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

Last updated on November 24, 2023 at 01:28pm

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‘**Unupervised Discretization** section: “$+\inf$ and ‘$-\inf, respectively$” should be “$-\inf$ and ‘$+\inf, respectively$”.

4. Page 71, second paragraph: “$\sigma_A = \sum_{i=1}^m |x_i - \mu|$” 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 $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.

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: “$\langle \mathbf{x}, \mathbf{y} \rangle$” should be “$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$”

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”

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

Chapter 3

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

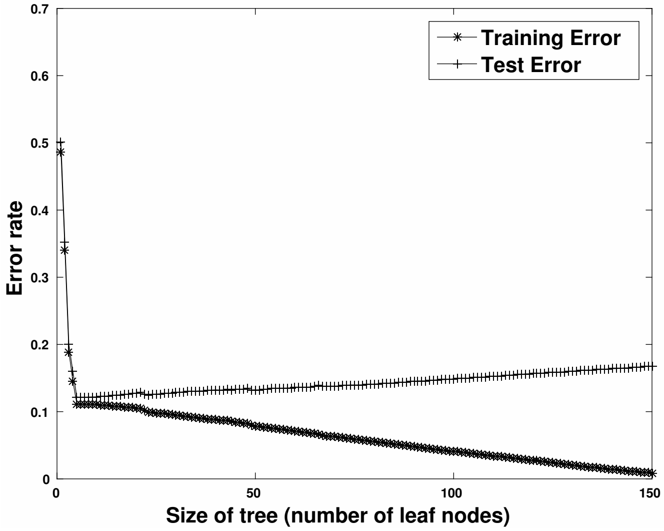
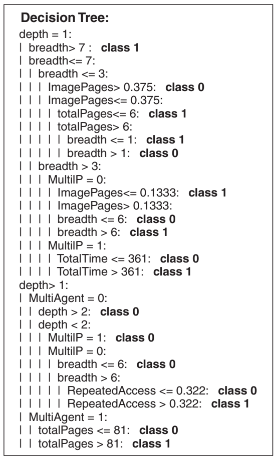


Figure 3.23b Varying tree size from 1 to 150.

2. Page 141, Figure 3.16: “width $>$ 3” should be “breadth $>$ 3”:



```
Decision Tree:
depth = 1:
| breadth> 7 :  class 1
| breadth<= 7:
| | breadth <= 3:
| | | ImagePages> 0.375:  class 0
| | | ImagePages<= 0.375:
| | | | totalPages<= 6:  class 1
| | | | totalPages> 6:
| | | | | breadth <= 1:  class 1
| | | | | breadth > 1:  class 0
| | breadth > 3:
| | | MultiIP = 0:
| | | | ImagePages<= 0.1333:  class 1
| | | | ImagePages> 0.1333:
| | | | breadth <= 6:  class 0
| | | | breadth > 6:  class 1
| | | MultiIP = 1:
| | | | TotalTime <= 361:  class 0
| | | | TotalTime > 361:  class 1
depth> 1:
| MultiAgent = 0:
| | depth > 2:  class 0
| | depth < 2:
| | | MultiIP = 1:  class 0
| | | MultiIP = 0:
| | | | breadth <= 6:  class 0
| | | | breadth > 6:
| | | | | RepeatedAccess <= 0.322:  class 0
| | | | | RepeatedAccess > 0.322:  class 1
| MultiAgent = 1:
| | totalPages <= 81:  class 0
| | totalPages > 81:  class 1
```

Figure 3.16 Decision tree model for web robot detection.

