Regole associative

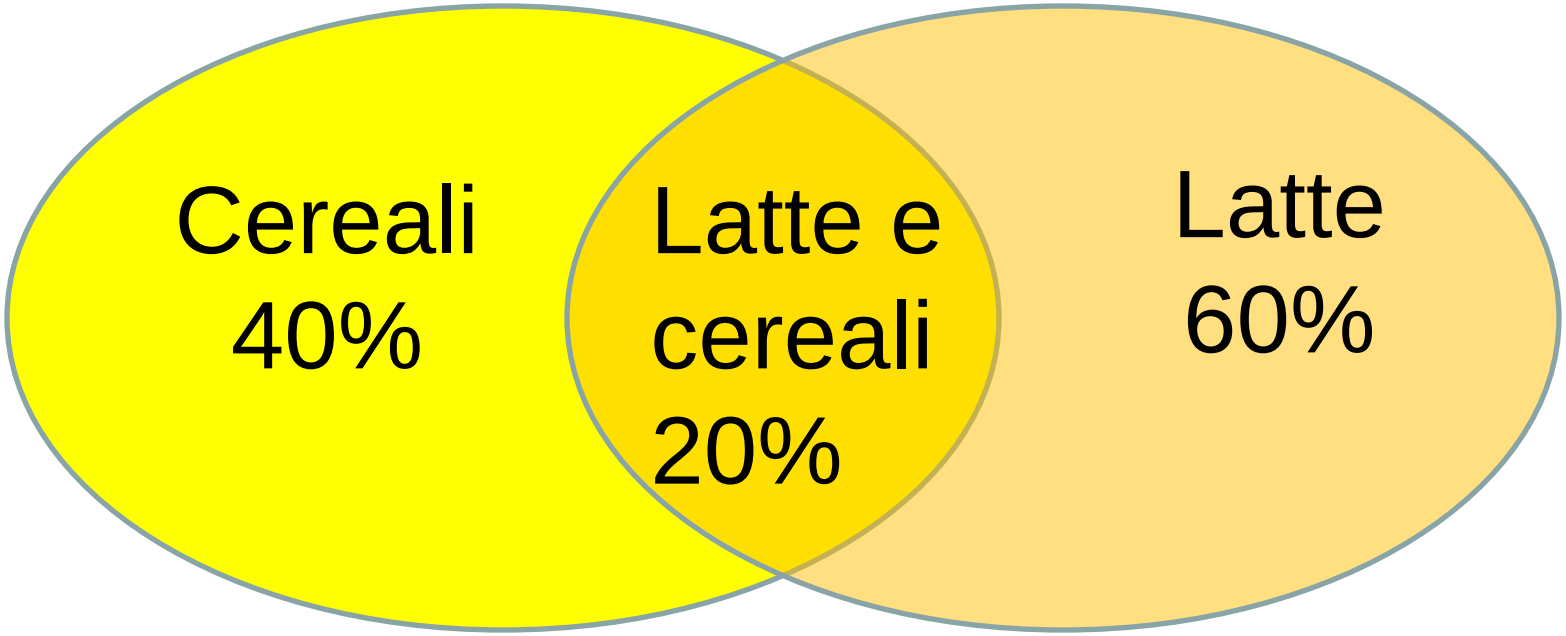
Introduzione

- Le regole associative si collocano tra i metodi di apprendimento non supervisionato e sono volte all'identificazione di regolarità e ricorrenze tra i dati.
- Sono semplici ed intuitive e trovano frequente applicazione nelle analisi di transazioni commerciali (*market basket analysis*).
- Vedremo alcuni esempi, i principali indicatori di valutazione e alcuni metodi.

Market basket analysis

- Ogni volta che un cliente effettua degli acquisti in un punto vendita, l'operazione viene registrata.
- Per ciascuna transazione vengono memorizzate la lista degli articoli acquistati, il prezzo, la data,…
- L'analisi di questi dati può produrre informazioni importanti per determinare regole ricorrenti che pongono in relazione l'acquisto di uno o più prodotti con altri.
  - ➢ queste informazioni si possono progettare azioni promozionali o per posizionare gli articoli sugli scaffali.

