Errata for *Introduction to Data Mining* by Tan, Steinbach, and Kumar.

Last updated on April 2, 2007 at 12:41pm

Please send all error reports to dmbook@cs.umn.edu

Preface

Page x, last sentence of first paragraph: The email address for reporting errata has an error. Please use the one given above.

Chapter 2

1. Page 23: The title “What Is an attribute?” should be “What is an Attribute?”.
2. Page 60, equation in the last paragraph: “$e_i = \sum_{i=1}^{k} p_{ij} \log_2 p_{ij}$” should be “$e_i = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$”.
3. Page 69, fourth line from bottom: “of x and y” should be “of $\mathbf{x}$ and $\mathbf{y}$”.
4. Page 70, second line from bottom: “$d(\mathbf{x}, \mathbf{x}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$” should be “$d(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$”.
5. Page 75, second equation before the last paragraph: $||\mathbf{y}||$ should be 2.45, not 2.24.
6. Page 78, last sentence of the first paragraph: “$x_k = y_k^2$” should be “$y_k = x_k^2$”.
7. Page 91, Exercise 14: “what sort of similarity measure” should be “what sort of proximity measure”.

Chapter 3

1. Page 100, Table 3.1: The number of freshman should be 200 and the number of seniors should be 110, as shown in Table 1.
2. Page 112, Figure 3.5: For stem 4, there should only be 4 occurrences of leaf 4.
3. Page 126, Example 3.21: “Figure 3.25 is another parallel coordinates plot of the same data,” should be “Figure 3.26 is another parallel coordinates plot of the same data,”

Table 1. Class size for students in a hypothetical college.

Class	Size	Frequency
freshman	200	0.33
sophomore	160	0.27
junior	130	0.22
senior	110	0.18

4. Page 132, Figure 3.30: The vertical axis should be labeled "Petal Length" not "Petal Width".

Chapter 4

1. Page 160, second line from the bottom of the second paragraph from the bottom: "the Gini index for attribute B is 0.375" should be "'the Gini index for attribute B is 0.371".

2. Page 161, Figure 4.14, bottom right table. "Gini = 0.375" should be "Gini = 0.371".

3. Page 173, second from bottom line: "Figure 4.23(b) shows the training and test error rates" should be "Figure 4.23 shows the training and test error rates".

4. Page 189, sixth from bottom line, the equation should be:

$$P(X = v) = \binom{N}{v} p^v (1-p)^{N-v}.$$

5. Page 192, Equation 4.17:

$$d_t^{cv} = 0.05 \pm 1.70 \times 0.002.$$

6. Page 193, Table 4.6. Column headings are given in Table 2.

7. Page 198, Exercise 3(a): "What is the entropy of this collection of training examples with respect to the positive class?" should be "What is the entropy of this collection of training examples with respect to the class attribute?".

8. Page 200, Exercise 5(c) both instances of "monotonously" should be "monotonically".

Table 2. Probability table for t-distribution.

	$(1-\alpha)$				
$k-1$	0.90	0.95	0.975	0.99	0.995
1	3.08	6.31	12.7	31.8	63.7
2	1.89	2.92	4.30	6.96	9.92
4	1.53	2.13	2.78	3.75	4.60
9	1.38	1.83	2.26	2.82	3.25
14	1.34	1.76	2.14	2.62	2.98
19	1.33	1.73	2.09	2.54	2.86
24	1.32	1.71	2.06	2.49	2.80
29	1.31	1.70	2.04	2.46	2.76

Chapter 5

1. Page 208, sixth from top line: "and *op* is a logical operator chosen" should be "and *op* is a comparison operator chosen".

2. Page 213, Algorithm 5.1 line 8: "$R \longrightarrow R \vee r$" should be "$R \leftarrow R \vee r$".

3. Page 218, tenth from bottom line: "rules r_1 and r_2 given in the preceding example are 43.12 and 2" should be "rules r_1 and r_2 given in the preceding example are 63.87 and 2.83".

