Errata

Preface

1. p. xxiv, paragraph 7, last line: Change “on data” to “in data”.

Chapter 1

1. Bibliography, p. 42, paragraph 1, line 11: Remove “with Java Implementations”.
2. Bibliography, p. 43, paragraph 3, line 9: Change “M. Ross” to “Ross”.
3. Bibliography, p. 43, paragraph 4, line 9: Change “Symposium” to “Conference” [from Peixiang Zhao].
4. Bibliography, p. 44, paragraph 2, last 2 lines: Change “Dob05” to “Dob01”. Change “JW05” to “JW02”.
5. Bibliography, p. 44, paragraph 2, line 4: Insert “and” in front of “Stork”.
6. Bibliography, p. 44, paragraph 5, line 3: Change “ML” to “ICML” [from Peixiang Zhao].

Chapter 2

1. p. 47, line 3 from the bottom, “eneration of concept hierarchies from...” should be “generation of concept hierarchies from...” [from Samsideen Olamide Mustapha]
2. p. 69, paragraph 5, line 2: Change “needed to” to “needs to” [from Peixiang Zhao].

Chapter 3

1. p. 119, starting at the 7th line from the bottom, change “For example, count() can be computed for a data cube by first partitioning the cube into a set of subcubes, computing count() for each subcube, and then summing up the counts obtained for each subcube. Hence, count() is a distributive aggregate function.” to “For example, sum() can be computed for a data cube by first partitioning the cube into a set of subcubes, computing sum() for each subcube, and then summing up the sums obtained for all of the subcubes. Hence, sum() is a distributive aggregate function.”

2. p. 119, 4th line from the bottom, “For the same reason, sum().” is changed to “For the same reason, count().”, and add footnote “2” after “count()” as follows:

Footnote 2: "By treating the count value of each nonempty base cell as 1 by default, count() of any cell in a cube can be viewed as the sum of the count values of all of its corresponding child cells in its subcube. Thus, count() is distributive."

3. p. 131, line 5: “OLEDB(Open Linking and Embedding for Databases)” is changed to “OLEDB(Object Linking and Embedding, Database)” [from Narendra Kumar]

4. p. 155, paragraph 3, line 13: Remove "bottom-up".

Chapter 4

1. p. 160, line 29 "minimum support(min_sup)" should be "minimum support (min_sup)" (Note: There is a space missing) [from Chris Ariagno].

2. p. 161 paragraph 4, Change the last sentence: "That is, out of 2^101-6 distinct aggregate cells, only 3 really offer new information." into "That is, out of 2^101-4 distinct base and aggregate cells, only three really offer new information." [from Rick English].

3. p. 167, bottom line: "relationa" should be changed to "relational" [from Chris Ariagno].

4. p. 167 parag. 3, line 5 "40 X 1000 (for one row of the AC plane)" should be changed to "10 X 4000 (for one row of the AC plane)" [from Desheng Xu].

5. p. 170, Figure 4.6 line (2) of the algorithm, “WriteAncestors(input[0], dim)” should be changed to "WriteDescendants(input[0], dim)” [Omar Khan]

6. p. 170, Figure 4.6 line (12) of the algorithm, “BUC(input[k…k+c], d+1)” should be changed to “BUC(input[k..k+c-1], d+1)” [Steve Leighton]

7. p. 172, line 2, “of the tuple's ancestor group-by's" should be changed to “of the tuple's descendant group-by's” [Omar Khan]

8. p. 177 ACD/A-tree of Figure 4.11, change "a_1CD/a_1:1" to "a_1CD/a_1:1_2" [from Ganesh Agarwal].

9. p. 177 Figure 4.11, change "ADB/AB-tree" to "ABD/AB-tree" [from Rick English].

10. p. 178 top right of Figure 4.12, change "1_8" to "2_7" [from Chris Ariagno].

11. p. 186, paragraph 3, line 9: Change “1-D fragments” to “2-D fragments” [from Peixiang Zhao].

12. p. 197, 2nd line from the bottom, “gradient_contraint_threshold” should be “gradient_constraint_threshold” [Chris Ham]

13. p. 209, equation 4.1: the denominator should be count(q_i)” [from Peixiang Zhao].

14. p. 210. Example 4.26, line 6: [t: 45, 00%] -> [t: 45.00%] [Chris Ham]

15. p. 215, paragraph 4, line 2: change “Rule (4.6)” to “Rule (4.5)” [from Peixiang Zhao].
16. p. 217, paragraph below (4.8), 2nd line: "the conditions are ORed to *from* a disjunct" -> "form" [Chris Ham]

17. p. 220, Exercise 4.1(d): Replace bold italic font for cells d and c with italic font (8 occurrences).

