Università di Pisa – A.A. 2003-2004

Analisi dei dati ed estrazione di conoscenza – Corso di Laurea Specialistica in *Informatica per l'Economia e per l'Azienda*

Tecniche di Data Mining – Corsi di Laurea Specialistica in *Informatica e Tecnologie Informatiche*

Verifica del 31 Marzo 2004

Esercizio 1 (5 punti) Sia dato il seguente database:

ID Transazione	Items
1	{f,a,d,b}
2	{d,a,c,e,b}
3	{c,a,b,e}
4	{b,a,d}

Fissati il supporto minimo s = 60% e la confidenza minima g = 80%

a) Indicare quali tra questi itemset sono frequenti.
   1) {a}
   2) {c}
   3) {b,c}
   4) {b,d}
   5) {a,b,d}
   6) {a,b,e}

b) Indicare quali tra queste regole sono valide
   1) {a}=>{b}
   2) {a}=>{d}
   3) {d}=>{a}
   4) {d}=>{a,b}
   5) {a,b}=>{d}

   {a,d}=>{b}

Esercizio 2 (5 punti)

Supponiamo di avere un database con le transazioni corrispondenti ad un anno di vendite. Assumiamo anche che le transazioni siano raggruppate in trimestri e all'interno di ogni trimestre siano raggruppate in singoli mesi.
Indicare quali delle seguenti affermazioni sono sempre verificate

.

1) Se l'itemset {a,b,c} è frequente separatamente nei mesi di gennaio, febbraio e marzo, allora è frequente anche nell'intero trimestre gennaio-marzo.
2) Se l'itemset {a,b} è frequente nell'intero anno allora è frequente anche nel singolo mese di aprile.
3) Assumendo che "Frutta" sia una generalizzazione di "Mele" e che "Latticini" sia una generalizzazione di "Formaggio", allora:

   Supp ({Frutta}=>{Latticini}) ³ Supp ({Mele}=>{Formaggio})
4) Se nell'intero anno la regola {a,b}=>{c,d} è valida, allora nello stesso periodo di tempo è valida anche la regola {a,b,c}=>{d}

Esercizio 3 (5 punti) Si considerino i seguenti vincoli su itemset indicano per ciascuno di questi se si tratta di un vincolo monotono, anti-monotono o nessuna delle due possibilità. Si ricorda che un vincolo *V* è *monotono* (risp., *anti-monotono*) se dal fatto che un itemset *I* soddisfa *V* possiamo concludere che qualunque *sovrainsieme* (risp., *sottoinsieme*) di I continua a soddisfare *V*.

1) L'itemset contiene almeno un cellulare Nokia
2) La somma dei prezzi degli item è minore di 45 euro
3) La somma dei prezzi degli item è almeno di 120 euro
4) La media dei prezzi degli item è compresa fra 10 e 50 euro

Esercizio 4 (5 punti) Si consideri il seguente training set relativo alle promozioni per una carta di credito:

Range di Reddito	Promozione Assicurazione vita	Assicurazione Carta di credito	Sesso	Età
40-50K	No	No	M	45
30-40K	Si	No	F	40
40-50K	No	No	M	42
30-40K	Si	Si	M	43
50-60K	Si	No	F	38
20-30K	No	No	F	55
30-40K	Si	Si	M	35
20-30K	No	No	M	27
30-40K	No	No	M	43
30-40K	Si	No	F	41
40-50K	Si	No	F	43
20-30K	Si	No	M	29
50-60K	Si	No	F	39
40-50K	No	No	M	55
20-30K	Si	Si	F	19

Si costruisca un albero di decisione per la variabile target "Promozione Assicurazione vita" selezionando ad ogni nodo la variabile di Split in base a considerazioni intuitive di miglior separazione dei dati. Si calcoli infine l'accuratezza (rispetto al training set) dell'albero costruito. Si noti che la variabile "Età" può essere utilizzata in uno split binario scegliendo opportunatamente un valore di soglia.

Esercizio 5 (4 punti)

Dato il cluster calcolato utilizzando l'algoritmo K-means standard formato dai seguenti vettori:
A1={4,5,5,0}, A2={0,2,3,5}, A3={2,0,5,4}, A4={5,2,0,1}, A5={4,1,2,5}.
Qual è il centroide corretto del cluster?

a) C = {4,2,3,4}
b) C = {3,2,3,3}
c) C = {4,2,5,5}
d) C = {2,0,5,4}

Esercizio 6 (5 punti)

Supponiamo di voler eseguire un clustering su un dataset contenente i dati di studenti universitari.

1) Codice identificativo dello studente
2) Età
3) Nome e Cognome
4) Corso di laurea
5) Numero di esami superati
6) Data di immatricolazione
7) Stato civile
8) Indirizzo
9) Comune di nascita
10) Reddito familiare
11) Numero dei componenti del nucleo familiare
12) Studi precedenti (es. Liceo scientifico, ITI, ITC ecc.)

Per ogni attributo indicare la potenziale utilità nel calcolo del clustering scegliendo tra le seguenti ipotesi:

b) può essere utile
c) può essere utile se discretizzato
d) può essere utile se soddisfa determinate condizioni (indicare quali)
e) probabilmente inutile

Esercizio 7 (5 punti) Si considerino i seguenti sei punti sul piano cartesiano:

Osservazione	**X**	**Y**
P1	1.0	1.5
P2	1.0	4.5
P3	2.0	1.5
P4	2.0	3.5
P5	3.0	2.5
P6	5.0	6.0

Si simuli in modo grafico l'esecuaione dell'algoritmo l'algoritmo K-means con K=2 a partire dai centroidi iniziali P1 e P3, evidenziando i centrodi e i cluster ad ognuna delle iterazioni.