4. Page 233, Equation 5.16 should be:

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left[- \frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right].$$

5. Page 264, sixth and seventh from bottom line, equations should be:

$$\begin{aligned} w_1 &= \sum_i \lambda_i y_i x_{i1} = 65.5261 \times 1 \times 0.3858 + 65.5261 \times -1 \times 0.4871 = -6.64. \\ w_2 &= \sum_i \lambda_i y_i x_{i2} = 65.5261 \times 1 \times 0.4687 + 65.5261 \times -1 \times 0.611 = -9.32. \end{aligned}$$

6. Page 271, Equation (5.55):

$$\Phi : (x_1, x_2) \longrightarrow (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, 1).$$

In the transformed space, we can find the parameters $\mathbf{w} = (w_0, w_1, \ldots, w_5)$ such that:

$$w_5x_1^2 + w_4x_2^2 + w_3\sqrt{2}x_1 + w_2\sqrt{2}x_2 + w_1\sqrt{2}x_1x_2 + w_0 = 0.$$

7. Page 271, tenth from bottom line: “all the circles are located in the lower right-hand side of the diagram” should be “all the circles are located in the lower left-hand side of the diagram”.

8. Page 273, second from top line: “instance z can be classified” should be “instance $\mathbf{z}$ can be classified”.

9. Page 273, Equation (5.60):

$$\begin{aligned}\Phi(\mathbf{u})\cdot\Phi(\mathbf{v}) &= (u_1^2, u_2^2, \sqrt{2}u_1, \sqrt{2}u_2, \sqrt{2}u_1u_2, 1)\cdot(v_1^2, v_2^2, \sqrt{2}v_1, \sqrt{2}v_2, \sqrt{2}v_1v_2, 1)\\ &= u_1^2v_1^2 + u_2^2v_2^2 + 2u_1v_1 + 2u_2v_2 + 2u_1u_2v_1v_2 + 1\\ &= (\mathbf{u}\cdot\mathbf{v}+1)^2.\end{aligned}$$

10. Page 274, second line in the second paragraph: “A test instance $\mathbf{x}$ is classified” should be “A test instance $\mathbf{z}$ is classified”.

11. Page 288, Equation 5.69 should be:

$$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} \times \begin{cases} e^{-\alpha_j} & \text{if } C_j(\mathbf{x_i}) = y_i\\ e^{\alpha_j} & \text{if } C_j(\mathbf{x_i}) \neq y_i \end{cases},$$

12. Page 315, Exercise 1(a): “exclustive” should be “exclusive”.

13. Page 317, Exercise 5(d) and 5(e): “examples covered by R1 are discarded)” should be “examples covered by R1 are discarded”.

14. Page 323, Exercise 17(c) and 17(d): “part (c)” should be “part (b)”.

Chapter 6

1. Page 331, Equation 6.3: This derivation assumes that neither of the itemsets of a rule are empty.

2. Page 331, Second line of paragraph 2: “From Equation 6.2” should be “From Equation 6.1”.

3. Page 356, caption in Figure 6.17: “(with minimum support count equal to 40%” should be “(with minimum support equals to 40%)”.

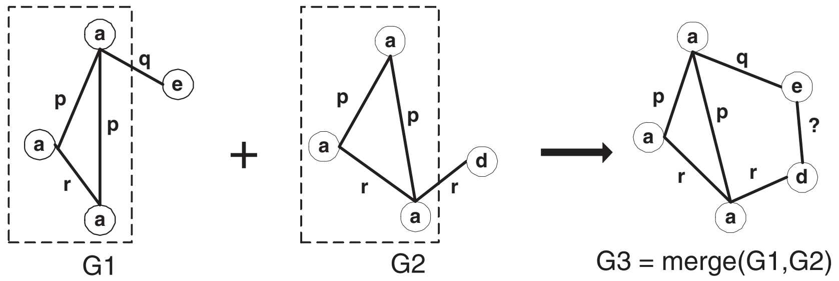
4. Page 382, last line of first paragraph: “but are invariant under the null addition operation” should be deleted. Many asymmetric measures are not null invariant.

5. Page 405, Exercise 4(a): “$c\big(\{p_k\} \longrightarrow \{p_1, p_3 \ldots, p_{k-1}\}\big)$” should be “$c\big(\{p_k\} \longrightarrow \{p_1, p_2 \ldots, p_{k-1}\}\big)$”

