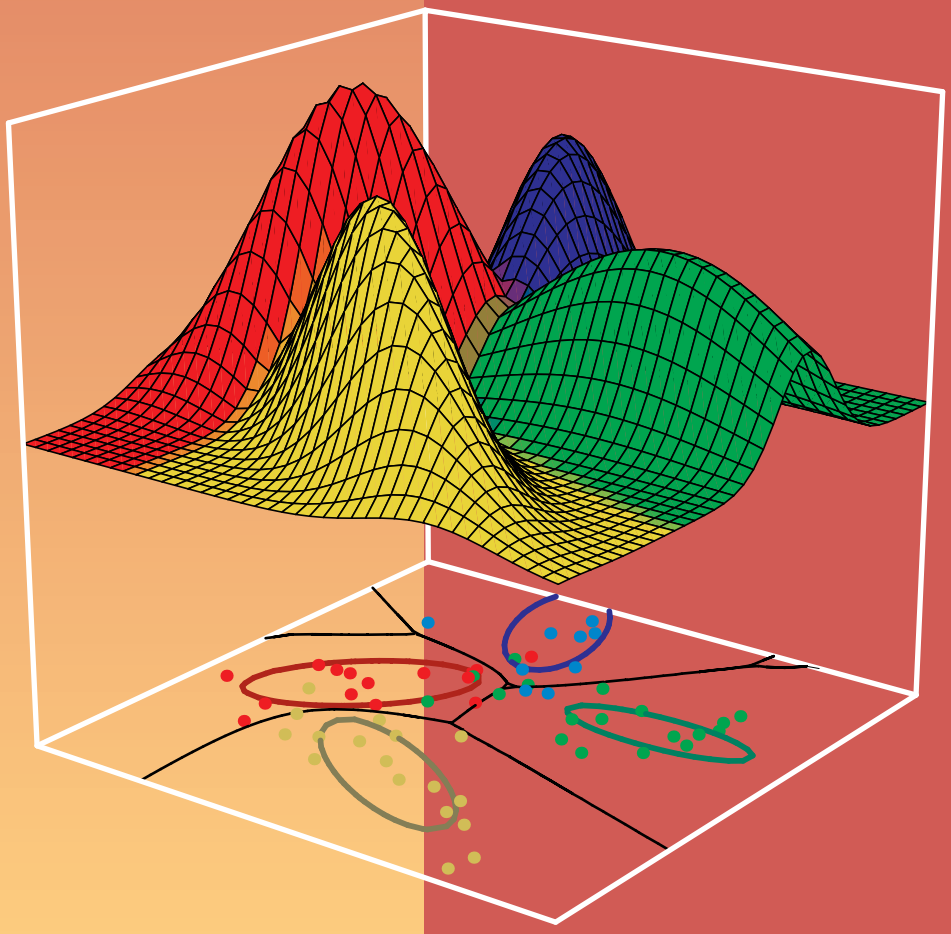
Solution Manual
to accompany

Pattern Classification

(2nd ed.)

David G. Stork



Solution Manual to accompany

Pattern Classification (2nd ed.)

by R. O. Duda, P. E. Hart and D. G. Stork

David G. Stork

June 18, 2003

Preface

In writing this **Solution Manual** I have learned a very important lesson. As a student, I thought that the best way to master a subject was to go to a superb university and study with an established expert. Later, I realized instead that the best way was to teach a course on the subject. Yet later, I was convinced that the best way was to write a detailed and extensive textbook. Now I know that all these years I have been wrong: in fact the best way to master a subject is to write the **Solution Manual**.

In solving the problems for this **Manual** I have been forced to confront myriad technical details that might have tripped up the unsuspecting student. Students and teachers can thank me for simplifying or screening out problems that required pages of unenlightening calculations. Occasionally I had to go back to the text and delete the word "easily" from problem references that read "it can easily be shown (Problem ...)." Throughout, I have tried to choose data or problem conditions that are particularly instructive. In solving these problems, I have found errors in early drafts of this text (as well as errors in books by other authors and even in classic refereed papers), and thus the accompanying text has been improved for the writing of this **Manual**.

I have tried to make the problem solutions self-contained and self-explanatory. I have gone to great lengths to ensure that the solutions are correct and clearly presented — many have been reviewed by students in several classes. Surely there are errors and typos in this manuscript, but rather than editing and rechecking these solutions over months or even years, I thought it best to distribute the **Manual**, however flawed, as early as possible. I accept responsibility for these inevitable errors, and humbly ask anyone finding them to contact me directly. (Please, however, do not ask me to *explain* a solution or help you solve a problem!) It should be a small matter to change the **Manual** for future printings, and you should contact the publisher to check that you have the most recent version. Notice, too, that this **Manual** contains a list of known typos and errata in the text which you might wish to photocopy and distribute to students.

I have tried to be thorough in order to help students, even to the occassional fault of verbosity. You will notice that several problems have the simple "explain your answer in words" and "graph your results." These were added for students to gain intuition and a deeper understanding. Graphing per se is hardly an intellectual challenge, but if the student graphs functions, he or she will develop intuition and remember the problem and its results better. Furthermore, when the student later sees graphs of data from dissertation or research work, the link to the homework problem and the material in the text will be more readily apparent. Note that due to the vagaries of automatic typesetting, figures may appear on pages after their reference in this **Manual**; be sure to consult the full solution to any problem.

I have also included worked examples and so sample final exams with solutions to

cover material in text. I distribute a list of important equations (without descriptions) with the exam so students can focus understanding and using equations, rather than memorizing them. I also include on every final exam one problem *verbatim* from a homework, taken from the book. I find this motivates students to review carefully their homework assignments, and allows somewhat more difficult problems to be included. These will be updated and expanded; thus if you have exam questions you find particularly appropriate, and would like to share them, please send a copy (with solutions) to me.

It should be noted, too, that a set of overhead transparency masters of the figures from the text are available to faculty adopters. I have found these to be invaluable for lecturing, and I put a set on reserve in the library for students. The files can be accessed through a standard web browser or an ftp client program at the Wiley STM ftp area at:

`ftp://ftp.wiley.com/public/sci_tech_med/pattern/`

or from a link on the Wiley Electrical Engineering software supplements page at:

`http://www.wiley.com/products/subject/engineering/electrical/software_supplem_elec_eng.html`

I have taught from the text (in various stages of completion) at the University of California at Berkeley (Extension Division) and in three Departments at Stanford University: Electrical Engineering, Statistics and Computer Science. Numerous students and colleagues have made suggestions. Especially noteworthy in this regard are Sudeshna Adak, Jian An, Sung-Hyuk Cha, Koichi Ejiri, Rick Guadette, John Heumann, Travis Kopp, Yaxin Liu, Yunqian Ma, Sayan Mukherjee, Hirobumi Nishida, Erhan Oztop, Steven Rogers, Charles Roosen, Sergio Bermejo Sanchez, Godfried Toussaint, Namrata Vaswani, Mohammed Yousuf and Yu Zhong. Thanks too go to Dick Duda who gave several excellent suggestions.

I would greatly appreciate notices of any errors in this **Manual** or the text itself. I would be *especially* grateful for solutions to problems not yet solved. Please send any such information to me at the below address. I will incorporate them into subsequent releases of this **Manual**.

This **Manual** is for the use of educators and must not be distributed in bulk to students in any form. Short excerpts may be photocopied and distributed, but only in conjunction with the use of **Pattern Classification** (2nd ed.).

I wish you all the best of luck in teaching and research.

David G. Stork

Ricoh Innovations, Inc.
2882 Sand Hill Road Suite 115
Menlo Park, CA 94025-7022 USA
`stork@rii.ricoh.com`

Contents

Chapter 1

Introduction

Problem Solutions

There are neither problems nor computer exercises in Chapter 1.

Chapter 2

Bayesian decision theory

Problem Solutions

Section 2.1

1. Equation 7 in the text states

$$P(error|x) = \min[P(\omega_1|x), P(\omega_2|x)].$$

(a) We assume, without loss of generality, that for a given particular x we have $P(\omega_2|x) \geq P(\omega_1|x)$, and thus $P(error|x) = P(\omega_1|x)$. We have, moreover, the normalization condition $P(\omega_1|x) = 1 - P(\omega_2|x)$. Together these imply $P(\omega_2|x) > 1/2$ or $2P(\omega_2|x) > 1$ and

$$2P(\omega_2|x)P(\omega_1|x) > P(\omega_1|x) = P(error|x).$$

This is true at every x, and hence the integrals obey

$$\int 2P(\omega_2|x)P(\omega_1|x)dx \geq \int P(error|x)dx.$$

In short, $2P(\omega_2|x)P(\omega_1|x)$ provides an upper bound for $P(error|x)$.

(b) From part (a), we have that $P(\omega_2|x) > 1/2$, but in the current conditions not greater than $1/\alpha$ for $\alpha < 2$. Take as an example, $\alpha = 4/3$ and $P(\omega_1|x) = 0.4$ and hence $P(\omega_2|x) = 0.6$. In this case, $P(error|x) = 0.4$. Moreover, we have

$$\alpha P(\omega_1|x)P(\omega_2|x) = 4/3 \times 0.6 \times 0.4 < P(error|x).$$

This does not provide an upper bound for all values of $P(\omega_1|x)$.

(c) Let $P(error|x) = P(\omega_1|x)$. In that case, for all x we have

$$\begin{aligned} P(\omega_2|x)P(\omega_1|x) &< P(\omega_1|x)P(error|x) \\ \int P(\omega_2|x)P(\omega_1|x)dx &< \int P(\omega_1|x)P(error|x)dx, \end{aligned}$$

and we have a lower bound.

(d) The solution to part (b) also applies here.

Section 2.2

2. We are given that the density is of the form $p(x|\omega_i) = ke^{-|x-a_i|/b_i}$.

(a) We seek k so that the function is normalized, as required by a true density. We integrate this function, set it to 1.0,

$$k\left[\int_{-\infty}^{a_i} \exp[(x-a_i)/b_i]dx + \int_{a_i}^{\infty} \exp[-(x-a_i)/b_i]dx\right] = 1,$$

which yields $2b_ik = 1$ or $k = 1/(2b_i)$. Note that the normalization is independent of a_i, which corresponds to a shift along the axis and is hence indeed irrelevant to normalization. The distribution is therefore written

$$p(x|\omega_i) = \frac{1}{2b_i}e^{-|x-a_i|/b_i}.$$

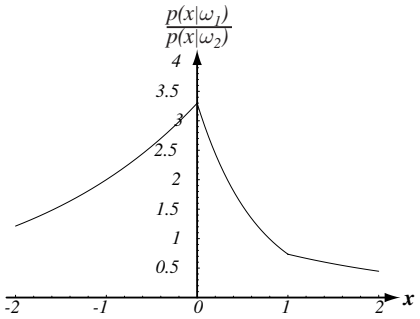
(b) The likelihood ratio can be written directly:

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} = \frac{b_2}{b_1}\exp\left[-\frac{|x-a_1|}{b_1} + \frac{|x-a_2|}{b_2}\right].$$

(c) For the case $a_1 = 0$, $a_2 = 1$, $b_1 = 1$ and $b_2 = 2$, we have the likelihood ratio is

$$\frac{p(x|\omega_2)}{p(x|\omega_1)} = \begin{cases} 2e^{(x+1)/2} & x \leq 0 \\ 2e^{(1-3x)/2} & 0 < x \leq 1 \\ 2e^{(-x-1)/2} & x > 1, \end{cases}$$

as shown in the figure.



Section 2.3

3. We are are to use the standard zero-one classification cost, that is $\lambda_{11} = \lambda_{22} = 0$ and $\lambda_{12} = \lambda_{21} = 1$.

(a) We have the priors $P(\omega_1)$ and $P(\omega_2) = 1 - P(\omega_1)$. The Bayes risk is given by Eqs. 12 and 13 in the text:

$$R(P(\omega_1)) = P(\omega_1) \int_{\mathcal{R}_2} p(x|\omega_1)dx + (1 - P(\omega_1)) \int_{\mathcal{R}_1} p(x|\omega_2)dx.$$

To obtain the prior with the minimum risk, we take the derivative with respect to $P(\omega_1)$ and set it to 0, that is

$$\frac{d}{dP(\omega_1)} R(P(\omega_1)) = \int_{\mathcal{R}_2} p(x|\omega_1)dx - \int_{\mathcal{R}_1} p(x|\omega_2)dx = 0,$$

which gives the desired result:

$$\int_{\mathcal{R}_2} p(x|\omega_1)dx = \int_{\mathcal{R}_1} p(x|\omega_2)dx.$$

(b) This solution is not always unique, as shown in this simple counterexample. Let $P(\omega_1) = P(\omega_2) = 0.5$ and

$$\begin{aligned} p(x|\omega_1) &= \begin{cases} 1 & -0.5 \le x \le 0.5 \\ 0 & \text{otherwise} \end{cases} \\ p(x|\omega_2) &= \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

It is easy to verify that the decision regions $\mathcal{R}_1 = [-0.5, 0.25]$ and $\mathcal{R}_1 = [0, 0.5]$ satisfy the equations in part (a); thus the solution is not unique.

4. Consider the minimax criterion for a two-category classification problem.

(a) The total risk is the integral over the two regions $\mathcal{R}_i$ of the posteriors times their costs:

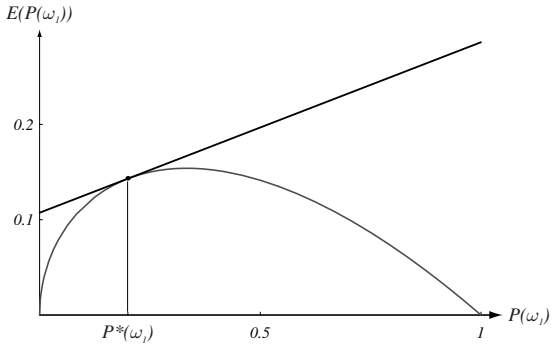
$$\begin{aligned} R &= \int_{\mathcal{R}_1} [\lambda_{11} P(\omega_1) p(\mathbf{x}|\omega_1) + \lambda_{12} P(\omega_2) p(\mathbf{x}|\omega_2)] \; d\mathbf{x} \\ &\quad + \int_{\mathcal{R}_2} [\lambda_{21} P(\omega_1) p(\mathbf{x}|\omega_1) + \lambda_{22} P(\omega_2) p(\mathbf{x}|\omega_2)] \; d\mathbf{x}. \end{aligned}$$

We use $\int_{\mathcal{R}_2} p(\mathbf{x}|\omega_2) \; d\mathbf{x} = 1 - \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) \; d\mathbf{x}$ and $P(\omega_2) = 1 - P(\omega_1)$, regroup to find:

$$\begin{aligned} R &= \lambda_{22} + \lambda_{12} \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) \; d\mathbf{x} - \lambda_{22} \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) \; d\mathbf{x} \\ &\quad + P(\omega_1) \Bigg[(\lambda_{11} - \lambda_{22}) + \lambda_{11} \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) \; d\mathbf{x} - \lambda_{12} \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) \; d\mathbf{x} \\ &\quad + \lambda_{21} \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) \; d\mathbf{x} + \lambda_{22} \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) \; d\mathbf{x} \Bigg] \end{aligned}$$

$$
\begin{aligned}
= \quad & \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)\, d\mathbf{x} \\
& + P(\omega_1) \Bigg[(\lambda_{11} - \lambda_{22}) + (\lambda_{11} + \lambda_{21}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)\, d\mathbf{x} \\
& \qquad + (\lambda_{22} - \lambda_{12}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)\, d\mathbf{x} \Bigg].
\end{aligned}
$$

(b) Consider an arbitrary prior $0 < P^*(\omega_1) < 1$, and assume the decision boundary has been set so as to achieve the minimal (Bayes) error for that prior. If one holds the same decision boundary, but changes the prior probabilities (i.e., $P(\omega_1)$ in the figure), then the error changes *linearly*, as given by the formula in part (a). The true Bayes error, however, must be less than or equal to that (linearly bounded) value, since one has the freedom to change the decision boundary at each value of $P(\omega_1)$. Moreover, we note that the Bayes error is 0 at $P(\omega_1) = 0$ and at $P(\omega_1) = 1$, since the Bayes decision rule under those conditions is to always decide ω_2 or ω_1, respectively, and this gives zero error. Thus the curve of Bayes error rate is concave down for all prior probabilities.



