Errata: Learning Kernel Classifiers—Theory and Algorithms

Page xvii, line -3 "Since the book uses a very rigorous notation systems, ..." should read "Since the book uses a very rigorous notation system, ..." (thanks to Neil Lawrence).

Page 8, line 5 "... reader should refer Sutton and Barto (1998)" should read "... reader is referred to Sutton and Barto (1998) ..." (thanks to Thore Graepel).

Page 19, line -12 "... (see Definition A.39)." should read "... (see Definition A.34)." (thanks to Marco Krüger).

Page 33, line 15 "$x \in \mathcal{K}$" should read "$x \in \mathcal{X}$" (thanks to Arthur Gretton).

Page 33, line -3 "... *where* $\mathbf{U} = \left(\mathbf{u}_1'; \ldots; \mathbf{u}_r'\right)$ *is* ..." should read "... *where* $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n) = \left(\mathbf{v}_1'; \ldots; \mathbf{v}_r'\right)$ *is* ...". Then, page 34, line 2 changes to

$$\boldsymbol{\phi}(x_i) = \boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{v}_i ,$$

and line 4 on the same page changes to

$$\mathbf{G}_{ij} = \left\langle \boldsymbol{\phi}(x_i), \boldsymbol{\phi}\left(x_j\right) \right\rangle = \left(\boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{v}_i\right)' \left(\boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{v}_j\right) = \mathbf{v}_i' \boldsymbol{\Lambda} \mathbf{v}_j = \mathbf{K}_{ij} .$$

(thanks to Malte Kuss).

Page 34, line 5 "... and a mapping $\boldsymbol{\Lambda}$ into it ..." should read "... and a mapping $\boldsymbol{\phi}$ into it ..." (thanks to Petra Philips).

Page 34, line 9 "*the nth mapped object* x_n *is*" should read "*the point* $\sum_{i=1}^{r} \mathbf{U}_{in} \boldsymbol{\phi}(x_i) = \boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{U}' \mathbf{u}_n$ *is*". Accordingly, the following line changes to

$$\left\| \boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{U}' \mathbf{u}_n \right\|^2 = \mathbf{u}_n' \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}' \mathbf{u}_n = \mathbf{e}_n' \boldsymbol{\Lambda} \mathbf{e}_n = \lambda_n < 0 .$$

(thanks to Malte Kuss).

Page 42, equation (2.26) The second case term should read "$k_r(\boldsymbol{u}, \boldsymbol{v}) + \sum_{j=1}^{|\boldsymbol{v}|} \lambda^2 \cdot k_r'(u_1 \boldsymbol{u}, \boldsymbol{v}[j : |\boldsymbol{v}|])$" (thanks to Michael Davy).

Page 43, equation (2.30) The term $\lambda^{|\boldsymbol{v}|-j}$ should read $\lambda^{|\boldsymbol{v}|-t}$ (thanks to Vikas Sindhwani).

Page 63, line -11 "... is tighter for less sparse solutions." should read "... is tighter for more sparse solutions." (thanks to Diego Andres Alvarez Marin).

Page 77, Example 3.5 "$\mathbf{P}_{\mathsf{X}} = \text{Binomial}\,(n,\, p)$" should read "$\mathbf{P}_{\mathsf{X}|\mathsf{P}=p} = \text{Binomial}\,(n,\, p)$" (thanks to Jaz Kandola).

Page 84, line -5 "$C\,(x, \tilde{x}) = \left\langle \mathbf{x}, \tilde{\mathbf{x}} \right\rangle + \sigma_t^2 \mathbf{I}_{x \neq \tilde{x}} = k\,(x, \tilde{x}) + \sigma_t^2 \mathbf{I}_{x \neq \tilde{x}}$" should read "$C\,(x, \tilde{x}) = \left\langle \mathbf{x}, \tilde{\mathbf{x}} \right\rangle + \sigma_t^2 \mathbf{I}_{x=\tilde{x}} = k\,(x, \tilde{x}) + \sigma_t^2 \mathbf{I}_{x=\tilde{x}}$" (thanks to Arthur Gretton).

