Uppsala University
Department of Information Technology
Kjell Orsborn

Errata for the book: Introduction to Data Mining, by Tan, Steinbach and Kumar, Pearson, 2006.

2009-09-30

Errors found beyond the official errata list:

1. Page 218, 2nd paragraph from bottom:

   Reads: The FOIL's information gain for rules r_1 and r_2 given in the preceding example are 43.12 and 2, respectively.

   Should read: The FOIL's information gain for rules r_1 and r_2 given in the preceding example are 63.87 and 2.83, respectively.

2. Page 221, Figure 5.5:

   Reads: r3: (P=Yes,Q=No) ==> +

   Should read: r3: (P=Yes,R=No) ==> +

   Comment: The third rule in the right-hand side of Figure 5.5 corresponds to following the P_{right} and R_{left} branch (i.e. $P \rightarrow R \rightarrow +$) of the tree in the figure.

3. Page 332, last paragraph:

   Reads: Such an approach can be very expensive because it requires $O(Nmw)$ comparisons, where N is the number of transactions, $M = 2^k - 1$ is the number of candidate itemsets, and w is the maximum transaction width.

   Should read: Such an approach can be very expensive because it requires $O(Nm\frac{w^2}{2})$ comparisons, where N is the number of transactions, $M = 2^k - 1$ is the number of candidate itemsets, and w is the maximum transaction width.

   Comment: The worst case scenario should at least include $w * w/2$, taking that w is the maximum width of transactions and $w/2$ is the average width of the candidates as well. One could probably restrict this further to $w/2 * w/2$ since we would find matching items in $w/2$ comparisons on average.

4. Page 432, after 1st paragraph:

Reads:

1-sequences: $< i_1 >, < i_2 >, \cdots, < i_n >$

2-sequences: $< \{i_1, i_2\} >, < \{i_1, i_3\} >, \cdots, < \{i_{n-1}, i_n\} >$
$< \{i_1\}, \{i_1\} >, < \{i_1\}, \{i_2\} >, \cdots, < \{i_{n-1}\}, \{i_n\} >$

3-sequences: $< \{i_1, i_2, i_3\} >, < \{i_1, i_2, i_4\} >, \cdots, < \{i_1, i_2\}\{i_1\} >, \cdots,$
$< \{i_1\}, \{i_1, i_2\} >, \cdots, < \{i_1\}\{i_1\}\{i_1\} >, \cdots, < \{i_n\}\{i_n\}\{i_n\} >$

Should read:

1-sequences: $< i_1 >, < i_2 >, \cdots, < i_n >$

2-sequences: $< \{i_1, i_2\} >, < \{i_1, i_3\} >, \cdots, < \{i_1, i_n\} >,$
$< \{i_2, i_3\} >, < \{i_2, i_4\} >, \cdots, < \{i_2, i_n\} >,$
$< \cdots >, < \cdots >, \cdots, < \cdots >,$
$< \{i_{n-1}, i_n\} >$

$< \{i_1\}, \{i_1\} >, < \{i_1\}, \{i_2\} >, \cdots, < \{i_1\}, \{i_n\} >,$
$< \{i_2\}, \{i_1\} >, < \{i_2\}, \{i_2\} >, \cdots, < \{i_2\}, \{i_n\} >,$
$< \cdots >, < \cdots >, \cdots, < \cdots >,$
$< \{i_n\}, \{i_1\} >, < \{i_n\}, \{i_2\} >, \cdots, < \{i_n\}, \{i_n\} >$

3-sequences: $< \{i_1\}\{i_1\}\{i_1\} >, < \{i_1\}\{i_1\}\{i_2\} >, \cdots, < \{i_1\}\{i_1\}\{i_n\} >, \cdots,$
$< \{i_1\}\{i_2\}\{i_1\} >, < \{i_1\}\{i_2\}\{i_2\} >, \cdots, < \{i_1\}\{i_2\}\{i_n\} >, \cdots,$
$< \{i_1\}\{i_n\}\{i_1\} >, < \{i_1\}\{i_n\}\{i_2\} >, \cdots, < \{i_1\}\{i_n\}\{i_n\} >, \cdots,$
$< \{i_2\}\{i_1\}\{i_1\} >, < \{i_2\}\{i_1\}\{i_2\} >, \cdots, < \{i_2\}\{i_1\}\{i_n\} >, \cdots,$
$< \{i_2\}\{i_2\}\{i_1\} >, < \{i_2\}\{i_2\}\{i_2\} >, \cdots, < \{i_2\}\{i_2\}\{i_n\} >, \cdots,$
$< \{i_2\}\{i_n\}\{i_1\} >, < \{i_2\}\{i_n\}\{i_2\} >, \cdots, < \{i_2\}\{i_n\}\{i_n\} >, \cdots,$
$< \{i_n\}\{i_1\}\{i_1\} >, < \{i_n\}\{i_1\}\{i_2\} >, \cdots, < \{i_n\}\{i_1\}\{i_n\} >, \cdots,$
$< \{i_n\}\{i_2\}\{i_1\} >, < \{i_n\}\{i_2\}\{i_2\} >, \cdots, < \{i_n\}\{i_2\}\{i_n\} >, \cdots,$
$< \{i_n\}\{i_n\}\{i_1\} >, < \{i_n\}\{i_n\}\{i_2\} >, \cdots, < \{i_n\}\{i_n\}\{i_n\} >, \cdots,$