(c) According to the general minimax equation in part (a), for our case (i.e., $\lambda_{11} = \lambda_{22} = 0$ and $\lambda_{12} = \lambda_{21} = 1$) the decision boundary is chosen to satisfy

$$\int_{\mathcal{R}_2} p(x|\omega_1)\, dx = \int_{\mathcal{R}_1} p(x|\omega_2)\, dx.$$

We assume that a *single* decision point suffices, and thus we seek to find x^* such that

$$\int_{-\infty}^{x^*} N(\mu_1, \sigma_1^2)\, dx = \int_{x^*}^{\infty} N(\mu_2, \sigma_2^2)\, dx,$$

where, as usual, $N(\mu_i, \sigma_i^2)$ denotes a Gaussian. We assume for definiteness and without loss of generality that $\mu_2 > \mu_1$, and that the single decision point lies between the means. Recall the definition of an error function, given by Eq. 96

in the Appendix of the text, that is,

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

We can rewrite the above as

$$\mathrm{erf}[(x^* - \mu_1)/\sigma_1] = \mathrm{erf}[(x^* - \mu_2)/\sigma_2].$$

If the values of the error function are equal, then their corresponding arguments must be equal, that is

$$(x^* - \mu_1)/\sigma_1 = (x^* - \mu_2)/\sigma_2$$

and solving for x^* gives the value of the decision point

$$x^* = \left(\frac{\mu_2\sigma_1 + \mu_1\sigma_2}{\sigma_1 + \sigma_2} \right).$$

(d) Because the mimimax error rate is independent of the prior probabilities, we can choose a particularly simple case to evaluate the error, for instance, $P(\omega_1) = 0$. In that case our error becomes

$$E = 1/2 - \mathrm{erf}[(x^* - \mu_1)/\sigma_1] = 1/2 - \mathrm{erf}\left[\frac{\mu_2\sigma_1 - \mu_1\sigma_2}{\sigma_1(\sigma_1 + \sigma_2)} \right].$$

(e) We substitute the values given in the problem into the formula in part (c) and find

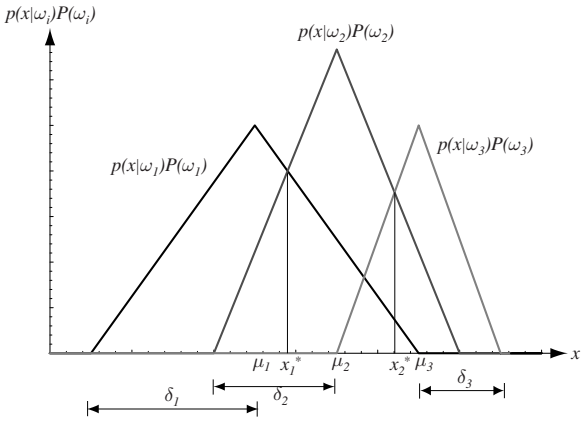
$$x^* = \frac{\mu_2\sigma_1 + \mu_1\sigma_2}{\sigma_1 + \sigma_2} = \frac{1/2 + 0}{1 + 1/2} = 1/3.$$

The error from part (d) is then

$$E = 1/2 - \mathrm{erf}\left[\frac{1/3 - 0}{1 + 0} \right] = 1/2 - \mathrm{erf}[1/3] = 0.1374.$$

(f) Note that the distributions have the same form (in particular, the same variance). Thus, by symmetry the Bayes error for $P(\omega_1) = P^*$ (for some value P^*) must be the same as for $P(\omega_2) = P^*$. Because $P(\omega_2) = 1 - P(\omega_1)$, we know that the curve, analogous to the one in part (b), is symmetric around the point $P(\omega_1) = 0.5$. Because the curve is concave down, therefore it must *peak* at $P(\omega_1) = 0.5$, that is, equal priors. The tangent to the graph of the error versus $P(\omega_1)$ is thus horizontal at $P(\omega_1) = 0.5$. For this case of equal priors, the Bayes decision point for this problem can be stated simply: it is the point midway between the means of the two distributions, that is, $x^* = 5.5$.

5. We seek to generalize the notion of minimax criteria to the case where *two* independent prior probabilities are set.



(a) We use the triangle distributions and conventions in the figure. We solve for the decision points as follows (being sure to keep the signs correct, and assuming that the decision boundary consists of just two points):

$$P(\omega_1)\left(\frac{\delta_1-(x_1^*-\mu_1)}{\delta_1^2}\right)=P(\omega_2)\left(\frac{\delta_2-(\mu_2-x_1^*)}{\delta_2^2}\right),$$

which has solution

$$x_1^*=\frac{P(\omega_1)\delta_2^2\delta_1+P(\omega_1)\delta_2^2\mu_1-P(\omega_2)\delta_1^2\delta_2+P(\omega_2)\mu_2\delta_1^2}{P(\omega_1)\delta_2^2+P(\omega_2)\delta_1^2}.$$

An analogous derivation for the other decision point gives:

$$P(\omega_2)\left(\frac{\delta_2-(x_2^*-\mu_2)}{\delta_2^2}\right)=P(\omega_3)\left(\frac{\delta_3-(x_2^*-\mu_3)}{\delta_3^2}\right),$$

which has solution

$$x_2^*=\frac{-P(\omega_2)\delta_3^2\mu_2+P(\omega_2)\delta_3^2\delta_2+P(\omega_3)\delta_2^2\delta_3+P(\omega_3)\delta_2^2\mu_3}{P(\omega_2)\delta_3^2+P(\omega_3)\delta_2^2}.$$

(b) Note that from our normalization condition, $\sum_{i=1}^{3}P(\omega_i)=1$, we can express all priors in terms of just *two* independent ones, which we choose to be $P(\omega_1)$ and $P(\omega_2)$. We could substitute the values for x_i^* and integrate, but it is just a bit simpler to go directly to the calculation of the error, E, as a function of priors $P(\omega_1)$ and $P(\omega_2)$ by considering the four contributions:

$$\begin{aligned}E &= P(\omega_1)\frac{1}{2\delta_1^2}\left[\mu_1+\delta_1-x_1^*\right]^2\\ &\quad+P(\omega_2)\frac{1}{2\delta_2^2}\left[\delta_2-\mu_2+x_1^*\right]^2\\ &\quad+P(\omega_2)\frac{1}{2\delta_2^2}\left[\mu_2+\delta_2-x_2^*\right]^2\\ &\quad+\Big[\underbrace{1-P(\omega_1)-P(\omega_2)}_{P(\omega_3)}\Big]\frac{1}{2\delta_3^2}\left[\delta_3-\mu_3+x_2^*\right]^2.\end{aligned}$$

To obtain the minimax solution, we take the two partial and set them to zero. The first of the derivative equations,

$$\frac{\partial E}{\partial P(\omega_1)} = \quad 0,$$

yields the equation

$$\begin{aligned}\left(\frac{\mu_1+\delta_1-x_1^*}{\delta_1}\right)^2 &= \left(\frac{\delta_3-\mu_3+x_2^*}{\delta_3}\right)^2 \text{ or}\\ \frac{\mu_1+\delta_1-x_1^*}{\delta_1} &= \frac{\delta_3-\mu_3+x_2^*}{\delta_3}.\end{aligned}$$

Likewise, the second of the derivative equations,

$$\frac{\partial E}{\partial P(\omega_2)} = 0,$$

yields the equation

$$\left(\frac{\delta_2-\mu_2+x_1^*}{\delta_2}\right)^2+\left(\frac{\mu_2+\delta_2-x_2^*}{\delta_2}\right)^2=\left(\frac{\delta_3-\mu_3+x_2^*}{\delta_3}\right)^2.$$

These simultaneous quadratic equations have solutions of the general form:

$$x_i^* = \frac{b_i+\sqrt{c_i}}{a_i} \qquad i=1,2.$$

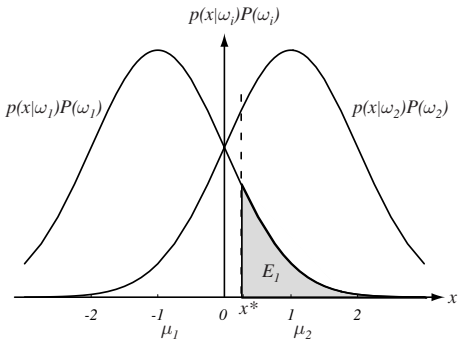
After a straightforward, but very tedious calculation, we find that:

$$\begin{aligned}
a_1 &= \delta_1^2-\delta_2^2+\delta_3^2,\\
b_1 &= -\delta_1^2\delta_2-\delta_1\delta_2^2-\delta_1\delta_2\ \delta_3-\delta_2^2\mu_1+\delta_3^2\mu_1+\delta_1^2\mu_2-\delta_1\delta_3\mu_2+\delta_1\delta_3\mu_3,\\
c_1 &= \delta_1^2\big(2\delta_1\delta_2^3+2\delta_2^4+2\delta_1\delta_2^2\delta_3+2\delta_2^3\delta_3+\delta_1\delta_2^2\mu_1+2\delta_2^3\mu_1\\
&\quad +2\,\delta_1\,\delta_2\,\delta_3\,\mu_1-2\,\delta_2\,{\delta_3}^2\,\mu_1+{\delta_2}^2\,{\mu_1}^2-{\delta_3}^2\,{\mu_1}^2-2\,{\delta_1}^2\,\delta_2\,\mu_2-2\,\delta_1\,{\delta_2}^2\,\mu_2\\
&\quad +2\,{\delta_2}^2\,\delta_3\,\mu_2+2\,\delta_2\,{\delta_3}^2\,\mu_2-2\,{\delta_2}^2\,\mu_1\,\mu_2+2\,\delta_1\,\delta_3\,\mu_1\,\mu_2+2\,{\delta_3}^2\,\mu_1\,\mu_2\\
&\quad -{\delta_1}^2\,{\mu_2}^2+2\,{\delta_2}^2\,{\mu_2}^2-2\,\delta_1\,\delta_3\,{\mu_2}^2-{\delta_3}^2\,{\mu_2}^2+2\,{\delta_1}^2\,\delta_2\,\mu_3-2\,{\delta_2}^3\,\mu_3\\
&\quad -2\,\delta_1\,\delta_2\,\delta_3\,\mu_3-2\,{\delta_2}^2\,\delta_3\,\mu_3-2\,\delta_1\,\delta_3\,\mu_1\,\mu_3+2\,{\delta_1}^2\,\mu_2\,\mu_3-2\,{\delta_2}^2\,\mu_2\,\mu_3\\
&\quad +2\,\delta_1\,\delta_3\,\mu_2\,\mu_3-{\delta_1}^2\,{\mu_3}^2+{\delta_2}^2\,{\mu_3}^2\big).
\end{aligned}$$

An analogous calculation gives:

$$\begin{aligned}
a_2 &= \delta_1^2-\delta_2^2+\delta_3^2,\\
b_2 &= \delta_1\,\delta_2\,\delta_3+\delta_2^2\,\delta_3+2\,\delta_2\,\delta_3^2+\delta_1\,\delta_3\,\mu_1-\delta_1\,\delta_3\,\mu_2+\delta_3^2\,\mu_2+\delta_1^2\,\mu_3-\delta_2^2\,\mu_3,\\
c_2 &= \big(\delta_1^2-\delta_2^2+\delta_3^2\big)\times\\
&\quad \big(\delta_2^2\,\delta_3^2+2\,\delta_2\,\delta_3^2\,\mu_1+\delta_3^2\,\mu_1^2-2\,\delta_3^2\,\mu_1\,\mu_2+2\,\delta_3^2\,\mu_2^2+2\,\delta_1\,\delta_2\,\delta_3\,\mu_3\\
&\quad +2\,\delta_2^2\,\delta_3\,\mu_3+2\,\delta_1\,\delta_3\,\mu_1\,\mu_3-2\,\delta_1\,\delta_3\,\mu_2\,\mu_3+\delta_1^2\,\mu_3^2-\delta_2^2\,\mu_3^2\big).
\end{aligned}$$

(c) For $\{\mu_i,\delta_i\}=\{0,1\},\{.5,.5\},\{1,1\}$, for $i=1,2,3$, respectively, we substitute into the above equations to find $x_1^*=0.2612$ and $x_2^*=0.7388$. It is a simple matter to confirm that indeed these two decision points suffice for the classification problem, that is, that no more than two points are needed.



6. We let x^* denote our decision boundary and $\mu_2 > \mu_1$, as shown in the figure.

(a) The error for classifying a pattern that is actually in ω_1 as if it were in ω_2 is:

$$\int_{\mathcal{R}_2} p(x|\omega_1)P(\omega_1)\ dx = \frac{1}{2}\int_{x^*}^{\infty} N(\mu_1, \sigma_1^2)\ dx \leq E_1.$$

Our problem demands that this error be less than or equal to E_1. Thus the bound on x^* is a function of E_1, and could be obtained by tables of cumulative normal distributions, or simple numerical integration.

(b) Likewise, the error for categorizing a pattern that is in ω_2 as if it were in ω_1 is:

$$E_2 = \int_{\mathcal{R}_1} p(x|\omega_2)P(\omega_2)\ dx = \frac{1}{2}\int_{-\infty}^{x^*} N(\mu_2, \sigma_2^2)\ dx.$$

(c) The total error is simply the sum of these two contributions:

$$\begin{aligned} E &= E_1 + E_2 \\ &= \frac{1}{2}\int_{x^*}^{\infty} N(\mu_1, \sigma_1^2)\ dx + \frac{1}{2}\int_{-\infty}^{x^*} N(\mu_2, \sigma_2^2)\ dx. \end{aligned}$$

(d) For $p(x|\omega_1) \sim N(-1/2, 1)$ and $p(x|\omega_2) \sim N(1/2, 1)$ and $E_1 = 0.05$, we have (by simple numerical integration) $x^* = 0.2815$, and thus

$$\begin{aligned} E &= 0.05 + \frac{1}{2}\int_{-\infty}^{0.2815} N(\mu_2, \sigma_2^2)\ dx \\ &= 0.05 + \frac{1}{2}\int_{-\infty}^{0.2815} \frac{1}{\sqrt{2\pi}0.05}\exp\left[-\frac{(x-0.5)^2}{2(0.5)^2}\right] dx \\ &= 0.168. \end{aligned}$$

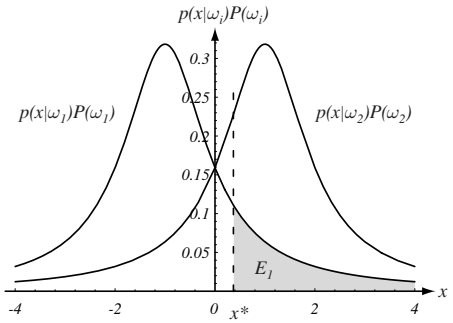
(e) The decision boundary for the (minimum error) Bayes case is clearly at $x^* = 0$. The Bayes error for this problem is:

$$E_B = 2\int_0^{\infty} \frac{1}{2} N(\mu_1, \sigma_1^2)\ dx$$

$$= \int_0^\infty N(1,1)\, dx = \text{erf}[1] = 0.159,$$

which of course is lower than the error for the Neyman-Pearson criterion case. Note that if the Bayes error were lower than $2 \times 0.05 = 0.1$ in this problem, we would use the Bayes decision point for the Neyman-Pearson case, since it too would ensure that the Neyman-Pearson criteria were obeyed *and* would give the lowest total error.

7. We proceed as in Problem 6, with the figure below.



(a) Recall that the Cauchy density is

$$p(x|\omega_i) = \frac{1}{\pi b} \frac{1}{1 + \left(\frac{x - a_i}{b}\right)^2}.$$

If we denote our (single) decision boundary point as x^*, and note that $P(\omega_i) = 1/2$, then the error for misclassifying a ω_1 pattern as ω_2 is:

$$\begin{aligned} E_1 &= \int_{x^*}^\infty p(x|\omega_1) P(\omega_1)\, dx \\ &= \frac{1}{2} \int_{x^*}^\infty \frac{1}{\pi b} \frac{1}{1 + \left(\frac{x - a_1}{b}\right)^2}\, dx. \end{aligned}$$

We substitute $(x - a_1)/b = y$, and $\sin\theta = 1/\sqrt{1 + y^2}$ to get:

$$\begin{aligned} E_1 &= \frac{1}{2\pi} \int_{\theta = \tilde{\theta}}^{\theta = 0} d\theta \\ &= \frac{1}{2\pi} \sin^{-1}\left[\frac{b}{\sqrt{b^2 + (x^* - a_1)^2}}\right], \end{aligned}$$

where $\tilde{\theta} = \sin^{-1}\left[\frac{b}{\sqrt{b^2 + (x^* - a_1)^2}}\right]$. Solving for the decision point gives

$$x^* = a_1 + b\sqrt{\frac{1}{\sin^2[2\pi E_1]} - 1} = a_1 + b/\tan[2\pi E_1].$$

(b) The error for the converse case is found similarly:

$$\begin{aligned} E_2 &= \frac{1}{\pi b}\int_{-\infty}^{x^*} \frac{1}{1+\left(\frac{x-a_2}{b}\right)^2} P(\omega_2)\, dx \\ &= \frac{1}{2\pi}\int_{\theta=-\pi}^{\theta=\tilde{\theta}} d\theta \\ &= \frac{1}{2\pi}\left\{\sin^{-1}\left[\frac{b}{\sqrt{b^2+(x^*-a_2)^2}}\right]+\pi\right\} \\ &= \frac{1}{2}+\frac{1}{\pi}\sin^{-1}\left[\frac{b}{\sqrt{b^2+(x^*-a_2)^2}}\right], \end{aligned}$$

where $\tilde{\theta}$ is defined in part (a).