Page 86, line 17 "... on the found local maximum." should read "... on the local maximum found ." (thanks to Thore Graepel).

Page 86, line 20 "...is of the ability of ..." should read "...is the ability of ..." (thanks to Thore Graepel).

Page 87, Figure 3.2 "... of the 6 observations ..." should read "... of the 7 observations ...". Furthermore, "This local maxima ..." should read "This local maximum ..." (thanks to Thore Graepel).

Page 87, line -12 "$\boldsymbol{\sigma} \in \mathbb{R}^+$" should read "$\boldsymbol{\sigma} \in \left(\mathbb{R}^+\right)^N$".

Page 92, Figure 3.5 "... this liklihood is not normalizable." should read "... this likelihood is not normalizable." (thanks to Neil Lawrence).

Page 99, line -7 "... out be the single weight vector ..." should read "... out by the single weight vector ..." (thanks to Matthias Heiler).

Page 101, line 5 "... we first run learning algorithm to find ..." should read "... we first run a learning algorithm to find ..." (thanks to Matthias Heiler).

Page 110, line -11 Closing parenthesis at MacKay (1998) missing.

Page 110, line -6 Closing parenthesis at Barber and Williams (1997) missing.

Page 122, equation (4.7) The subscript should read "$R_{\text{emp}}\,[h, \mathbf{z}] = 0$" instead of "$R_{\text{emp}}\,[h] = 0$" (thanks to Petra Philips).

Page 123, line 13 "... real-valued loss functions conceptually similar ..." should read "... real-valued loss functions is conceptually similar ..." (thanks to Petra Philips).

Page 123, line -8 "... *for all* $\delta \in (0, 1]$*, and all training samples sizes m, with probability at least* $1 - \delta$ *over the random draw of the training sample* $\mathbf{z} \in \mathcal{Z}^m$ *we have* ..." should read "... *for the zero-one loss* l_{0-1} *given by equation (2.10) and all* $\varepsilon > 0$ *we have* ..." as in Theorem 4.7 (thanks to Simon Hill).

Page 124, line 10 "... (see equation (4.7)) ..." should read "... (see equation (4.8)) ..." (thanks to Jürgen Schweiger).

Page 125, line 5 "It we denote the maximum number ..." should read "If we denote the maximum number ..." (thanks to Petra Philips).

Page 126, enumeration 1 "If the function $\mathcal{N}_{\mathcal{H}}$ fulfills $\mathcal{N}_{\mathcal{H}}(m) = 2^m$ ..." should read "If the function $\mathcal{N}_{\mathcal{H}}$ fulfills $\mathcal{N}_{\mathcal{H}}(2m) = 2^{2m}$ ..." (thanks to Petra Philips).

Page 127, line 2: "*The best constants that can be achieved are* 2 *as a coefficient of and* 1 *in the exponent of the exponential term, respectively.*" should read "*The best constants that can be achieved for the coefficients of the exponent in the exponential term are* 2 *and* 1*, respectively.*" (thanks to Jaz Kandola).

Page 131, line 3 "Clearly, all functions in h are monotonically ..." should read "Clearly, all functions in $\mathcal{H}$ are monotonically ..." (thanks to Petra Philips).

Page 135, line -1 In the statement, "$R[h]$" should read "$R[h] - R_{\text{emp}}[h, z]$" (thanks to Petra Philips).

Page 137, line 12 The function ω is typed $\omega : \mathbb{N} \times \mathbb{R} \times [0, 1]$ rather than $\omega : \mathbb{R} \times [0, 1]$ because it depends on the training sample size m. Hence, the line

$$\mathbf{P}_{Z^{2m}}\left(\exists h \in \mathcal{H} : \ell_L(\mathbf{Z}, h) > \omega\left(L\left(\left(\mathsf{Z}_1, \ldots, \mathsf{Z}_m\right), h\right), \delta\right)\right) \leq \delta$$

should read

$$\mathbf{P}_{Z^{2m}}\left(\exists h \in \mathcal{H} : \ell_L(\mathbf{Z}, h) > \omega\left(m, L\left(\left(\mathsf{Z}_1, \ldots, \mathsf{Z}_m\right), h\right), \delta\right)\right) \leq \delta\,.$$

This change also appears in line 18, -6 and -3 on page 137, line 2 on page 138, line 17 on page 139 and line 7 on page 140 (thanks to Petra Philips!).

