere le cose principali

Chapter 6

Classificazione: Tecniche avanzate

6.1 Tipi di classificatori

Per distinguere diversi tipi di classificatori si può fare riferimento alle diverse caratteristiche dei loro **output**

Classificatori Binari e Multi-class

I classificatori binari assegnano ad ogni istanza una fra 2 possibili etichet-te, solitamente $+1$ e -1. Inoltre solitamente la classe che risulta di interesse e' quella positiva. Differentemente da quelli binari, quelli Multi-class assegnano ad ogni istanza una fra k possibili etichette, dove k e' il numero di classi.

Classificatori Deterministici e Probabilistici

Come sottolinea il nome quelli deterministici producono un valore discreto in output per ogni istanza che viene classificata. Mentre quelli robabilistici producono una probabilita' che l'istanza appartenga ad una certa classe. Tale probabilita' e' rappresentata da un valore continuo compreso tra 0 e 1.

Classificatori Lineari e Non-Lineari

I primi usano un iperpiano per separare linearmente le istanze di classi differenti. Al contrario quelli non-lineari usano una costruzione piu' complessa che gli permette di separare anche non linearmente le classi.

Classificatori Globali e Locali

I classificatori globali adattano un singolo modello all'intero set di dati. A meno che il modello non sia altamente non lineare, questa strategia unica può non essere efficace quando la relazione tra gli attributi e le etichette di classe efficace quando la relazione tra gli attributi e le etichette della classe varia nello spazio di input. Al contrario quelli locali suddividono lo spazio di ingresso in regioni più piccole e adatta un modello distinto alle istanze di training in ciascuna regione.

Classificatori Generativi e Discriminativi

Data un istanza x normalmente i classificatori producono un etichetta y per classificarla. Al contrario i classificatori generativi producono un istanza attendibile a pratire dalla sua etichetta. Al contrario i classificatori discriminativi producono un etichetta a partire dalla sua istanza.

6.2 Classificatori KNN

In una visione ad alto livello, l'algoritmo computa la distanza (o similarita') fra ogni istanza di test (X', y') e di train $(X, y) \in D$ per determinare la lista di istanze piu simili ad essa. Una volta calcolata, all'istanza nuova, presa dal set di test verra' assegnata un etichetta in base alla maggioranza delle etichette delle k istanze piu' simili ad essa.

La "votazione per maggioranza" si puo calcolare anche come:

$$y' = \arg\max_{v} \sum_{i=1}^{k} \mathbb{I}_{v=y_i}$$

dove v e' un'etichetta di qualche classe, y_i l'etichetta della classe di uno dei k vicini piu' simili a X' e $\mathbb{I}_{v=y_i}$ e' un indicatore che vale 1 se $v = y_i$ e 0 altrimenti.

Si puo' inoltre estendere la formula aggiungendo un peso w_i a ciascun vicino i:

$$y' = \arg\max_{v} \sum_{i=1}^{k} w_i \mathbb{I}_{v=y_i}$$

6.2.1 Caratteristiche di un classificatore KNN

1. Questo tipo di classificazione e' parte di una tecnica molto piu' generalizzata detta "instance-based learning", la quale non costruisce un modello generale ma memorizza le istanze di training e usa la similarita' fra esse per classificare nuove istanze.

2. La classificazione di una nuova istanza puo' risultare costosa, dato che e' necessario calcolare la similarita' fra essa e tutte le istanze di training.

3. Generalmente i classificatori KNN producono le loro predizioni basandosi su un contesto locale delle istanze, al posto di creare delle regole piu' generalizzate.

4. Essi producono dei "bordi di decisione" anche complessi, rendendo questo classificatore piu' flessibile rispetto ad alberi di decisione e classificatori basati su regole di separazione.

5. Esiste una difficolta' non trascurabile nel trattare i casi di "missing values" sia in fase di training che in fase di testing.

6. Essi possono gestire bene i casi di attributi che interagiscono fra loro o che sono fra essi correlati. Ad esempio attributi che hanno piu' potere predittivo assieme rispetto a prenderli singolarmente.

7. Se esistono degli attributi irrilevanti con alta frequenza, essi possono distorcere il risultato della classificazione, in quanto cambiano il modo in cui calcola la similarita'.