Market basket analysis



- P(Cereali) = 0.4, P(Latte) = 0.6
- P(Cereali $\Rightarrow$ Latte) = (0.4 * 0.2) = 0.08
- P(Latte $\Rightarrow$ Cereali) = (0.6 * 0.2) = 0.12

Struttura e valutazione

- Date due proposizioni Y e Z, una *regola* è una implicazione del tipo $Y => Z$.
- Y => Z, significa che se Y è vera, allora anche Z è vera.
- Una regola si dice probabilistica se la validità di Z è associata ad una probabilità p: se Y è vera, allora anche Z è vera, con probabilità p.
- Talvolta le regole prodotte si limitano a rispecchiare circostanze evidenti, mentre talvolta capovolgono il legame causale di una relazione.
  - Es. Gli acquirenti di una polizza assicurativa acquistano anche un'autovettura con probabilità 0.98.

Dataset

ID T_i	Transazione
001	{ a (pane), c (cereali) }
002	{ a (pane), b (latte), d (caffè) }
003	{ b (latte), c (cereali) }
004	{ b (latte), d (caffè) }
005	{ a (pane), b (latte), c (cereali) }
006	{ b (latte), c (cereali) }
007	{ a (pane), c (cereali) }
008	{ a (pane), b (latte), e (the) }
009	{ a (pane), b (latte), c (cereali), e (the) }
010	{ a (pane), e (the) }

Matrice degli item

ID T_i	a	b	c	d	e
001	1	0	1	0	0
002	1	1	0	1	0
003	0	1	0	1	0
004	0	1	0	1	0
005	1	1	1	0	0
006	0	1	1	0	0
007	1	0	1	0	0
008	1	1	0	0	1
009	1	1	1	0	1
010	1	0	0	0	1

Regole associative

- La *frequenza empirica* $f(L)$ di un insieme L di oggetti è il numero di transazioni che contengono L:

$$f(L) = \text{card}\{T_i : L \supset T_i,\ i=1,\ldots,m\}.$$

- La *confidenza* di una regola è:

$$p = \text{conf}\{L \Rightarrow H\} = \frac{f(L \cup H)}{f(L)}$$

  - L viene detto corpo della regola, H viene detto testa.

- Il *supporto* di una regola è:

$$s = supp\{L \Rightarrow H\} = \frac{f(L \cup H)}{m}$$

Esempio

ID T_i	a	b	c	d	e
001	1	0	1	0	0
002	1	1	0	1	0
003	0	1	0	1	0
004	0	1	0	1	0
005	1	1	1	0	0
006	0	1	1	0	0
007	1	0	1	0	0
008	1	1	0	0	1
009	1	1	1	0	1
010	1	0	0	0	1

- L={a, c} H={b}
- *Frequenze empiriche*:

$f(\mathrm{L})=4, \quad f(\mathrm{H})=7, \quad f(\mathrm{L} \cup \mathrm{H})=2$

- Confidenza di $L \Rightarrow H$:

$$p = \mathrm{conf}\{L \Rightarrow H\} = \frac{f(L \cup H)}{f(L)} = \frac{2}{4} = 0.5$$

- Supporto di $L \Rightarrow H$:

$$s = \mathrm{supp}\{L \Rightarrow H\} = \frac{f(L \cup H)}{m} = \frac{2}{10} = 0.2$$

Generazione degli itemset frequenti

- Il supporto di una regola dipende solo dall'insieme $L \cup H$ (*itemset*) e non dall'effettiva distribuzione degli oggetti tra corpo e testa.
- Ad esempio, le sei regole che si possono generare per gli oggetti {a, b, c} hanno tutte supporto s = 2/10 = 0.2:

$$\{a,b\} \Rightarrow \{c\}, \quad \{a,c\} \Rightarrow \{b\},$$

$$\{b,c\} \Rightarrow \{a\}, \quad \{a\} \Rightarrow \{b,c\},$$

$$\{b\} \Rightarrow \{a,c\}, \quad \{c\} \Rightarrow \{a,b\}.$$

- Se il valore di soglia per il supporto è s_{min}=0.25, le sei regole vanno eliminate, sulla base dell'analisi dell'insieme unione.