$< \{i_1\}, \{i_1, i_2\} >, < \{i_1\}, \{i_1, i_3\} >, \cdots, < \{i_1\}, \{i_1, i_n\} >, \cdots,$
$< \{i_1\}, \{i_2, i_3\} >, < \{i_1\}, \{i_2, i_4\} >, \cdots, < \{i_1\}, \{i_2, i_n\} >, \cdots,$
$< \{i_1\}, \{i_{n-1}, i_n\} >, \cdots,$

$< \{i_2\}, \{i_1, i_2\} >, < \{i_2\}, \{i_1, i_3\} >, \cdots, < \{i_2\}, \{i_1, i_n\} >, \cdots,$
$< \{i_2\}, \{i_2, i_3\} >, < \{i_2\}, \{i_2, i_4\} >, \cdots, < \{i_2\}, \{i_2, i_n\} >, \cdots,$
$< \{i_2\}, \{i_{n-1}, i_n\} >, \cdots,$

$< \{i_n\}, \{i_1, i_2\} >, < \{i_n\}, \{i_1, i_3\} >, \cdots, < \{i_n\}, \{i_1, i_n\} >, \cdots,$
$< \{i_n\}, \{i_2, i_3\} >, < \{i_n\}, \{i_2, i_4\} >, \cdots, < \{i_n\}, \{i_2, i_n\} >, \cdots,$
$< \{i_n\}, \{i_{n-1}, i_n\} >, \cdots,$

$$< \{i_1, i_2\}, \{i_1\} >, < \{i_1, i_3\}, \{i_1\} >, \cdots, < \{i_1, i_n\}, \{i_1\} >, \cdots,$$
$$< \{i_2, i_3\}, \{i_1\} >, < \{i_2, i_4\}, \{i_1\} >, \cdots, < \{i_2, i_n\}, \{i_1\} >, \cdots,$$
$$< \{i_{n-1}, i_n\}, \{i_1\} >, \cdots,$$

$$< \{i_1, i_2\}, \{i_2\} >, < \{i_1, i_3\}, \{i_2\} >, \cdots, < \{i_1, i_n\}, \{i_2\} >, \cdots,$$
$$< \{i_2, i_3\}, \{i_2\} >, < \{i_2, i_4\}, \{i_2\} >, \cdots, < \{i_2, i_n\}, \{i_2\} >, \cdots,$$
$$< \{i_{n-1}, i_n\}, \{i_2\} >, \cdots,$$

$$< \{i_1, i_2\}, \{i_n\} >, < \{i_1, i_3\}, \{i_n\} >, \cdots, < \{i_1, i_n\}, \{i_n\} >, \cdots,$$
$$< \{i_2, i_3\}, \{i_n\} >, < \{i_2, i_4\}, \{i_n\} >, \cdots, < \{i_2, i_n\}, \{i_n\} >, \cdots,$$
$$< \{i_{n-1}, i_n\}, \{i_n\} >, \cdots,$$

$$< \{i_1, i_2, i_3\} >, < \{i_1, i_2, i_4\} >, \cdots, < \{i_1, i_2, i_n\} >, \cdots,$$
$$< \{i_2, i_3, i_4\} >, < \{i_2, i_3, i_5\} >, \cdots, < \{i_2, i_3, i_n\} >, \cdots,$$
$$< \{i_{n-2}, i_{n-1}, i_n\} >$$

5. Page 432, 3rd paragraph (item 1):

Reads: On the other hand, there are many candidate 2-sequences, such as $< \{i_1, i_2\} >, < \{i_1\}, \{i_2\} >, < \{i_2\}, \{i_1\} >$ and $< \{i_1, i_1\} >$, that can be generated.

Should read: On the other hand, there are many candidate 2-sequences, such as $< \{i_1, i_2\} >, < \{i_1\}, \{i_2\} >, < \{i_2\}, \{i_1\} >, < \{i_1\}, \{i_1\} >$ and $< \{i_2\}, \{i_2\} >$, that can be generated.