Università di Pisa – A.A. 2004-2005

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 7 Aprile 2005

Esercizio 1 (5 punti) Si consideri il seguente database di transazioni:

TID	Items	Count
1	{a, b, c}	10
2	{a, b, d}	15
3	{a, c}	25
4	{b, d}	15
5	{a, b, d, e}	15
6	{b, e}	20

Fissati il supporto minimo s = 35% e la confidenza minima g = 60%

a) Elencare gli itemset frequenti.
b) Elencare le regole valide.
c) Calcolare l'interesse (o correlazione o lift) delle regole valide e discuterne il risultato utilizzando la matrice di contingenza per tali regole.

Soluzione

Itemset Frequenti:

({a},65),
({b,},75),
({c},35),
({d},45),
({e},35)
({a,b},40)
({a,c},35)
({b,e},35)
({b,d},45)

Regole Valide:

a => b, c=0.61, I= 0.82 I= P(A U B)/P(A)*P(B)
c => a, c=1, I= 1.54
b => d, c=0.6, I= 1.3
d => b, c=1, I = 1.3
e => b, c=1.3, I=1.3

Matrice di contingenza per a => b se confrontato a **not a => b C=1**

	a	Not a	
b	40	35	75
Not b	25		25
	65	35	100

Esercizio 2 (5 punti)
Si consideri seguente dataset 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	40
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

Fissati il supporto minimo s = 25% e la confidenza minima g = 80% determinare le regole associative valide della forma:

Età(X,A) Ù RangeReddito(X,B) => PromozioneAssicurazioneVita(X,C)

Soluzione

Effettuando una discretizzazione sull'attributo età ad es. 19-29, 30-39, 40-49, 50-55 si
Età(40-49,A) Ù RangeReddito(30-40,B) => PromozioneAssicurazioneVita(SI,C) con supporto 4/15 (0.26) e confidenza 4/5 (0.8)

Esercizio 3 (3 punti)
Nell'algoritmo Apriori gli itemset frequenti L_k sono generati sulla base degli itemset candidati C_k. Gli itemset L_{k-1}, L_{k-2},.... , L_1 sono necessari in questo passo dell'algoritmo? Si giustifichi la risposta.

Soluzione

Il testo dell'esercizio si è rivelato ambiguo a molti. L'intenzione era quella di chiedere se gli itemset L_{k-1}, L_{k-2},.... , L_1 sono necessari allo scopo di generare L_k a partire dagli itemset candidati C_k: la risposta a questa domanda è SI, ma solo L_{k-1} in quanto per ogni candidato, prima di andare ad effettuare il conteggio sul DB, conviene controllare che tutti i suoi sottoinsiemi $_{k-1}$ sino frequenti e cioè in L_{k-1}.

Esercizio 4 (5 punti)

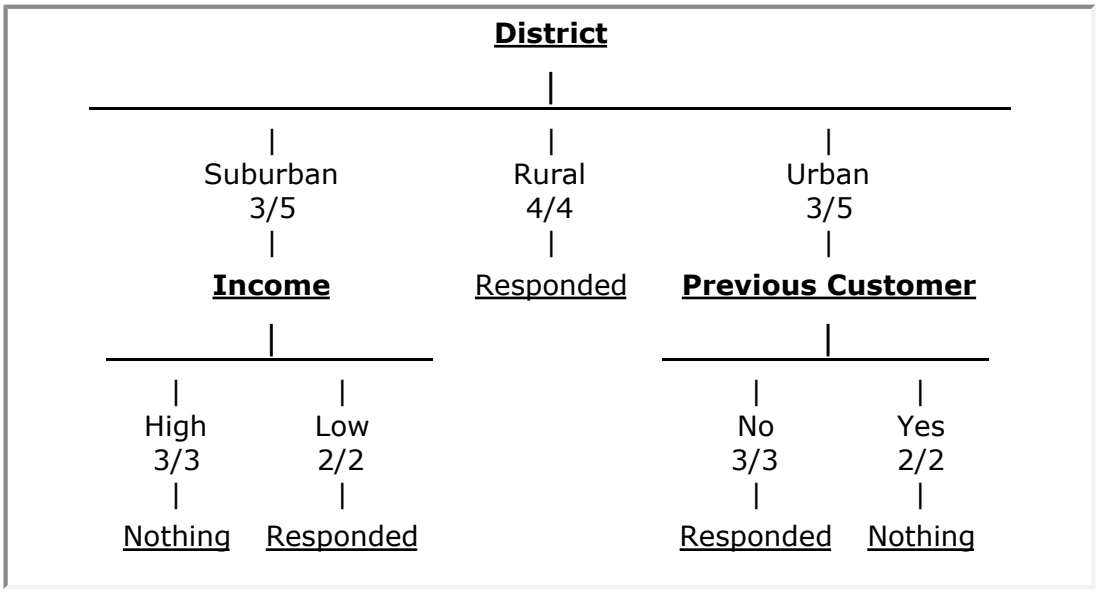
Si consideri il seguente training set relativo ai risultati di una campagna di marketing: una azienda ha inviato una qualche promozione alle famiglie, registrando alcune informazioni per ciascuna famiglia (distretto di residenza, tipo di abitazione, reddito, se cliente in precedenza o meno) insieme con il risultato della promozione, ovvero se la famiglia ha risposto o meno.

