Università di Pisa	A.A. 2008- 2009

Data Mining

Appello del 22 giugno 2010
Soluzioni

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

Si consideri la seguente sequenza di input:

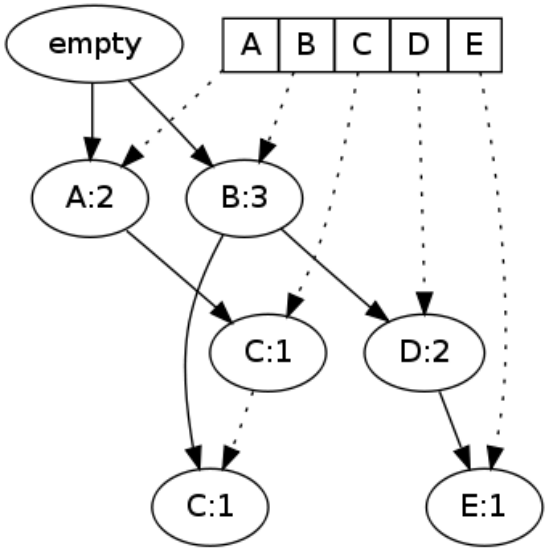
<	{A,C}	{C,D,E}	{F}	{A,H}	{B,C,D}	{E}	{A,B,D}	>
	t=0	t=1	t=2	t=3	t=4	t=5	t=6	

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 = 2* (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=2*
es.: <{C}{H}{C}>	**<0,3,4> <1,3,4>**	**<1,3,4>**
w_1 = <{A} {B} >	**<0,4> <0,6> <3,4> <3,6>**	**<3,4>**
w_2 = <{C}{C}{E}>	**<0,1,5> <0,4,5><1,4,5>**	**nessuna**
w_2 = <{A}{E}>	**<0,1> <0,5> <3,5>**	**<0,1> <3,5>**

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 min_supp=1.



Soluzione: idem, rimuovendo i nodi D ed E (opzionalmente anche C, dato che è implicito), e mettendo tutte le frequenze a 1.

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

vConsiderare la seguente tabella di transazioni:

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

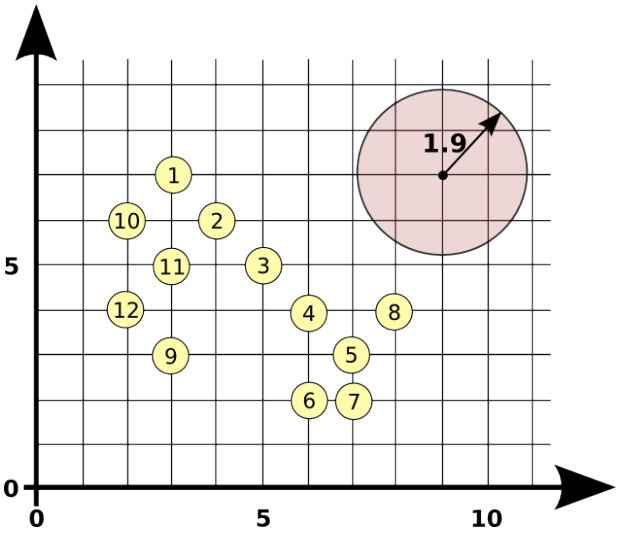
ID	ITEMS
6	B C
7	C D
8	A B E
9	A B C E
10	A B C

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

(m=massimale)
A (70.0)
B (50.0)
C (80.0)
D (30.0)
E (20.0)
A B (40.0)
A C (50.0)
A E (20.0)
B C (30.0)
B E (20.0)
m C D (30.0)
m A B C (20.0)
m A B E (20.0)

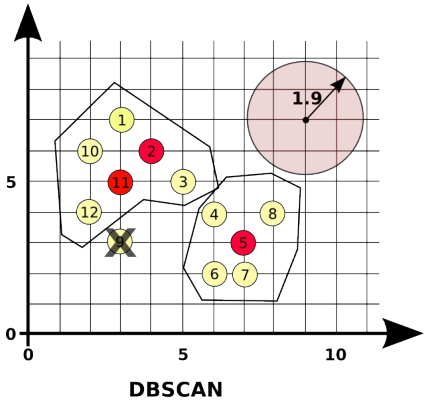
Esercizio 4 - Clustering **(10 punti)**

Sul seguente dataset:

