

Data Mining

Appello del 13 luglio 2010

Soluzioni

Esercizio 1 - Sequential Patterns (4 punti)

Si consideri la seguente sequenza di input:

$$\begin{array}{cccccccc} < & \{A,C\} & \{C,D\} & \{F,H\} & \{A,B\} & \{B,C,D\} & \{E\} & \{A,B,D\} & \{F\} & > \\ & t=0 & t=1 & t=2 & t=3 & t=4 & t=5 & t=6 & t=7 & \end{array}$$

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 $min-gap = 1$ (colonna destra). Per brevità, si rappresenti ogni occorrenza tramite la corrispondente ennupla di tempi nella sequenza di input, es.: $\langle 0,2,3 \rangle = \langle t=0, t=2, t=3 \rangle$.

	Occorrenze	Occorrenze con $min-gap=1$
es.: $\langle \{C\} \{H\} \{C\} \rangle$	$\langle 0,2,4 \rangle \langle 1,2,4 \rangle$	$\langle 0,2,4 \rangle$
$w_1 = \langle \{A\} \{B\} \rangle$	$\langle 0,3 \rangle \langle 0,4 \rangle \langle 0,6 \rangle$ $\langle 3,4 \rangle \langle 3,6 \rangle$	$\langle 0,3 \rangle \langle 0,4 \rangle \langle 0,6 \rangle$ $\langle 3,6 \rangle$
$w_2 = \langle \{C\} \{D\} \{E\} \rangle$	$\langle 0,1,5 \rangle \langle 0,4,5 \rangle \langle 1,4,5 \rangle$	nessuna
$w_2 = \langle \{F\} \{B\} \{F\} \rangle$	$\langle 2,3,7 \rangle \langle 2,4,7 \rangle \langle 2,6,7 \rangle$	$\langle 2,4,7 \rangle$

Esercizio 2 – Closed Itemset (2 punti)

Da un database di transazioni si sono estratti gli itemset frequenti e quelli closed con supporto minimo del 20%. Gli itemset closed trovati, con relativo supporto, sono i seguenti:

C (100.0) A C (70.0) B C (80.0) C D (30.0) A B C (50.0) A C D (20.0) Si elenchino tutti gli itemset frequenti, con relativo supporto.	Soluzione: A (70.0) B (80.0) C (100.0) D (30.0) A B (50.0) A C (70.0) A D (20.0) B C (80.0) C D (30.0) A B C (50.0) A C D (20.0)
-------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------

Esercizio 3 – Itemset Frequenti (6 punti)

Considerare la seguente tabella di transazioni:

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

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

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

(m=massimale) A (70.0) B (70.0) C (70.0) D (30.0) m E (20.0) AB (40.0) AC (50.0) m AD (20.0) BC (40.0) m BD (20.0) m CD (20.0) m ABC (20.0)