Generazione delle regole

- Dopo aver generato tutti gli itemset frequenti, vanno identificate le regole forti, vale a dire quelle regole che superano una soglia di confidenza fissata.
- Gli oggetti in ciascun itemset vanno separati secondo tutte le combinazioni di corpo e testa per verificare se la confidenza della regola supera una soglia p_{min}.
- Se l'itemset è frequente, si possono quindi ricavare le regole, ma solo alcune saranno forti.

Lift di una regola

- Non sempre le regole forti sono significative e potenzialmente interessanti.
- Esempio. Su 1000 transazioni, di cui 600 includono fotocamere, 750 includono stampanti e 400 includono entrambi.
- Per s_{min} = 0.3 e p_{min} = 0.6, tra le regole forti viene selezionata anche $\{fotocamera\} \Rightarrow \{stampante\}$ che ha s=.40 e p=0.66.
- La probabilità di acquistare una stampante è 0.75, maggiore della regola selezionata. Inoltre, se compro una fotocamera, probabilmente non comprerò una stampante…

Lift di una regola

- Per valutare la significatività di una regola si usa l'indice di lift:

$$l = \text{lift}\{L \Rightarrow H\} = \frac{\text{conf}(L \Rightarrow H)}{f(H)} = \frac{f(L \cup H)}{f(L)f(H)}$$

- Valori di l > 1 indicano che la regola è più efficace nel predire la probabilità che la testa sia contenuta in una generica transazione, di quanto lo sia la sua frequenza.
- Valori di l < 1 indicano che la regola che nega la testa, è più efficace della regola iniziale.

Algoritmo Apriori

- Un dataset formato a partire da n oggetti può contenere fino a 2^n-1 itemset frequenti.
- Nelle applicazioni pratiche n è almeno nell'ordine di alcune decine e un metodo di generazione esaustivo degli itemset è impraticabile.
- Il presupposto teorico su cui si basa l'algoritmo Apriori consiste nella proprietà:

Teorema. *Se un insieme di oggetti (itemset) è frequente, allora anche tutti i suoi sottoinsiemi sono frequenti.*

  - Se un itemset non è frequente, allora neanche gli insiemi che lo contengono sono frequenti.

Generazione degli itemset frequenti

- L'algoritmo Apriori affronta la fase di generazione degli itemset frequenti per approssimazioni successive, a partire dagli itemset con un solo elemento.
- Il numero delle iterazioni è k_{max}+1, dove k_{max} indica la cardinalità massima di un itemset frequente.

Algoritmo Apriori

1. Si calcola la frequenza relativa di ciascun oggetto e si scartano quelli con frequenza minore della soglia di supporto s_{min}. Si pone $k = 2$.
2. Si generano gli itemset di ordine k a partire da quelli di ordine k - 1.
3. Si calcola il supporto di ciascun itemset e si scartano quelli con supporto inferiore ad s_{min}.
4. L'algoritmo si arresta se non è stato generato alcun k-itemset, altrimenti si pone $k = k + 1$ e si procede al passo 2.

Esempio algoritmo Apriori (1/3)

1. s_{min}=.2

Itemset	Frequenza relativa	Stato
a	7/10 = 0.7	frequente
b	7/10 = 0.7	frequente
c	5/10 = 0.5	frequente
d	3/10 = 0.3	frequente
e	3/10 = 0.3	frequente

k = 2.