18. p. 222, Exercise 4.11(a) part (ii) of Output, last 2 lines (in parentheses): Change “this” to “This” and add a period after “results”.

Chapter 5

1. p. 245, line 1: Change “I1” to “I4” [from Jeff Huang].

2. p. 245, line 1: remove the sentence "Notice that although I5 follows I4 in the first branch, there is no need to include I5 in the analysis here because any frequent patterns involving I5 is analyzed in the examination of I5" [from Omar Khan].

3. p. 257, 2nd parag., 4th line from bottom: "... no greater *that* sup" -> "than" [Chris Ham]

4. p. 258 line 7, "area" should be "are" [from Chris Ariagno]

5. p. 270 line 5 from the bottom, “Specifically, such a set must consist of a nonempty set of items whose price is no less than $500. It is of the form $S$, where $S \neq \Phi$ is a subset of the set of all those items with prices no less than $500$, and $S_2$, possibly empty, is a subset of the set of all those items with prices no greater than $500$.” should be changed to “Specifically, such a set must be a subset of the set of all those items whose price is no less than $500$. It is of the form $S$, where $S \neq \Phi$ is a subset of the set of all those items with prices no less than $500$.” [from Chris Ariagno]

6. p. 278, Exercise 5.12(b), line 1: Remove “FP-tree”. Insert “FP-tree-based” after "proposed".

Chapter 6

1. p. 286, first line of paragraph 3, "How does classification work? Should be changed to "How does classification work?" Note: There is an open quote but no close quote. [from Chris Ariagno]

   a. p. 297, line 1 below equation (6.1), “where $p_i$ is the probability that …” should be changed to “where $p_i$ is the non-zero probability that …” [Clodoveu Davis]

   b. p. 299, line 4 below Table 6.1, “+ 4/14 X (− 4/4 log_2 4/4 − 0/4 log_2 0/4)” should be changed to “+ 4/14 X (− 4/4 log_2 4/4)” [Clodoveu Davis]

   c. p. 301, Example 6.2, 3rd to last line: Change 0.926 to 1.557 for SplitInfo_A(D) [from Tianyi Wu].

   d. p. 301, Example 6.2, last line: Change “0.029/0.926=0.031” to “0.029/1.557=0.019” [from Tianyi Wu].

   e. p. 301, Equation (6.6): change “SplitInfo(A)” to “SplitInfo_A(D)” [from Peixiang
Zhao].

f. p. 301, Example 6.2, line 5: Change “SplitInfo_A(D)” to “SplitInfo_{income}(D)” [from Peixiang Zhao].

g. p. 303, Example 6.3, 3rd and 4th lines for calculations of \( Gini_{income \{low,medium\}}(D) \) should be [from Marcel Bieler and Tianyi Wu]:

\[
\frac{10}{14}(1 - \frac{7}{10}^2 - \frac{3}{10}^2) + \frac{4}{14}(1 - \frac{2}{4}^2 - \frac{2}{4}^2) = 0.443
\]

h. p. 303 Example 6.3, text following calculations for \( Gini_{income \{low,medium\}}(D) \) should be [from Marcel Bieler and Tianyi Wu]:

2. Similarly, the Gini index values for splits on the remaining subsets are 0.458 (for the subsets \{low, high\} and \{medium\}) and 0.450 (for the subsets \{medium, high\} and \{low\}). Therefore, the best binary split for attribute income is on \{low, medium\} (or \{high\}) because it minimizes the Gini index. Evaluating age, we obtain \{youth, senior\} (or \{middle_aged\}) as the best split for age with a Gini index of 0.357; the attributes student and credit_rating are both binary, with Gini index values of 0.367 and 0.429, respectively.

3. The attribute age and splitting subset \{youth, senior\} therefore give the minimum Gini index overall, with a reduction in impurity of 0.459 – 0.357 = 0.102. The binary split “age \{youth, senior\}? results in the maximum reduction in impurity of the tuples in D and is returned as the splitting criterion. Node N is labeled with the criterion, two branches are grown from it, and the tuples are partitioned accordingly. [Authors' note: For the expression, “age \{youth, senior\}?” use the mathematical symbol for “element of” (not available here) in place of “IN”.

