| 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 3 settembre 2009

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

Si consideri il seguente dataset di sequenze:

< {A,C} {E, C} {B} {C,D} {A,H} {B,C} >
< {B} {B,C, D,E} {E} {E} {H} {A,B}>
< {B} {D,E} {E, C} {E,H} {H} {A} >
< {A,B} {A,C} {D,E} {B, C} {E} {H} {A} >

Si indichi il supporto delle seguenti sotto-sequenze senza considerare vincoli temporali (colonna sinistra) e considerando il vincolo temporale *max-span = 3* (colonna destra):

| | *supporto* | *supporto con max-span=3* |
|---|---|---|
| $w_1$ = <{A,C}{B,C}> | | |
| $w_2$ = <{C,D} {H} > | | |
| $w_3$ = <{B}{A}> | | |

### Esercizio 2 Itemset Frequenti (**10 punti**)

Considerare I seguenti 5 prodotti alimentari: {pane, pasta, latte, yogurt, aranciata} e la seguente tabella di transazioni:

| ID | PRODOTTI ACQUISTATI |
|---|---|
| 1 | pane , aranciata |
| 2 | pasta , aranciata , latte |
| 3 | pane , pasta , aranciata , yogurt , latte |
| 4 | aranciata , pane, pasta |
| 5 | pasta , latte , yogurt , aranciata |
| 6 | latte , aranciata , pasta |
| 7 | latte |
| 8 | yogurt , latte , pasta , aranciata |
| 9 | pasta , latte |
| 10 | yogurt , pasta , latte , pane |

A) Eseguire l'algoritmo *Apriori* per l'estrazione di itemset frequenti con min_sup = 50%, mostrando le varie fasi dell'algoritmo.
B) Determinare le regole associative con confidenza minima 80% formate da almeno tre item.
C) Indicare quali sono gli itemset frequenti massimali.

## Esercizio 3 Proprietà itemset frequenti (**4 punti**)

Siano "a", "b", "c" e "d" . Dire quali delle seguenti affermazioni è sempre vera:

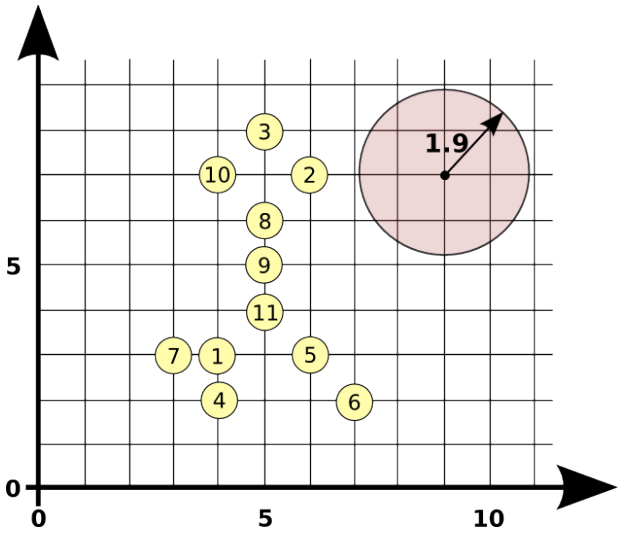
A) confidenza(a,b=>c) ≥ confidenza(a => c)
B) confidenza(a,b=>c) > confidenza(a,c => b)
C) se supporto(a) > supporto(b), allora: confidenza(a=>b) < confidenza(b => a)
D) se supporto(a) =1, allora confidenza(a,b=>c) = confidenza(b => c)

## 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 4 passi, quanti e quali cluster si ottengono? **(4 punti)**



Esercizio 5 Classificazione**(17 punti)**

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

| EtaBin | Esperienza | Bomber | Assistman | Aggressivo tipo | CLASSE |
|---|---|---|---|---|---|
| Vecchio | No | Si | Si | Medio | A |
| Vecchio | No | No | Si | Basso | B |
| Vecchio | Si | No | No | Basso | B |
| Vecchio | No | Si | Si | Basso | A |
| Giovane | No | Si | No | Medio | A |
| Giovane | Si | No | No | Basso | A |
| Vecchio | Si | Si | No | Basso | A |
| Vecchio | No | No | No | Alto | B |
| Vecchio | No | No | No | Alto | B |
| Vecchio | Si | No | No | Alto | A |
| Glovane | No | No | Si | Basso | B |
| Giovane | Si | Si | No | Basso | A |
| Vecchio | Si | Si | No | Basso | A |

A) Si costruisca su tale dataset un albero di decisione per la variabile CLASSE , utilizzando il criterio di split basato su misclassification rate , espandendo i nodi dell'albero fino a che la precisione non è più migliorabile. **(10 punti)**
B) Calcolare la matrice di confusione dell'albero ottenuto al punto A), sia sul training set che sul test set riportato qui sotto. Confrontare le due matrici e commentare il risultato. **(7 punti)**

| EtaBin | Esperienza | Bomber | Assistman | Aggressivo tipo | CLASSE |
|---|---|---|---|---|---|
| Vecchio | No | Si | No | Alto | B |
| Vecchio | Si | Si | Si | Medio | A |
| Glovane | Si | No | No | Alto | A |
| Vecchio | Si | No | No | Medio | B |
| Giovane | No | Si | No | Medio | B |
| Vecchio | No | No | Si | Basso | A |
| Vecchio | No | Si | Si | Basso | A |
| Glovane | No | No | No | Alto | B |
| Giovane | Si | Si | No | Basso | A |
| Vecchio | No | No | Si | Alto | B |

Esercizio 6 Classificazione**(15 punti)**

Si consideri il seguente insieme di transazioni con attributi sia discreti che continui:

| A | B | Class |
|---|---|---|
| Si | 15 | 1 |
| No | 15 | 1 |
| Si | 19 | 0 |
| No | 45 | 1 |
| No | 79 | 1 |
| Si | 91 | 0 |
| Si | 97 | 0 |
| No | 126 | 0 |

A) Si costruisca un albero di decisione per la variabile target Class , terminando la costruzione quando la precisione dell'albero non è più migliorabile. **(10 punti)**
B) Prima ancora di costruire l'albero di decisione, era possibile capire che l'albero risultante avrebbe avuto precisione pari al 100%. Come e perché? **(2 punti)**
C) Si calcoli la matrice di confusione dell'albero di decisione sul seguente test set. **(3 punti)**:

| A | B | Class |
|---|---|---|
| Si | 2 | 1 |
| No | 90 | 0 |
| No | 13 | 0 |
| Si | 50 | 1 |
| Si | 79 | 1 |
| Si | 30 | 0 |
| No | 45 | 1 |
| No | 300 | 0 |