The significance of changes is denoted by a scale of one to three stars:

★ ★ ★ --- an error which could mislead an informed reader;
★ ★ --- an error which many readers would notice and correct;
★ --- a small change to enhance clarity or consistency, or to correct grammar.

The color of the stars is just a reminder to the author, and is not otherwise meaningful.

★ Page 5, Section 2.3 (Data), second paragraph
It is sometimes useful to consider the data as produced by a two-step process, in one of two ways:
by drawing \( Y \) from marginal distribution \( P(Y) \) on \( f(X) \) and then drawing a corresponding feature vector \( X \) from conditional distribution \( P(X|Y) \) on \( X \); or by drawing feature vector \( X \) from marginal distribution \( P(X) \) on \( X \) and then drawing a corresponding \( Y \) from conditional distribution \( P(Y|X) \) on \( f(X) \).

★ Page 10, Exercise 2.2
After “… given losses \( L(1,2) > 0 \) and \( L(2,1) > 0 \)” add “(assume \( L(1,1) = L(2,2) = 0 \)).”

★★ Page 19, Section 3.4 (Properties of Fitted Values)
The first sentence should refer to Exercise 3.2, not Exercise 3.1.

★★ Page 27, Section 3.9 (Feature Transformations, Expansions, and Interactions)
In the second displayed equation, the limits on the double sum should be \( \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \).

★★★ Page 51, Section 4.4.5 (Histograms), first paragraph
The histogram density estimate is
\[
\hat{P}(x|Y = c) = \frac{n_c}{V(R_c)} \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{1}{V(R_c)},
\]
where \( V(R_c) = h_{c,1} \cdots h_{c,m} \) is the volume of the rectangular cells for class \( c \). This situation is illustrated in Figure 4.9.

★★★ Page 52, Section 4.4.5 (Histograms), Figure 4.9

★★★ Page 54, Section 4.4.5 (Histograms), Exercise 4.8
The last two displayed equations should be
argmin_{c=1, \ldots, C} \sum_{d=1}^C \frac{L(d, c) n_{d}^{R_d(x)}}{V(R_d)}

and

argmax_{c=1, \ldots, C} \frac{n_{c}^{R_c(x)}}{V(R_c)}.

★★ Page 60, Section 4.6 (Logistic Regression)
In the second displayed equation on page 60, the vector \( \theta \) is

\[
\begin{bmatrix}
\theta_1 \\
\theta_2 \\
\vdots \\
\theta_C \\
\end{bmatrix}
\]

★★ Page 61, Section 4.6 (Logistic Regression)
The second sentence on page 61 should have some inconsistent subscripts. It should read: “The training data are assumed to be independent, which means that the class label \( Y_i \) is a single draw from the multinomial distribution with probability parameter \( P(Y = 1|X = x_i) \),

\[
Y_i|X_i = x_i \sim \text{Multinomial} \left( 1, \left( P(Y = 1|X = x_i), \ldots, P(Y = C|X = x_i) \right) \right).
\]

★★ Page 73, Section 4.7.4 (Logistic Regression and Neural Networks), Figure 4.21
In a perfect world, the labels at right would be consistent with those in Figure 4.17.

★★ Page 84, Section 4.9.1 (Support Vector Machine Classifiers), second paragraph
\( \theta_* = \sum_{i: (x_i, y_i) \text{is a support vector}} y_i x_i \cdot \)

★★ Page 86, Section 4.9.1 (Support Vector Machine Classifiers), first paragraph
In the second-to-last sentence, starting “As the distance from correctly classified data…”, this sentence should be consistently either about data (plural) or a datum (singular).

★ Page 88, Section 4.9.1 (Support Vector Machine Classifiers), Figure 4.30
The picture would show what’s going on more clearly with a small triangle added as shown:
The following should be appended to the caption: “The black triangle in the right-hand cell encloses a region in feature space where the orange-vs-blue SVM predicts orange, the orange-vs-purple SVM predicts purple, and the blue-vs-purple SVM predicts blue. The R function `svm()` has chosen (arbitrarily) to predict orange here.”

★★★ Page 93, Section 4.10 (Postscript: Example Problem Revisited)
On the fifth line from the bottom of the page, the text should read “those with risk 0.212 or below”.

★★★ Page 99, Section 5.1 (Squared-Error Loss), Figure 5.1 caption, first sentence
The number 15 should be changed to 17.

