Università di Pisa	A.A. 2009- 2010

Data Mining

Appello del 9 settembre 2010
Soluzioni

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

Si consideri la seguente sequenza di input:

<	{A,C}	{C,D}	{F,H}	{A,B}	{B,C,D}	{E}	{A,B,D}	{F}	>
	t=0	t=1	t=2	t=3	t=4	t=5	t=6	t=7	

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 *min-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 min-gap=1*
es.: <{C}{H}{C}>	**<0,2,4> <1,2,4>**	**<0,2,4>**
w_1 = <{A} {F} >	**<0,2> <0,7>** **<3,7> <6,7>**	**<0,2> <0,7>** **<3,7>**
w_2 = <{A}{A}{D}>	**<0,3,4> <0,3,6>**	**<0,3,6>**
w_2 = <{A}{A,B}{F}>	**<0,3,7> <0,6,7> <3,6,7>**	**<0,3,7>**

Esercizio 2 – Lift e confidenza di regole associative (**2 punti**)

La regola associativa "a → b" ha supporto pari a 0.6, confidenza = 0.8 e lift = 2. Si calcolino i valori del supporto di "a" e di "b".

Soluzione:

confidenza = s(a,b)/s(a) => s(a) = s(a,b) / confidenza = 0.6 / 0.8 = 0.75

lift=s(a,b)/[s(a)s(b)] => s(b) = s(a,b) / [s(a) lift] = 0.6 / [0.75 x 2] = 0.4

Si può notare, però, che il valore ottenuto per s(b) < s(a,b), il ché mostra che i valori di partenza del problema sono incoerenti, cioè non esiste alcun dataset D da cui essi possano emergere.

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

Considerare la seguente tabella di transazioni:

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

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

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)

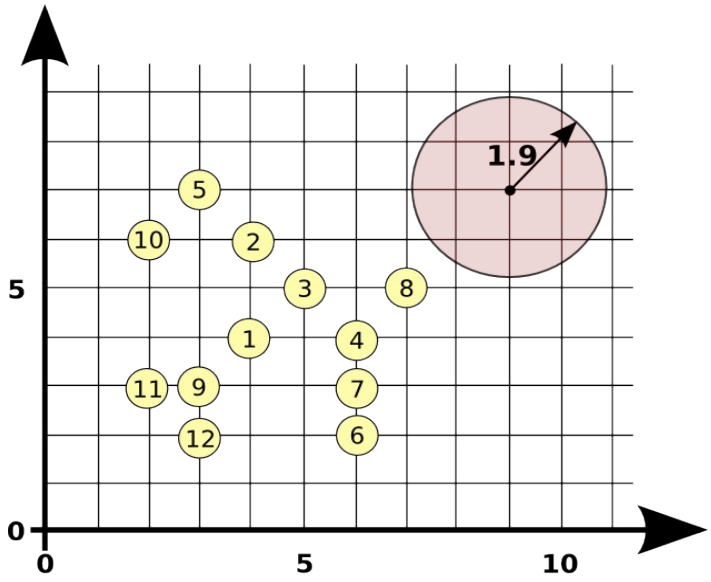
A (60.0)
B (70.0)
C (90.0)
D (40.0)
A B (30.0)
A C (60.0)
A D (20.0)
B C (60.0)
B D (30.0)
C D (40.0)
m A B C (30.0)
m A C D (20.0)
m B C D (30.0)

