

**Data Mining**

Appello del 22 giugno 2010

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

Si consideri la seguente sequenza di input:

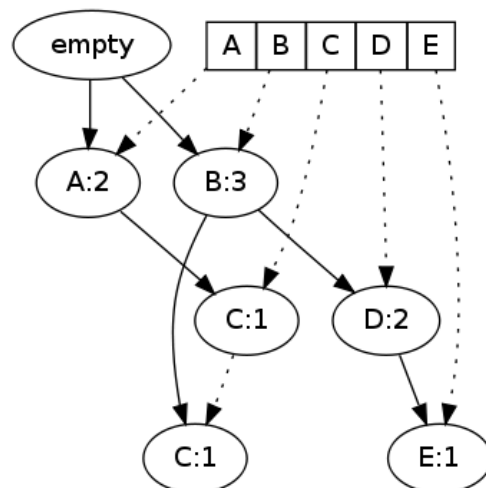
$$\begin{array}{ccccccc} < & \{A,C\} & \{C,D,E\} & \{F\} & \{A,H\} & \{B,C,D\} & \{E\} & \{A,B,D\} & > \\ & t=0 & t=1 & t=2 & t=3 & t=4 & t=5 & t=6 & \end{array}$$

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  $max\text{-}gap = 2$  (colonna destra). Per brevità, si rappresenti ogni occorrenza tramite la corrispondente ennupla di tempi nella sequenza di input, es.:  $\langle 0,2,3 \rangle = \langle t=0, t=2, t=3 \rangle$ .

	<i>Occorrenze</i>	<i>Occorrenze con max-gap=2</i>
<i>es.:</i> $\langle \{C\}\{H\}\{C\} \rangle$	$\langle 0,3,4 \rangle \langle 1,3,4 \rangle$	$\langle 1,3,4 \rangle$
$w_1 = \langle \{A\} \{B\} \rangle$		
$w_2 = \langle \{C\}\{C\}\{E\} \rangle$		
$w_2 = \langle \{A\}\{E\} \rangle$		

Esercizio 2 – FP-tree (2 punti)

L' FP-tree in figura rappresenta un dataset di transazioni (itemset). Si costruisca l'FP-tree che si ottiene proiettando il dataset/FP-tree per il suffisso "C". Si assuma che  $\text{min\_supp}=1$ .



Esercizio 3 – Itemset Frequenti (6 punti)

Considerare la seguente tabella di transazioni:

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

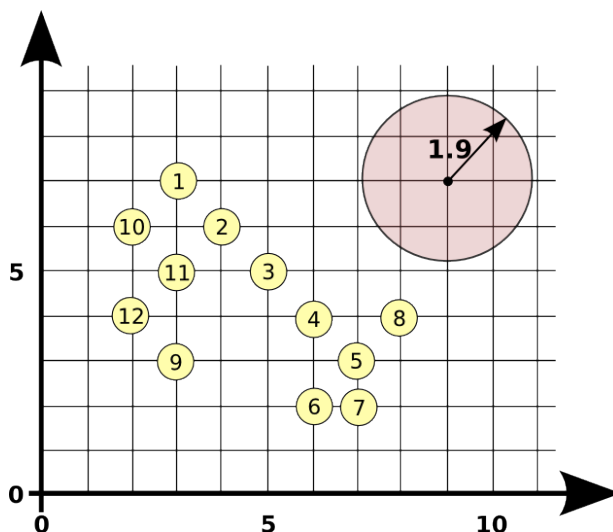
ID	ITEMS
6	B C
7	C D
8	A B E
9	A B C E
10	A B C

- A) Elencare gli itemset frequenti nel caso di  $\text{min\_sup} = 20\%$  ed indicare il loro supporto.
- B) Quali itemset frequenti sono anche massimali?

Esercizio 4 - Clustering (10 punti)

Sul seguente dataset:

- A) Si utilizzi l'algoritmo di clustering density-based DBSCAN, con raggio ( $\epsilon$ ) pari a 1.9, e  $\text{minPts}$  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) Simulare l'esecuzione dell'algoritmo k-means sullo stesso insieme di punti, con  $k=2$  e centri iniziali  $c_1=(3,7)$  e  $c_2=(8,2)$ . (5 punti)



## Esercizio 5 – Classificazione (10 punti)

---

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

Training set:

nFigli	Status	Lavoro	Sussidio
3	Celibe	No	Si
2	Celibe	No	Si
0	Celibe	Stabile	No
4	Sposato	Precario	Si
1	Sposato	Precario	No
1	Sposato	No	No
1	Celibe	Precario	Si
2	Sposato	No	No
6	Celibe	No	Si
3	Sposato	Stabile	No

- A) Si costruisca su tale dataset un albero di decisione per la variabile “Sussidio”, 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:

nFigli	Status	Lavoro	Sussidio
1	Sposato	Stabile	No
4	Sposato	Precario	No
1	Celibe	Precario	Si
6	Sposato	No	No
5	Sposato	No	Si
3	Celibe	Precario	Si
1	Celibe	Stabile	No
5	Celibe	Precario	Si