Page 140, line 5 "... in Appendix C.4 that $L(z, h) = -\vartheta_{\text{eff}}(z)$ ..." should read "... in Appendix C.4 that $L(z, h) = -\vartheta_{\mathcal{H}}(z)$ ..." (thanks to Petra Philips).

Page 175, Figure 5.1 "... with (solid line) and without (dashed line) ..." should read "... with (dashed line) and without (solid line) ..." (thanks to Dongwei Cao).

Page 182, Definition 5.20 The update function $\mathcal{U}$ maps from $\mathcal{Y} \times \mathcal{X} \times \mathcal{H} \to \mathcal{H}$ rather than $\mathcal{Y} \times \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$ as stated in the text. This changes the definition of an online learning algorithm. As a consequence, the sentence preceding the displayed equation ("... and the prediction of the current hypothesis $h_j \in \mathcal{H}$ ...") changes to "... and the current hypothesis $h_j \in \mathcal{H}$ ..." (thanks to Thore Graepel!).

Page 183, first equation "$\mathcal{U}(x, y, h(x)) = h$" changes to "$\mathcal{U}(x, y, h) = h$" (thanks to Thore Graepel).

Page 183, line 15 Similar to p. 177, line -5, "(compression function $\mathcal{C}_{\mathbf{i}}$)" should read "(compression function $\mathcal{C}_{|\mathbf{i}|}$)".

Page 184, line -8: The fourth point should read "*If all training examples are correctly classified, it outputs C and classifies according to (5.18).*". Furthermore, the last two sentences of this example are wrong and should be deleted (thanks to Thore Graepel).

Page 187, equation (5.19) This is a definition (similar to (2.14) at page 29) and not an equation.

Page 194, line 19 "preceeding Shawe-Taylor and Williamson (1997, p. 4)" should read "preceding Shawe-Taylor and Williamson (1997)".

Page 210, line -5 "... we also write $\mathbf{Y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$." should read "... we also write $\mathbf{X} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$." (thanks to Jian Huang).

Page 211, line -8 "... is a Gaussian measures, ..." should read "... are Gaussian measures, ..." (thanks to Petra Philips).

Page 217, line 5 In the definition of ℓ_p^n the second case should read $\max_{i=1,\ldots,n} |x_i| < \infty$ rather than $\max_{i=1,\ldots,n} |x_i|$ (thanks to Arthur Gretton).

Page 222, line 6 "... is the smallest number $\varepsilon > 0$ such that ..." should read "... is the smallest number $\varepsilon \geq 0$ such that ..." (thanks to Petra Philips).

Page 240, line -5: "$a \neq 0$" should read "$a \geq 0$" because otherwise exponentiation of $1 + a/x$ with x invalidates the inequality (thanks to Peter Bollmann-Sdorra).

Page 257, line 6 The braces must not include the summation over t (thanks to Vikas Sindhwani).

Page 282, line 3 In the statement on the l.h.s. $\mathbf{P}(A)$ should read $\mathbf{P}_{\mathsf{Z}}(A)$ (thanks to Petra Philips).

Page 283, Lemma C.2 There is a major flaw in the proof of this lemma. At the bottom of this page, it is argued that Theorem A.116 proves that the probability that a binomially distributed variable with mean of at least ε exceeds a value of $\frac{m\varepsilon}{2}$ is at least $\frac{1}{2}$. This is wrong. In order to prove this statement we need a different theorem. Thus, we will add the following theorem between Theorem A.116 and Definition A.117 at page 250.

