| 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 30 giugno 2009

## Esercizio 1 - Sequential Patterns (6 punti)

Si consideri la seguente sequenza *W* di input:

$$W = <\ \{A,B\}\ \{A,C\}\ \{D,E\}\ \{B, C\}\ \{E\}\ \{H\}\ \{A\}\ >$$

Si indichi quali delle seguenti sequenze sono sotto-sequenze semplici di *W* (senza vincoli temporali), e quali rispettano anche il vincolo temporale *max-span = 3*:

| | $w_i \leq W$ | *max-span=3* |
|---|---|---|
| $w_1$ = <{A,C}{B,C}> | | |
| $w_2$ = <{C,D} {H} > | | |
| $w_3$ = <{B}{A}> | | |
| $w_4$ = <{A}{B}{H}> | | |
| $w_5$=<{B}{E}{A}> | | |

## Esercizio 2 Regole associative (10 punti)

Si consideri il seguente insieme di transazioni:

| Transazioni | Item Acquistati |
|---|---|
| 1 | {A,C} |
| 2 | {A,C,D} |
| 3 | {C,D} |
| 4 | {A,B,E} |
| 5 | {B,D,E,F} |
| 6 | {B,E,F} |
| 7 | {C,D,F} |
| 8 | {A,D,F} |
| 9 | {B,D,E,F} |
| 10 | {B,D,E} |

A) ) Eseguire l'algoritmo *Apriori* per l'estrazione di itemset frequenti con min_sup = 30%, mostrando le varie fasi dell'algoritmo. **(8 punti)**

B) Quali sono gli itemset frequenti massimali? Quali sono gli itemset frequenti *closed*, ovvero tali per cui ogni loro sovra-insieme ha supporto strettamente minore?
(Formalmente: closed(I) = $\forall J.\ J \supset I \Rightarrow supp(J) < supp(I)$ ) **(2 punti)**

## Esercizio 3 Regole associative (4 punti)

Quali delle seguenti affermazioni sono vere? Giustificare la risposta.

A) Se $\{e_1, e_2\}$ e $\{e_3, e_4\}$ sono frequenti, allora anche $\{e_1, e_2, e_3, e_4\}$ lo è.
B) La confidenza di $\{e_1, e_2\} \Rightarrow \{e_3, e_4\}$ è minore di quella di $\{e_1, e_2\} \Rightarrow \{e_4\}$
C) Il supporto dell'itemset vuoto, denotato $\{\}$ è uguale a zero
D) La confidenza di $\{\} \rightarrow \{e_1, e_2\}$ è uguale a quella di $\{e_1\} \rightarrow \{e_2\}$

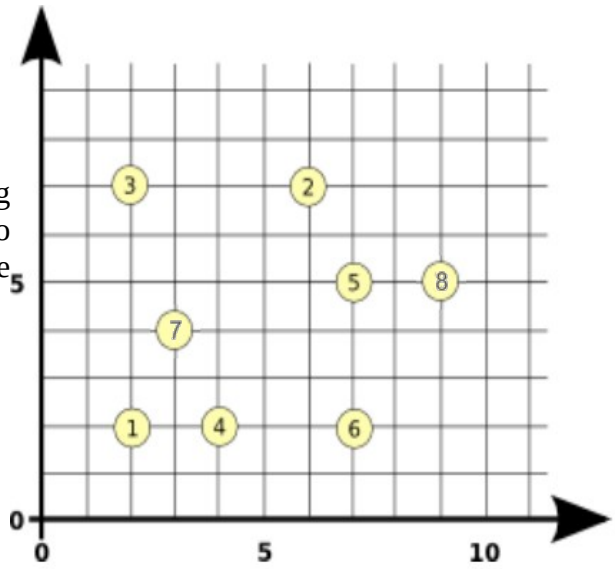
## Esercizio 4 - Clustering **(12 punti)**

Nel seguente dataset:

A) si calcoli la matrice delle distanze euclidee.
**(2 punti)**

B) Si esegua passo l'algoritmo di clustering gerarchico agglomerativo Min-link, mostrando ad ogni passo le matrici delle distanze aggiornate ai cluster attuali.
**(7 punti)**

C) Tracciare il corrispondente dendogramma.
**(3 punti)**



Esercizio 5 Classificazione**(17 punti)**

Si consideri il seguente insieme di transazioni.

| Temperature | Outlook | Humidity | Wind | OUTCOME |
|---|---|---|---|---|
| Cool | Rain | High | Strong | N |
| Hot | Sunny | High | Strong | N |
| Hot | Sunny | High | Light | N |
| Mild | Rain | High | Strong | N |
| Mild | Sunny | High | Light | N |
| Cool | Overcast | Normal | Strong | S |
| Cool | Rain | Normal | Light | S |
| Cool | Sunny | Normal | Light | S |
| Hot | Overcast | Normal | Light | S |
| Hot | Overcast | High | Light | S |
| Mild | Overcast | High | Strong | S |
| Mild | Rain | Normal | Light | S |
| Mild | Rain | High | Light | S |
| Mild | Sunny | Normal | Strong | S |

A) Si costruisca su tale dataset un albero di decisione per la variabile target "OUTCOME" selezionando ad ogni nodo la variabile di split in base al criterio di splitting l'*indice di Gini*. Indicarne l accuratezza.**(10 punti)**

B) Potare dall'albero tutti i nodi a profondità maggiore di 2, facendo diventare foglie i nodi che si trovano a profondità due. Nota: una profondità pari a 2 corrisponde a 2 split consecutivi. In riferimento al seguente test-set, si calcoli la matrice di confusione, sia per l'albero completo che per quello semplificato . Confrontare le due matrici e commentare il risultato.**(7 punti)**

| Temperature | Outlook | Humidity | Wind | OUTCOME |
|---|---|---|---|---|
| Cool | Overcast | High | Strong | S |
| Cool | Rain | High | Strong | N |
| Cool | Rain | Normal | Light | S |
| Cool | Sunny | Normal | Light | S |
| Hot | Overcast | High | Light | S |
| Hot | Overcast | Normal | Light | N |
| Hot | Sunny | High | Light | N |
| Hot | Sunny | High | Strong | S |
| Mild | Rain | High | Strong | N |
| Mild | Sunny | High | Light | N |

Esercizio 6 Classificazione**(15 punti)**

Si consideri il seguente insieme di transazioni con attributi continui (eccetto la classe Difettosi ):

A) Si costruisca un albero di decisione, utilizzando il *misclassification rate* come misura di correttezza, e non eseguendo split oltre profondità due. Nota: come nell'esercizio sopra, questo significa the ogni percorso radice-foglia dell'albero contiene non più di due split. **(6 punti)**
B) Supponiamo di applicare le seguenti strategie di discretizzazione agli attributi continui del data set:
   B1: partiziona il range di ogni attributo continuo in 4 intervalli della stessa ampiezza (*equal-size bin*)
   B2: partiziona il range di ogni attributo continuo in 4 intervalli con lo stesso numero di transazioni (*natural distribution bin*)
C) Per ogni strategia costruire una versione binarizzata del data set
D) Costruire quindi un albero di decisione sul nuovo dataset ottenuto al punto B1, anch'esso basato su misclassification rate. Confrontare l'accuratezza dei due alberi di decisione ottenuti. **(6 punti)**