District	House Type	Income	Previous Customer?	Outcome

Suburban	Detached	High	No	Nothing
Suburban	Detached	High	Yes	Nothing
Rural	Detached	High	No	Responded
Urban	Semi-detached	High	No	Responded
Urban	Semi-detached	Low	No	Responded
Urban	Semi-detached	Low	Yes	Nothing
Rural	Semi-detached	Low	Yes	Responded
Suburban	Terrace	High	No	Nothing
Suburban	Semi-detached	Low	No	Responded
Urban	Terrace	Low	No	Responded
Suburban	Terrace	Low	Yes	Responded
Rural	Terrace	High	Yes	Responded
Rural	Detached	Low	No	Responded
Urban	Terrace	High	Yes	Nothing

Si costruisca un albero di decisione per la variabile target "Outcome" selezionando ad ogni nodo la variabile di split in base a considerazioni intuitive di miglior separazione dei dati. Si richiede che le foglie dell'albero di decisione siano omogenee, ovvero relative ad osservazioni tutte della stessa classe (*Responded* oppure *Nothing*): in altre parole, è richiesta la costruzione di un albero di decisione con una accuratezza del 100%.

Soluzione



Esercizio 5 (3 punti)

L'algoritmo di clustering *K-means* è fortemente dipendente dalla scelta del parametro *K*, che indica appunto il

numero dei cluster in cui partizionare il dataset. Si proponga un metodo iterativo per scegliere un valore di K in grado di migliorare rispetto alla scelta casuale, discutendone pregi e difetti.

Soluzione

Un approccio possibile e quello di iterare il calcolo del *K-means* con valori crescenti di K, a partire da $K = 2$ o $K = 3$, fino a raggiungere un valore di K in cui la qualità del clustering prodotto non migliora ulteriormente, ovvero la qualità del clustering con $K+1$ non migliora più di una certa soglia minima prefissata. La qualità di un clustering può essere misurata con una funzione che tiene conto del grado di similarità intra-cluter (es. la media delle distanze dei punti di ciascun cluster con il centroide del cluster stesso) e di dissimilarità inter-cluster (es. la media delle distanze di ciascun punto con i centroidi degli altri cluster).