Theorem 0.1 (Binomial mean deviation bound) *Let $\mathsf{X}_1, \ldots, \mathsf{X}_n$ be independent random variables such that, for all $i \in \{1, \ldots, n\}$, $\mathbf{P}_{\mathsf{X}_i}(\mathsf{X}_i = 1) = 1 - \mathbf{P}_{\mathsf{X}_i}(\mathsf{X}_i = 0) = \mathbf{E}_{\mathsf{X}_i}[\mathsf{X}_i] = \mu$. Then, for all $\alpha \in [0, 1]$ and for all $\lambda \in \mathbb{N}$ we*

have

$$\mathbf{P}_{X^n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\mu\right) > \frac{\lambda-1}{\lambda},$$

provided that $n\mu \geq \lambda\,(1-\alpha)^{-2}$.

Proof First note that the statement in the theorem is equivalent to

$$\mathbf{P}_{X^n}\left(\sum_{i=1}^{n} X_i \geq \alpha n\mu\right) > \frac{\lambda-1}{\lambda}.$$

By Theorem A.110 we know that

$$\mathbf{P}_{X^n}\left(\sum_{i=1}^{n} X_i \geq n\mu - \sqrt{\lambda n\mu\,(1-\mu)}\right) > \frac{\lambda-1}{\lambda},$$

because $\mathrm{Var}\left(\sum_i X_i\right) = n\mu\,(1-\mu)$ by the independence assumption and Theorem A.17. Thus, it suffices to show that $n\mu - \sqrt{\lambda n\mu\,(1-\mu)} \geq \alpha n\mu$ given that $n\mu \geq \lambda\,(1-\alpha)^{-2}$. However, by definition $\mu \in [0,1]$ and thus

$$\begin{aligned} n\mu &\geq \frac{\lambda}{(1-\alpha)^2}\,(1-\mu) \\ \sqrt{n\mu} &\geq \sqrt{\frac{\lambda}{(1-\alpha)^2}\,(1-\mu)} \\ (1-\alpha)\,n\mu &\geq \sqrt{\lambda n\mu\,(1-\mu)}. \end{aligned}$$

The theorem is proven. ■

Note that the statement at the bottom of page 283 now follows by setting $\alpha = \frac{1}{2}$ and $\lambda = 2$ provided the slightly stronger condition of $m\varepsilon \geq 8$ holds which is ensured in every application of this lemma throughout the book (thanks to Ulrich Kockelkorn, Mingrui Wu and Vu Ha for pointing this out mistake and helping with the additional proof).

Page 283, line 13 "... $\epsilon\,(z) \wedge \mathbf{v}_z\,(A\,(z)) = 0$ if such a set exists ..." should read "... $\epsilon \wedge \mathbf{v}_z\,(A\,(z)) = 0$ if such a set exists ..." (thanks to Petra Philips).

Page 285, line -8 "... given in equation (22) and ..." should read "... given in equation (2.11) and ..." (thanks to Petra Philips).

Page 302, Section C.8 The section heading should read ”A PAC-Bayesian Margin Bound” rather than ”A PAC-Bayesian Marin Bound” (thanks to Thore Graepel).

Page 303, line 4 In the numerator, the integration is up to π rather than 2π (thanks to John Shawe-Taylor).

Page 340 The entry Bartlett and Shawe-Taylor (1998) has the wrong year and is identical to Bartlett and Shawe-Taylor (1999).

Page 340 In the entry Bennett (1998) there is one extra “19” in the paper title (thanks to Jaz Kandola).

Page 347 In the entry Lauritzen (1981) “t. n. thiele” should read “T. N. Thiele” (thanks to Jaz Kandola).

Page 349 In the entry Neal (1997b) “Technical Report” is spelt twice (thanks to Jaz Kandola).

Page 350 In the entry Robert (1994) “Ney York” should read “New York” (thanks to Jaz Kandola).

Page 354 The entry Watkins (1998) contains (almost) the same content as Watkins (2000) and will be eliminated in future editions.

Page 355 In the entry Williams and Seeger (2001), “nystrom” should read “Nyström” (thanks to Jaz Kandola).