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 18 gennaio 2010

Esercizio 1 - Sequential Patterns (**6 punti**)

Si consideri il seguente dataset di sequenze:

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

Si indichi il supporto delle seguenti sotto-sequenze senza considerare vincoli temporali (colonna sinistra) e considerando il vincolo temporale *max-gap = 2* (colonna destra):

	supporto	*supporto con max-gap=2*
w_1 = <{C}{B}{C}>		
w_2 = <{E} {H} >		
w_3 = <{B}{A,B}>		

Esercizio 2 – Itemset Frequenti (**12 punti**)

Considerare la seguente tabella di transazioni:

ID	ITEMS
1	A B
2	A D
3	C
4	B D
5	A C D

ID	ITEMS
6	A
7	A D
8	A C
9	B C D
10	B C D

A) Disegnare il reticolo (*lattice*) degli itemset generabili a partire da tali transasizioni.

B) Indicare (per iscritto o graficamente sul reticolo) gli itemset frequenti nel caso di un supporto minimo min_sup = 20%.

C) Indicare (per iscritto o graficamente sul reticolo) gli itemset **infrequenti** che l'algoritmo apriori andrebbe comunque a generare (ovvero quelli che non è in grado di scartare senza una scansione ulteriore del dataset).

D) Indicare (per iscritto o graficamente sul reticolo) gli itemset frequenti massimali.

Esercizio 3 – Proprietà itemset frequenti / regole associative (**2 punti**)

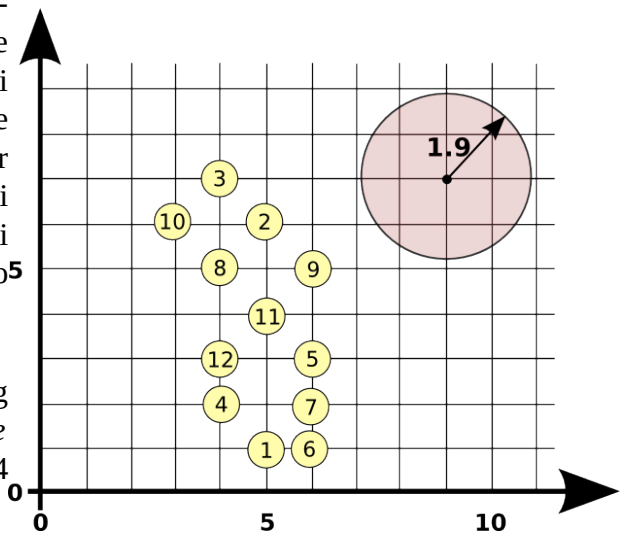
Sia dato un itemset A, 3 dataset contenenti rispettivamente X, 100 e 200 transazioni. Sappiamo che A ha un supporto del 50% sui primi due dataset, ma non sappiamo nulla del suo supporto nel terzo dataset. Quale valore minimo deve avere X per aver la certezza che A raggiunga un supporto del 40% sul dataset totale, ovvero quello ottenuto unendo i tre dataset e formato da X+300 transazioni?

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 MIN-link (o *Single linkage*), fermando la computazione dopo 4 passi, quali cluster si ottengono? **(4 punti)**



Esercizio 5 – Classificazione **(17 punti)**

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

Consumo	**Nuova**	**Accelerazione**	**Risposta**	**Guasto**
Alto	No	Media	Lenta	Si
Alto	No	Bassa	Lenta	Si
Alto	Si	Media	Lenta	No
Alto	No	Alta	Veloce	No
Alto	No	Media	Lenta	No
Basso	No	Bassa	Veloce	Si
Alto	No	Media	Veloce	Si
Basso	No	Bassa	Veloce	No
Basso	Si	Bassa	Lenta	No
Alto	Si	Media	Veloce	No
Alto	No	Media	Veloce	No
Alto	Si	Bassa	Veloce	Si

A) Si costruisca su tale dataset un albero di decisione per la variabile “Guasto”, utilizzando il criterio di split basato su “misclassification rate”, espandendo i nodi dell'albero fino a che la precisione non è più migliorabile localmente (ovvero nessuno split da' un guadagno). **(10 punti)**

B) Si valuti l'accuratezza dell'albero ottenuto al punto A) tramite matrice di confusione, calcolata sia sul training set che sul test set riportato qui sotto. Confrontare i risultati. **(7 punti)**

Consumo	Nuova	Accelerazione	Risposta	Guasto
Alto	No	Bassa	Lenta	Si
Basso	Si	Bassa	Veloce	No
Alto	Si	Alta	Veloce	No
Alto	No	Alta	Veloce	Si
Basso	No	Alta	Lenta	No
Alto	No	Media	Lenta	No
Basso	Si	Media	Lenta	Si
Alto	Si	Media	Lenta	No
Basso	No	Alta	Lenta	Si
Alto	Si	Alta	Veloce	No

Esercizio 6 – Classificazione **(15 punti)**

Si consideri il seguente insieme di transazioni con attributi sia discreti che continui:

Cost	Duration	Class
High	31	No
High	36	No
High	40	No
High	20	Yes
Low	45	No
Low	10	Yes
Low	13	Yes
Low	35	Yes

A) Si costruisca un albero di decisione per la variabile target “Class”, utilizzando come criterio di split il “Misclassification Rate” e terminando la costruzione quando la precisione dell'albero non è più migliorabile. **(10 punti)**

B) Si sostituisca la variabile “Duration” con “DurDisc”, che ha valore “L” quando “Duration<36” e “H” in tutti gli altri casi. Si costruisca un albero di decisione per il nuovo dataset, e si confronti con quello ottenuto al punto A, commentando le differenze. **(5 punti)**