Università di Pisa	A.A. 2009- 2010

Data Mining

II verifica intermedia – 25 maggio 2010

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

Si consideri la seguente sequenza di input:

<	{A}	{A,B,C}	{D}	{H}	{B, E}	{A,B,D}	>
	t=0	t=1	t=2	t=3	t=4	t=5	

Si indichi quali sono le occorrenze delle seguenti sotto-sequenze nella sequenza di input, senza considerare vincoli temporali (colonna sinistra) e considerando il vincolo temporale *max-gap = 1* (colonna destra). Per brevità, si rappresenti ogni occorrenza tramite la corrispondente ennupla di tempi nella sequenza di input, es.: <0,2,3> = <t=0, t=2, t=3>.

	Occorrenze	*Occorrenze con max-gap=1*
es.: <{A}{D}{H}>	**<0,2,3> <1,2,3>**	**<1,2,3>**
w_1 = <{A} {B} {D} >		
w_2 = <{A}{H}{B}>		
w_2 = <{A}{C}{E}>		

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

Considerare la seguente tabella di transazioni:

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

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

A) Elencare gli itemset frequenti nel caso di un supporto minimo min_sup = 25% ed indicare il loro supporto.

B) Quali itemset frequenti sono anche closed? Quali sono massimali?

Esercizio 3 – FP-tree (**2 punti**)

Si disegni l' FP-tree corrispondente al seguente dataset:

ID	Itemset
1	A
2	B D E
3	A C
4	B D
5	B C

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

Sul 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) Si disegni il dendogramma ottenuto con un algoritmo di clustering agglomerativo MIN-link (o *Single linkage*). **(4 punti)**