3. Page 164, Figure 3.32: "width $>$ 3" should be "breadth $>$ 3":

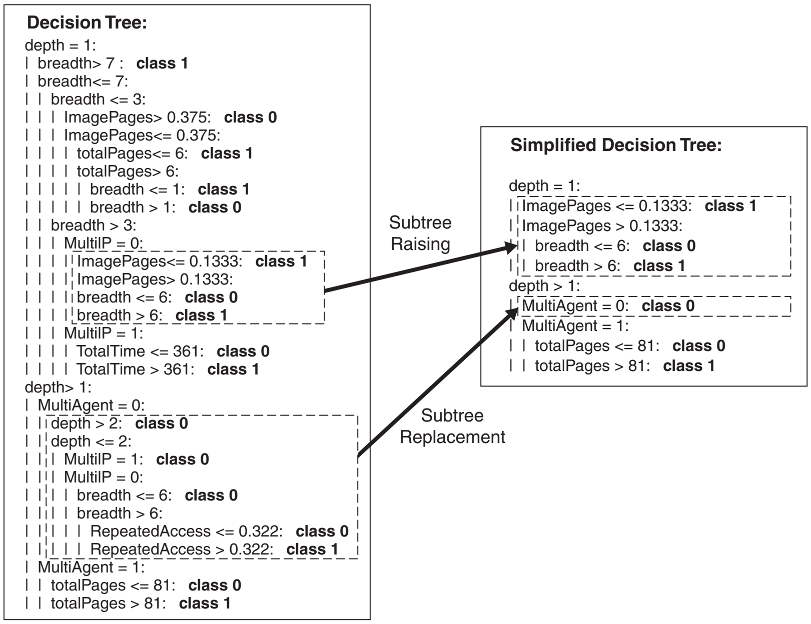

Decision Tree:
depth = 1:
| breadth> 7 : class 1
| breadth<= 7:
| | breadth <= 3:
| | | ImagePages> 0.375: class 0
| | | ImagePages<= 0.375:
| | | | totalPages<= 6: class 1
| | | | totalPages> 6:
| | | | | breadth <= 1: class 1
| | | | | breadth > 1: class 0
| | breadth > 3:
| | | MultiIP = 0:
| | | | ImagePages<= 0.1333: class 1
| | | | ImagePages> 0.1333:
| | | | breadth <= 6: class 0
| | | | breadth > 6: class 1
| | | MultiIP = 1:
| | | | TotalTime <= 361: class 0
| | | | TotalTime > 361: class 1
depth> 1:
| MultiAgent = 0:
| | depth > 2: class 0
| | depth <= 2:
| | | MultiIP = 1: class 0
| | | MultiIP = 0:
| | | | breadth <= 6: class 0
| | | | breadth > 6:
| | | | | RepeatedAccess <= 0.322: class 0
| | | | | RepeatedAccess > 0.322: class 1
| MultiAgent = 1:
| | totalPages <= 81: class 0
| | totalPages > 81: class 1
Subtree Raising
Subtree Replacement
Simplified Decision Tree:
depth = 1:
| ImagePages <= 0.1333: class 1
| ImagePages > 0.1333:
| | breadth <= 6: class 0
| | breadth > 6: class 1
depth > 1:
| MultiAgent = 0: class 0
| MultiAgent = 1:
| | totalPages <= 81: class 0
| | totalPages > 81: class 1

Figure 3.32 Post-pruning of the decision tree for web robot detection.

Chapter 4

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

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

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:

$$\text{Weighted accuracy} = \frac{w_1\text{TP} + w_4\text{TN}}{w_1\text{TP} + w_2\text{FP} + w_3\text{FN} + w_4\text{TN}}. \tag{1}$$

The relationship between weighted accuracy and other performance measures is summarized in the following table:

Measure	w_1	w_2	w_3	w_4
Recall	1	0	1	0
Precision	1	1	0	0
F_β	β^2+1	β^2	1	0
Accuracy	1	1	1	1

Chapter 5

1. Page 359, Last line of the page should be the following:
   An itemset X is called frequent if $s(X)$ is greater than or equal to some user-defined threshold, *minsup*.

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

Algorithm Procedure `ap-genrules`(f_k, H_m).

```
$k = |f_k|$    {size of frequent itemset.}
$m = size\ of\ itemsets\ in\ H_m$    {size of rule consequent.}
{Generate rules with consequent of size m.}
if $k \geq m + 1$ then
  for each $h_m \in H_m$ do
    $conf = \sigma(f_k)/\sigma(f_k - h_m)$.
    if $conf \geq minconf$ then
      output the rule $(f_k - h_m) \longrightarrow h_m$.
    else
      delete $h_m$ from $H_m$.
    end if
  end for
end if
{Recursively call ap-genrules to generate rules with larger consequents.}
if $k > m + 1$ then
  $H_{m+1}$ = candidate-gen($H_m$).
  $H_{m+1}$ = candidate-prune($H_{m+1}$,$H_m$).
  call ap-genrules($f_k, H_{m+1}$.)
end if
```

3. Page 445, Exercise 13(b)(iii): The formula for Interest should be the following:
   Interest$(X \longrightarrow Y) = \frac{P(X,Y)}{P(X)P(Y)}$.

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`"

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: "$p_1, p_2, p_3, p_4, andp_5$" should be "$p_1, p_2, p_3, p_4,$ and p_5".

4. Page 592, Example 7.16: "$L2 = \{p_3, p_4, p_5\}$" 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.