Università di Pisa	A.A. 2009- 2010

Data Mining

Appello del 1 giugno 2010

Esercizio 1 - Sequential Patterns (**4 punti**)

Si consideri la seguente sequenza di input:

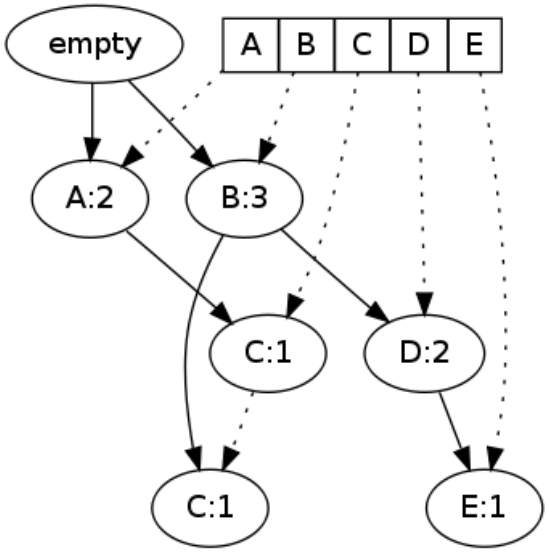
<	{A}	{A,B,C}	{C,D,E}	{A,E,H}	{B}	{A,B,D} >
	t=0	t=1	t=2	t=3	t=4	t=5

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 = <{B} {E} {D} >		
w_2 = <{A}{A,B}>		
w_2 = <{A}{C}{E}>		

Esercizio 2 – FP-tree (**2 punti**)

Si ricostruisca il dataset di transazioni (itemset) da cui il seguente FP-tree è stato ottenuto.



Esercizio 3 – Itemset Frequenti (**6 punti**)

Considerare la seguente tabella di transazioni:

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

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

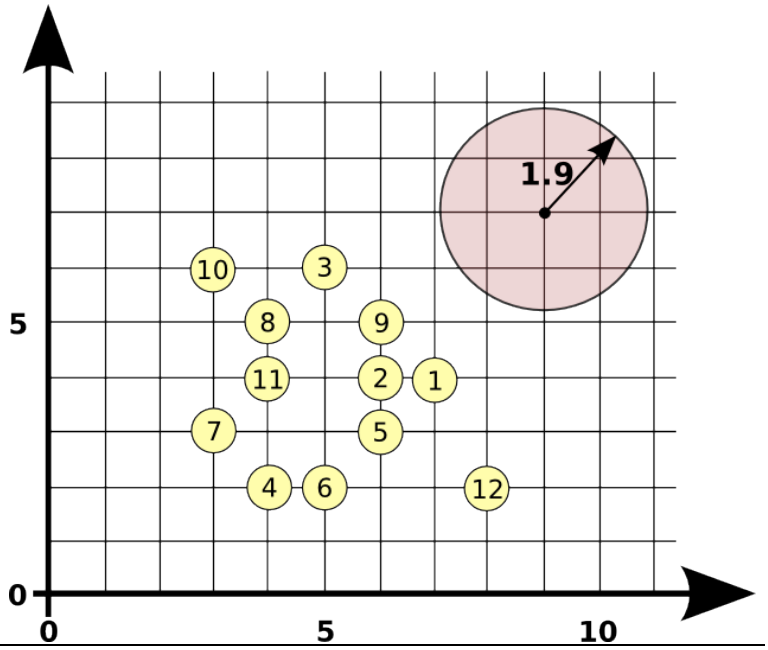
A) Elencare gli itemset frequenti nel caso di un supporto minimo min_sup = 20% ed indicare il loro supporto.
B) Quali itemset frequenti sono anche massimali?

Esercizio 4 - Clustering **(10 punti)**

Sul dataset mostrato in figura:

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).
(1) per ogni punto dire se si tratta di un *core point*, *border point* o *rumore;*
(2) indicare la composizione dei cluster ottenuti. **(5 punti)**

B) Si simuli una esecuzione dell' algoritmo k-means sullo stesso insieme di punti, con k=2 e centri iniziali c_1=(2,4) e c_2=(9,4).
(5 punti)



Esercizio 5 – Classificazione **(10 punti)**

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

Training set:

Età	Contratto	Sesso	Aumento
50	Temporaneo	F	Si
35	Stabile	M	No
50	Temporaneo	F	Si
40	Stabile	M	No
25	Temporaneo	F	Si
30	Temporaneo	F	No
50	Stabile	F	No
40	Stabile	F	Si
55	Temporaneo	M	No
55	Stabile	F	No
25	Temporaneo	M	Si
40	Stabile	M	No

A) Si costruisca su tale dataset un albero di decisione per la variabile "Aumento", 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). **(7 punti)**

B) Si mostrino accuratezza e matrice di confusione dell'albero ottenuto al punto A), calcolati sia sul training set che sul test set riportato qui sotto. Confrontare i risultati. **(3 punti)**

Test set:

Età	Contratto	Sesso	Aumento
35	Stabile	F	No
45	Temporaneo	M	No
30	Stabile	M	No
25	Stabile	F	Si
40	Temporaneo	M	No
55	Stabile	F	Si
30	Temporaneo	F	Si
35	Temporaneo	M	Si