(c) The total error is merely the sum of the component errors:

$$E = E_1 + E_2 = E_1 + \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left[\frac{b}{\sqrt{b^2+(x^*-a_2)^2}}\right],$$

where the numerical value of the decision point is

$$x^* = a_1 + b/\tan[2\pi E_1] = 0.376.$$

(d) We add the errors (for $b = 1$) and find

$$E = 0.1 + \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left[\frac{b}{\sqrt{b^2+(x^*-a_2)^2}}\right] = 0.2607.$$

(e) For the Bayes case, the decision point is midway between the peaks of the two distributions, i.e., at $x^* = 0$ (cf. Problem 6). The Bayes error is then

$$E_B = 2\int_0^{\infty} \frac{1}{1+\left(\frac{x-a}{b}\right)^2} P(\omega_2)\, dx = 0.2489.$$

This is indeed lower than for the Neyman-Pearson case, as it must be. Note that if the Bayes error were lower than $2 \times 0.1 = 0.2$ in this problem, we would use the Bayes decision point for the Neyman-Pearson case, since it too would ensure that the Neyman-Pearson criteria were obeyed *and* would give the lowest total error.

8. Consider the Cauchy distribution.

(a) We let k denote the integral of $p(x|\omega_i)$, and check the normalization condition, that is, whether $k = 1$:

$$k = \int_{-\infty}^{\infty} p(x|\omega_i)\, dx = \frac{1}{\pi b}\int_{-\infty}^{\infty} \frac{1}{1+\left(\frac{x-a_i}{b}\right)^2}\, dx.$$

We substitute $(x - a_i)/b = y$ into the above and get

$$k = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{1+y^2}\, dy,$$

and use the trigonometric substition $1/\sqrt{1+y^2} = \sin\,\theta$, and hence $dy = d\theta/\sin^2\theta$ to find

$$k = \frac{1}{\pi} \int_{\theta=-\pi}^{\theta=0} \frac{\sin^2\theta}{\sin^2\theta}\, d\theta = 1.$$

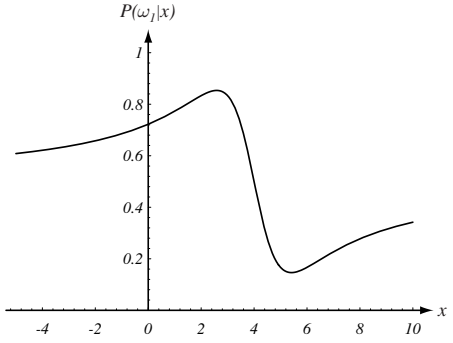
Indeed, $k = 1$, and the distribution is normalized.

(b) We let x^* denote the decision boundary (a single point) and find its value by setting $p(x^*|\omega_1)P(\omega_1) = p(x^*|\omega_2)P(\omega_2)$. We have then

$$\frac{1}{\pi b}\frac{1}{1+\left(\frac{x^*-a_1}{b}\right)^2}\frac{1}{2} = \frac{1}{\pi b}\frac{1}{1+\left(\frac{x^*-a_2}{b}\right)^2}\frac{1}{2},$$

or $(x^* - a_1) = \pm(x^* - a_2)$. For $a_1 \neq a_2$, this implies that $x^* = (a_1 + a_2)/2$, that is, the decision boundary is midway between the means of the two distributions.

(c) For the values $a_1 = 3, a_2 = 5$ and $b = 1$, we get the graph shown in the figure.



(d) We substitute the form of $P(\omega_i|x)$ and $p(x|\omega_i)$ and find

$$\begin{aligned}\lim_{x\to\infty} P(\omega_i|x) &= \lim_{x\to\infty} \frac{\frac{1}{2}\left[\frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_i}{b}\right)^2}\right]}{\left[\frac{1}{2}\left[\frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2}\right] + \frac{1}{2}\left[\frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}\right]\right]} \\ &= \lim_{x\to\infty} \frac{b^2 + (x-a_i)^2}{b^2 + (x-a_1)^2 + b^2 + (x-a_2)^2} = \frac{1}{2},\end{aligned}$$

and likewise, $\lim_{x\to-\infty} P(\omega_i|x) = 1/2$, as can be confirmed in the figure.

9. We follow the terminology in Section 2.3 in the text.

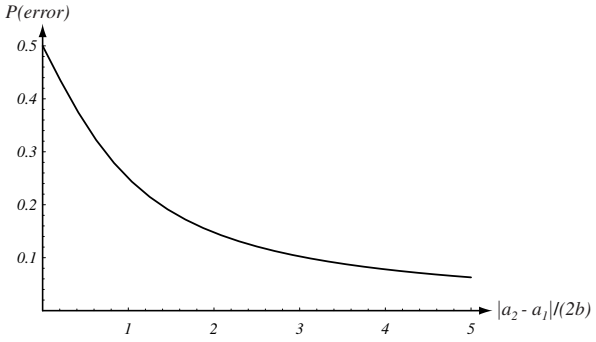
(a) Without loss of generality, we assume that $a_2 > a_1$, note that the decision boundary is at $(a_1 + a_2)/2$. The probability of error is given by

$$\begin{aligned} P(error) &= \int_{-\infty}^{(a_1+a_2)/2} p(\omega_2|x)dx + \int_{(a_1+a_2)/2}^{\infty} p(\omega_1|x)dx \\ &= \frac{1}{\pi b}\int_{-\infty}^{(a_1+a_2)/2} \frac{1/2}{1+\left(\frac{x-a_2}{b}\right)^2}\,dx + \frac{1}{\pi b}\int_{(a_1+a_2)/2}^{\infty} \frac{1/2}{1+\left(\frac{x-a_1}{b}\right)^2}\,dx \\ &= \frac{1}{\pi b}\int_{-\infty}^{(a_1-a_2)/2} \frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}\,dx = \frac{1}{\pi}\int_{-\infty}^{(a_1-a_2)/2} \frac{1}{1+y^2}\,dy, \end{aligned}$$

where for the last step we have used the trigonometric substitution $y = (x-a_2)/b$ as in Problem 8. The integral is a standard form for $\tan^{-1} y$ and thus our solution is:

$$\begin{aligned} P(error) &= \frac{1}{\pi}\left[\tan^{-1}\left|\frac{a_1-a_2}{2b}\right| - \tan^{-1}[-\infty]\right] \\ &= \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left|\frac{a_2-a_1}{2b}\right|. \end{aligned}$$

(b) See figure.



(c) The maximum value of the probability of error is $P_{max}(\frac{a_2-a_1}{2b}) = 1/2$, which occurs for $|\frac{a_2-a_1}{2b}| = 0$. This occurs when either the two distributions are the same, which can happen because $a_1 = a_2$, or even if $a_1 \neq a_2$ because $b = \infty$ and both distributions are flat.

10. We use the fact that the conditional error is

$$P(error|x) = \begin{cases} P(\omega_1|x) \text{ if we decide } \omega_2 \\ P(\omega_2|x) \text{ if we decide } \omega_1. \end{cases}$$

(a) Thus the decision as stated leads to:

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)dx.$$

Thus we can write the probability of error as

$$\begin{aligned} P(error) &= P(x < \theta \text{ and } \omega_1 \text{ is the true state}) \\ &\quad + P(x > \theta \text{ and } \omega_2 \text{ is the true state}) \\ &= P(x < \theta|\omega_1)P(\omega_1) + P(x > \theta|\omega_2)P(\omega_2) \\ &= P(\omega_1)\int_{-\infty}^{\theta} p(x|\omega_1)\ dx + P(\omega_2)\int_{\theta}^{\infty} p(x|\omega_2)\ dx. \end{aligned}$$

(b) We take a derivative with respect to θ and set it to zero to find an extremum, that is,

$$\frac{dP(error)}{d\theta} = P(\omega_1)p(\theta|\omega_1) - P(\omega_2)p(\theta|\omega_2) = 0,$$

which yields the condition

$$P(\omega_1)p(\theta|\omega_1) = P(\omega_2)p(\theta|\omega_2),$$

where we have used the fact that $p(x|\omega_i) = 0$ at $x \to \pm\infty$.

(c) No, this condition does not uniquely define θ.

1. If $P(\omega_1)p(\theta|\omega_1) = P(\omega_2)p(\theta|\omega_2)$ over a *range* of θ, then θ would be unspecified throughout such a range.
2. There can easily be multiple values of x for which the condition hold, for instance if the distributions have the appropriate multiple peaks.

(d) If $p(x|\omega_1) \sim N(1, 1)$ and $p(x|\omega_2) \sim N(-1, 1)$ with $P(\omega_1) = P(\omega_2) = 1/2$, then we have a *maximum* for the error at $\theta = 0$.

11. The deterministic risk is given by Bayes' Rule and Eq. 20 in the text

$$R = \int R(\alpha_i(\mathbf{x})|\mathbf{x})\ d\mathbf{x}.$$

(a) In a random decision rule, we have the *probability* $P(\alpha_i|\mathbf{x})$ of deciding to take action α_i. Thus in order to compute the full probabilistic or randomized risk, R_{ran}, we must integrate over all the conditional risks weighted by their probabilities, i.e.,

$$R_{ran} = \int \left[\sum_{i=1}^{a} R(\alpha_i(\mathbf{x})|\mathbf{x})P(\alpha_i|\mathbf{x})\right] p(\mathbf{x})\ d\mathbf{x}.$$

(b) Consider a fixed point $\mathbf{x}$ and note that the (deterministic) Bayes minimum risk decision at that point obeys

$$R(\alpha_i(\mathbf{x})|\mathbf{x}) \geq R(\alpha_{max}(\mathbf{x})|\mathbf{x}).$$

Therefore we have the risk in the randomized case

$$\begin{aligned} R_{ran} &= \int \left[\sum_{i=1}^{a} R(\alpha_i(\mathbf{x})|\mathbf{x}) P(\alpha_i|\mathbf{x})\right] p(\mathbf{x}) d\mathbf{x} \\ &\geq \int R(\alpha_{max}|\mathbf{x}) \left[\sum_{i=1}^{a} P(\alpha_i|\mathbf{x})\right] p(\mathbf{x}) d\mathbf{x} \\ &= \int R(\alpha_{max}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= R_B, \end{aligned}$$

the Bayes risk. Equality holds if and only if $P(\alpha_{max}(\mathbf{x})|\mathbf{x}) = 1$.

12. We first note the normalization condition

$$\sum_{i=1}^{c} P(\omega_i|\mathbf{x}) = 1$$

for all $\mathbf{x}$.

(a) If $P(\omega_i|\mathbf{x}) = P(\omega_j|\mathbf{x})$ for all i and j, then $P(\omega_i|\mathbf{x}) = 1/c$ and hence $P(\omega_{max}|\mathbf{x}) = 1/c$. If one of the $P(\omega_i|\mathbf{x}) < 1/c$, then by our normalization condition we must have that $P(\omega_{max}|\mathbf{x}) > 1/c$.

(b) The probability of error is simply 1.0 minus the probability of being correct, that is,

$$P(error) = 1 - \int P(\omega_{max}|\mathbf{x}) p(\mathbf{x})\ d\mathbf{x}.$$

(c) We simply substitute the limit from part (a) to get

$$\begin{aligned} P(error) &= 1 - \int \underbrace{P(\omega_{max}|\mathbf{x})}_{=g \geq 1/c} p(\mathbf{x})\ d\mathbf{x} \\ &= 1 - g \int p(\mathbf{x})\ d\mathbf{x} = 1 - g. \end{aligned}$$

Therefore, we have $P(error) \leq 1 - 1/c = (c-1)/c$.

(d) All categories have the same prior probability and each distribution has the same form, in other words, the distributions are indistinguishable.

Section 2.4

13. If we choose the category ω_{max} that has the maximum posterior probability, our risk at a point $\mathbf{x}$ is:

$$\lambda_s \sum_{j \neq max} P(\omega_j|\mathbf{x}) = \lambda_s[1 - P(\omega_{max}|\mathbf{x})],$$

whereas if we reject, our risk is λ_r. If we choose a non-maximal category ω_k (where $k \neq max$), then our risk is

$$\lambda_s \sum_{j \neq k} P(\omega_j|\mathbf{x}) = \lambda_s[1 - P(\omega_k|\mathbf{x})] \geq \lambda_s[1 - P(\omega_{max}|\mathbf{x})].$$

This last inequality shows that we should never decide on a category other than the one that has the maximum posterior probability, as we know from our Bayes analysis. Consequently, we should either choose ω_{max} or we should reject, depending upon which is smaller: $\lambda_s[1 - P(\omega_{max}|\mathbf{x})]$ or λ_r. We reject if $\lambda_r \leq \lambda_s[1 - P(\omega_{max}|\mathbf{x})]$, that is, if $P(\omega_{max}|\mathbf{x}) \geq 1 - \lambda_r/\lambda_s$.

14. Consider the classification problem with rejection option.

(a) The minimum-risk decision rule is given by:

$$\begin{aligned} \text{Choose } \omega_i \text{ if } P(\omega_i|\mathbf{x}) &\geq P(\omega_j|\mathbf{x}), \text{for all } j \\ \text{and if } P(\omega_i|\mathbf{x}) &\geq 1 - \frac{\lambda_r}{\lambda_s}. \end{aligned}$$

This rule is equivalent to

$$\begin{aligned} \text{Choose } \omega_i \text{ if } p(\mathbf{x}|\omega_i)P(\omega_i) &\geq p(\mathbf{x}|\omega_j)P(\omega_j) \text{ for all } j \\ \text{and if } p(\mathbf{x}|\omega_i)P(\omega_i) &\geq \left(1 - \frac{\lambda_r}{\lambda_s}\right) p(\mathbf{x}), \end{aligned}$$

where by Bayes' formula

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})}.$$

The optimal discriminant function for this problem is given by

$$\text{Choose } \omega_i \text{ if } g_i(\mathbf{x}) \geq g_j(\mathbf{x}) \text{ for all } i = 1, \ldots, c, \text{ and } j = 1, \ldots, c+1.$$

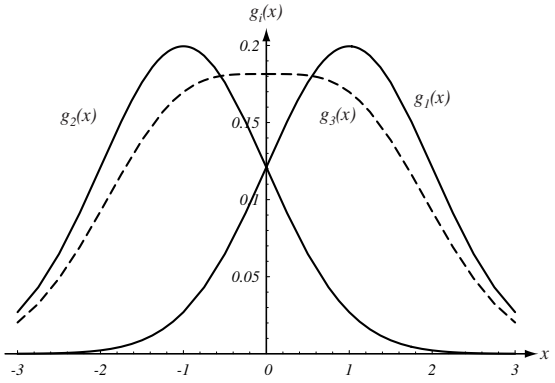
Thus the discriminant functions are:

$$\begin{aligned} g_i(\mathbf{x}) &= \begin{cases} p(\mathbf{x}|\omega_i)P(\omega_i), & i = 1, \ldots, c \\ \left(\frac{\lambda_s - \lambda_r}{\lambda_s}\right) p(\mathbf{x}), & i = c+1, \end{cases} \\ &= \begin{cases} p(\mathbf{x}|\omega_i)P(\omega_i), & i = 1, \ldots, c \\ \frac{\lambda_s - \lambda_r}{\lambda_s} \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j), & i = c+1. \end{cases} \end{aligned}$$

(b) Consider the case $p(x|\omega_1) \sim N(1,1), p(x|\omega_2) \sim N(-1,1), P(\omega_1) = P(\omega_2) = 1/2$ and $\lambda_r/\lambda_s = 1/4$. In this case the discriminant functions in part (a) give

$$\begin{aligned} g_1(x) &= p(x|\omega_1)P(\omega_1) = \frac{1}{2}\frac{e^{-(x-1)^2/2}}{\sqrt{2\pi}} \\ g_2(x) &= p(x|\omega_2)P(\omega_2) = \frac{1}{2}\frac{e^{-(x+1)^2/2}}{\sqrt{2\pi}} \\ g_3(x) &= \left(1 - \frac{\lambda_r}{\lambda_s}\right)[p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2)] \\ &= \left(1 - \frac{1}{4}\right)\left[\frac{1}{2}\frac{e^{-(x-1)^2/2}}{\sqrt{2\pi}} + \frac{1}{2}\frac{e^{-(x+1)^2/2}}{\sqrt{2\pi}}\right] \\ &= \frac{3}{8\sqrt{2\pi}}\left[e^{-(x-1)^2/2} + e^{-(x+1)^2/2}\right] = \frac{3}{4}[g_1(x) + g_2(x)]. \end{aligned}$$

as shown in the figure.

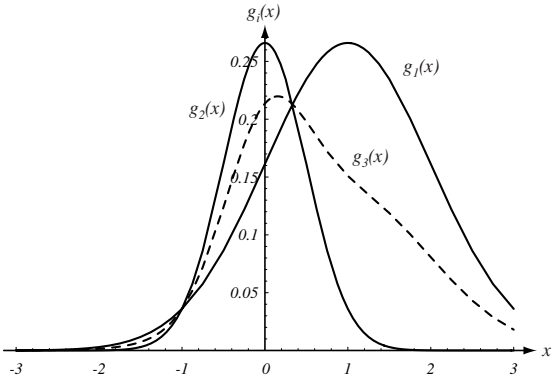


(c) If $\lambda_r/\lambda_s = 0$, there is no cost in rejecting as unrecognizable. Furthermore, $P(\omega_i|\mathbf{x}) \geq 1 - \lambda_r/\lambda_s$ is never satisfied if $\lambda_r/\lambda_s = 0$. In that case, the decision rule will always reject as unrecognizable. On the other hand, as $\lambda_r/\lambda_s \rightarrow 1, P(\omega_i|\mathbf{x}) \geq 1-\lambda_r/\lambda_s$ is always satisfied (there is a high cost of not recognizing) and hence the decision rule is the Bayes decision rule of choosing the class ω_i that maximizes the posterior probability $P(\omega_i|\mathbf{x})$.

