



Errata for *Mining of Massive Datasets - Second Edition*

For errata in the first edition, please see [The Errata Sheet for the First Edition](#).

Page numbers refer to the pages in the book's hardcopy edition, not the downloads. We shall endeavor to keep the downloads up-to-date.

Section	Location	Problem	Reported By	Date Reported
2.3	p. 29, l. 8; p. 33, l. 6, 17; p. 34, l. 6, 14, -11; p. 35, l. 14; p. 36, l. 17, -12	At each of these points, the boldface x should have been The Map Function:		10/28/15
2.3.9	p. 36, l. 24-25	remove the second right parenthesis in the parenthesized expressions on these lines.	John Phillips	1/1/16
2.4.2	p. 40, l. -14	We should explain that therefore P is the transitive closure of E.	Rafi Kamal	7/6/15
2.6.6	p. 58, l. 13-14	"upper bound" is "lower bound" on l. 13, and on l. 14, the result is a lower bound on replication rate.	Saman Haratizadeh	2/23/15
2.6.7	p. 60, l. 14	The last two equations on this line should be $q = 2n^2/g$ and $r = 2n^2/q$.	Saman Haratizadeh	2/23/15
2.6.7	p. 61, l. 8	The sentence "This algorithm..." is not exactly right. Section 2.3.9 describes an algorithm that uses reducers of size $2n$, with each reducer responsible for all the elements having a fixed value of j . However, we could have used reducers of size 2 if we had created one reducer for each (i,j,k) and sent it only the elements m_{ij} and n_{jk} .	Saman Haratizadeh	10/20/15
2.6.7	p. 63, l. 2	n^2 should be n^3 .	Saman Haratizadeh	2/28/15
2.6.7	p. 63, l. 7	"replication rate" should be "communication cost".	Saman Haratizadeh	2/28/15
3.3.6	p. 81, l. 10-11	the functions on both these lines should be computed modulo 5.	Hitesh Shetty	3/24/15
3.4.2	p. 84, l. 14	"in all rows of any of the bands" should be "in at least one row of each band".	Hitesh Shetty	3/27/15
3.9.3	p. 113, l. 16-	"upper" should be "lower" on l. 16, and "distance" should be	Zhang	11/25/14

	17	"similarity" on l. 17.	JunFeng	
3.9.3	p. 113, l. -3	"distance" should be "similarity".	Jeff Hwang	2/19/16
3.9.5	p. 115, l. 15	"distance" should be "similarity".		2/19/16
3.9.6	p. 118, l. 11-13	For the case $i=2$, the value of p is 8, and the proper constraint is $q \geq 9$. The inequality that must be satisfied is $9/(q+j+1) \geq 0.8$, which does have a solution $q=8$ and $j=1$. However, we already knew that from the study of the case $p \geq q$. Also, for the case $i=3$, we have $p=7$ and $q \geq 8$. The resulting inequality is $8/(q+j+2) \geq 0.8$, which has no solution for j a positive integer.	Hitesh Shetty	7/6/15
4.2.2	p. 129, l. 12	Triple quotes should be double.	Yunan Luo	12/22/14
4.5.3	p. 139	There are five occurrences of the expression $E(2 * X.value + 1)$. They should all be $E(n * (2X.value - 1))$.	Saman Haratizadeh	11/17/14
6.1.1	p. 193, Fig. 6.2	7 should be added to the sets in the column for "and" and the rows for "dog" and "cat".	Yokila Arora	1/15/18
6.2.3	p. 202, top	Example 6.7 mistates the frequent itemsets in Example 6.1. In fact, there are five frequent pairs, including {cat, a}, and one frequent tripleton: {cat, dog, a}. As a result, the maximal frequent itemsets are {training}, {cat, and}, {dog, and}, and {cat, dog, a}.	Hitesh Shetty	11/2/15/14
6.5.2	p. 223, l. 5	$2/c$ should be $1/2c$.	Timon Ruban	1/20/17
7.5.4	p. 256, l. 8	Delete one "from"	Saman Haratizadeh	11/30/14
8.4.6	p. 282, l. 10	$B/(N-i)$ should be $B/(N-i+1)$.		1/20/17
9.1.2	295, 6 lines below box	"Niether" should be "Neither".	Yunan Luo	12/22/14
10.4.5	p. 349, l. 7	"node 6" should be "node 5".		1/20/17
10.8.1	p. 367, l. -12	v_1 should be v_i .	Hua Feng	3/12/16
11.1.3	p. 388, l. -6	"eigenvector" should be "eigenvalue"	Rch Seiter	11/17/14
11.1.3	p. 389, item (2)(a) and l. -4	We need to limit what we say to symmetric matrices M . In particular, in item (2)(a) the statement "but even..." is not necessarily true for an arbitrary matrix, and at the bottom of the page the discussion of eigenvectors being orthonormal likewise is guaranteed only for symmetric matrices.		7/10/15
11.1.3	p. 389, second line of Example 11.4	The order of the vector and its transpose needs to be reversed. That is, [0.447, 0.894] should be moved right, to the point just before the = sign. Moreover, in the third line, 2.601 should be 1.601. And in addition, each of the entries in these matrices is off by about 0.002.	Bob Resendes	3/1/15
11.3.2	p. 398, second line below Fig. 11.6	"last two rows" should be "last two columns".		1/20/17

11.3.3	p. 401, l. 6	"corresponding s rows" should be "corresponding s columns".		1/20/17
11.3.6	p. 404, l. -6	"on the left" should be "on the right".		1/20/17
11.4.2	p. 409, l. 13, 21-24	There are a number of arithmetic errors. On l. 13, 0.430 should be 0.608, and the column is [0,0,0,6.38,8.22,3.29]. On l. 21, 0.454 should be 0.642, and on l. 22, 0.556 should be 0.786. On l. 23, 11.01 should be 7.79, and on l. 24, 8.99 should be 6.36,		10/12/15
11.5	p. 412, l. -6	"second-smallest" should be "second-largest".	David Z. Liu	3/12/16
12.1.2	p. 417, l. -7	The point (0,2) should be (1,2).	Harizo Rajaona	6/14/15
12.2.1	p. 426, l. 12	"or" should be "of".	Marcus Gemeinder	10/17/15
12.2.8	p. 435, l. 3	cyx_i should be ηyx_i .	Marcus Gemeinder	10/17/15
12.3.1	p. 437, l. -20, -19	On each line, the $w.x$ should be followed by "+b".	Marcus Gemeinder	10/17/15
12.4.3	p. 449, l. -7	"(" needed before the "3".	Yunan Luo	12/22/14