Analisi dei dati ed estrazione della conoscenza

Esercizi

Gianluca Amato

14 aprile 2005

1. Definire le seguenti funzionalità di un sistema di data mining: *caratterizzazione*, *discriminazione*, *associazione*.

2. Supponi che un data warehouse consista di tre dimensioni *ora*, *dottore* e *paziente*, e di due misure *conteggio* e *tariffa* dove *tariffa* è l'ammontare che il dottore richiede al paziente per una visita.

   - enumera tre tipi di schemi che sono comuni per modellare un data warehouse;
   - disegna uno schema per il presente data warehouse, scegliendo uno dei tre tipi al punto precedente;
   - partendo dal cuboide base giorno, dottore, paziente, quali operazioni OLAP devono essere eseguite per ottenere l'elenco dei proventi ottenuti da ogni dottore nell'anno 2000?
   - per ottenere lo stesso risultato, scrivi una query SQL assumendo che il dato memorizzato in un database relazionale secondo il seguente schema:

     proventi (giorno, mese, anno, dottore, ospedale, paziente, contatore, tariffa) .

3. Con riferimento al calcolo delle misure in un data cube:

   - enumera tre tipi di misure, sulla base del tipo di funzioni aggregate usate per il calcolo del data cube;
   - per un data cube con tre dimensioni *tempo*, *luogo* e *prodotto*, a quale categoria appartiene la misura *varianza*?
     Suggerimento: la formula per il calcolo della varianza è $\frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \bar{x_i}^2$ dove $\bar{x_i}$ è la media degli x_i.

4. Supponi che i dati per una analisi includono l'attributo *età*. I valori per questo attributo presenti nelle istanze, ordinati in ordine crescente, sono: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 36, 40, 45, 46, 52, 70.

- usare la tecnica di *equidepth binning* con profondità (depth) uguale a 3 per *ammorbidire* i dati in input.
- che altri metodi si possono usare in alternativa all'equidepth binning?

5. Usando gli stessi dati dell'esercizio precedente, mostra degli esempi di campioni estratti usando una di queste due tecniche: campionamento stratificato senza rimpiazzo, campionamento semplice con rimpiazzo. Usare campioni di dimensione 5 e gli strati $età \leq 20$, $21 \leq età \leq 40$ e $età \geq 41$.

6. Supponete che la seguente tabella venga derivata da una procedura di *attribute-oriented induction*.

classe	luogonascita	conteggio
	Canada	180
Programmatore	altro	120
	Canada	20
DBA	altro	80

   - Trasforma la tabella in una crosstab, mostrando i t-weight e i d-weight corrispondenti.
   - Scrivere una *regola descrittiva quantitativa* per la classe Programmatore.

7. Si considerino gli stessi dati per l'attributo età dell'esercizio n. 4.

   - qual è la *media* dei dati? quale la mediana?
   - qual è la *moda* (o quali sono le *mode*) ?
   - qual è il *midrange* ?
   - quali sono il primo quartile e il terzo quartile?
   - dai il cosiddetto *five number summary* per i dati
   - disegna un *boxplot* per i dati

8. Un database ha 4 transazioni. Siano $min\_sup = 60\%$ e $min\_conf = 80\%$.

TID	**data**	**items**
T100	15/10/1999	{K,A,D,B}
T200	15/10/1999	{D,A,C,E,B}
T300	19/10/1999	{C,A,B,E}
T400	22/10/1999	{B,A,D}

   - Trova tutti gli itemset frequenti usando l'algoritmo Apriori. Mostrare chiaramente gli effetti dei passi Join e Prune dell'algoritmo.
   - Elencare tutte le regole associative forti, e i rispettivi valori di supporto e confidenza, che rispettino la seguente meta-regola:

     $\forall x \in transactions, buys(x, item_1) \wedge buys(x, item_2) \Rightarrow buys(x, item_3)[s, c]$

     dove $item_i$ è una variabile rappresentante oggetti (item).

9. La seguente tabella di contingenza riassume le transazioni di un supermercato, dove *hotdogs* si riferisce alle transazioni contenenti hot dog e $\overline{hotdogs}$ si riferisce alle transazioni che non contengono hot dog (analogamente per *hamburgers* e $\overline{hamburgers}$.

	$hotdogs$	$\overline{hotdogs}$	$\sum_{righe}$
$hamburgers$	2000	500	2500
$\overline{hamburgers}$	1000	1500	2500
$\sum_{colonne}$	3000	2000	5000

   - Considera la regola associativa "$hotdogs \Rightarrow hamburgers$". Data le soglie di supporto minimo del 25% e confidenza minima del 50%, si tratta di una regola forte?
   - In base ai dati forniti, *hotdogs* è indipendente da *hamburgers*? Se no, che tipo di *correlazione* c'è tra di essi?

10. (difficile) Il prezzo degli oggetto in un negozio è sempre positivo o nullo. Il direttore del negozio è interessato solo a determinare le regole del tipo "*un oggetto gratuito spinge all'acquisto di almeno 200 euro di merci in totale*", in cui l'obiettivo è determinare quali sono gli oggetti in questione. Spiegare come determinare velocemente questi oggetti.