6. Page 405, Exercise 5: Assume that neither of the itemsets of a rule are empty.
7. Page 408, Exercise 9(b): "Use the visited leaf nodes in part (b)" should be "Use the visited leaf nodes in part (a)".
8. Page 410, Exercise 12, part (b), subpart ii: "Interest($X \longrightarrow Y$) $= \frac{P(X,Y)}{P(X)} P(Y)$" should be "Interest($X \longrightarrow Y$) $= \frac{P(X,Y)}{P(X)P(Y)}$".
9. Page 410, Exercise 12, part (b), subpart v: "Klosgen($X \longrightarrow Y$) $= \sqrt{P(X,Y)} \times (P(Y|X) - P(Y))$" should be "Klosgen($X \longrightarrow Y$) $= \sqrt{P(X,Y)} \times \max(P(Y|X) - P(Y), P(X|Y) - P(X))$".
10. Page 414, Exercise 11: The last sentence before part (a) should end with a period instead of a question mark. The sentence in part (e) should end with a question mark instead of a period.
11. Page 411, Exercise 15(b): "$P(A,B) \times P(A,\overline{B}) = P(A,\overline{B}) \times P(\overline{A},B)$" should be "$P(A,B) \times P(\overline{A},\overline{B}) = P(A,\overline{B}) \times P(\overline{A},B)$".
12. Page 413, Exercise 17: "If the support" should be "Assume the support".
13. Page 413, Exercise 17(d)(i): $c(\{\overline{a}\} \rightarrow \{b\}) > c(\{\overline{a}\} \rightarrow \{b\})$ should be $c(\{\overline{a}\} \rightarrow \{b\}) > c(\{a\} \rightarrow \{b\})$.

Chapter 7

1. Page 421, the rule "$R_{12}^{(4)}$: Age $\in [20, 24) \rightarrow$ Chat Online = No" should be "$R_{12}^{(4)}$: Age $\in [20, 24) \rightarrow$ Chat Online = Yes".
2. Page 433, list of potential candidate 2-sequences following the first paragraph: "$< \{i_{n-1}\}\{i_n\} >$" should be "$< \{i_n\}\{i_n\} >$".
3. Page 433, last sentence of item 1, which begins "An item can": "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 be "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".
4. Page 437, fourth from bottom line: "events in one element must occur immediately after the events" should be "events in one element must occur after the events".
5. Page 449, Figure 7.13 should be as shown in Figure 1.
6. Page 439, bottom line: "$maxgap = 1$" should be "$maxgap = 2$"



$$M_{G1} = \begin{pmatrix} 0 & p & p & q \\ p & 0 & r & 0 \\ p & r & 0 & 0 \\ q & 0 & 0 & 0 \end{pmatrix} \quad M_{G2} = \begin{pmatrix} 0 & p & p & 0 \\ p & 0 & r & 0 \\ p & r & 0 & r \\ 0 & 0 & r & 0 \end{pmatrix} \quad M_{G3} = \begin{pmatrix} 0 & p & p & q & 0 \\ p & 0 & r & 0 & 0 \\ p & r & 0 & 0 & r \\ q & 0 & 0 & 0 & ? \\ 0 & 0 & r & ? & 0 \end{pmatrix}$$

Figure 1. Vertex-growing strategy.

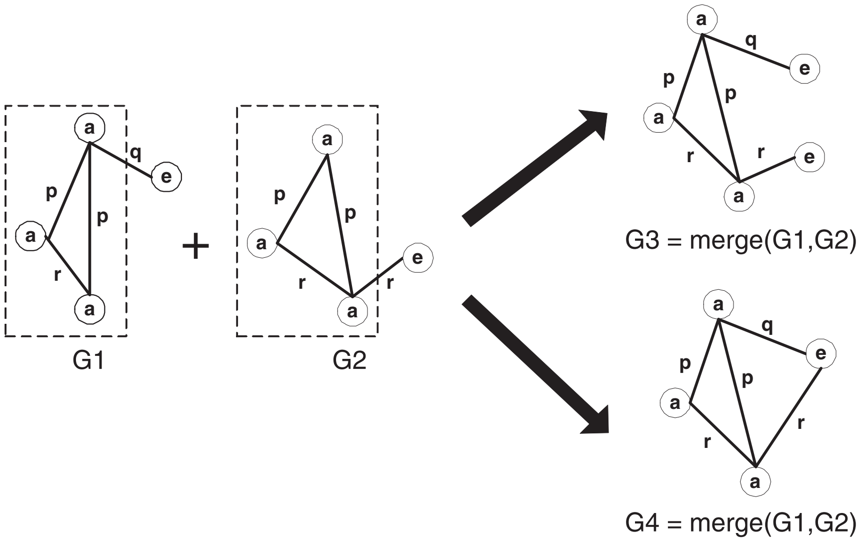


Figure 2. Edge-growing strategy.