Esempio algoritmo Apriori (2/3)

Itemset	Frequenza relativa	Stato
{a, b}	4/10 = 0.4	frequente
{a, c}	4/10 = 0.4	frequente
{a, d}	1/10 = 0.1	non frequente
{a, e}	3/10 = 0.3	frequente
{b, c}	3/10 = 0.3	frequente
{b, d}	3/10 = 0.3	frequente
{b, e}	2/10 = 0.2	frequente
{c, d}	0	non frequente
{c, e}	1/10 = 0.1	non frequente
{d,e}	0	non frequente

Esempio algoritmo Apriori (3/3)

Itemset	Frequenza relativa	Stato
{a, b, c}	2/10 = 0.2	frequente
{a, b, e}	2/10 = 0.2	frequente

- Il numero di operazioni richieste dall'algoritmo cresce esponenzialmente con il numero degli oggetti n. Per 100 oggetti:

$$\sum_{h=1}^{100}\binom{100}{h} = 2^{100} - 1 \approx 10^{30}$$

Generazione delle regole forti

1. Si effettua una scansione della lista degli itemset frequenti generati nella prima fase. Se la lista è vuota si arresta, altrimenti sia B il successivo itemset, che viene tolto dalla lista.
2. Si suddivide l'itemset B in L e $H = B - L$ in tutte le combinazioni possibili.
3. Per ciascuna regola candidata $L \Rightarrow H$ si calcola la confidenza:
$$p = \text{conf}\{L \Rightarrow H\} = \frac{f(B)}{f(L)}$$
4. Se $p \geq p_{min}$ la regola viene inserita nella lista delle regole forti, altrimenti viene eliminata.

Esempio di generazione di regole forti

Itemset	Regola	Frequenza relativa	Stato
{a, b}	$\{a \rangle b\}$	p = 4/7 = 0.57	forte
{a, b}	$\{b \rangle a\}$	p = 4/7 = 0.57	forte
{a, c}	$\{a \rangle c\}$	p = 4/7 = 0.57	forte
{a, c}	$\{c \rangle a\}$	p = 4/5 = 0.80	forte
{a, e}	$\{a \Rightarrow e\}$	p = 3/7 = 0.43	non forte
{a, e}	$\{e \Rightarrow a\}$	p = 3/3 = 1.00	forte
{b, c}	$\{b \Rightarrow c\}$	p = 3/7 = 0.43	non forte
{b, c}	$\{c \Rightarrow b\}$	p = 3/5 = 0.60	forte
{b, d}	$\{b \Rightarrow d\}$	p = 3/7 = 0.43	non forte
{b, d}	$\{d \Rightarrow b\}$	p = 3/3 = 1.00	forte
{b, e}	$\{b \Rightarrow e\}$	p = 2/7 = 0.29	non forte
{b, e}	$\{e \Rightarrow b\}$	p = 2/3 = 0.67	forte
{a, b, c}	$\{a, b \Rightarrow c\}$	p = 2/4 = 0.50	non forte
{a, b, c}	$\{c \Rightarrow a, b\}$	p = 2/5 = 0.40	non forte

Esempio di generazione di regole forti

Itemset	Regola	Frequenza relativa	Stato
{a, b, c}	$\{a, c \rangle b\}$	p = 2/4 = 0.50	non forte
{a, b, c}	$\{b \rangle a, c\}$	p = 2/7 = 0.29	non forte
{a, b, c}	$\{b, c \rangle a\}$	p = 2/3 = 0.67	forte
{a, b, c}	$\{a \rangle b, c\}$	p = 2/7 = 0.29	non forte
{a, b, e}	$\{a, b \rangle e\}$	p = 2/4 = 0.50	non forte
{a, b, e}	$\{e \rangle a, b\}$	p = 2/3 = 0.67	forte
{a, b, e}	$\{a, e \rangle b\}$	p = 2/3 = 0.67	forte
{a, b, e}	$\{b \rangle a, e\}$	p = 2/7 = 0.29	non forte
{a, b, e}	$\{b, e \rangle a\}$	p = 2/2 = 1.0	forte
{a, b, e}	$\{a \rangle b, e\}$	p = 2/7 = 0.29	non forte

Miglioramenti

- Strutture di rappresentazione dei dati quali dizionari ed alberi binari.
- Partizioni delle transazioni in sottoinsiemi disgiunti mediante teoria dei grafi o clustering.
- Estrazioni di campioni significativi casuali.

Esercizi

- Determinare le regole forti per il dataset supermarket.arff degli esempi di Weka e discutere i risultati.
- Determinare il lift di ciascuna regola.