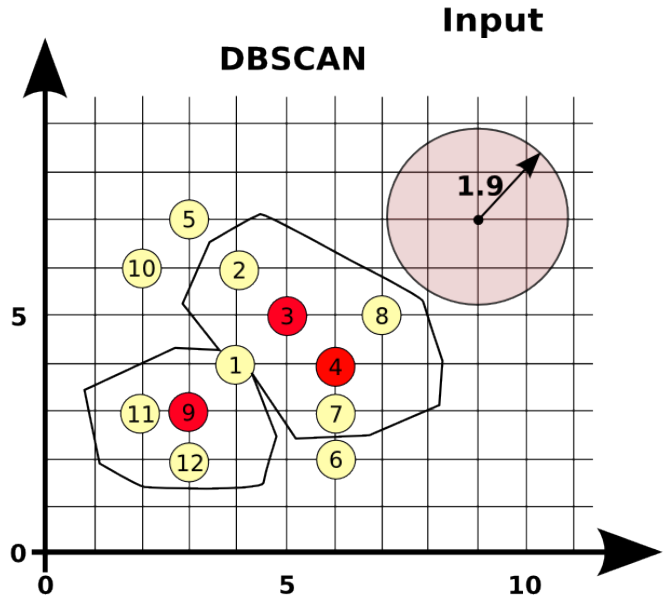
Esercizio 4 - Clustering **(10 punti)**

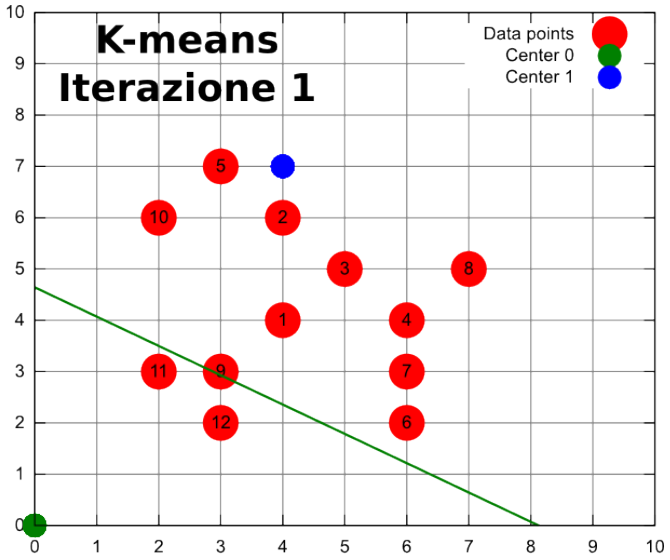
Sul seguente dataset:

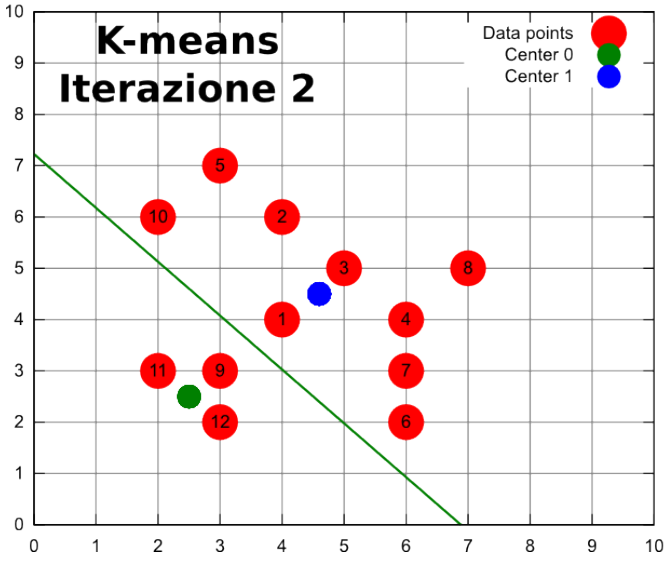
A) Si utilizzi l'algoritmo di clustering density-based DBSCAN, con raggio (ε) pari a 1.9, e minPts pari a 4 (=3 vicini + il punto di cui si calcola la densità).
   (1) per ogni punto dire se si tratta di un *core point*, *border point* o *rumore;*
   (2) indicare la composizione dei cluster ottenuti.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

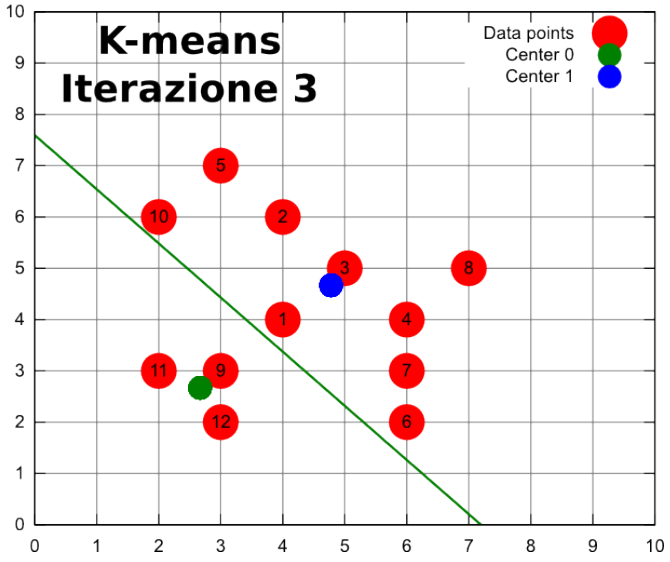
B) Simulare l'esecuzione dell'algoritmo k-means sullo stesso insieme di punti, con k=2 e centri iniziali c_1=(0,0) e c_2=(4,7). **(5 punti)**


