ERRATA (R EDITION)

노트북: 네이버
만든 날짜: 2021-12-11 오후 11:04 **수정한 날…** 2021-12-11 오후 11:06
작성자: sunryul23@naver.com
URL: https://www.dataminingbook.com/content/errata-r-edition

ERRATA (R EDITION)

To view tables and graphs referred to in the errata, please log in.

p. 22, footnote 1	Quotation marks should start before "Zestimates..." not before "Harney..."
p. 22	The URL for the dataset no longer works. Instead, go to **https://data.boston.gov/dataset/property-assessment** and choose Property Assessment FY2014
Tables 2.11-2.13	The regression model should not include TAX as a predictor. See the **R code** for the correct output.
p. 48 box by Herb Edelstein	Copyright year should be 2017
Fig 3.1, R code	Code line: plot(housing.df$MEDV.... , xlab = "MDEV",...) should read: plot(housing.df$MEDV ~ housing.df$LSTAT, xlab = "LSTAT", ylab = "MEDV")
Fig 3.9, R code	Name of dataset should be Amtrak.csv (not Amtrak data.csv): Amtrak.df <- read.csv("Amtrak.csv")
p.80-81 and Fig 3.14 caption	p.80: Remove the text "Circle size represents the number of transactions that the node (seller or buyer) was involved in within this network. Line width represents the number of auctions that the bidder--seller pair interacted in." Fig 3.14 caption: remove "Circle size represents the node's number of transactions. Line width represents the number of transactions between that pari of seller-buyer"
Fig 5.2, R code	Code line: validation <- sample(toyota.corolla.df$Id, 400) should read: validation <- sample(setdiff(toyota.corolla.df$Id, training), 400)
p. 126 top	First paragraph should read: "The top-right cell gives the number of class 1 members that were misclassified as 0's... lower-left cell gives the number of class 0 members that were misclassified as 1s (25 such records)."
p. 138, Fig 5.6 caption	replace "top" with "left", and "bottom" with "right"
Fig 5.7, caption	add: (Note: Percentiles do not match deciles exactly due to the small sample of discrete data, with multiple records sharing the same decile boundary)

Fig 5.10, Fig 5.11	"Classify as 'x'" should be at bottom and "Classify as 'o'" should be at top
p. 148, Problem 5.6	In problem 5.6, text should read "The global mean is about $2500" In part (b), text should read "roughly double the sales effort"
Table 6.2, R code	Commented out text should read: # use lm() to run a linear regression of Price on all the predictors in the # training set (it will automatically turn Fuel_Type into dummies).
p. 166	Paragraph before last: delete "In comparison... any predictor". Instead: "The results for forward selection (Table 6.7) and stepwise selection..."
p. 167 Table 6.6	Ignore comment "set directions..."

Table 6.7	The output is incorrect. It should be identical to the output in Table 6.8.
	Add code lines:
	# create model with no predictors
	car.lm.null <- lm(Price~1, data = train.df)
	# use step() to run forward regression.
	car.lm.step <- step(car.lm.null,
	scope=list(lower=car.lm.null, upper=car.lm), direction = "forward")
	summary(car.lm.step) # Which variables were added?
	car.lm.step.pred <- predict(car.lm.step, valid.df)
	accuracy(car.lm.step.pred, valid.df$Price)

p. 169	Problem 6.1 part (c), ignore the final text "What is the prediction error?"
p. 178, 5th line from bottom	should read "We would choose k=8, which maximizes our accuracy..."
p. 184, prob 7.2 (a)	should read: "Consider the following customer: Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education = 2, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1, and Credit Card = 1. Perform a k-NN classification with all predictors except ID and ZIP code using k = 1. Remember to define categorical predictors with more than two categories as factors (for k-NN, to automatically handle categorical predictors, use library class, rather than FNN). Use the default cutoff value of 0.5. How would this customer be classified?"
p.185 prob 7.2(d)	should read: "Education = 2"
p. 206	last paragraph (before IF) should read "The values below a white node are the counts of the two classes (0 and 1) in that node"
p. 209	[This is a clarification] Addition: "As with k-nearest-neighbors, a predictor with m categories (m>2) should be factored into m dummies (not m-1). In addition, whether

