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

Last updated on January 6, 2023 at 12:50pm

Please send all error reports to dmbook@umn.edu

Preface

Page viii, last sentence of Section entitled, **Support Materials**: The email address for reporting errata has been updated to be dmbook@umn.edu. However, the old address dmbook@cs.umn.edu should still work.

Chapter 2

1. Page 27: The title “What Is an attribute?” should be “What is an Attribute?”.

2. Page 40, Figure 2.4(c): In the y-axis label, “celsius” should be capitalized, i.e., “Celsius”.

3. Page 65, The last sentence before the Unsupervised Discretization section: “+ inf and ‘− inf, respectively” should be “− inf and ‘+ inf, respectively”.

4. Page 71, second paragraph: “σ_A = \sum_{i=1}^{m} |x_i - \mu|” should be “\sigma_A = \frac{1}{m} \sum_{i=1}^{m} |x_i - \mu|”.

5. Page 77: In the properties of a metric, condition 1(b) should be d(x, y) = 0 if and only if x = y.

6. Page 89, The first sentence after equation (2.15): “I(X, Y) = I(Y)” should be “I(X, Y) = I(Y, X)”

7. Page 93, the first line: “(x, y)” should be “(\phi(x), \phi(y))”

8. Page 93, 2 lines before equation 2.19: “then these two” should be “then these three”

9. Page 93, Example 2.24, First sentence: “presented in the previous section” should be “discussed above”
2 Errata

10. Page 94, Equation 2.24: The inner product should be a sum, not a tuple, so equation 2.24 should be
\[ \kappa(x, y) = (x'y + c)^2 = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 + 2cx_1y_1 + 2cx_2y_2 + c^2 = \langle \varphi(x), \varphi(y) \rangle \]

Chapter 3

1. Page 148, Figure 3.23b should be as follows:

![Figure 3.23b](image)

**Figure 3.23b** Varying tree size from 1 to 150.
2. Page 141, Figure 3.16: “width > 3” should be “breadth > 3”:

![Decision Tree Model](image)

**Figure 3.16** Decision tree model for web robot detection.
Chapter 4

1. Page 251, Equation 4.48: This equation should be as follows:

\[ \hat{y} = \begin{cases} 1, & \text{if } w^T x + b > 0. \\ -1, & \text{otherwise}. \end{cases} \]

2. Page 312, second paragraph: the out of bag sample is 37% of the base classifiers, not 27%.

3. Page 322, Table just above Section 4.11.3: This table should be as follows:

\[ \text{Weighted accuracy} = \frac{w_1 TP + w_4 TN}{w_1 TP + w_2 FP + w_3 FN + w_4 TN}. \]
The relationship between weighted accuracy and other performance measures is summarized in the following table:

<table>
<thead>
<tr>
<th>Measure</th>
<th>w_1</th>
<th>w_2</th>
<th>w_3</th>
<th>w_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Precision</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F_\beta</td>
<td>\beta^2 + 1</td>
<td>\beta^2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Accuracy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Chapter 5

1. Page 382, Algorithm 5.3: This algorithm should be revised as follows:

Algorithm  Procedure \texttt{ap-genrules}(f_k, H_m).

1: $k = \lvert f_k \rvert$  \{size of frequent itemset.\}
2: $m = \text{size of itemsets in } H_m$  \{size of rule consequent.\}
3: \{Generate rules with consequent of size m.\}
4: \textbf{if} $k \geq m + 1$ \textbf{then}
5: \hspace{1em} \textbf{for each } $h_m \in H_m$ \textbf{do}
6: \hspace{2em} $conf = \sigma(f_k)/\sigma(f_k - h_m)$.
7: \hspace{2em} \textbf{if} $conf \geq \text{minconf}$ \textbf{then}
8: \hspace{3em} \textbf{output} the rule $(f_k - h_m) \rightarrow h_m$.
9: \hspace{2em} \textbf{else}
10: \hspace{3em} delete $h_m$ from $H_m$.
11: \hspace{2em} \textbf{end if}
12: \hspace{1em} \textbf{end for}
13: \textbf{end if}
14: \{Recursively call \texttt{ap-genrules} to generate rules with larger consequents.\}
15: \textbf{if} $k > m + 1$ \textbf{then}
16: \hspace{1em} $H_{m+1} = \text{candidate-gen}(H_m)$.
17: \hspace{1em} $H_{m+1} = \text{candidate-prune}(H_{m+1}, H_m)$.
18: \hspace{1em} call \texttt{ap-genrules}(f_k, H_{m+1}).
19: \textbf{end if}

Chapter 6

1. Page 452, 1st paragraph: “as well as nominal attributes such as Level of Education and State” should be “as well as categorical attributes such as Level of Education and State”
6  Errata

2. Page 487, line 9 of Algorithm 6.2. The comment should say, “Identify all candidates contained in $g$.”

Chapter 7

1. Page 586, the second sentence of Example 7.11, which is in parentheses: This sentence should be “(The data for this figure consists of the six two-dimensional points given in Table 7.3.)”

2. Page 587, the caption for Table 7.7: This caption should be “Cophenetic distance matrix for single link and data in Table 7.3 on page 557.”

3. Page 592, Example 7.16: “$p_1, p_2, p_3, p_4, and p_5$” should be “$p_1, p_2, p_3, p_4, and p_5$”.

4. Page 592, Example 7.16: “$L_2 = \{p_3, p_4, p_5\}$” should be “$L_2 = \{p_3, p_4, p_5\}$”.

5. Page 610, Exercise 29. This exercise should be as follows: Prove that $\sum_{i=1}^{K} \sum_{x \in C_i} (x - c_i)(c - c_i) = 0$. This fact was used in the proof that TSS = SSE + SSB on page 578 in Section 7.5.2.