Università di Pisa	A.A. 2008- 2009

Data Mining - *Corso di Laurea Specialistica in Informatica per l economia e l Azienda*
Tecniche Data Mining - *Corsi di Laurea Specialistica in Informatica e Tecnologie Informatiche*

PARTE A = Esercizi 1-4 **PARTE B = Esercizi 5-6**

Appello del 21 luglio 2009

Esercizio 1 - Sequential Patterns (6 punti)

Si consideri il seguente dataset di sequenze:

< {A,B} {A,C} {D,E} {B, C} {E} {H} {A} >
< {A,C} {E, C} {B} {A,H} {B,C} >
< {B} {B, D,E} {E} {H} {A,B} >
< {B} {D,E} {E, C} {E,H} {H} {A} >

Si indichi il supporto delle seguenti sotto-sequenze senza vincoli temporali (), e quali rispettano anche il vincolo temporale *max-span = 3*:

	supporto	*supporto con max-span=3*
w_1 = <{A,C}{B,C}>		
w_2 = <{C,D} {H} >		
w_3 = <{B}{A}>		

Esercizio 2 Itemset Frequenti (**10** punti)

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

Si applichi la strategia di discretizzazione (*equal-size bin*) agli attributi continui del data set in figura, partizionando il range di ogni attributo continuo in 3 intervalli della stessa ampiezza.
Inoltre:
a) si costruisca la versione binarizzata del data set ottenuto
b) si derivino gli itemesets frequenti con MinSupp>= 30%

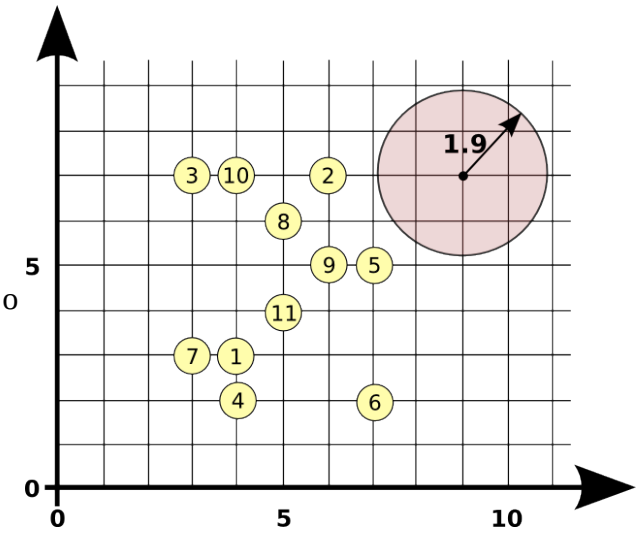
Esercizio 3 Similarity function (4 punti)

Definire la funzione di similarità di Jaccard fra vettori binari e descrivere le situazioni in cui assume valore massimo e minimo.

Esercizio 4 - Clustering **(12 punti)**

Nel seguente dataset:

A) Si utilizzi l'algoritmo di clustering density-based DBSCAN, con raggio (ε) pari a 1.9, e minPts pari a 4 (=3 vicini + il punto di cui si calcola la densità). Si richiede di (1) indicare il numero di cluster che si ottengono; (2) per ogni punto indicare il cluster di appartenenza; (3) per ogni punto dire se si tratta di un core point , border point o rumore **(8 punti)**
B) Se si utilizza un algoritmo di clustering gerarchico agglomerativo MAX-link, fermando la computazione dopo 4 passi, quanti e quali cluster si ottengono? **(4 punti)**



Esercizio 5 Classificazione**(17 punti)**

Si consideri il seguente insieme di transazioni (*training set*).

Price	Weight	Owner	Bid	Success
High	Regular	Private	Stable	N
High	Exceed	Private	Increasing	N
Medium	Regular	Private	Stable	N
High	Exceed	Private	Increasing	N
High	Exceed	Company	Stable	N
Low	Exceed	Company	Increasing	N
Medium	Exceed	Private	Stable	N
Low	Regular	Company	Stable	Y
Low	Exceed	Private	Stable	Y
Medium	Exceed	Private	Increasing	Y
Medium	Regular	Company	Increasing	Y
High	Regular	Company	Increasing	Y
Medium	Exceed	Company	Increasing	Y
Low	Regular	Private	Increasing	Y

A) Si costruisca su tale dataset un albero di decisione per la variabile Success , utilizzando il criterio di split basato su misclassification rate , espandendo i nodi dell'albero fino a che la precisione non è più migliorabile ed assicurando che tutte le foglie contengono almeno 2 transazioni del training set (ovvero: evitando espansioni di nodi che porterebbero ad avere uno o più nodi/foglie con solo 1 transazione). **(10 punti)**

B) Calcolare la matrice di confusione dell'albero ottenuto al punto A), sia sul training set che sul test set riportato qui sott . Confrontare le due matrici e commentare il risultato. **(7 punti)**

Price	Weight	Owner	Bid	Success
High	Exceed	Private	Stable	Y
High	Exceed	Company	Stable	N
Low	Regular	Company	Increasing	N
Medium	Exceed	Private	Stable	N
Medium	Regular	Company	Stable	Y
Low	Exceed	Private	Stable	N
Medium	Exceed	Private	Increasing	Y
Medium	Regular	Company	Increasing	Y

Esercizio 6 Classificazione**(15 punti)**

Si consideri il seguente insieme di transazioni con attributi continui:

A	B	Class
23	61	0
43	6	0
51	66	0
60	3	1
80	52	1
83	3	1
90	88	0
98	26	1

A) Si costruisca un albero di decisione per la variabile target Class , terminando la costruzione quando l'albero ha precisione 100% sul training set. **(8 punti)**
B) Si discretizzino le due variabili continue in 2 intervalli di uguale ampiezza, e si costruisca un albero di decisione sul dataset discretizzato. **(5 punti)**
C) Si valuti la precisione dei due alberi di decisione sul seguente test set **(2 punti)**:

A	B	Class
39	58	1
20	7	1
19	70	0
32	18	1
7	55	0
13	92	1
82	57	0
60	43	1