(d) Consider the case $p(x|\omega_1) \sim N(1,1), \quad p(x|\omega_2) \sim N(0,1/4), P(\omega_1) = 1/3, P(\omega_2) = 2/3$ and $\lambda_r/\lambda_s = 1/2$. In this case, the discriminant functions of part (a) give

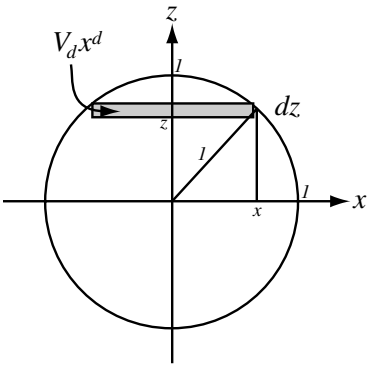
$$\begin{aligned}
g_1(x) &= p(x|\omega_1)P(\omega_1) = \frac{2}{3}\frac{e^{-(x-1)^2/2}}{\sqrt{2\pi}} \\
g_2(x) &= p(x|\omega_2)P(\omega_2) = \frac{1}{3}\frac{2e^{-2x^2}}{\sqrt{2\pi}} \\
g_3(x) &= \left(1 - \frac{\lambda_r}{\lambda_s}\right)[p(x|\omega_1)P(\omega_1) + p(x|\omega_2)P(\omega_2)] \\
&= \frac{1}{2}\cdot\frac{2}{3}\left[\frac{e^{-(x-1)^2/2}}{\sqrt{2\pi}} + \frac{e^{-2x^2}}{\sqrt{2\pi}}\right] \\
&= \frac{1}{3\sqrt{2\pi}}\left[e^{-(x-1)^2/2} + e^{-2x^2}\right] = \frac{1}{2}[g_1(x) + g_2(x)].
\end{aligned}$$

Note from the figure that for this problem we should never reject.



Section 2.5

15. We consider the volume of a d-dimensional hypersphere of radius 1.0, and more generally radius x, as shown in the figure.



(a) We use Eq. 47 in the text for $d =$ odd, that is, $V_d = 2^d \pi^{(d-1)/2} (\frac{d-1}{2})!/d!$. When applied to $d = 1$ (a line) we have $V_1 = 2^1 \pi^0 1 = 2$. Indeed, a line segment $-1 \leq x \leq +1$ has generalized volume (length) of 2. More generally, a line of "radius" x has volume of $2x$.

(b) We use Eq. 47 in the text for $d =$ even, that is, $V_d = \pi^{(d/2)}/(d/2)!$. When applied to $d = 2$ (a disk), we have $V_2 = \pi^1/1! = \pi$. Indeed, a disk of radius 1 has generalized volume (area) of π. More generally, a disk of radius x has volume of πx^2.

(c) Given the volume of a line in $d = 1$, we can derive the volume of a disk by straightforward integration. As shown in the figure, we have

$$V_2 = 2 \int_0^1 \sqrt{1 - z^2}\, dz = \pi,$$

as we saw in part (a).

(d) As can be seen in the figure, to find the volume of a generalized hypersphere in $d + 1$ dimensions, we merely integrate along the z (new) dimension the volume of a generalized hypersphere in the d-dimensional space, with proper factors and limits. Thus we have:

$$V_{d+1} = 2 \int_0^1 V_d (1 - z^2)^{d/2} dz = \frac{V_d \sqrt{\pi}\, \Gamma(d/2 + 1)}{\Gamma(d/2 + 3/2)},$$

where for integer k the gamma function obeys

$$\Gamma(k + 1) = k! \text{ and } \Gamma(k + 1/2) = 2^{-2k+1} \sqrt{\pi}(2k - 1)!/(k - 1)!.$$

(e) Using this formula for $d = 2k$ even, and V_d given for even dimensions, we get that for the next higher (odd) dimension d^*:

$$V_{d^*} \quad = \quad V_{d+1} = \frac{2\pi^{d/2}}{(d/2)!} \left[\frac{\sqrt{\pi}}{2} \frac{(d/2)!}{\Gamma(d/2 + 3/2)} \right]$$

$$= \frac{\pi^{d/2}\ k!\ 2^{2k+1}}{(2k+1)!}$$
$$= \frac{\pi^{(d^*-1)/2}(\frac{d^*-1}{2})!\ 2^{d^*}}{(d^*)!},$$

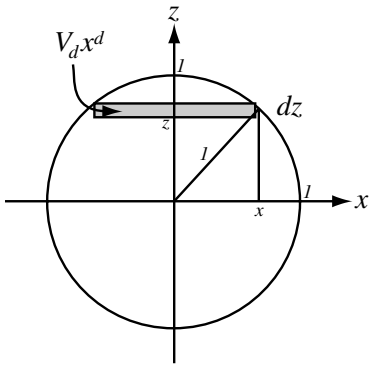
where we have used $2k = d$ for some integer k and $d^* = d + 1$. This confirms Eq. 47 for odd dimension given in the text.

(f) We repeat the above steps, but now use V_d for d odd, in order to derive the volume of the hypersphere in an even dimension:

$$V_{d+1} = V_d \frac{\sqrt{\pi}}{2} \frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+\frac{3}{2})}$$
$$= \frac{2^d\ \pi^{(d-1)/2}\ (\frac{d-1}{2})!}{d!}\ \frac{\sqrt{\pi}}{2}\ \frac{\Gamma((k+1)+\frac{1}{2})}{(k+1)!}$$
$$= \frac{\pi^{d^*/2}}{(d^*/2)!},$$

where we have used that for odd dimension $d = 2k + 1$ for some integer k, and $d^* = d+1$ is the (even) dimension of the higher space. This confirms Eq. 47 for even dimension given in the text.

16. We approach the problem analogously to problem 15, and use the same figure.



(a) The "volume" of a line from $-1 \le x \le 1$ is indeed $V_1 = 2$.

(b) Integrating once for the general case (according to the figure) gives

$$V_{d+1} = 2\int_0^1 V_d(1-z^2)^{d/2}dz = \frac{V_d\sqrt{\pi}\Gamma(d/2+1)}{\Gamma(d/2+3/2)},$$

where for integer k the gamma function obeys

$$\Gamma(k+1) = k! \text{ and } \Gamma(k+1/2) = 2^{-2k+1}\sqrt{\pi}(2k-1)!/(k-1)!.$$

Integrating again thus gives:

$$V_{d+2} = \underbrace{V_d\left[\frac{\sqrt{\pi}\Gamma(d/2+1)}{\Gamma(d/2+3/2)}\right]}_{V_{d+1}}\left[\frac{\sqrt{\pi}\Gamma((d+1)/2+1)}{\Gamma((d+1)/2+3/2)}\right]$$

$$\begin{aligned} &= V_d \pi \frac{\Gamma(d/2+1)}{\Gamma((d+1)/2+3/2)} \\ &= V_d \frac{\pi \Gamma(d/2+1)}{(d/2+1)\Gamma(d/2+1)} = V_d \frac{\pi}{d/2+1}. \end{aligned}$$

This is the central result we will use below.

(c) For d odd, we rewrite $\pi/(d/2+1)$ as $2\pi/(d+2)$. Thus we have

$$\begin{aligned} V_d &= V_{d-2}\left[\frac{2\pi}{(d-2)+2}\right] = V_{d-4}\left[\frac{2\pi}{(d-4)+2}\right]\left[\frac{2\pi}{(d-2)+2}\right] = \cdots \\ &= V_1 \underbrace{\left[\frac{2\pi}{(d-(d-1))+2}\right]\left[\frac{2\pi}{(d-(d-3))+2}\right] \times \cdots \times \left[\frac{2\pi}{(d-4)+2}\right]\left[\frac{2\pi}{(d-2)+2}\right]}_{(d-1)/2 \ terms} \\ &= \pi^{(d-1)/2} 2^{(d+1)/2} \left[\frac{1}{3}\right]\left[\frac{1}{5}\right] \times \cdots \times \left[\frac{1}{d-2}\right]\left[\frac{1}{d}\right] \\ &= \pi^{(d-1)/2} 2^{(d+1)/2} \left[\frac{1}{d!!}\right] = \frac{\pi^{(d-1)/2} 2^d}{d!}. \end{aligned}$$

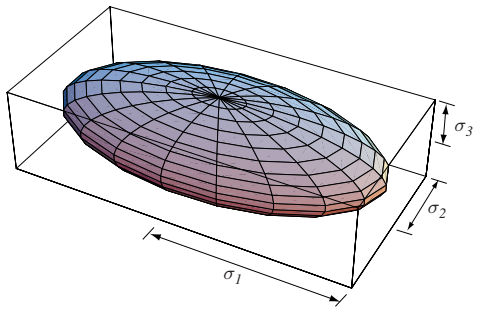
We have used the fact that $V_1 = 2$, from part (a), and the notation $d!! = d \times (d-2) \times (d-4) \times \cdots$, read "$d$ double factorial."

(d) Analogously to part (c), for d even we have

$$\begin{aligned} V_d &= V_{d-2}\left[\frac{\pi}{(d-2)/2+1}\right] = V_{d-4}\left[\frac{\pi}{(d-4)/2+1}\right]\left[\frac{\pi}{(d-2)/2+1}\right] = \cdots \\ &= \underbrace{\overbrace{\left[\frac{\pi}{(d-d)/2+1}\right]}^{V_2=\pi}\left[\frac{\pi}{(d-(d-2))/2+1}\right] \times \cdots \times \left[\frac{\pi}{(d-4)/2+1}\right]\left[\frac{\pi}{(d-2)/2+1}\right]}_{d/2 \ terms} \\ &= \pi^{d/2}\left[\frac{1}{1}\right]\left[\frac{1}{2}\right] \times \cdots \times \left[\frac{1}{d/2-1}\right]\left[\frac{1}{d/2}\right] \\ &= \frac{\pi^{d/2}}{(d/2)!}. \end{aligned}$$

(e) The central mathematical reason that we express the formulas separately for even and for odd dimensions comes from the gamma function, which has different expressions as factorials depending upon whether the argument is integer valued, or half-integer valued. One could express the volume of hyperspheres in *all* dimensions by using gamma functions, but that would be computationally a bit less elegant.

17. Consider the minimal rectangular bounding box in d dimensions that contains the hyperellipsoid, as shown in the figure. We can work in a coordinate system in which the principal axes of the ellipsoid are parallel to the box axes, that is, the covariance matrix is diagonal ($\hat{\mathbf{\Sigma}} = \text{diag}[1/\sigma_1^2, 1/\sigma_2^2, \ldots, 1/\sigma_d^2]$). The squared Mahalanobis distance from the origin to the surface of the ellipsoid is given by r^2,



that is, points $\mathbf{x}$ that satisfy

$$\begin{aligned} r^2 &= \mathbf{x}^t \tilde{\mathbf{\Sigma}} \mathbf{x} \\ &= \mathbf{x}^t \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_d^2 \end{pmatrix} \mathbf{x}. \end{aligned}$$

Thus, along each of the principal axes, the distance obeys $x_i^2 = \sigma_i^2 r^2$. Because the distance across the rectangular volume is twice that amount, the volume of the rectangular bounding box is

$$V_{rect} = (2x_1)(2x_2)\cdots(2x_d) = 2^d r^d \prod_{i=1}^{d} \sigma_i = 2^d r^d |\tilde{\mathbf{\Sigma}}|^{1/2}.$$

We let V be the (unknown) volume of the hyperellipsoid, V_d the volume of the unit hypersphere in d dimension, and V_{cube} be the volume of the d-dimensional cube having length 2 on each side. Then we have the following relation:

$$\frac{V}{V_{rect}} = \frac{V_d}{V_{cube}}.$$

We note that the volume of the hypercube is $V_{cube} = 2^d$, and substitute the above to find that

$$V = \frac{V_{rect} V_d}{V_{cube}} = r^d |\tilde{\mathbf{\Sigma}}|^{1/2} V_d,$$

where V_d is given by Eq. 47 in the text. Recall that the determinant of a matrix is unchanged by rotation of axes ($|\tilde{\mathbf{\Sigma}}|^{1/2} = |\mathbf{\Sigma}|^{1/2}$), and thus the value can be written as

$$V = r^d |\mathbf{\Sigma}|^{1/2} V_d.$$

18. Let $X_1, \dots, X_n$ be a random sample of size n from $N(\mu_1, \sigma_1^2)$ and let $Y_1, \dots, Y_m$ be a random sample of size m from $N(\mu_2, \sigma_2^2)$.

(a) Let $Z = (X_1 + \cdots + X_n) + (Y_1 + \cdots + Y_m)$. Our goal is to show that Z is also normally distributed. From the discussion in the text, if $\mathbf{X}_{d\times 1} \sim N(\boldsymbol{\mu}_{d\times 1}, \mathbf{\Sigma}_{d\times d})$

and $\mathbf{A}$ is a $k \times d$ matrix, then $\mathbf{A}^t\mathbf{X} \sim N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$. Here, we take

$$\mathbf{X}_{(n+m)\times 1} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \\ Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix}.$$

Then, clearly $\mathbf{X}$ is normally distributed in $(n+m)\times 1$ dimensions. We can write Z as a particular matrix $\mathbf{A}^t$ operating on $\mathbf{X}$:

$$Z = X_1 + \cdots + X_n + Y_1 + \cdots + Y_m = \mathbf{1}^t\mathbf{X},$$

where $\mathbf{1}$ denotes a vector of length $n+m$ consisting solely of 1's. By the above fact, it follows that Z has a univariate normal distribution.

(b) We let μ_3 be the mean of the new distribution. Then, we have

$$\begin{aligned} \mu_3 &= \mathcal{E}(Z) \\ &= \mathcal{E}[(X_1 + \cdots + X_n) + (Y_1 + \cdots + Y_m)] \\ &= \mathcal{E}(X_1) + \cdots + \mathcal{E}(X_n) + \mathcal{E}(Y_1) + \cdots + \mathcal{E}(Y_m) \\ &\quad \text{(since } X_1, \ldots, X_n, Y_1, \ldots, Y_m \text{ are independent)} \\ &= n\mu_1 + m\mu_2. \end{aligned}$$

(c) We let σ_3^2 be the variance of the new distribution. Then, we have

$$\begin{aligned} \sigma_3^2 &= \mathrm{Var}(Z) \\ &= \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) + \mathrm{Var}(Y_1) + \cdots + \mathrm{Var}(Y_m) \\ &\quad \text{(since } X_1, \ldots, X_n, Y_1, \ldots, Y_m \text{ are independent)} \\ &= n\sigma_1^2 + m\sigma_2^2 \end{aligned}$$

(d) Define a column vector of the samples, as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \\ \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix}.$$

Then, clearly $\mathbf{X}$ is $[(nd+md)\times 1]$-dimensional random variable that is normally distributed. Consider the linear projection operator $\mathbf{A}$ defined by

$$\mathbf{A}^t = (\underbrace{\mathbf{I}_{d\times d}\ \mathbf{I}_{d\times d} \cdots \mathbf{I}_{d\times d}}_{(n+m)\ times}).$$

Then we have

$$\mathbf{Z} = \mathbf{A}^t\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_n + \mathbf{Y}_1 + \cdots + \mathbf{Y}_m,$$

which must therefore be normally distributed. Furthermore, the mean and variance of the distribution are

$$\begin{aligned}
\boldsymbol{\mu}_3 = \mathcal{E}(\mathbf{Z}) &= \mathcal{E}(\mathbf{X}_1) + \cdots + \mathcal{E}(\mathbf{X}_n) + \mathcal{E}(\mathbf{Y}_1) + \cdots + \mathcal{E}(\mathbf{Y}_m) \\
&= n\boldsymbol{\mu}_1 + m\boldsymbol{\mu}_2. \\
\boldsymbol{\Sigma}_3 = \mathrm{Var}(\mathbf{Z}) &= \mathrm{Var}(\mathbf{X}_1) + \cdots + \mathrm{Var}(\mathbf{X}_n) + \mathrm{Var}(\mathbf{Y}_1) + \cdots + \mathrm{Var}(\mathbf{Y}_m) \\
&= n\boldsymbol{\Sigma}_1 + m\boldsymbol{\Sigma}_2.
\end{aligned}$$

19. The entropy is given by Eq. 37 in the text:

$$H(p(x)) = -\int p(x)\ \mathrm{ln}p(x)dx$$

with constraints

$$\int b_k(x)p(x)dx = a_k \qquad \text{for } k = 1, \ldots, q.$$

(a) We use Lagrange factors and find

$$\begin{aligned}
H_s &= \int p(x)\mathrm{ln}p(x)dx + \sum_{k=1}^{q}\left[\int b_k(x)p(x)dx - a_k\right] \\
&= -\int p(x)\left[\mathrm{ln}p(x) - \sum_{k=0}^{q}\lambda_k b_k(x)\right] - \sum_{k=0}^{q} a_k\lambda_k.
\end{aligned}$$

From the normalization condition $\int p(x)dx = 1$, we know that $a_0 = b_0 = 1$ for all x.