4. p. 303, last sentence in Example 6.3: Remove this sentence, which begins “Hence, ...”.

5. p. 313, Example 6.4, the 9th line from the bottom: “PX|Ci)” should be “P(X|Ci)” [from Ziang Song]

6. p. 315, Example 6.5, line 5: Change “999/1000” to ”990/1000)” [from Marcel Bieler]

7. p. 322, last paragraph, line 3: Change “C” to “C_i” [from Peixiang Zhao].

8. p. 330, Figure 6.16 the ‘’ at the end of the (3) of Method should be removed. [from Ziang Song]

9. p. 339, “(a)” and “(b)” should be added to the two diagrams as captions [from Peixiang Zhao].

10. p. 345, line 2 from the bottom of the page, “CBA (Classification-Based Association)” should be changed to CBA (Classification-Based on Associations) [from Khurram Shehzad, Cardiff University]

11. p. 349, line 10, “... computed for attributes that not numeric,” should be changed to “... computed for attributes that are not numeric,” [from Chris Ariagno]
12. p. 339, paragraph 5, line 8 change "then twice" to "than twice" [from Peixiang Zhao].

13. p. 357, last line before section 6.11.2, "(see references above.)" should be changed to "(see references above)." [from Chris Ariagno]

14. p. 360, Figure 6.27, the total recognition rate should be changed from 95.52% to 95.42%. [from Wai-Shing Ho (Hong Kong Univ.)]

15. p. 360, line 2 of Figure 6.27 description, "an entry is row i and column j" should be "an entry in row i and column j" [from Chris Ariagno]

16. p. 360, last line (footnote), "negatives" should be "negative" [from Chris Ariagno]

17. p. 360, line 8 from the bottom, "... with the rest of the entries being close to zero" should be "... with the rest of the entries being zero or close to zero" [from Chris Ariagno]

18. p. 361, the first line after equation (6.57), "where ..." should be changed to "where ..." (note: remove the extra space) [from Chris Ariagno]

19. p. 362, line 2 of section 6.12.2, "($\begin{bmatrix} X \end{bmatrix}^2, y_2$)" should be "($\begin{bmatrix} X \end{bmatrix}^2, y_2$)" (i.e., there is a space in front of $y_2$ [from Chris Ariagno]

20. p. 363, the first line after equation (6.64), the question "$\bar{y} = \frac{\sum_{i=1}^{t}y_i}{d}$" should be "$\bar{y} = \frac{\sum_{i=1}^{t}y_i}{d}$". [from Chris Ariagno]

21. p. 365, last line of second paragraph in section 6.13.3 "632 bootstrap." should be "632 bootstrap." [from Chris Ariagno]

22. p. 365, equation (6.65), add 1/k in front of $\Sigma_{i=1}^{k}$. [from Wai-Shing Ho (Hong Kong Univ.)]

23. p. 368, equation (6.66), "summation from j to d" should be "summation from j=1 to d" [from Chris Ariagno]

24. p. 372, line 6 from the bottom "... curve cases off ..." should be "... curve eases off ..." [from Chris Ariagno]

25. p. 373, line 2 from the bottom "... Bayes, theo-" should be "... Bayes' theo-" [from Chris Ariagno]

26. p. 374, line 12 "... data in a higher dimension ..." should be "... data into a higher dimension ..." [from Chris Ariagno]

27. p. 381, 7th line from bottom: Change "texts" to "texts by".

Chapter 7

1. p. 392, the first line after equation (7.12), "where ..." should be changed to "where ..." (note: remove the extra space) [from Chris Ariagno]

2. p. 411 line 12, "Farthest-neighbor algorithms tend to minimize the increase in diameter of the clusters at each iteration as little as possible." should be changed to "Farthest-neighbor algorithms tend to minimize the increase in diameter of the clusters at each
iteration." [from Chris Ariagno].

3. p. 414, 2nd paragraph, line 6, “if the size of the memory that is needed for storing the CF tree is larger than the size of the main memory, then a smaller threshold can be specified and the CF tree is rebuilt.” should be changed to “if the size of the memory that is needed for storing the CF tree is larger than the size of the main memory, then a larger threshold can be specified and the CF tree is rebuilt.” [from Wubulikasimu Aisikaer, Linköpings universitet, Sweden].

4. p. 418, line 5 from the bottom, “for $1 \le i \le n$” should be changed to “for $1 \le i < n$”. [from Chris Ariagno].

5. p. 423, 2nd to last paragraph, lines 4-5: change $0 < i < k$ to $0 < i <= k$. [from Chris Ariagno].