★★★ Page 99, Section 5.1 (Squared-Error Loss)
The one paragraph on this page should read as follows:

“The bias and variance of three different methods of approximating function $f(x) = \frac{3}{4}x + \sin(\pi x)$ on the interval [-1,1] are illustrated in Figure 5.1. In the examples illustrated, $n$ one-dimensional feature vectors $X_1, ..., X_n$ are drawn uniformly from [-1,1], and unioned with feature vectors $X_{n+1} = -1$ and $X_{n+2} = 1$. Then, for $i = 1, ..., n + 2$, response $Y_i|X_i$ is drawn from a Gaussian distribution with mean $f(X_i)$ and unit variance (so the intrinsic risk with respect to squared-error loss is 1).”

★★★ Page 100, Section 5.1 (Squared-Error Loss), first paragraph and footnote 2
The number 15 should be changed to 17 in the three places it occurs.

★★★ Page 100, Section 5.1 (Squared-Error Loss), fifth paragraph
The phrase “For each data size $|S| \in \{15,150,1500,15000,150000\}$” should read “For each data size $|S| \in \{15,150,1500,15000,150000\} + 2$”.

★★★ Page 101, Section 5.1 (Squared-Error Loss), Table 5.1 caption
The number 15 should be changed to 17.

★★★ Page 101, Section 5.1 (Squared-Error Loss), Table 5.1
Each number in the left-hand column should have “+2” after it. For example, the first three numbers in the left-hand column should be “15+2”, “150+2”, and “1500+2”. Additionally, Table 5.1 would be easier to read if horizontal lines separated it into three blocks: one where the values...
of Degree (second column) are all 0, one where the values are all 3, and one where the values are all 6.

**Page 109, Section 6.1 (Ensembles), Figure 6.2 caption**
The following should be appended to the caption: “As noted in Figure 4.30, the R function svm( ) has chosen (arbitrarily) to break ties by predicting orange.”

**Page 112, Section 6.3 (Bagging)**
The second sentence should have the phrase “, independently and uniformly” appended, reading: “A bootstrap sample of a dataset of size n is obtained by sampling the set with replacement n times, independently and uniformly.”

**Page 112, Section 6.3 (Bagging), Exercise 6.3**
The second sentence should have the phrase “and uniformly” inserted, beginning: “Show that if a dataset of size n is sampled independently and uniformly with replacement […]”

**Page 112, Section 6.3 (Bagging), Exercise 6.3**
The phrase “a given bootstrap sample is about” should read “a given bootstrap sample (ρ = 1) is about”.

**Page 113, Section 6.3 (Bagging), Figure 6.3**
The risk of the 1-nearest neighbor classifier is 0.248, not 0.247.

**Page 116, Section 6.5 (Random Forests)**
In the first sentence, “combination earlier” should be “combination of earlier”.

**Page 117, Section 6.3 (Random Forests), Exercise 6.5, third paragraph, first sentence**
“Breiman defined the random forest proximity of two data points (x_i, y_i) and (x_j, y_j) to be the proportion of trees in a random forest with the property that (x_i, y_i) and (x_j, y_j) are in the same terminal node (leaf).”

**Page 119, Section 6.6 (Boosting), caption to Figure 6.7, second sentence**
It would have been more helpful if the second sentence read “Modifications for the case C > 2 and a reason for the particular choice of weight 1 - R_i / R_l are described in the text and exercises.”

**Page 130, Section 7.2.3 (Size of Training, Validation, and Test Sets), first paragraph**
The phrase ‘The answer depends balancing’ should be “The answer depends on balancing”.

**Page 131, Section 7.2.4 (Exercise 7.6 part (A))**
The parenthetical comment should have the word “to” added between the words “test” and “have”, reading: “(this is what it means for the test to have level α – see Chapter 12).”

**Page 134, Section 7.3 (Cross-Validation)**
The reference to Exercise 7.2 should have been a reference to Exercise 7.3.

**Page 138, Section 7.6 (Akaike’s Information Criterion)**
In the first paragraph, the following sentence should be inserted before the last sentence: “AIC decreases the better the model fits the observed data, and increases the more complex the model is, where complexity is measured by the number of parameters, k.”

The second paragraph should be replaced by the following:
“Akaike’s criterion is an extension of the maximum likelihood principle. Suppose there are L models under consideration, M_1, ..., M_L, and that model M_i has a vector parameter θ_i of dimension k_i, and let θ̂_i denote an estimate of θ_i derived in some way from a dataset S = \{(X_1, Y_1), ..., (X_n, Y_n)\}. The maximum likelihood principle states that, given observed data (x_1, y_1), ..., (x_n, y_n), the “best” choice of model M_i and corresponding parameter θ_i is the one that maximizes the likelihood of the observed responses or classes,

arg\underset{\theta_i}{\text{max}} P(y_1, ..., y_n | x_1, ..., x_n, \theta_i, M_i).