(b) In order to find the maximum or minimum value for H (having constraints), we take the derivative of H_s (having no constraints) with respect to $p(x)$ and set it to zero:

$$\frac{\partial H_s}{\partial p(x)} = -\int\left[\mathrm{ln}p(x) - \sum_{k=0}^{q}\lambda_k b_k(x) + 1\right]dx = 0.$$

The argument of the integral must vanish, and thus

$$\mathrm{ln}p(x) = \sum_{k=0}^{q}\lambda_k b_k(x) - 1.$$

We exponentiate both sides and find

$$p(x) = \exp\left[\sum_{k=0}^{q}\lambda_k b_k(x) - 1\right],$$

where the $q+1$ parameters are determined by the constraint equations.

20. We make use of the result of Problem 19, that is, the maximum-entropy distribution $p(x)$ having constraints of the form

$$\int b_k(x)p(x)dx = a_k$$

for $k = 1, \ldots, q$ is

$$p(x) = \exp\left[\sum_{k=0}^{q} \lambda_k b_k(x) - 1\right].$$

(a) In the current case, the normalization constraint is

$$\int_{x_l}^{x_h} p(x)dx = 1,$$

and thus

$$p(x) = \exp\left[\sum_{k=0}^{q} \lambda_k b_k(x) - 1\right] = \exp(\lambda_0 - 1).$$

In order to satisfy this constraint equation, we demand

$$\int_{x_l}^{x_h} p(x)dx = \exp(\lambda_0 - 1)(x_u - x_l) = 1,$$

and thus the distribution is uniform,

$$p(x) = \begin{cases} 1/|x_u - x_l| & x_l \leq x \leq x_u \\ 0 & \text{otherwise.} \end{cases}$$

(b) Here the constraint equations are

$$\begin{aligned} \int_0^\infty p(x)dx &= 1 \\ \int_0^\infty xp(x)dx &= \mu. \end{aligned}$$

Thus, using the above result, the density is of the form

$$p(x) = \exp[\lambda_0 + \lambda_1 x - 1]$$

and we need to solve for λ_0 and λ_1. The normalization constraint gives

$$\begin{aligned} \int_0^\infty \exp[\lambda_0 + \lambda_1 x - 1]dx &= e^{\lambda_0 - 1}\frac{e^{\lambda_1 x}}{\lambda_1}\Big|_0^\infty \\ &= e^{\lambda_0 - 1}\left(-\frac{1}{\lambda_1}\right) = 1. \end{aligned}$$

Likewise, the mean constraint gives

$$\int_0^\infty e^{\lambda_1} e^{\lambda_0 - 1} x dx = e^{\lambda_0 - 1}\left(\frac{1}{\lambda_1^2}\right) = \mu.$$

Hence $\lambda_1 = -1/\mu$ and $\lambda_0 = 1 - \ln\mu$, and the density is

$$p(x) = \begin{cases} (1/\mu)e^{-x/\mu} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(c) Here the density has three free parameters, and is of the general form

$$p(x) = \exp[\lambda_0 - 1 + \lambda_1 x + \lambda_2 x^2],$$

and the constraint equations are

$$\int_{-\infty}^{\infty} p(x)dx = 1 \qquad (*)$$

$$\int_{-\infty}^{\infty} xp(x)dx = \mu \qquad (**)$$

$$\int_{-\infty}^{\infty} x^2 p(x)dx = \sigma^2. \qquad (***)$$

We first substitute the general form of $p(x)$ into $(*)$ and find

$$\frac{1}{\sqrt{-\lambda_2}} \frac{\sqrt{\pi}}{2} e^{\lambda_0 - 1 - \lambda_1^2/(4\lambda_2)} \text{erf}\left[\sqrt{-\lambda_2}x - \frac{\lambda_1}{2\sqrt{-\lambda_2}}\right]\Bigg|_{-\infty}^{\infty} = 1.$$

Since $\lambda_2 < 0$, $\text{erf}(\infty) = 1$ and $\text{erf}(-\infty) = -1$, we have

$$\frac{\sqrt{\pi}\exp[\lambda_0 - 1 - \lambda_1^2/(4\lambda_2)]}{\sqrt{-\lambda_2}} = 1.$$

Likewise, next substitute the general form of $p(x)$ into $(**)$ and find

$$-\frac{\lambda_1\sqrt{\pi}\exp[\lambda_0 - 1 - \lambda_1^2/(4\lambda_2)]}{4\lambda_2\sqrt{-\lambda_2}} \text{erf}\left[\sqrt{-\lambda_2}x - \lambda_2/(2\sqrt{-\lambda_2})\right]\Bigg|_{-\infty}^{\infty} = \mu,$$

which can be simplified to yield

$$\frac{\lambda_1\sqrt{\pi}}{2\lambda_2\sqrt{-\lambda_2}}\exp[\lambda_0 - 1 - \lambda_1^2/(4\lambda_2)] = -\mu.$$

Finally, we substitute the general form of $p(x)$ into $(***)$ and find

$$\frac{\sqrt{\pi}}{2\lambda_2\sqrt{-\lambda_2}}\exp[\lambda_0 - 1 - \lambda_1^2/(4\lambda_2)] = -\sigma^2.$$

We combine these three results to find the constants:

$$\begin{aligned}\lambda_0 &= 1 - \frac{\mu^2}{2\sigma_2} + \ln[1/(\sqrt{2\pi}\sigma)] \\ \lambda_1 &= \mu/\sigma^2 \\ \lambda_2 &= -1/(2\sigma^2).\end{aligned}$$

We substitute these values back into the general form of the density and find

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right],$$

that is, a Gaussian.

21. A Gaussian centered at $x = 0$ is of the form

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-x^2/(2\sigma^2)].$$

The entropy for this distribution is given by Eq. 37 in the text:

$$\begin{aligned}H(p(x)) &= -\int_{-\infty}^{\infty} p(x)\ln p(x)dx \\ &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp[-x^2/(2\sigma^2)]\ln\left[\frac{1}{\sqrt{2\pi}\sigma}\exp[-x^2/(2\sigma^2)]\right]dx \\ &= \ln[\sqrt{2\pi}\sigma] + 1/2 = \ln[\sqrt{2\pi e}\sigma].\end{aligned}$$

For the uniform distribution, the entropy is

$$H(p(x)) = -\int_{x_l}^{x_u} \frac{1}{|x_u - x_l|}\ln\left[\frac{1}{|x_u - x_l|}\right]dx = -\ln\left[\frac{1}{|x_u - x_l|}\right] = \ln[x_u - x_l].$$

Since we are given that the mean of the distribution is 0, we know that $x_u = -x_l$. Further, we are told that the variance is σ^2, that is

$$\int_{x_l}^{x_h} x^2 p(x)dx = \sigma^2$$

which, after integration, implies

$$x_u^2 + x_u x_l + x_l^2 = 3\sigma^2.$$

We put these results together and find for the uniform distribution $H(p(x)) = \ln[2\sqrt{3}\sigma]$.

We are told that the variance of the triangle distribution centered on 0 having half-width w is σ^2, and this implies

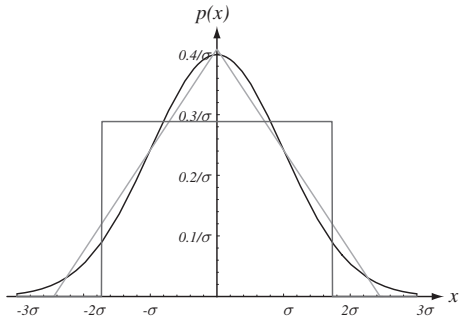
$$\int_{-w}^{w} x^2\frac{w-|x|}{w^2}dx = \int_{0}^{w} x^2\frac{w-x}{w^2}dx + \int_{-w}^{0} x^2\frac{w+x}{w^2}dx = w^2/6 = \sigma^2.$$

The entropy of the triangle distribution is then

$$\begin{aligned} H(p(x)) &= -\int_{-w}^{w} \frac{w-|x|}{w^2} \ln\left[\frac{w-|x|}{w^2}\right] dx \\ &= \int_{0}^{w} \frac{w-x}{w^2} \ln\left[\frac{w-x}{w^2}\right] dx - \int_{-w}^{0} \frac{w+x}{w^2} \ln\left[\frac{w+x}{w^2}\right] dx \\ &= \ln w + 1/2 = \ln[\sqrt{6}\sigma] + 1/2 = \ln[\sqrt{6e}\sigma], \end{aligned}$$

where we used the result $w = \sqrt{6}\sigma$ from the variance condition.

Thus, in order of decreasing entropy, these equal-variance distributions are Gaussian, uniform then triangle, as illustrated in the figure, where each has the same variance σ^2.



22. As usual, we denote our multidimensional Gaussian distribution by $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

According to Eq. 37 in the text, the entropy is

$$\begin{aligned} H(p(\mathbf{x})) &= -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \\ &= -\int p(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \underbrace{\ln\left[(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}\right]}_{indep.\ of\ \mathbf{x}}\right] d\mathbf{x} \\ &= \frac{1}{2}\int \left[\sum_{i=1}^{d}\sum_{j=1}^{d}(x_i-\mu_i)[\boldsymbol{\Sigma}^{-1}]_{ij}(x_j-\mu_j)\right] d\mathbf{x} + \frac{1}{2}\ln[(2\pi)^d|\boldsymbol{\Sigma}|] \\ &= \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\int (x_j-\mu_j)(x_i-\mu_i) \underbrace{[\boldsymbol{\Sigma}^{-1}]_{ij}}_{indep.\ of\ \mathbf{x}} d\mathbf{x} + \frac{1}{2}\ln[(2\pi)^d|\boldsymbol{\Sigma}|] \\ &= \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}[\boldsymbol{\Sigma}]_{ji}[\boldsymbol{\Sigma}^{-1}]_{ij} + \frac{1}{2}\ln[(2\pi)^d|\boldsymbol{\Sigma}|] \\ &= \frac{1}{2}\sum_{j=1}^{d}[\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}]_{jj} + \frac{1}{2}\ln[(2\pi)^d|\boldsymbol{\Sigma}|] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}\underbrace{\sum_{j=1}^{d}[\mathbf{I}]_{jj}}_{d}+\frac{1}{2}\ln[(2\pi)^d|\mathbf{\Sigma}|]\\
&= \frac{d}{2}+\frac{1}{2}\ln[(2\pi)^d|\mathbf{\Sigma}|]\\
&= \frac{1}{2}\ln[(2\pi e)^d|\mathbf{\Sigma}|],
\end{aligned}$$

where we used our common notation of $\mathbf{I}$ for the d-by-d identity matrix.

23. We have $p(\mathbf{x}|\omega)\sim N(\boldsymbol{\mu},\mathbf{\Sigma})$, where

$$\boldsymbol{\mu}=\begin{pmatrix}1\\2\\2\end{pmatrix} \text{ and } \mathbf{\Sigma}=\begin{pmatrix}1&0&0\\0&5&2\\0&2&5\end{pmatrix}.$$

(a) The density at a test point $\mathbf{x}_o$ is

$$p(\mathbf{x}_o|\omega)=\frac{1}{(2\pi)^{3/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}_o-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_o-\boldsymbol{\mu})\right].$$

For this case we have

$$\begin{aligned}
|\mathbf{\Sigma}| &= \begin{vmatrix}1&0&0\\0&5&2\\0&2&5\end{vmatrix}=1\begin{vmatrix}5&2\\2&5\end{vmatrix}=21,\\
\mathbf{\Sigma}^{-1} &= \begin{pmatrix}1&0&0\\0&5&2\\0&2&5\end{pmatrix}^{-1}=\begin{pmatrix}1&0&0\\0&\begin{pmatrix}5&2\\2&5\end{pmatrix}^{-1}\end{pmatrix}=\begin{pmatrix}1&0&0\\0&5/21&-2/21\\0&-2/21&5/21\end{pmatrix},
\end{aligned}$$

and the squared Mahalanobis distance from the mean to $\mathbf{x}_o=(.5,0,1)^t$ is

$$\begin{aligned}
&(\mathbf{x}_o-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_o-\boldsymbol{\mu})\\
&= \left[\begin{pmatrix}.5\\0\\1\end{pmatrix}-\begin{pmatrix}1\\2\\2\end{pmatrix}\right]^t\begin{pmatrix}1&0&0\\0&5/21&-2/21\\0&-2/21&5/21\end{pmatrix}^{-1}\left[\begin{pmatrix}.5\\0\\1\end{pmatrix}-\begin{pmatrix}1\\2\\2\end{pmatrix}\right]\\
&= \begin{bmatrix}-0.5\\-8/21\\-1/21\end{bmatrix}^t\begin{bmatrix}-0.5\\-2\\-1\end{bmatrix}=0.25+\frac{16}{21}+\frac{1}{21}=1.06.
\end{aligned}$$

We substitute these values to find that the density at $\mathbf{x}_o$ is:

$$p(\mathbf{x}_o|\omega)=\frac{1}{(2\pi)^{3/2}(21)^{1/2}}\exp\left[-\frac{1}{2}(1.06)\right]=8.16\times10^{-3}.$$

(b) Recall from Eq. 44 in the text that $\mathbf{A}_w=\mathbf{\Phi}\mathbf{\Lambda}^{-1/2}$, where $\mathbf{\Phi}$ contains the normalized eigenvectors of $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. The characteristic equation, $|\mathbf{\Sigma}-\lambda\mathbf{I}|=0$, in this case is

$$\begin{aligned}
\begin{vmatrix}1-\lambda&0&0\\0&5-\lambda&2\\0&2&5-\lambda\end{vmatrix} &= (1-\lambda)\left[(5-\lambda)^2-4\right]\\
&= (1-\lambda)(3-\lambda)(7-\lambda)=0.
\end{aligned}$$

The three eigenvalues are then $\lambda = 1, 3, 7$ can be read immediately from the factors. The (diagonal) $\mathbf{\Lambda}$ matrix of eigenvalues is thus

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 7 \end{pmatrix}.$$

To find the eigenvectors, we solve $\mathbf{\Sigma x} = \lambda_i \mathbf{x}$ for $(i = 1, 2, 3)$:

$$\mathbf{\Sigma x} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 5 \end{pmatrix} \mathbf{x} = \begin{pmatrix} x_1 \\ 5x_2 + 2x_3 \\ 2x_2 + 5x_3 \end{pmatrix} = \lambda_i \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \qquad i = 1, 2, 3.$$

The three eigenvectors are given by:

$$\begin{aligned}
\lambda_1 &= 1: \quad \begin{pmatrix} x_1 \\ 5x_2 + 2x_3 \\ 2x_2 + 5x_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \Rightarrow \boldsymbol{\phi}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \\
\lambda_2 &= 3: \quad \begin{pmatrix} x_1 \\ 5x_2 + 2x_3 \\ 2x_2 + 5x_3 \end{pmatrix} = \begin{pmatrix} 3x_1 \\ 3x_2 \\ 3x_3 \end{pmatrix} \Rightarrow \boldsymbol{\phi}_2 = \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}, \\
\lambda_3 &= 7: \quad \begin{pmatrix} x_1 \\ 5x_2 + 2x_3 \\ 2x_2 + 5x_3 \end{pmatrix} = \begin{pmatrix} 7x_1 \\ 7x_2 \\ 7x_3 \end{pmatrix} \Rightarrow \boldsymbol{\phi}_3 = \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}.
\end{aligned}$$

Thus our final $\mathbf{\Phi}$ and $\mathbf{A}_w$ matrices are:

$$\mathbf{\Phi} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

and

$$\begin{aligned}
\mathbf{A}_w = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 7 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{6} & 1/\sqrt{14} \\ 0 & -1/\sqrt{6} & 1/\sqrt{14} \end{pmatrix}.
\end{aligned}$$

We have then, $\mathbf{Y} = \mathbf{A}_w^t(\mathbf{x} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{I})$.

(c) The transformed point is found by applying $\mathbf{A}$, that is,

$$\begin{aligned}
\mathbf{x}_w &= \mathbf{A}_w^t(\mathbf{x}_o - \boldsymbol{\mu}) \\
&= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{6} & 1/\sqrt{14} \\ 0 & -1/\sqrt{6} & 1/\sqrt{14} \end{pmatrix} \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix} = \begin{pmatrix} -0.5 \\ -1/\sqrt{6} \\ -3/\sqrt{14} \end{pmatrix}.
\end{aligned}$$

(d) From part (a), we have that the squared Mahalanobis distance from $\mathbf{x}_o$ to $\boldsymbol{\mu}$ in the original coordinates is $r^2 = (\mathbf{x}_o - \boldsymbol{\mu})^t \mathbf{\Sigma}^{-1} (\mathbf{x}_o - \boldsymbol{\mu}) = 1.06$. The Mahalanobis distance from $\mathbf{x}_w$ to $\mathbf{0}$ in the transformed coordinates is $\mathbf{x}_w^t \mathbf{x}_w = (0.5)^2 + 1/6 + 3/14 = 1.06$. The two distances are the same, as they must be under any linear transformation.