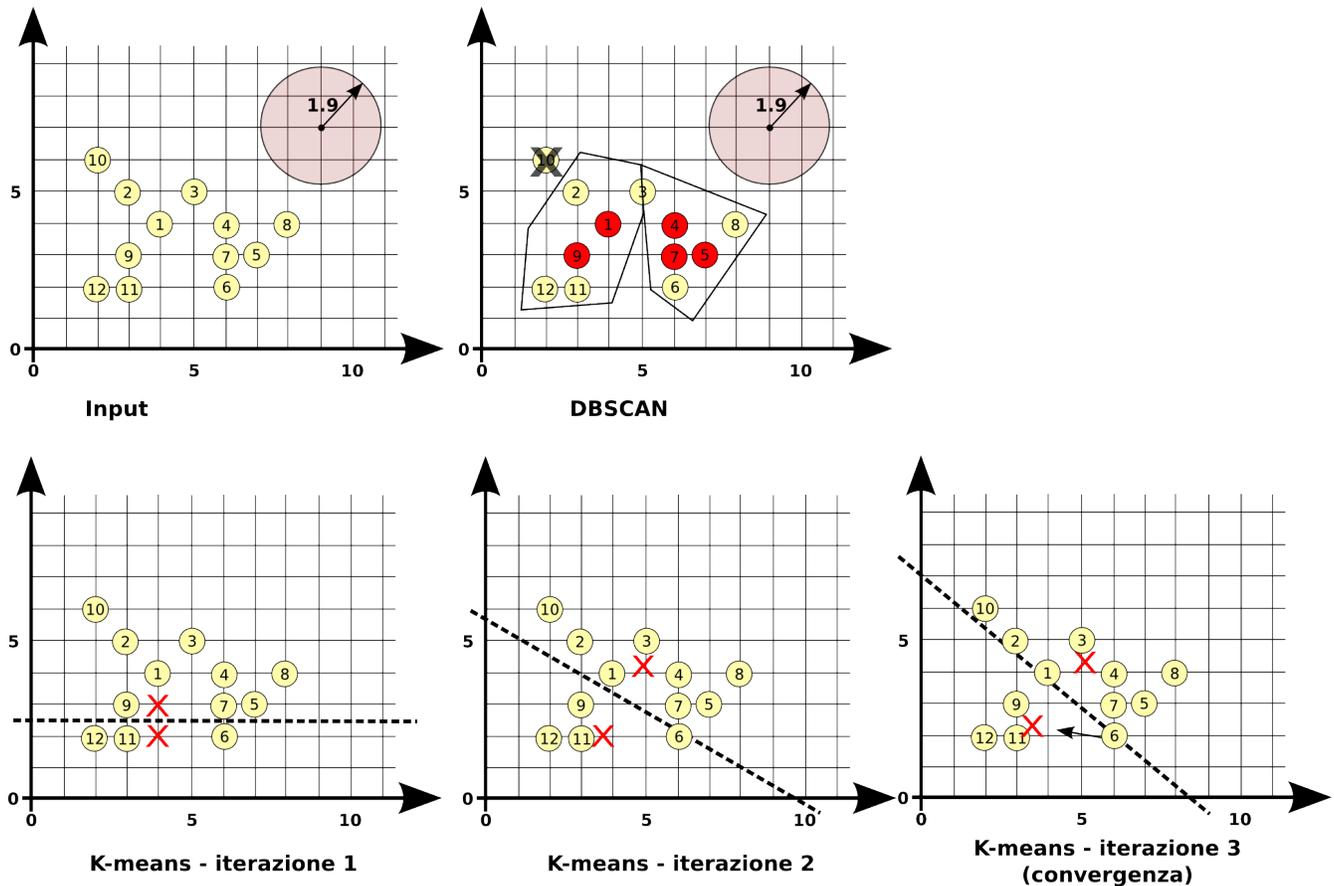
Esercizio 4 - Clustering (10 punti)

Sul seguente dataset:

A) Si utilizzi l'algorithmo 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'algorithmo k-means sullo stesso insieme di punti, con $k=2$ e centri iniziali $c_1=(4,2)$ e $c_2=(4,3)$. (5 punti)

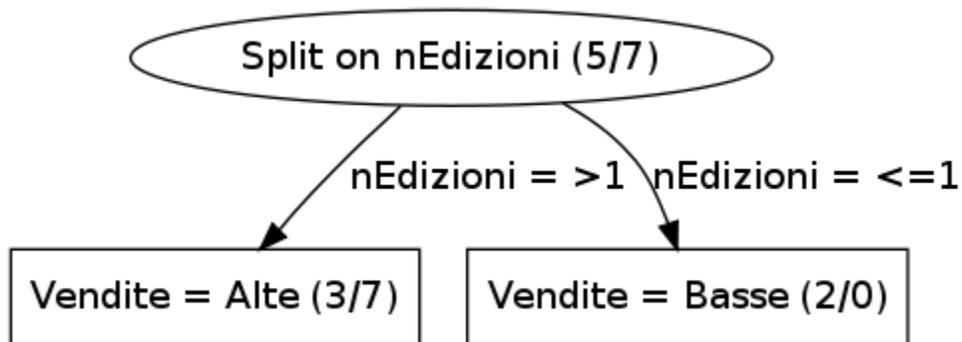


Esercizio 5 – Classificazione (10 punti)

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

nEdizioni	Rilegato	Autori	Vendite
6	Si	Singolo	Alte
6	Si	Singolo	Basse
3	Si	Singolo	Alte
4	No	Singolo	Basse
1	No	Singolo	Basse
3	Si	Singolo	Alte
2	No	Multipli	Alte
1	No	Singolo	Basse
6	No	Singolo	Alte
6	No	Multipli	Alte
6	Si	Multipli	Basse
4	Si	Multipli	Alte

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

	nEdizioni	Rilegato	Autori	Vendite
Test set:	2	Si	Singolo	Alte
	5	Si	Multipli	Basse
	2	Si	Singolo	Basse
	2	Si	Singolo	Basse
	1	No	Multipli	Basse
	1	Si	Singolo	Basse
	4	Si	Multipli	Alte
	1	Si	Singolo	Basse

Matrici di confusione e accuratezza:

Train:

	Alte	Basse
Alte	7	0
Basse	3	2

(Reale)

(Predetta)

Accuracy:

75,00%

Test:

	Alte	Basse
Alte	2	0
Basse	3	3

(Reale)

(Predetta)

Accuracy:

62,50%