Note that the maximization is taken over all models M_i and all possible parameters θ_i for each model M_i. Exercise 7.1 indicates a potential problem when using the maximum likelihood
principle for model selection. In contrast to the maximum likelihood principle, Akaike (1973) begins with the idea that the “best” choice of model $M_i$ is the one that maximizes the expected log-likelihood of a response or class which has not yet been observed,

$$
\arg\max_{M_1, \ldots, M_L} E_{S,X,Y} [\log P(Y|X, \hat{\theta}, M_i)],
$$

where the parameter estimate $\hat{\theta}_i$ for model $M_i$ is estimated from the data set $S$. Akaike shows that, for large $n$, the model which maximizes the expected log-likelihood is approximately the model which maximizes the formula for AIC given at the start of this section. In the case of regression with additive, Gaussian-distributed noise, the expected log-likelihood is

$$
E_{S,X,Y} [\log P(Y|X, \hat{\theta}_i, M_i)] = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} E_{S,X,Y} \left[ (Y - (1,X)\hat{\theta}_i)^2 \right],
$$

so selecting the model which maximizes the expected log-likelihood is equivalent to selecting the model which minimizes risk with respect to squared-error loss (see Section 3.6). In the case of classification by logistic regression and some neural networks, selecting the model which maximizes the expected log-likelihood is equivalent to selecting the model which minimizes risk with respect to cross-entropy loss (Sections 4.6 and 4.7, Exercise 4.11).”

★★★ Page 139, Section 7.7 (Schwartz’s Bayesian Information Criterion)
Immediately before the second displayed expression, the words “the logarithm of the integral above” should be “the logarithm of the expression above”.

★★ Page 140, Section 7.8 (Rissanen’s Minimum Description Length Crit.), second paragraph
The phrase “observed features $x_1, \ldots, x_n$” should be “observed feature vectors $x_1, \ldots, x_n$”.

★ Page 140, Section 7.8 (Rissanen’s Minimum Description Length Crit.), 2nd and 3rd paragraphs
The references to Rissanen (1978) should be to Rissanen (1978, 1983).

★★ Page 140, Section 7.9 ($R^2$ and Adjusted $R^2$)
In the two displayed equations on this page, the expression “n x” should appear immediately before $\hat{r}_{\text{train}}(f)$.

★ Page 141, Section 7.9 ($R^2$ and Adjusted $R^2$)
In the first displayed equation on this page, it would add to notational consistency to write

$$
\text{MSS} = \sum_{i=1}^{n} (\bar{y} - \bar{y})^2 = \sum_{i=1}^{n} (\bar{y} - \bar{y})^2.
$$

★★★ Page 141, Section 7.9 ($R^2$ and Adjusted $R^2$), last paragraph, 2nd sentence
“After all, RSS is a decreasing function of the training error” should be “After all, RSS is $n$ times $\hat{r}_{\text{train}}(f)$”.

★★ Page 142, Section 7.10 (Stepside Model Selection), Figure 7.2
In step (5), the subscript should refer to steps (3) and (4), not steps (2) and (3).

★ Page 159, Chapter 10 introduction, second paragraph
The phrase “Chapters 4 and 9” should be “Chapters 3, 4, 6, and 9”.

★ Page 172, Section 10.6.3 (Optimization from Multiple Starting Points)
The author should have included the following paragraph as a second paragraph to this section.

“It is also good practice, when an optimization algorithm converges to an approximate minimizer $\omega^*$, to restart the algorithm at $\omega^*$. Either the algorithm does not find a better minimizer than $\omega^*$, in which case it typically declares convergence at $\omega^*$ after little work, relative to the work performed to find $\omega^*$ in the first place; or it does find a better minimizer than $\omega^*$.”

★ Page 217, Introduction to Chapter 12, first paragraph
In the first sentence, the first “the” should be omitted from “in the one or more of the”.

★★ Page 220, Section 12.2 (Terminology for Binary Decisions), Table 12.2
In the right most column, the entry “risk” should be changed to “risk (under 0-1 loss)”.

★ Page 226, Section 12.6.1 (Control the Familywise Error), first paragraph
The first sentence would be clearer if a mathematical expression for the familywise error rate were given.