(e) A Gaussian distribution is written as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

Under a general linear transformation $\mathbf{T}$, we have that $\mathbf{x}' = \mathbf{T}^t\mathbf{x}$. The transformed mean is

$$\boldsymbol{\mu}' = \sum_{k=1}^{n}\mathbf{x}'_k = \sum_{k=1}^{n}\mathbf{T}^t\mathbf{x}_k = \mathbf{T}^t\sum_{k=1}^{n}\mathbf{x}_k = \mathbf{T}^t\boldsymbol{\mu}.$$

Likewise, the transformed covariance matrix is

$$\begin{aligned}\mathbf{\Sigma}' &= \sum_{k=1}^{n}(\mathbf{x}'_k-\boldsymbol{\mu}')(\mathbf{x}'_k-\boldsymbol{\mu}')^t\\ &= \mathbf{T}^t\left[\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})\right]\mathbf{T}\\ &= \mathbf{T}^t\mathbf{\Sigma}\mathbf{T}.\end{aligned}$$

We note that $|\mathbf{\Sigma}'| = |\mathbf{T}^t\mathbf{\Sigma}\mathbf{T}| = |\mathbf{\Sigma}|$ for transformations such as translation and rotation, and thus

$$p(\mathbf{x}_o|N(\boldsymbol{\mu},\mathbf{\Sigma})) = p(\mathbf{T}^t\mathbf{x}_o|N(\mathbf{T}^t\boldsymbol{\mu},\mathbf{T}^t\mathbf{\Sigma}\mathbf{T})).$$

The volume element is proportial to $|\mathbf{T}|$ and for transformations such as scaling, the transformed covariance is proportional to $|\mathbf{T}|^2$, so the transformed normalization contant contains $1/|\mathbf{T}|$, which exactly compensates for the change in volume.

(f) Recall the definition of a whitening transformation given by Eq. 44 in the text: $\mathbf{A}_w = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2}$. In this case we have

$$\mathbf{y} = \mathbf{A}_w^t\mathbf{x} \sim N(\mathbf{A}_w^t\boldsymbol{\mu}, \mathbf{A}_w^t\mathbf{\Sigma}\mathbf{A}_w),$$

and this implies that

$$\begin{aligned}\mathrm{Var}(\mathbf{y}) &= \mathbf{A}_w^t(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{A}_w\\ &= \mathbf{A}_w^t\mathbf{\Sigma}\mathbf{A}\\ &= (\mathbf{\Phi}\mathbf{\Lambda}^{-1/2})^t\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^t(\mathbf{\Phi}\mathbf{\Lambda}^{-1/2})\\ &= \mathbf{\Lambda}^{-1/2}\mathbf{\Phi}^t\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^t\mathbf{\Phi}\mathbf{\Lambda}^{-1/2}\\ &= \mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2}\\ &= \mathbf{I},\end{aligned}$$

the identity matrix.

24. Recall that the general multivariate normal density in d-dimensions is:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}\ \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

(a) Thus we have if $\sigma_{ij} = 0$ and $\sigma_{ii} = \sigma_i^2$, then

$$\begin{aligned}\boldsymbol{\Sigma} &= \text{diag}(\sigma_1^2, \ldots, \sigma_d^2) \\ &= \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_d^2 \end{pmatrix}.\end{aligned}$$

Thus the determinant and inverse matrix are particularly simple:

$$\begin{aligned}|\boldsymbol{\Sigma}| &= \prod_{i=1}^{d} \sigma_i^2, \\ \boldsymbol{\Sigma}^{-1} &= \text{diag}(1/\sigma_1^2, \ldots, 1/\sigma_d^2).\end{aligned}$$

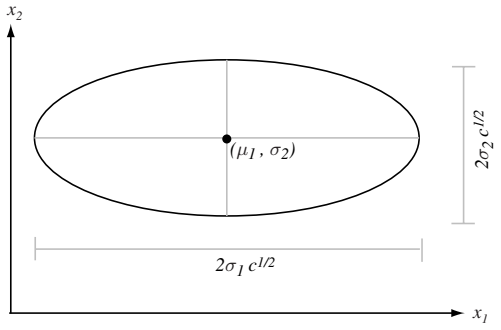
This leads to the density being expressed as:

$$\begin{aligned}p(\mathbf{x}) &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \left[\text{diag}(1/\sigma_1^2, \ldots, 1/\sigma_d^2)\right](\mathbf{x}-\boldsymbol{\mu})\right] \\ &= \frac{1}{\prod_{i=1}^{d} \sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right].\end{aligned}$$

(b) The contours of constant density are concentric ellipses in d dimensions whose centers are at $(\mu_1, \ldots, \mu_d)^t = \boldsymbol{\mu}$, and whose axes in the ith direction are of length $2\sigma_i\sqrt{c}$ for the density $p(\mathbf{x})$ held constant at

$$\frac{e^{-c/2}}{\prod_{i=1}^{d} \sqrt{2\pi}\sigma_i}.$$

The axes of the ellipses are parallel to the coordinate axes. The plot in 2 dimensions ($d = 2$) is shown:



(c) The squared Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$ is:

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= (\mathbf{x}-\boldsymbol{\mu})^t \begin{pmatrix} 1/\sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_d^2 \end{pmatrix}(\mathbf{x}-\boldsymbol{\mu}) \\ &= \sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2.\end{aligned}$$

Section 2.6

25. A useful discriminant function for Gaussians is given by Eq. 52 in the text,

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln\, P(\omega_i).$$

We expand to get

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}\left[\mathbf{x}^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \mathbf{x}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right] + \ln\, P(\omega_i) \\ &= -\frac{1}{2}\Big[\underbrace{\mathbf{x}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}}_{\text{indep. of } i} - 2\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\Big] + \ln\, P(\omega_i). \end{aligned}$$

We drop the term that is independent of i, and this yields the equivalent discriminant function:

$$\begin{aligned} g_i(\mathbf{x}) &= \boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln\, P(\omega_i) \\ &= \mathbf{w}_i^t\mathbf{x} + w_{io}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{w}_i &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \\ w_{io} &= -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln\, P(\omega_i). \end{aligned}$$

The decision boundary for two Gaussians is given by $g_i(\mathbf{x}) = g_j(\mathbf{x})$ or

$$\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln\, P(\omega_i) = \boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln\, P(\omega_j).$$

We collect terms so as to rewrite this as:

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln\,\frac{P(\omega_i)}{P(\omega_j)} = 0$$

$$(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}\left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) + \frac{\ln\,[P(\omega_i)/P(\omega_j)](\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}\right]$$

$$\underbrace{-\frac{1}{2}\boldsymbol{\mu}_j^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j}_{=0} = 0.$$

This is the form of a linear discriminant

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_o) = 0,$$

where the weight and bias (offset) are

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

and

$$\mathbf{x}_o = \frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j) - \frac{\ln\,[P(\omega_i)/P(\omega_j)](\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)},$$

respectively.

26. The densities and Mahalanobis distances for our two distributions with the same covarance obey

$$\begin{aligned} p(\mathbf{x}|\omega_i) &\sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \\ r_i^2 &= (\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \end{aligned}$$

for $i = 1, 2$.

(a) Our goal is to show that $\nabla r_i^2 = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$. Here ∇r_i^2 is the gradient of r_i^2, that is, the (column) vector in d-dimensions given by:

$$\begin{pmatrix} \frac{\partial r_i^2}{\partial x_1} \\ \vdots \\ \frac{\partial r_i^2}{\partial x_d} \end{pmatrix}.$$

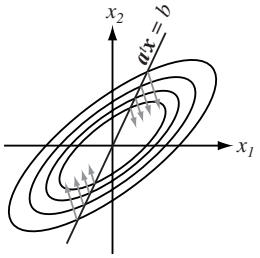
We carefully write out the components of ∇r_i^2 for $j = 1, \ldots, d$, and have:

$$\begin{aligned} \frac{\partial r_i^2}{\partial x_j} &= \frac{\partial}{\partial x_j}\left[(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right] \\ &= \frac{\partial}{\partial x_j}\Big[\mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\mathbf{x} + \underbrace{\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i}_{indep.\ of\ \mathbf{x}}\Big] \\ &= \frac{\partial}{\partial x_j}\left[\sum_{k=1}^{d}\sum_{l=1}^{d} x_k x_l \xi_{kl} - 2\sum_{k=1}^{d}\sum_{l=1}^{d}\mu_{i,k}\xi_{kl}x_l\right] \quad (\text{where } \boldsymbol{\Sigma}^{-1} = [\xi_{kl}]_{d\times d}) \\ &= \frac{\partial}{\partial x_j}\left[\sum_{k=1}^{d} x_k^2\xi_{kk} + \sum_{k=1}^{d}\sum_{\substack{l=1\\k\neq l}}^{d} x_k x_l \xi_{kl} - 2\sum_{k=1}^{d}\mu_{i,k}\sum_{l=1}^{d} x_l\xi_{kl}\right] \\ &= 2x_j\xi_{jj} + \sum_{\substack{k=1\\k\neq j}}^{d} x_k\xi_{kj} + \sum_{\substack{l=1\\l\neq j}}^{d} x_l\xi_{jl} - 2\sum_{k=1}^{d}\mu_{i,k}\xi_{kj} \\ &= 2\left[x_j\xi_{jj} + \sum_{\substack{k=1\\k\neq j}}^{d} x_k\xi_{kj} - 2\sum_{k=1}^{d}\mu_{i,k}\xi_{kj}\right] \qquad (\text{note that } \boldsymbol{\Sigma}^{-1} \text{ is symmetric}) \\ &= 2\left[\sum_{k=1}^{d}(x_k-\mu_{i,k})\xi_{kj}\right] \\ &= 2\times j^{th} \text{ component of } \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i). \end{aligned}$$

Since this is true for each component j, we have

$$\nabla r_i^2 = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i).$$

(b) From part (a) we know that $\nabla r_i^2 = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$. We can work in the translated frame of reference where the mean is at the origin. This does not change the derivatives. Moreover, since we are dealing with a single Gaussian, we can dispense with the needless subscript indicating the category. For notational



simplicity, we set the constant matrix $\mathbf{\Sigma}^{-1} = \mathbf{A}$. In this translated frame, the derivative is then $\nabla r^2 = 2\mathbf{A}\mathbf{x}$. For points along a line through the origin, we have $\mathbf{a}^t\mathbf{x} = b$ for some contant vector $\mathbf{a}$, which specifies the direction. Different points have different offsets, b. Thus for all points $\mathbf{x}$ on the line, we can write

$$\mathbf{x} = \frac{b\mathbf{a}}{\|\mathbf{a}\|^2}.$$

Thus our derivative is

$$\nabla r^2 = 2\mathbf{A}\mathbf{x} = 2\mathbf{A}\frac{b\mathbf{a}}{\|\mathbf{a}\|^2}.$$

We are interested in the *direction* of this derivative vector, not its magnitude. Thus we now normalize the vector to find a unit vector:

$$\begin{aligned}\frac{2\mathbf{A}\frac{b\mathbf{a}}{\|\mathbf{a}\|^2}}{\sqrt{[2\mathbf{A}\mathbf{x}]^t[2\mathbf{A}\mathbf{x}]}} &= \frac{\mathbf{A}\frac{b\mathbf{a}}{\|\mathbf{a}\|^2}}{\sqrt{\left(\frac{b\mathbf{a}}{\|\mathbf{a}\|^2}\right)^t\mathbf{A}^t\mathbf{A}\frac{b\mathbf{a}}{\|\mathbf{a}\|^2}}} \\ &= \frac{\mathbf{A}\mathbf{a}}{\mathbf{a}^t\mathbf{A}^t\mathbf{A}\mathbf{a}}.\end{aligned}$$

Note especially that this normalized vector is *independent* of b, and thus independent of the point $\mathbf{x}$ on the line. Note too that in the very special case where the Gaussian in hyperspherical (i.e., $\mathbf{\Sigma} = q\mathbf{I}$, the identity matrix and q a scalar), then $\mathbf{A}\mathbf{a} = 1/q\mathbf{I}\mathbf{a} = 1/q\mathbf{a}$, and $\mathbf{A}^t\mathbf{A} = 1/q^2\mathbf{I}^t\mathbf{I} = 1/q^2\mathbf{I}$. In this case, the derivative depends only upon $\mathbf{a}$ and the scalar q.

(c) We seek now to show that ∇r_1^2 and ∇r_2^2 point in opposite direction along the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$. As we saw above, ∇r_i^2 points in the same direction along any line. Hence it is enough to show that ∇r_1^2 and ∇r_2^2 point in the opposite directions at the point $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$. From above we know that $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are parallel to each other along the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$. Also, ∇r_i^2 points in the same direction along any line through $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Thus, ∇r_1^2 points in the same direction along a line through $\boldsymbol{\mu}_1$ as ∇r_2^2 along a line through $\boldsymbol{\mu}_2$. Thus ∇r_1^2 points in the same direction along the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$ as ∇r_2^2 along the line from $\boldsymbol{\mu}_2$ to $\boldsymbol{\mu}_1$. Therefore ∇r_1^2 and ∇r_2^2 point in opposite directions along the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$.

(d) From Eqs. 60 and 62 the text, we know that the separating hyperplane has normal vector $\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. We shall show that the vector perpendicular to surfaces of constant probabilities is in the same direction, $\mathbf{w}$, at the point of intersecton of the line connecting $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and the discriminanting hyperplane. (This point is denoted $\mathbf{x}_o$.)

As we saw in part (a), the gradient obeys $\nabla r_1^2 = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x}_o - \boldsymbol{\mu}_1)$, Thus, because of the relationship between the Mahalanobis distance and probability density in a Gaussian, this vector is also in the direction normal to surfaces of constant density. Equation 65 in the text states

$$\mathbf{x}_o = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{\ln\left[P(\omega_1)/P(\omega_2)\right]}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Thus ∇r_1^2 evaluated at $\mathbf{x}_o$ is:

$$\begin{aligned}
\nabla r_1^2\Big|_{\mathbf{x}=\mathbf{x}_0} &= 2\boldsymbol{\Sigma}^{-1}(\mathbf{x}_o - \boldsymbol{\mu}_1) \\
&= 2\boldsymbol{\Sigma}^{-1}\left[\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{\ln\left[P(\omega_1)/P(\omega_2)\right](\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\right] \\
&= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\underbrace{\left[1 - \frac{2\ln\left[P(\omega_1)/P(\omega_2)\right]}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\right]}_{=scalar\ constant} \\
&\propto 2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\
&= \mathbf{w}.
\end{aligned}$$

Indeed, these two vectors are in the same direction.

(e) True. We are given that $P(\omega_1) = P(\omega_2) = 1/2$. As described in the text, for these conditions the Bayes decision boundary is given by:

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_o) = 0$$

where

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

and

$$\begin{aligned}
\mathbf{x}_o &= \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{\overbrace{\ln\left[P(\omega_1)/P(\omega_2)\right]}^{=0}}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\mathbf{x} &= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= \frac{1}{2}\left[\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \underbrace{\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}_{=0}\right].
\end{aligned}$$

Consequently the Bayes decision boundary is the set of points $\mathbf{x}$ that satisfy the following equation:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\mathbf{x} = \frac{1}{2}\left[\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right].$$

Equal squared Mahalanobis distances imply the following equivalent equations:

$$\begin{aligned} r_1^2 &= r_2^2 \\ (\mathbf{x}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) &= (\mathbf{x}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \\ \mathbf{x}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-2\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\mathbf{x} &= \mathbf{x}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_2^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2-2\boldsymbol{\mu}_2^t\boldsymbol{\Sigma}^{-1}\mathbf{x} \\ \boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 &= 2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\mathbf{x} \\ (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\mathbf{x} &= \frac{1}{2}(\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2). \end{aligned}$$

From these results it follows that the Bayes decision boundary consists of points of equal Mahalanobis distance.