1.9
5
10
2
3
8
1
4
11
9
7
12
6
0
5
10
0
5


Input
DBSCAN
1.9
5
10
2
3
8
1
4
11
9
7
12
6
5
0
0
5
10


K-means
Iterazione 1
Data points
Center 0
Center 1


K-means
Iterazione 2
Data points
Center 0
Center 1

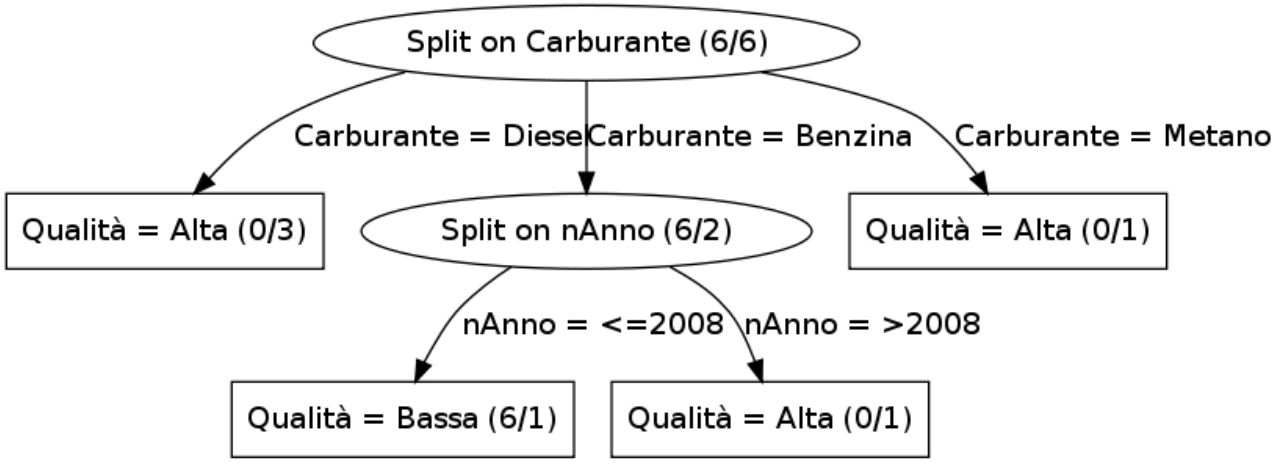

K-means
Iterazione 3
Data points
Center 0
Center 1

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

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

nAnno	Usato	Carburante	Qualità
2004	Si	Benzina	Bassa
2008	No	Metano	Alta
2003	No	Diesel	Alta
2006	Si	Benzina	Alta
2001	No	Benzina	Bassa
2001	Si	Benzina	Bassa
2008	No	Benzina	Bassa
2006	Si	Diesel	Alta
2001	No	Diesel	Alta
2002	Si	Benzina	Bassa
2007	Si	Benzina	Bassa
2009	Si	Benzina	Alta

A) Si costruisca su tale dataset un albero di decisione per la variabile "Qualità", 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:

nAnno	Usato	Carburante	Qualità
2001	No	Metano	Bassa
2008	No	Diesel	Alta
2010	No	Benzina	Bassa
2006	Si	Diesel	Bassa
2005	No	Metano	Alta
2010	No	Benzina	Bassa
2007	Si	Diesel	Bassa
2001	No	Diesel	Bassa

Matrici di confusione e accuratezza:

Train:

	Alte	Basse
Alte	5	1
Basse	0	6

(Predetta) Accuracy: 91,67%

(Reale)

Test:

	Alta	Bassa
Alta	0	0
Bassa	6	2

(Predetta) Accuracy: 25,00%

(Reale)