7. Page 450, Figure 7.14 should be as shown in Figure 2.

8. Page 480, Exercise 12(c): "$w = \langle\{A\}\{B, C, D\}\{A\}\rangle$" should be "$w = \langle\{A\}\{A, B, C, D\}\{A\}\rangle$".

9. Page 483, Exercise 19(a): "join the two undirected and unweighted subgraphs shown in Figure 19a" should be "join the two undirected and unweighted subgraphs shown below".

Chapter 8

1. Page 519: The numbers in Tables 8.3 and 8.4 were rounded to two decimal places. Thus, if the x and y coordinates of the points given in Table 8.3 are used to compute the pairwise distances, the results don't quite match those shown in Table 8.4. The original, more precise values are given in Tables 3 and 4.

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table 3. X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table 4. Distance Matrix for Six Points

2. Page 517, the fifth line of the first paragraph: "see Section 8.1.2" should be "see Section 8.1.3".

3. Page 522, the fourth line from the bottom: "$dist(\{3,6,4\},\{2,5\}) = (0.15+0.28+0.25+0.39+0.20+0.29)/(6*2)$" should be "$dist(\{3,6,4\},\{2,5\}) = (0.15+0.28+0.25+0.39+0.20+0.29)/(3*2)$"

4. Page 529, last sentence to first sentence on 530. The terms, "low" and "high", are interchanged. "If the *Eps* threshold is low enough that DBSCAN finds C and D as clusters, then A and B and the points surrounding them will become a single cluster. If the *Eps* threshold is high enough that DBSCAN finds A and B as separate clusters, and the points surrounding them are marked as noise, then C and D and the points surrounding them will also be marked as noise." should be "For

a fixed $MinPts$, if the Eps threshold is high enough that DBSCAN finds C and D as clusters, then A and B and the points surrounding them will become a single cluster. If the Eps threshold is low enough that DBSCAN finds A and B as separate clusters, and the points surrounding them are marked as noise, then C and D and the points surrounding them will also be marked as noise."

5. Page 549, the third line of the paragraph with the heading, **Entropy**: "for cluster j we compute p_{ij}" should be "for cluster i we compute p_{ij}".

Chapter 9

1. Page 584, Equation 9.7: "x_i" should be "x".
2. Page 586, in Equations 9.9 and 9.10, as well as in the first line below Equation 9.10, u should be μ.
3. Page 596, the first line after Equation 9.16: "the difference, $\mathbf{p}(t) - \mathbf{m}_j(t)$, between the centroid, $\mathbf{m}_j(t)$, and the current object, $\mathbf{p}(t)$" should be "the difference, $\mathbf{p}(t) - \mathbf{m}_j(t)$, between the current object, $\mathbf{p}(t)$, and the centroid, $\mathbf{m}_j(t)$".
4. Page 605, Figure 9.11: "(c) View in the xy plane" should be "(c) View in the xz plane"; "(d) View in the xy plane" should be "(d) View in the yz plane".
5. Page 618, Equation 9.17: "$RC =$" should be "$RC(C_i, C_j) =$".
6. Page 619, Equation 9.18: "$RI =$" should be "$RI(C_i, C_j) =$".

   Page 637, the fourth line before Algorithm 9.14: "the total number of clusters is m/pq" should be "the total number of clusters is m/q".
7. Page 639, the first line: "Overall, m/pq clusters are produced" should be "Overall, m/q clusters are produced".
8. Page 639, the third line: "is not pq" should be "is not q".
9. Page 639, the fourth line: "m/pq of the intermediate clusters" should be "m/q of the intermediate clusters".

Chapter 10

1. Page 661, the first line below Equation 10.1: "$prob(|x|) \geq c = \alpha$" should be "$prob(|x| \geq c) = \alpha$".
2. Page 669, All occurrences of y should be bold ($\mathbf{y}$) in Equation 10.7.

Appendix A

1. Equation (A.4) should be as follows:

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}}{\|\mathbf{u}\|} \cdot \frac{\mathbf{v}}{\|\mathbf{v}\|}.$$

Page 700, first line of the bibliographic notes: “Stramg” should be “Strang”.

Appendix C

1. Page 727, eighth from bottom line: “variance $s(X) \times s(X)/N$” should be “variance $s(X) \times (1 - s(X))/N$”.
2. Page 727, fourth from bottom line: “variance $\mathit{minsup} \times \mathit{minsup}/N$” should be “variance $\mathit{minsup} \times (1 - \mathit{minsup})/N$”.