27. Our Gaussian distributions are $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and prior probabilities are $P(\omega_i)$. The Bayes decision boundary for this problem is linear, given by Eqs. 63–65 in the text $\mathbf{w}^t(\mathbf{x}-\mathbf{x}_o)=0$, where

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\ \mathbf{x}_o &= \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{\ln\left[P(\omega_1)/P(\omega_2)\right]}{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2). \end{aligned}$$

The Bayes decision boundary does *not* pass between the two means, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ if and only if $\mathbf{w}^t(\boldsymbol{\mu}_1-\mathbf{x}_o)$ and $\mathbf{w}^t(\boldsymbol{\mu}_2-\mathbf{x}_o)$ have the same sign, that is, either

$$\mathbf{w}^t(\boldsymbol{\mu}_1-\mathbf{x}_o)>0 \quad \text{and} \quad \mathbf{w}^t(\boldsymbol{\mu}_2-\mathbf{x}_o)>0$$

or

$$\mathbf{w}^t(\boldsymbol{\mu}_1-\mathbf{x}_o)<0 \quad \text{and} \quad \mathbf{w}^t(\boldsymbol{\mu}_2-\mathbf{x}_o)<0.$$

These conditions are equivalent to

$$\begin{aligned} \mathbf{w}^t(\boldsymbol{\mu}_1-\mathbf{x}_o) &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right)-\ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right] \\ &= \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right] \\ \mathbf{w}^t(\boldsymbol{\mu}_2-\mathbf{x}_o) &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right)-\ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right] \\ &= -\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right], \end{aligned}$$

and therefore we have:

$$\mathbf{w}^t(\boldsymbol{\mu}_1-\mathbf{x}_o)>0 \text{ and } \mathbf{w}^t(\boldsymbol{\mu}_2-\mathbf{x}_o)>0.$$

This last equation implies

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)>2\ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right]$$

and

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)<-2\ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right].$$

Likewise, the conditions can be written as:

$$\mathbf{w}^t(\boldsymbol{\mu}_1 - \mathbf{x}_o) < 0 \text{ and } \mathbf{w}^t(\boldsymbol{\mu}_2 - \mathbf{x}_o) < 0$$

or

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 2 \ln \left[\frac{P(\omega_1)}{P(\omega_2)}\right] \text{ and}$$
$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > -2 \ln \left[\frac{P(\omega_1)}{P(\omega_2)}\right].$$

In sum, the condition that the Bayes decision boundary does not pass between the two means can be stated as follows:

Case 1 : $P(\omega_1) \leq P(\omega_2)$. Condition: $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 2 \ln \left[\frac{P(\omega_1)}{P(\omega_2)}\right]$ and this ensures $\mathbf{w}^t(\boldsymbol{\mu}_1 - \mathbf{x}_o) > 0$ and $\mathbf{w}^t(\boldsymbol{\mu}_2 - \mathbf{x}_o) > 0$.

Case 2 : $P(\omega_1) > P(\omega_2)$. Condition: $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 2 \ln \left[\frac{P(\omega_1)}{P(\omega_2)}\right]$ and this ensures $\mathbf{w}^t(\boldsymbol{\mu}_1 - \mathbf{x}_o) < 0$ and $\mathbf{w}^t(\boldsymbol{\mu}_2 - \mathbf{x}_o) < 0$.

28. We use Eqs. 42 and 43 in the text for the mean and covariance.

(a) The covariance obeys:

$$\begin{aligned}
\sigma_{ij}^2 &= \mathcal{E}\left[(x_i - \mu_i)(x_j - \mu_j)\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \underbrace{p(x_i, x_j)}_{\substack{=p(x_i)p(x_j) \\ \text{by indep.}}} (x_i - \mu_i)(x_j - \mu_j) dx_i dx_j \\
&= \int_{-\infty}^{\infty} (x_i - \mu_i) p(x_i) dx_i \int_{-\infty}^{\infty} (x_j - \mu_j) p(x_j) dx_j \\
&= 0,
\end{aligned}$$

where we have used the fact that

$$\int_{-\infty}^{\infty} x_i p(x_i) dx_i = \mu_i \qquad \text{and} \qquad \int_{-\infty}^{\infty} p(x_i) dx_i = 1.$$

(b) Suppose we had a two-dimensional Gaussian distribution, that is,

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\right),$$

where $\sigma_{12} = \mathcal{E}[(x_1 - \mu_1)(x_2 - \mu_2)]$. Furthermore, we have that the joint density is Gaussian, that is,

$$p(x_1, x_2) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right].$$

If $\sigma_{12} = 0$, then $|\boldsymbol{\Sigma}| = |\sigma_1^2 \sigma_2^2|$ and the inverse covariance matrix is diagonal, that is,

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{pmatrix}.$$

In this case, we can write

$$\begin{aligned} p(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left[-\frac{1}{2}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right]\cdot\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right] \\ &= p(x_1)p(x_2). \end{aligned}$$

Although we have derived this for the special case of two dimensions and $\sigma_{12} = 0$, the same method applies to the fully general case in d dimensions and two arbitrary coordinates i and j.

(c) Consider the following discrete distribution:

$$x_1 = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2, \end{cases}$$

and a random variable x_2 conditioned on x_1 by

$$\begin{aligned} &\text{If } x_1 = +1, \qquad x_2 = \begin{cases} +1/2 & \text{with probability } 1/2 \\ -1/2 & \text{with probability } 1/2. \end{cases} \\ &\text{If } x_1 = -1, \qquad x_2 = 0 \text{ with probability 1.} \end{aligned}$$

It is simple to verify that $\mu_1 = \mathcal{E}(x_1) = 0$; we use that fact in the following calculation:

$$\begin{aligned} \text{Cov}(x_1, x_2) &= \mathcal{E}\left[(x_1-\mu_1)(x_2-\mu_2)\right] \\ &= \mathcal{E}[x_1x_2] - \mu_2\mathcal{E}[x_1] - \mu_1\mathcal{E}[x_2] - \mathcal{E}[\mu_1\mu_2] \\ &= \mathcal{E}[x_1x_2] - \mu_1\mu_2 \\ &= \frac{1}{2}P(x_1 = +1, x_2 = +1/2) + \left(-\frac{1}{2}\right)P(x_1 = +1, x_2 = -1/2) \\ &\quad +0\cdot P(x_1 = -1) \\ &= 0. \end{aligned}$$

Now we consider

$$\begin{aligned} \Pr(x_1 = +1, x_2 = +1/2) &= \Pr(x_1 = 1, x_2 = -1/2) \\ &= \Pr(x_2 = +1/2|x_1 = 1)\Pr(x_1 = 1) \\ &= \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}. \end{aligned}$$

However, we also have

$$\Pr(x_2 = +1/2) = \Pr(x+2 = +1/2|x_1 = 1)\Pr(x_1 = 1) = 1/4.$$

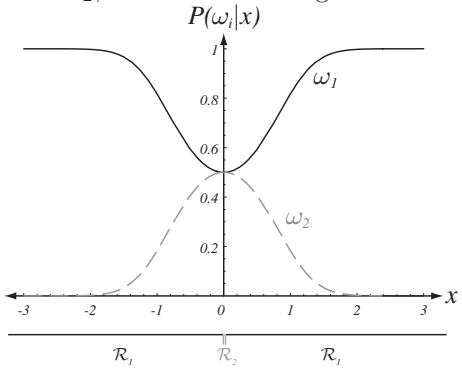
Thus we note the inequality

$$\Pr(x_1 = 1, x_2 = +1/2) = 1/4 \neq \Pr(x_1 = 1)\Pr(x_2 = +1/2) = 1/8,$$

which verifies that the features are independent. This inequality shows that even if the covariance is zero, the features need not be independent.

29. Figure 2.15 in the text shows a decision region that is a mere line in three dimensions. We can understand that unusual by analogy to a one-dimensional problem that has a decision region consisting of a single point, as follows.

(a) Suppose we have two one-dimensional Gaussians, of possibly different means and variances. The posteriors are also Gaussian, and the priors $P(\omega_i)$ can be adjusted so that the posteriors are equal at a single point, as shown in the figure. That is to say, the Bayes decision rule is to choose ω_1 *except at a single point*, where we should choose ω_2, as shown in the figure.



(b) Likewise, in the three-dimensional case of Fig. 2.15, the priors can be adjusted so that the decision region is a single *line segment*.

30. Recall that the decision boundary is given by Eq. 66 in the text, i.e., $g_1(\mathbf{x}) = g_2(\mathbf{x})$.

(a) The most general case of a hyperquadric can be written as

$$\mathbf{x}^t\mathbf{W}_1\mathbf{x} + \mathbf{w}_1^t\mathbf{x} + w_{10} = \mathbf{x}^t\mathbf{W}_2\mathbf{x} + \mathbf{w}_2^t\mathbf{x} + w_{20}$$

where $\mathbf{W}_i$ are positive definite d-by-d matrices, $\mathbf{w}_i$ are arbitrary vectors in d dimensions, and w_{i0} scalar offsets. The above equation can be written in the form

$$\frac{1}{2}\mathbf{x}^t(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1})\mathbf{x} + (\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\mathbf{x} + (w_{10} - w_{20}) = 0.$$

If we have $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$ and $P(\omega_i)$, then we can always choose $\mathbf{\Sigma}_i$ for the distribution to be $\mathbf{\Sigma}_i$ for the classifier, and choose a mean $\boldsymbol{\mu}_i$ such that $\mathbf{\Sigma}_i\boldsymbol{\mu}_i = \mathbf{w}_i$.

(b) We can still assign these values if $P(\omega_1) = P(\omega_2)$.

Section 2.7

31. Recall that from the Bayes rule, to minimize the probability of error, the decision boundary is given according to:

Choose ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise choose ω_2.

(a) In this case, since $P(\omega_1) = P(\omega_2) = 1/2$, we have $P(\omega_1|x) \propto p(x|\omega_1)$; and since the prior distributions $p(x|\omega_i)$ have the same σ, therefore the decision boundary is:

$$\text{Choose } \omega_1 \text{ if } |x-\mu_1| < |x-\mu_2|, \text{ otherwise choose } \omega_2.$$

Without loss of generality, we assume $\mu_1 < \mu_2$, and thus:

$$\begin{aligned} P(error) &= P(|x-\mu_1| > |x-\mu_2| \ |\omega_1)P(\omega_1) + P(|x-\mu_2| > |x-\mu_1| \ |\omega_2)P(\omega_2) \\ &= \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-u^2/2} du \end{aligned}$$

where $a = |\mu_2 - \mu_1|/(2\sigma)$.

(b) We note that a Gaussian decay overcomes an inverse function for sufficiently large argument, that is,

$$\lim_{a\to\infty} \frac{1}{\sqrt{2\pi}a} e^{-a^2/2} = 0$$

and

$$\begin{aligned} P_e &= \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-u^2/2} du \\ &\leq \frac{1}{\sqrt{2\pi}} \int_a^\infty \frac{u}{a} e^{-u^2/2} du \\ &= \frac{1}{\sqrt{2\pi}a} e^{-a^2/2}. \end{aligned}$$

With these results we can conclude that P_e goes to 0 as $a = |\mu_2 - \mu_1|/\sigma$ goes to infinity.

32. We note first that the categories have equal priors, $P(\omega_1) = P(\omega_2) = 1/2$, and spherical normal distributions, that is,

$$p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \sigma^2\mathbf{I}).$$

(a) The Bayes decision rule for this problem is given by:

$$\text{Choose } \omega_1 \text{ if } p(\omega_1|\mathbf{x}) > p(\omega_2|\mathbf{x}); \text{ otherwise choose } \omega_2.$$

Equal priors, $P(\omega_1) = P(\omega_2) = 1/2$, imply $P(\omega_i|\mathbf{x}) \propto p(\mathbf{x}|\omega_i)$ for $i = 1, 2$. So the decision rule is given by choosing ω_1 in any of the below functionally equivalent conditions:

$$\begin{aligned} p(\mathbf{x}|\omega_1) &> p(\mathbf{x}|\omega_2) \\ \frac{\exp\left[-1/2(\mathbf{x}-\boldsymbol{\mu}_1)^t(\sigma^2\mathbf{I})^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]}{(2\pi)^{d/2}|\sigma^2\mathbf{I}|^{d/2}} &> \frac{\exp\left[-1/2(\mathbf{x}-\boldsymbol{\mu}_2)^t(\sigma^2\mathbf{I})^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]}{(2\pi)^{d/2}|\sigma^2\mathbf{I}|^{d/2}} \\ (\mathbf{x}-\boldsymbol{\mu}_1)^t(\mathbf{x}-\boldsymbol{\mu}_1) &< (\mathbf{x}-\boldsymbol{\mu}_2)^t(\mathbf{x}-\boldsymbol{\mu}_2) \\ \mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_1^t\mathbf{x} + \boldsymbol{\mu}_1^t\boldsymbol{\mu}_1 &< \mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_2^t\mathbf{x} + \boldsymbol{\mu}_2^t\boldsymbol{\mu}_2 \\ (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t\mathbf{x} &< \frac{1}{2}(-\boldsymbol{\mu}_1^t\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^t\boldsymbol{\mu}_2), \end{aligned}$$

and otherwise choose ω_2.

The minimum probability of error P_e can be computed directly from the Bayes decision rule, but an easier way of computing P_e is as follows: Since the Bayes decision rule for the d-dimensional problem is determined by the one-dimensional random variable $\mathbf{Y} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t\mathbf{X}$, it is sufficient to consider the two-category one-dimensional problem:

$$P(\mathbf{Y}|\omega_i) \sim N(\tilde{\boldsymbol{\mu}}_i, \tilde{\sigma}^2),$$

with $P(\omega_1) = P(\omega_2) = 1/2$. In this case, the mean and variance are

$$\begin{aligned} \tilde{\mu}_i = \mathcal{E}(\mathbf{Y}|\omega_i) &= (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t\boldsymbol{\mu}_i \\ \tilde{\sigma}^2 = \mathrm{Var}(\mathbf{Y}|\omega_i) &= (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t\sigma^2\mathbf{I}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t \\ &= \sigma^2\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2. \end{aligned}$$

As shown in Problem 31, it follows that

$$P_e = \frac{1}{\sqrt{2\pi}}\int_a^\infty e^{-u^2/2}du,$$

where

$$\begin{aligned} a &= \frac{|\tilde{\mu}_2 - \tilde{\mu}_1|}{2\tilde{\sigma}} = \frac{|(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t\boldsymbol{\mu}_2 - (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t\boldsymbol{\mu}_1|}{2\sigma\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|} \\ &= \frac{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)}{2\sigma\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|} = \frac{\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2}{2\sigma\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|} \\ &= \frac{\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|}{2\sigma}. \end{aligned}$$

(b) We work in a translated coordinate system in which $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = (\mu_1, \ldots, \mu_d)^t$. We use the result

$$P_e \leq \frac{1}{\sqrt{2\pi}a}e^{-a^2/2} \rightarrow 0 \text{ if } a \rightarrow \infty,$$

where for this case we have

$$a = \frac{\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|}{2\sigma} = \frac{\|\boldsymbol{\mu}_2 - \mathbf{0}\|}{2\sigma} = \frac{1}{2\sigma}\left(\sum_{i=1}^d \mu_i^2\right)^{1/2}.$$

We conclude, then, that $P_e \rightarrow 0$ if $\sum_{i=1}^d \mu_i^2 \rightarrow \infty$ as $d \rightarrow \infty$.

(c) We have $P_e \rightarrow \infty$ as $d \rightarrow \infty$, provided $\sum_{i=1}^d \mu_i^2 \rightarrow \infty$. This implies that as $d \rightarrow \infty$, the two points $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ in d-dimensions become well separated (unless $\sum_{i=1}^\infty \mu_i^2 < \infty$ implies $\mu_i \rightarrow 0$ as $i \rightarrow \infty$). Consequently, with increasing d, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ can be easily distinguished and $P_e \rightarrow 0$.

33. Consider a line in d dimensions that is projected down to a line in the lower dimension space.

(a) The conditional error in d dimensions is

$$\begin{aligned} E_d &= \int_{\text{line}} \min[P(\omega_1)p(\mathbf{x}|\omega_1), P(\omega_2)p(\mathbf{x}|\omega_2)]\ dx \\ &= \int_{\Gamma_1} P(\omega_1)p(\mathbf{x}|\omega_1)\ dx + \int_{\Gamma_2} P(\omega_2)p(\mathbf{x}|\omega_2)\ dx, \end{aligned}$$

where Γ_1 and Γ_2 are the disjoint segments where the posterior probabilities of each of the distributions are minimum.

In the lower dimension space we have

$$E_{d-1} = \int_{\text{line}} \min\Big[P(\omega_1)p(\mathbf{x}|\omega_1)\ dx, P(\omega_2)p(\mathbf{x}|\omega_2)\Big]\ dx.$$

Let us suppose for definiteness that

$$\int_{\text{line}} P(\omega_1)p(\mathbf{x}|\omega_1)\ dx < \int_{\text{line}} P(\omega_2)p(\mathbf{x}|\omega_2)\ dx,$$

and thus

$$E_{d-1} = \int_{\text{line}} P(\omega_1)p(\mathbf{x}|\omega_1)\ dx.$$

If we let $\Gamma_1 + \Gamma_2$ denote the full line, we have

$$\begin{aligned} E_{d-1} &= \int_{\Gamma_1+\Gamma_2} P(\omega_1)p(\mathbf{x}|\omega_1)\ dx \\ &= \int_{\Gamma_1} P(\omega_1)p(\mathbf{x}|\omega_1)\ dx + \int_{\Gamma_2} P(\omega_1)p(\mathbf{x}|\omega_1)\ dx \\ &= \int_{\Gamma_1} P(\omega_1)p(\mathbf{x}|\omega_1)\ dx + \int_{\Gamma_2} [P(\omega_2)p(\mathbf{x}|\omega_2) + |f(x)|]\ dx \\ &= E_d + \int_{\Gamma_2} |f(x)|\ dx \\ &\geq E_d, \end{aligned}$$

where we have used the fact that $P(\omega_1)p(\mathbf{x}|\omega_1) = P(\omega_2)p(\mathbf{x}|\omega_2) + |f(x)|$ in Γ_2 for some unknown function f.