	predictors are numerical or categorical, it does not make any difference whether they are standardized (normalized) or not."
p. 235-6, Problem 9.3	(a): replace 100 with 30. (a)(ii): delete "What is happening with the training set predictions?" (a)(iii): replace text with "How might we achieve better validation predictive performance at the expense of training performance?" (a)(iv): replace text with "Create a less deep tree by leaving the arguments cp, minbucket, and maxdepth at their defaults. Compared to the deeper tree, what is the predictive performance on the validation set?" (b): replace last sentence "Keep the minimum..." with "As in the less deep regression tree, leave the arguments cp, minbucket, and maxdepth at their defaults" (b)(i) and (b)(ii): use the less deep RT for these questions.
p. 246	line 7 should read $e^{(0.03757)(100)}$ instead of $e^{(0.039)(100)}$
p. 251	Text should read "we see that Sundays and Tuesdays saw the largest proportion of delays"
p. 253	The reference to Figure 10.4 should be to Figure 10.6 (creating base categories)
p. 256, R code	Should be gain <- gains(valid.df$isDelay, pred, groups=100)
p. 278	For output node 6 the error is 0.481(1-0.481)(0-0.481)= -0.120
Table 11.6	Due to a change in caret package for confusionMatrix, make sure to first convert each variable into factors, e.g. confusionMatrix(as.factor(validation.class), as.factor(accidents.df[validation,]$MAX_SEV_IR))
Table 11.7	Table 11.7 is redundant and should be deleted (the same output appears in Table 11.6)
p. 302	-50.58 should be -51.58
Table 13.1	3rd line from bottom of the table, code should read "...data = train.df" instead of "...data = bank.df"
Table 13.2	Final confusion matrix (for boosting) should read Reference Prediction 0 1 0 1804 23 1 3 170 Accuracy : 0.987
p. 302	In line 2, replace "a sample of 1000 records was drawn" with "a reduced sample of 600 records was drawn (with categories combined so that most predictors are binary)"

p. 365	In Distance Measures for Categorical Data, replace "x_ij's" with "p measurements", and replace n with p in the table and in the Matching coefficient formula.
Ch 15-17	Several of the time series datasets used in the problems (souvenir sales, shampoo sales, Australian wine sales) have a new source reference: Hyndman, R., and Yang, Y. Z. (2018). tsdl: Time Series Data Library. v0.1.0. https://pkg.yangzhourang.com/tsdl/
Problem 15.2.c (p. 383)	Should read: "... with respect to the categorical variables (10 to 12)"
p. 391, Fig 16.1 code	Lines 2-3 of commented out text should read: # with monthly data, the frequency of periods per cycle is 12 (per year). # arguments start and end are (cycle [=year] number, seasonal period [=month] number) pairs.
p. 420, last para	Reference to Figure 17.6 should be Table 17.7 (AR(1) model)
p. 463	Closeness definition should be: This is measured by finding the shortest path from that node to all the other nodes, then taking the reciprocal of the sum of these path lengths.
Table 19.3	Values for betweenness and closeness should be: > betweenness(g) Dave Jenny Peter John Sam Albert 0 0 6 0 4 0 > closeness(g) Dave Jenny Peter John Sam Albert 0.12500000 0.12500000 0.16666667 0.12500000 0.12500000 0.08333333
Table 19.5	Caption should be "Computing Network Measures in R"
p. 528	Available Data should read: "Part of the historic information is available in the file bicup2006.csv. The file contains the historic information with known demand for a 3-week period, separated into 15-minute intervals, and dates and times for a future 3-day period (DEMAND = NaN), for which forecasts should be generated (as part of the 2006 competition)."
p. 533, Data Files Used in the Book	File Amtrak data.csv s