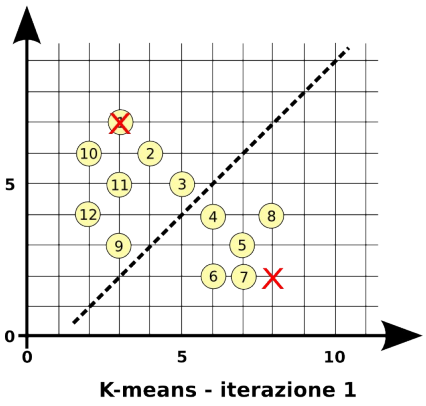


A) Si utilizzi l'algoritmo di clustering density-based DBSCAN, con raggio (ε) pari a 1.9, e minPts pari a 4 (=3 vicini + il punto di cui si calcola la densità).
(1) per ogni punto dire se si tratta di un *core point*, *border point* o *rumore;*
(2) indicare la composizione dei cluster ottenuti. **(5 punti)**

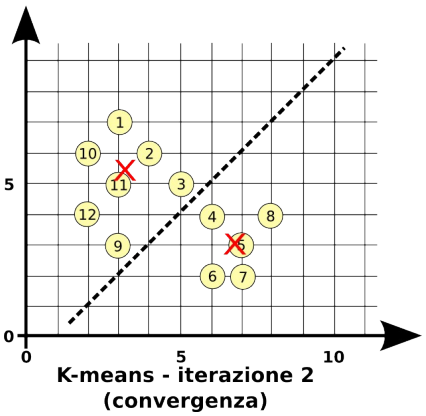
B) Simulare l'esecuzione dell'algoritmo k-means sullo stesso insieme di punti, con k=2 e centri iniziali c_1=(3,7) e c_2=(8,2). **(5 punti)**



DBSCAN



K-means - iterazione 1



K-means - iterazione 2 (convergenza)

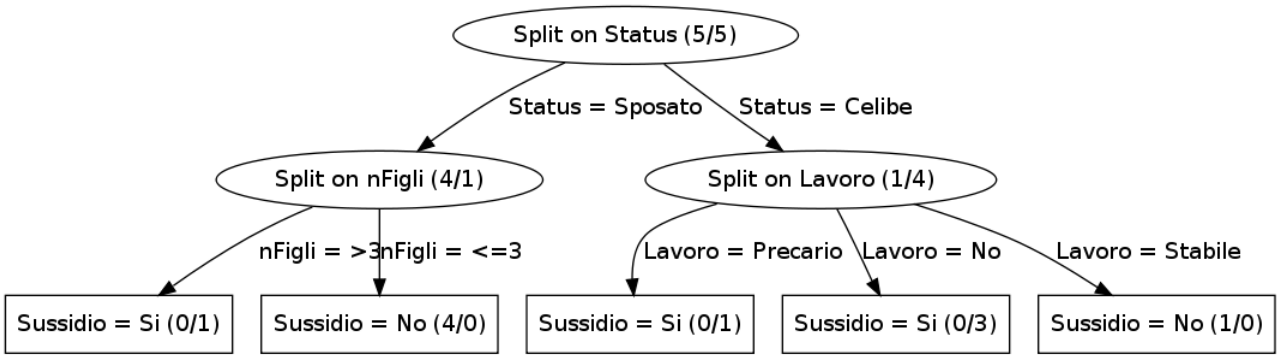
Esercizio 4 – Classificazione **(10 punti)**

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

Matrici di confusione e accuratezza:

Train:

	Si	No
Si	5	0
No	0	5

(Predetta) (Reale)

Accuracy: 100,00%

Test:

	Si	No
Si	4	0
No	2	2

(Predetta) (Reale)

Accuracy: 75,00%