6. p. 424, line 14, "dimensional" should be "dimensional" [from Chris Ariagno].

7. p. 444 line 7 from the bottom, there is a space missing between of and "knowledge" [from Chris Ariagno].

8. p. 454 footnote 12, at the end of line 2, “$P(3 – d_{min} \le x \le 3 + d_{min}) < – p_{ct}$” should be “$P(3 – d_{min} \le x \le 3 + d_{min}) < 1 – p_{ct}$” [from Chris Ariagno].

9. p. 457, the last paragraph, line 1, “If an object p is not a local outlier, LOF(p) is close to 1.” should be “If an object p is not a local outlier, LOF(p) is close to 0.” [from Allison N. Tegge]

10. p. 459. The recurring header "7.12 Outlier Analysis" should be changed to "7.11 Outlier Analysis" [from Allison N. Tegge]

11. p. 461 for the paragraph beginning "A constraint-based clustering method", eliminate the sentence, "For example, clustering with the existence of obstacle objects and clustering under user-specified constraints are typical methods of constraint-based clustering." [from Chris Ariagno].

12. p. 462, Exercise 7.3(c): Replace "q = 3" by "p = 3".

13. p. 462, Exercise 7.6 (a): Insert “of” in front of “execution”.

14. p. 462, Exercise 7.7, lines 1 and 3: Change “illustrate” to “summarize”.

15. p. 462, Exercise 7.10: Change “Given” to “Give”. Change “application examples” to “sample data sets”.


Chapter 8

1. p. 502, 2nd paragraph from bottom: Change “Note than” to “Note that” [from Maryam Karimzadehgan]

2. p. 503 (the bottom line) and p. 504 (line 1): Change “then compresses this information
into a frequent-pattern tree, or FP-tree. The FP-tree is used to generate to “then generates”.

3. p. 506, Table 8.2, third entry of the projected database of prefix <a> shall be changed from "<(b)(df)eb>" to "<(b)(df)cb>" [from Selim Mimaroglu@cs.umb.edu].

4. p. 507, line 6: In the term "<aa>{<(bc)(ac)d(cf)}>{<(e)}>", remove the 2nd "{" because it should be "<aa>{<(bc)(ac)d(cf)>,<(e)>}" [from Tianyi Wu].

5. p. 511, Example 8.13, line 7: Change “one or a set of events C” to “zero or more occurrences of event C” [from Govind Kabra].

6. p. 525, Equation (8.15): Change v_(i)(k) to v_k(i-1) [from Jianlin Feng].

Chapter 9

1. p. 573, Example 9.8. line 1, “Loan(L, _, _, _, payment >= 12, _)” should be “Loan(L, _, _, duration >= 12, _, _)”. [Jing Li]

2. p587, Exercise 9.10, 2nd sentence: Change to “For example, a student could form part of a class, a research project group, a family, a neighborhood, and so on.”

3. Bibliography, p. 589, line 2: Change "MMR05" to "MMR+05".

Chapter 10

1. p. 619, Example 10.9, the end-of-example blackbox should be moved two lines down to the end of the TF-IDF(d_4, t_6) equation.

2. p. 642, 2nd para, line 5: Change “three based” to “three popular” [from Govind Kabra].

3. p. 645, Exercise 10.15(b), 2nd line: Change “algorithms” to “algorithm”.

4. p. 646, paragraph 3, end of line 4 from bottom: Change “have” to “has”.

5. p. 647, paragraph 1, last line: Change “performed” to “presented” (since subject is “overview”).

6. p. 647, paragraph 2, lines 13 and 14: Change “Method” to “Methods” in line 13. Change “has been” to “have been” in line 14.

7. p. 648, paragraph 1, last line: remove one “the” in “the the”.

Chapter 11


http://www.cs.uiuc.edu/homes/hanj/bk2/errataindex.htm

Bibliography

1. p. 706, [BB02], Change “Mining molecular fragments: Finging relevant substructures of molecules” to "Mining molecular fragments: Finding relevant substructures of molecules". [from Desheng Xu].

2. p. 729: change "MMR04" to "MMR+05".

Back to the Home Page of Data Mining: Concepts and Techniques, 2nd ed.
Back to Jiawei Han's Home Page

Back to the Home Page of Data and Information Systems Research Laboratory, Computer Science, University of Illinois at Urbana-Champaign

Page maintained by: Jiawei Han
Last update: July 8, 2007