Università di Pisa	A.A. 2008- 2009

Data Mining - *Corso di Laurea Specialistica in Informatica per l'economia e l'Azienda*
Tecniche Data Mining - *Corsi di Laurea Specialistica in Informatica e Tecnologie Informatiche*

PARTE A = Esercizi 1-4 **PARTE B = Esercizi 5-6**

Appello del 18 febbraio 2010

Esercizio 1 - Sequential Patterns (**6 punti**)

Si consideri la seguente sequenza di input:

<	{A,B}	{A,C}	{D,E}	{A, B, C}	{B, E}	{H}	{A} >
	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.: <{D}{B}{A}>	**<2,3,6> <2,4,6>**	**<2,4,6>**
w_1 = <{A} {B} {E} >		
w_2 = <{B}{A}>		

Esercizio 2 – Itemset Frequenti (**12 punti**)

Considerare la seguente tabella di transazioni:

ID	ITEMS
1	X
2	Y
3	W X Z
4	W Y
5	X Y

ID	ITEMS
6	W X Y
7	Y Z
8	W X
9	W Y
10	W X Z

A) Elencare gli itemset frequenti nel caso di un supporto minimo min_sup = 20% ed indicare il loro supporto.

B) Oltre a quelli (frequenti) elencati sopra, per quali altri itemset l'algoritmo a-priori andrebbe a calcolare il supporto esatto tramite scansione del database?

Esercizio 3 – Proprietà itemset frequenti / regole associative (**2 punti**)

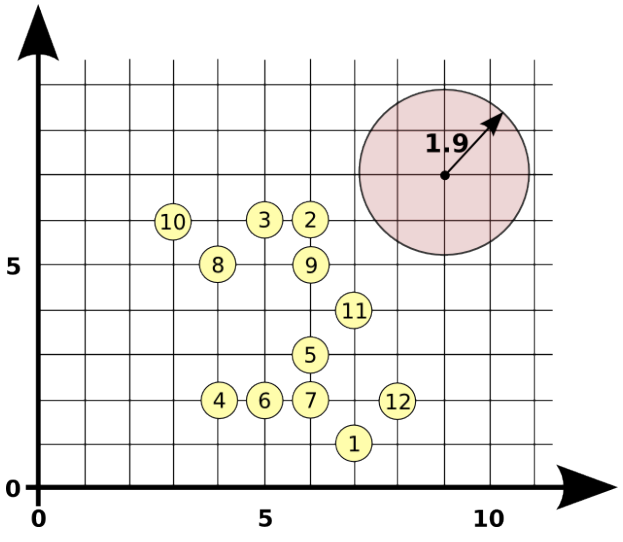
Sia dato un dataset di 1000 transazioni, sul quale la regola A=>B ha supporto 50% e confidenza 80%. Se aggiungiamo 1000 nuove transazioni al dataset di partenza, sulle quali A=>B ha supporto 30%, quale può essere la confidenza di A=>B sul dataset complessivo nel caso ottimo (confidenza massima) e nel caso pessimo (confidenza minima)?

Esercizio 4 - Clustering **(12 punti)**

Nel 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). Si richiede di (1) indicare il numero di cluster che si ottengono; (2) per ogni punto indicare il cluster di appartenenza; (3) per ogni punto dire se si tratta di un *core point*, *border point* o *rumore*. **(8 punti)**

B) Se si utilizza un algoritmo di clustering gerarchico agglomerativo MIN-link (o *Single linkage*), fermando la computazione dopo 5 passi, quali cluster si ottengono? **(4 punti)** (Nota: ad ogni passo si fondono 2 soli cluster. In caso di più scelte con la stessa distanza, la coppia di cluster da fondere è arbitraria)



Esercizio 5 – Classificazione **(20 punti)**

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

Altezza	**Peso**	**Età**	**Sesso**	**Malattia**
Bassa	Alto	Giovane	F	No
Bassa	Basso	Giovane	F	Si
Bassa	Basso	Anziano	M	No
Bassa	Medio	Giovane	M	Si
Bassa	Alto	Giovane	M	Si
Alta	Medio	Anziano	F	No
Bassa	Alto	Giovane	F	No
Alta	Basso	Anziano	M	Si
Alta	Basso	Anziano	M	Si
Bassa	Medio	Anziano	M	Si

A) Si costruisca su tale dataset un albero di decisione per la variabile “Malattia”, 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). **(12 punti)**

B) Si valuti l'accuratezza dell'albero ottenuto al punto A) tramite matrice di confusione, calcolata sia sul training set che sul test set riportato qui sotto. Confrontare i risultati. **(8 punti)**

Altezza	Peso	Età	Sesso	Malattia
Alta	Alto	Anziano	F	No
Bassa	Alto	Anziano	F	No
Bassa	Basso	Giovane	F	No
Alta	Basso	Giovane	M	Si
Alta	Medio	Anziano	M	Si
Alta	Basso	Giovane	M	No
Alta	Medio	Anziano	F	No
Alta	Medio	Anziano	F	No

Esercizio 6 – Classificazione con attributi continui **(12 punti)**

Si consideri il seguente insieme di transazioni con un unico attributo continuo (oltre alla classe):

Value	Class
7	No
12	No
20	No
25	Yes
31	Yes
40	Yes
45	No
80	No
81	No
85	Yes
90	Yes

Si costruisca un albero di decisione per la variabile target “Class”, utilizzando come criterio di split il “Misclassification Rate” e terminando la costruzione quando la precisione dell'albero non è più migliorabile.