“The familywise error rate of T statistical tests is the probability that any of the tests incorrectly rejects the null hypothesis,

\[ P \left( \bigcup_{i=1}^{T} \{ H_0 \text{ is true for the } i\text{th test and the } i\text{th test rejects } H_0 \} \right). \]

To make the texts parallel, then, in Page 227, Section 12.6.2 (Control the False Discovery Rate), first paragraph, the first sentence should be rewritten as follows.

“The false discovery rate of T statistical tests is probability that the null hypothesis is true for a test, conditional on that test rejecting the null hypothesis,

\[ P( H_0 \text{ is true}| H_0 \text{ is rejected}). \]

Page 228, Section 12.7 (Expert Systems)
The final statement by Client should have closing quotation marks.

Page 249, Section 14.9 (Histograms), third paragraph
The line

\[ HT^\text{iv} \text{ --- a vector } (V(R_1)^{-1}, \ldots, V(R_C)^{-1}) \text{ of reciprocals of cell volumes;} \]

should be added between the lines

\[ HT^\text{h} \text{ --- the } C \times m \text{ matrix of bandwidth parameters, } h; \]

and

\[ HT^\text{prior} \text{ --- the marginal distribution of class labels which occur in the training data.} \]

Page 250, Section 14.9 (Histograms), function hist.insert() The two lines

\[
\text{if(has.key(key,HT[[daty]])==FALSE) \{ HT[[daty]][key]=1; \} else \{ HT[[daty]][key]=values(HT[[daty]][key])+1; \}}
\]

should be replaced with the two lines

\[
\text{if(has.key(key,HT[[daty]])==FALSE) \{ HT[[daty]][key]=HT^\text{iv}[daty]; \} else \{ HT[[daty]][key]=values(HT[[daty]][key])+HT^\text{iv}[daty]; \}}
\]

Page 250, Section 14.9 (Histograms), function hist.train() The line

\[ HT^\text{iv} \leftarrow 1/apply(HT^\text{h},1,\text{prod}); \]

should be inserted between the line

\[ \text{for(cc in 1:nc) \{ HT[[cc]] \leftarrow \text{hash}(); \}} \]

and the line

\[ \text{apply(cbind(datx,daty),1,hist.insert);} \]

Page 264, Section 14.15 (Classification Trees)
The word “tree” should be inserted in the phrase “The estimated risk \( R \) of the one-leaf can be computed”, resulting in “The estimated risk \( \hat{R} \) of the one-leaf tree can be computed”.

Page 287, Solution to Exercise 6.1
The following should be inserted before the last sentence of the solution. “Since risk R grows without bound as any \( c_1 \) goes to \( \infty \), R has no maximum. Since \( R \geq \beta \), R has a minimum, and the critical point is the minimum.”

Page 289, Solution to Exercise 7.1
The one inequality symbol in the solution should be reversed: the statement \( \hat{R}_{\text{valid}}(f_{d+1}) \geq \hat{R}_{\text{valid}}(f_d) \) should be \( \hat{R}_{\text{valid}}(f_{d+1}) \leq \hat{R}_{\text{valid}}(f_d) \).

Page 311, References
Rissanen’s 1978 paper is in *Automatica* volume 14, not volume 15. Also, I should add the following reference, which is the source of the key ideas presented in Section 7.8 (Minimum Description Length):


★★ Solution to Exercise 3.4 (web solutions)

In the following equation, the expectation shown in red is missing and should be added:

\[ E[\hat{e}] = E[y - \hat{y}] = E[y] - E[\hat{y}] \]

★★★ Solution to Exercise 4.8 (web solutions)

**Exercise 4.8** The histogram classifier predicts

\[
\arg\min_{c=1,\ldots,C} \sum_{d=1}^{C} L(d, c) \frac{R_{d}(x)}{P_{x}(\gamma)} = \arg\min_{c=1,\ldots,C} \sum_{d=1}^{C} L(d, c) \frac{n_{d}}{V(R_{d})} = \arg\min_{c=1,\ldots,C} \sum_{d=1}^{C} L(d, c) \frac{n_{d}}{V(R_{d})}
\]

Under 0-1 loss, minimizing the sum of losses means maximizing the term left out, so the predicted class is

\[
\arg\max_{c=1,\ldots,C} \frac{n_{c}(x)}{V(R_{c})}
\]

★ Solution to Exercise 7.4 (web solutions)

The prime notation can be removed from the letter \( \mu \) where it occurs in the first sentence (\( \mu' \) is replaced by \( \mu \)).