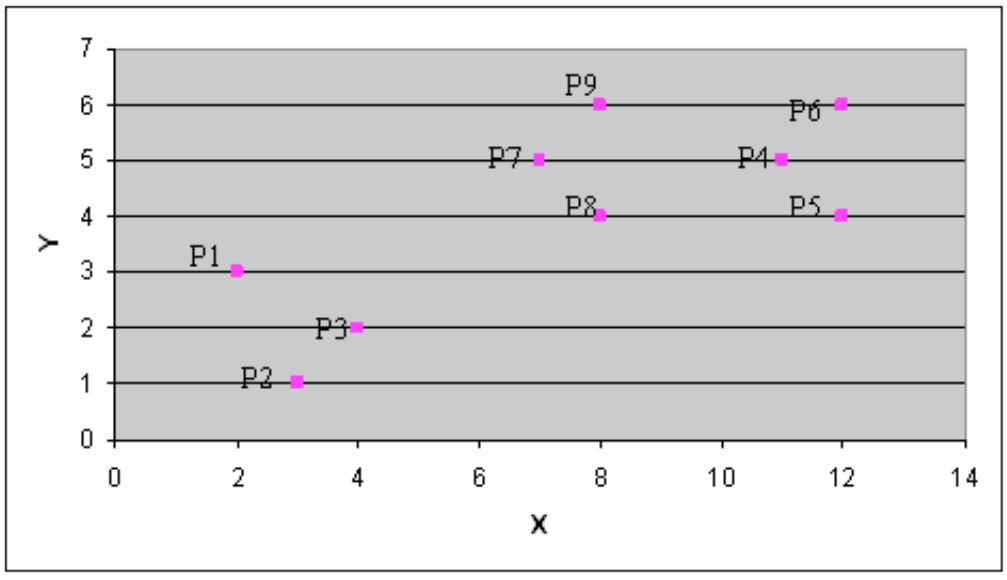
Esercizio 6 (5 punti) Si considerino i seguenti nove punti sul piano cartesiano:

Pt.	X	Y
P1	2	3
P2	3	1
P3	4	2
P4	11	5
P5	12	4
P6	12	6
P7	7	5
P8	8	4
P9	8	6

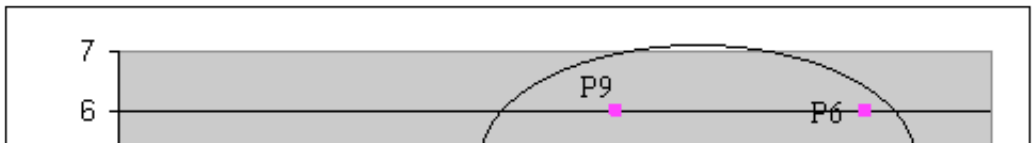
Si determinino due diverse condizioni iniziali per l'esecuzione dell'algoritmo *K-means* con K=2 in modo tale che i cluster ottenuti siano diversi nei due casi. Si usi la simulazione grafica dell'algoritmo, evidenziando, in ciascuna delle due esecuzioni, i centroidi e i cluster di ogni iterazione dell'algoritmo.

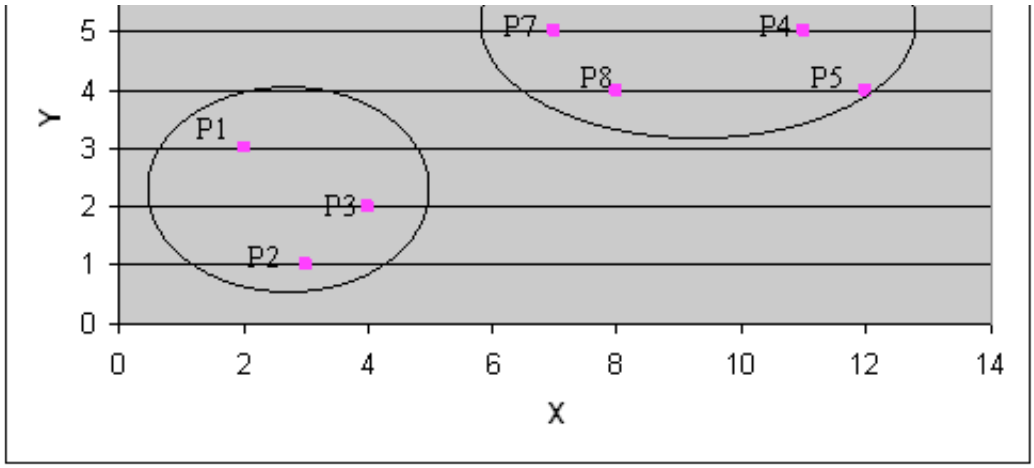
Soluzione

I punti sul piano cartesiano:



Caso 1) Centroidi iniziali K1 = P2 e K2 = P8. Si ottengono i seguenti clusters:





Caso 2) Centroidi iniziali K1 = P7 e K2 = P4. Si ottengono i seguenti clusters: