Università di Pisa – A.A. 2005-2006

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 6 aprile 2006

Esercizio 1 (4 punti**)**
Consideriamo il seguente insieme di 3-itemsets frequenti:

{1,2,3}, {1,2,4}, {1,2,5}, {1,3,4}, {1,3,5}, {2,3,4}, {2,3,5}, {3,4,5}

(assumiamo che ci siano solo 5 items nel dataset.)

A) Elencare tutti i 4-itemsets candidati ottenuti attraverso la procedura di *candidate generation* contenuta nell'algoritmo Apriori.
B) Elencare tutti i 4-itemsets che sopravvivono al passo di *candidate pruning* dell'algoritmo Apriori.

Esercizio 2 (4 punti)
Dato il seguente DB:

t_1	$ACDE$
t_2	BCE
t_3	$ABCE$
t_4	$BCDE$
t_5	$ABCE$
t_6	$ABCD$

a) Trovare i Frequent Itemsets con un valore di soglia per il supporto *minsup* = 2
b) Mostare tutte le regole associative che possono essere generate a partire dall'itemset {BCE}.

Esercizio 3 (4 punti)
Mostrare che il supporto di un itemset H che contiene sia un item h che un suo antenato $\hat{h}$ ha lo stesso supporto dell'itemset H – $\hat{h}$.

Esercizio 4 (4 punti**)**
Si consideri la seguente tabella di contingenza dei due item: caffè e the, dove *caffè* e *the* si riferiscono a transazioni che li contengono e *caffè* e *the* si riferiscono a transazioni che NON li contengono:

	The	The	Tot.
Caffè	2000	500	2500
Caffè	1000	1500	2500
Tot	3000	2000	5000

La regola **Caffè →The** con MinSupp= 25% e MinCon=50% è forte?
Gli acquisti di Caffè e Thè sono indipendenti? Quale tipo di correlazione esiste tra loro?

Esercizio 5 (6 punti**)**
Si consideri il seguente data set:

Table 7.4. Data set for Exercise 2.

TID	Temperature	Pressure	Alarm 1	Alarm 2	Alarm 3
1	95	1105	0	0	1
2	85	1040	1	1	0
3	103	1090	1	1	1
4	97	1084	1	0	0
5	80	1038	0	1	1
6	100	1080	1	1	0
7	83	1025	1	0	1
8	86	1030	1	0	0
9	101	1100	1	1	1

Supponiamo di applicare le seguenti strategie di discretizzazione agli attributi continui del data set:
D1: partiziona il range di ogni attributo continuo in 3 intervalli della stessa ampiezza (*equal-size bin*)
D2: partiziona il range di ogni attributo continuo in 3 intervalli con lo stesso numero di transazioni (*natural distribution bin*)

Per ogni strategia:

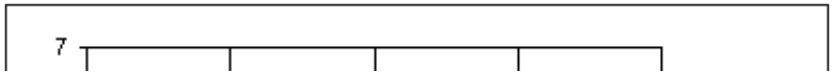
a) costruire una versione binarizzata del data set
b) derivare gli itemesets frequenti con MinSupp>= 30%

Esercizio 6 (6 punti)
Si consideri il seguente training set:

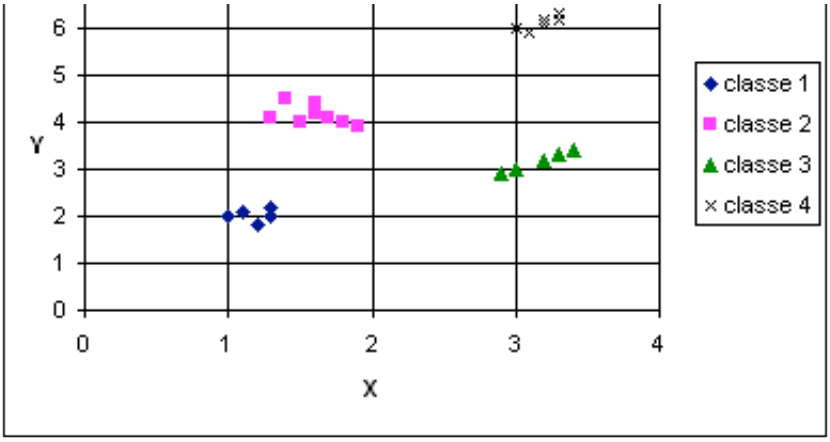
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Weak	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

A. si costruisca un albero di decisione per la variabile target "PlayTennis" selezionando ad ogni nodo la variabile di split in base al criterio di *misclassification rate*;
B. si calcoli l'errore di classificazione dell'albero costruito utilizzando sia il metodo ottimistico che quello pessimistico;
C. si valuti in modo intuitivo se la scelta dell'attributo di split alla radice dell'albero sia influenzata o meno dal criterio di split (misclassification rate, indice di Gini, entropia).

Esercizio 7 (4 punti)



Dato il seguente dataset di punti sul piano, costruire un albero di decisione



per l'attributo **classe**, scegliendo opportuni valori per le variabili X ed Y, in modo che l'errore di classificazione sul training set sia nullo.

X	Y	classe
1	2	1
1.2	1.8	1
1.1	2.1	1
1.3	2.2	1
1.3	2	1
1.5	4	2
1.3	4.1	2
1.4	4.5	2
1.6	4.4	2
1.6	4.2	2
1.7	4.1	2
1.8	4	2
1.9	3.9	2
3	3	3
3.1	3.2	3
3	2.9	3
3.2	3.3	3
3.3	3.4	3
3.1	3.3	3
3	6	4
3.1	5.9	4
3.2	6.1	4
3.2	6.2	4
3.3	6.3	4
3.3	6.2	4

Esercizio 8 (4 punti)
Dato il dataset di punti sul piano dell'esercizio 7 (senza considerare l'attributo **classe**) si discuta il risultato del calcolo dell'algoritmo *K*-means per *K*=2, 3 e 4, valutando in ciascun caso l'impatto della scelta iniziale dei centroidi.