(b) In actual pattern recognition problems we rarely have the true distributions, but have instead just *estimates*. The above derivation does not hold if the distributions used are merely estimates.

Section 2.8

34. Consider the Bhattacharyya and Chernoff bounds for $\mu_1 = -\mu_2 = \mu$ and $\sigma_1^2 = \sigma_2^2 = \mu^2$.

(a) In this case, the Chernoff bound for the one-dimensional problem is given by Eqs. 73–75 in the text,

$$\begin{aligned} k(\beta) &= \frac{\beta(1-\beta)}{2} 2\mu[\beta\mu^2 + (1-\beta)\mu^2]^{-1} 2\mu + \frac{1}{2}\ln\frac{\mu^2}{\mu^2} \\ &= 2\beta(1-\beta). \end{aligned}$$

We set the derivative to zero,

$$\frac{\partial k(\beta)}{\partial \beta} = 2(1-\beta) - 2\beta = 0,$$

to find the value of β that leads to the extreme value of this error. This optimizing value is $\beta^* = 0.5$, which gives $k(\beta^*) = 0.5$. The Chernoff bound is thus

$$P_{Cher}(error) \leq \sqrt{P(\omega_1)P(\omega_2)}e^{-1/2}.$$

In this case, the Bhattacharyya bound is the same as the Chernoff bound:

$$P_{Bhat}(error) \leq \sqrt{P(\omega_1)P(\omega_2)}e^{-1/2}.$$

(b) Recall the Bayes error given by Eqs. 5 and 7 in the text:

$$\begin{aligned} P(error) &= P(x \in \mathcal{R}_2|\omega_1)P(\omega_1) + P(x \in \mathcal{R}_1|\omega_2)P(\omega_2)) \\ &= \int_{\mathcal{R}_2} p(x|\omega_2)P(\omega_1)dx + \int_{\mathcal{R}_1} p(x|\omega_1)P(\omega_2)dx. \end{aligned}$$

In order to find the decision point, we set $p(x|\omega_1)P(\omega_1) = p(x|\omega_2)P(\omega_2)$ and find

$$\frac{1}{\sqrt{2\pi}\sigma}\exp[-(x-\mu)^2/(2\sigma^2)]P(\omega_2) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-(x+\mu)^2/(2\sigma^2)]P(\omega_1),$$

and thus

$$\ln P(\omega_2) - \frac{(x-\mu)^2}{2\sigma^2} = \ln P(\omega_1) - \frac{(x+\mu)^2}{2\sigma^2}.$$

The Bayes value x_B that satisfies this equation is

$$x_B = \frac{\sigma^2}{2\mu}\ln\frac{P(\omega_1)}{P(\omega_2)}.$$

We substitute x_B into the equation for the probability of error and find

$$P(error) = \int_{-\infty}^{x_B} p(x|\omega_2)P(\omega_2)dx + \int_{x_B}^{\infty} p(x|\omega_1)P(\omega_1)dx$$

$$\begin{aligned}
&= \int_{-\infty}^{x_B} P(\omega_2)\frac{1}{\sqrt{2\pi}\sigma}\exp[-(x-\mu)^2/(2\sigma^2)]dx + \int_{x_B}^{\infty} P(\omega_1)\frac{1}{\sqrt{2\pi}\sigma}\exp[-(x+\mu)^2/(2\sigma^2)]dx \\
&= 1-\text{erf}[(x_B-\mu)/\sigma)]P(\omega_2) + \text{erf}[(x_B+\mu)/\sigma]P(\omega_1) \\
&= 1+(P(\omega_1)-P(\omega_2))\text{erf}\left[\frac{\sigma}{2\mu}\ln\frac{P(\omega_1)}{P(\omega_2)} - \frac{\mu}{\sigma}\right].
\end{aligned}$$

(c) Problem not yet solved

(d) Problem not yet solved

35. We let $P_d(P(\omega_1), \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, P(\omega_2), \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, or simply P_d, denote the Bhattacharyya error bound if we consider the distributions restricted to d dimensions.

(a) We seek to show that $k(1/2)$ must increase as we go from d to $d+1$ dimensions. Let $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, k(1/2)$ be the relevant quantities in $(d+1)$ dimensions. Equation 75 in the text gives us:

$$\tilde{k}(1/2) = \frac{1}{8}(\tilde{\boldsymbol{\mu}}_2-\tilde{\boldsymbol{\mu}}_1)^t\left[\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}\right]^{-1}(\tilde{\boldsymbol{\mu}}_2-\tilde{\boldsymbol{\mu}}_1) + \frac{1}{2}\ln\frac{\left|\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}\right|}{\sqrt{|\tilde{\boldsymbol{\Sigma}}_1||\tilde{\boldsymbol{\Sigma}}_2|}}.$$

We make the following definitions:

$$\begin{aligned}
\mathbf{A} &= \frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2} \\
\tilde{\mathbf{A}} &= \frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2} \\
&= \begin{bmatrix} \mathbf{A}_{d\times d} & \mathbf{u}_{d\times 1} \\ \mathbf{u}^t_{1\times d} & c_{1\times 1} \end{bmatrix} \\
\tilde{\boldsymbol{\mu}}_2-\tilde{\boldsymbol{\mu}}_1 &= \begin{bmatrix} \boldsymbol{\theta}_{d\times 1} \\ \phi_{1\times 1} \end{bmatrix}, \qquad \text{where } \boldsymbol{\theta} = \boldsymbol{\mu}_2-\boldsymbol{\mu}_1.
\end{aligned}$$

Substituting, we then have

$$(\tilde{\boldsymbol{\mu}}_2-\tilde{\boldsymbol{\mu}}_1)^t\left[\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}\right]^{-1}(\tilde{\boldsymbol{\mu}}_2-\tilde{\boldsymbol{\mu}}_1) = (\boldsymbol{\theta}^t\ \phi)\begin{bmatrix} \mathbf{A} & \mathbf{u} \\ \mathbf{u}^t & c \end{bmatrix}^{-1}\begin{pmatrix} \boldsymbol{\theta}^t \\ \phi \end{pmatrix}.$$

Note that $\mathbf{A} = (\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)/2$ is invertible and $\tilde{\mathbf{A}} = (\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2)/2$ is the average of the appropriate covariance matrices,

$$\tilde{\mathbf{A}}^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{u} \\ \mathbf{u}^t & c \end{bmatrix}^{-1} = \begin{bmatrix} x\mathbf{A}^{-1}+\alpha\mathbf{A}^{-1}\mathbf{u}\mathbf{u}^t\mathbf{A}^{-1} & -\alpha\mathbf{A}^{-1}\mathbf{u} \\ -\alpha\mathbf{u}^t\mathbf{A}^{-1} & \alpha \end{bmatrix},$$

where $\alpha = c-\mathbf{u}^t\mathbf{A}^{-1}\mathbf{u}$. Thus, we have the following:

$$\begin{aligned}
&(\tilde{\boldsymbol{\mu}}_2-\tilde{\boldsymbol{\mu}}_1)\left(\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}\right)^{-1}(\tilde{\boldsymbol{\mu}}_2-\tilde{\boldsymbol{\mu}}_1) \\
= &\ (\boldsymbol{\theta}^t\ \phi)\begin{bmatrix} x\mathbf{A}^{-1}+\alpha\mathbf{A}^{-1}\mathbf{u}\mathbf{u}^t\mathbf{A}^{-1} & -\alpha\mathbf{A}^{-1}\mathbf{u} \\ -\alpha\mathbf{u}^t\mathbf{A}^{-1} & \alpha \end{bmatrix}\begin{pmatrix} \boldsymbol{\theta} \\ \phi \end{pmatrix}
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\theta}^t\mathbf{A}^{-1}\boldsymbol{\theta} + \alpha\boldsymbol{\theta}^t\mathbf{A}^{-1}\mathbf{u}\mathbf{u}^t\mathbf{A}^{-1}\boldsymbol{\theta} - \phi\alpha\mathbf{u}^t\mathbf{A}^{-1}\boldsymbol{\theta} - \phi\alpha\boldsymbol{\theta}^t\mathbf{A}^{-1}\mathbf{u} + \phi^2\alpha \\
&= \boldsymbol{\theta}^t\mathbf{A}^{-1}\boldsymbol{\theta} + \alpha(\phi - \boldsymbol{\theta}^t\mathbf{A}^{-1}\mathbf{u})^2 \\
&\geq \boldsymbol{\theta}^t\mathbf{A}^{-1}\boldsymbol{\theta} \\
&\quad \text{as } \alpha = c - \mathbf{u}^t\mathbf{A}^{-1}\mathbf{u} = \frac{|\tilde{\mathbf{A}}|}{|\mathbf{A}|} > 0 \text{ (and } \tilde{\mathbf{A}} \text{ and } \mathbf{A} \text{ are positive definite)} \\
&= (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right)^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).
\end{aligned}
$$

Consequently, we have

$$
(\tilde{\boldsymbol{\mu}}_2 - \tilde{\boldsymbol{\mu}}_1) \left(\frac{\tilde{\boldsymbol{\Sigma}}_1 + \tilde{\boldsymbol{\Sigma}}_2}{2}\right)^{-1} (\tilde{\boldsymbol{\mu}}_2 - \tilde{\boldsymbol{\mu}}_1) \geq (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right)^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1),
$$

where equality holds if an only if $\phi = \boldsymbol{\theta}^t\mathbf{A}\mathbf{A}^{-1}\mathbf{u}$. Now we let

$$
\tilde{\boldsymbol{\Sigma}}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{u}_i \\ \mathbf{u}_i^t & c_i \end{bmatrix}, \qquad i = 1, 2.
$$

Then $|\tilde{\boldsymbol{\Sigma}}_i| = \alpha_i|\boldsymbol{\Sigma}_i|$, where $\alpha_i = c_i - \mathbf{u}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{u}_i$. We have

$$
|\tilde{\mathbf{A}}| = \left|\frac{\tilde{\boldsymbol{\Sigma}}_1 + \tilde{\boldsymbol{\Sigma}}_2}{2}\right| = \left|\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right| \cdot \alpha = |\mathbf{A}|\alpha
$$

where $\alpha = c - \mathbf{u}\mathbf{A}^{-1}\mathbf{u}$. Now, we also expand our definition of $\tilde{\mathbf{A}}$:

$$
\tilde{\mathbf{A}} = \frac{\tilde{\boldsymbol{\Sigma}}_1 + \tilde{\boldsymbol{\Sigma}}_2}{2} = \left[\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right] = \begin{bmatrix} \frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2} & \frac{\mathbf{u}_1+\mathbf{u}_2}{2} \\ \frac{\mathbf{u}_1^t+\mathbf{u}_2^t}{2} & \frac{c_1+c_2}{2} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{u} \\ \mathbf{u}^t & c \end{bmatrix}.
$$

Thus, $\mathbf{u} = (\mathbf{u}_1 + \mathbf{u}_2)/2$ and $c = (c_1 + c_2)/2$. We substitute these values into the above equations and find

$$
\begin{aligned}
\ln\left(\frac{|\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}|}{\sqrt{|\tilde{\boldsymbol{\Sigma}}_1||\tilde{\boldsymbol{\Sigma}}_2|}}\right) &= \ln\left(\frac{|\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}|\alpha}{\sqrt{|\boldsymbol{\Sigma}_1\alpha_1||\boldsymbol{\Sigma}_2\alpha_2|}}\right) \\
&= \ln\left(\frac{|\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}\right) + \ln\frac{\alpha}{\sqrt{\alpha_1\alpha_2}}
\end{aligned}
$$

If $\mathbf{X}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, and $\tilde{\mathbf{X}}_i = (\mathbf{X}_i, X_{i,d+1}) \sim N(\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i)$ for $i = 1, 2$, then the conditional variance of $X_{i,d+1}$ given $\mathbf{X}_i$ is:

$$
\begin{aligned}
\alpha_i &= \text{Var}(X_{i,d+1}|\mathbf{X}_i), i = 1, 2. \\
\mathbf{Y} &= \frac{\mathbf{X}_1 + \mathbf{X}_2}{\sqrt{2}} \\
\tilde{\mathbf{Y}} &= (\mathbf{Y}, Y_{d+1}) = \frac{\tilde{\mathbf{X}}_1 + \tilde{\mathbf{X}}_2}{\sqrt{2}},
\end{aligned}
$$

and thus $\alpha = \text{Var}(Y_{d+1}|\mathbf{Y})$. Assume now that $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. In that case we have

$$
\alpha = \text{Var}(Y_{d+1}|\mathbf{Y}) = \text{Var}(1/\sqrt{2}\ (X_{1,d+1} + X_{2,d+1})|\mathbf{X}_1, \mathbf{X}_2)
$$

$$\begin{aligned} &= \frac{1}{2}\left[\mathrm{Var}(X_{1,d+1}\mathbf{X}_1) + \mathrm{Var}(X_{2,d+1}|\mathbf{X}_2)\right] \\ &= \frac{1}{2}(\alpha_1+\alpha_2). \end{aligned}$$

We substitute these values and find

$$\ln\left(\frac{\left|\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}\right|}{\sqrt{|\tilde{\boldsymbol{\Sigma}}_1|\,|\tilde{\boldsymbol{\Sigma}}_2|}}\right)_{d+1} = \ln\left(\frac{\left|\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}\right|}{\sqrt{|\tilde{\boldsymbol{\Sigma}}_1|\,|\tilde{\boldsymbol{\Sigma}}_2|}}\right)_{d} + \underbrace{\ln\frac{\frac{\alpha_1+\alpha_2}{2}}{\sqrt{\alpha_1\alpha_2}}}_{\substack{\geq 0 \ since \\ \frac{\alpha_1+\alpha_2}{2}\geq\sqrt{\alpha_1\alpha_2}}}$$

and thus

$$\ln\left(\frac{\left|\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}\right|}{\sqrt{|\tilde{\boldsymbol{\Sigma}}_1|\,|\tilde{\boldsymbol{\Sigma}}_2|}}\right)_{d+1} \geq \ln\left(\frac{\left|\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}\right|}{\sqrt{|\tilde{\boldsymbol{\Sigma}}_1|\,|\tilde{\boldsymbol{\Sigma}}_2|}}\right)_{d},$$

where equality holds if and only if $(\alpha_1+\alpha_2)/2 = \sqrt{\alpha_1\alpha_2}$, or equivalently if $\alpha_1 = \alpha_2$.

We put these results together and find that $\tilde{k}(1/2) \geq k(1/2)$. This implies

$$P_{d+1} = \sqrt{P(\omega_1)P(\omega_2)}e^{-\tilde{k}(1/2)} \leq \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)} = P_d.$$

That is, $P_{d+1} \leq P_d$, as desired.

(b) The above result was derived adding the $(d+1)^{th}$ dimension and does not depend on *which* dimension is added. This is because, by a permutation of the coordinates, any dimension that is added can be designated as the $(d+1)^{th}$ dimension. This gives a permutation of the coordinates of $\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, i = 1, 2$, but does not change the error bound.

(c) The "pathological case" where equality holds (that is, $P_{d+1} = P_d$) is when:

$$\begin{aligned} \phi &= \boldsymbol{\theta}^t\mathbf{A}^{-1}\mathbf{u} \\ \alpha_1 &= \alpha_2, \end{aligned}$$

where

$$\begin{aligned} \theta &= \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \\ \mathbf{A} &= \left(\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right) \\ \tilde{\mathbf{A}} &= \left(\frac{\tilde{\boldsymbol{\Sigma}}_1+\tilde{\boldsymbol{\Sigma}}_2}{2}\right) = \begin{bmatrix} \mathbf{A} & \mathbf{u} \\ \mathbf{u}^t & c \end{bmatrix} \\ \phi &= \mu_{2,d+1} - \mu_{1,d+1} \\ \alpha_i &= c_i - \mathbf{u}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{u}_i. \end{aligned}$$

For exmple, the condition of equality is satisfied if $\tilde{\boldsymbol{\mu}}_1 = \tilde{\boldsymbol{\mu}}_2$ and $\tilde{\boldsymbol{\Sigma}}_1 = \tilde{\boldsymbol{\Sigma}}_2$.

(d) No, the true error can never increase as we go to a higher dimension, as illustrated in Fig. 3.3 in the text.

(e) For non-pathological distributions, $P_d \propto e^{-k(1/2)}$ goes to zero as $d \to \infty$. This is because $k(1/2) \to \infty$ for $d \to \infty$.

(f) No. First note that

$$\begin{aligned} P_d(error) \leq P_d &= \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)} \\ P_{d+1}(error) \leq P_{d+1} &= \sqrt{P(\omega_1)P(\omega_2)}e^{-\tilde{k}(1/2)}. \end{aligned}$$

But, there is no clear relation between $P_{d+1}(error)$ and $P_d = \sqrt{P(\omega_1)(P(\omega_2)}e^{-k(1/2)}$. So, even if $\tilde{k}(1/2) > k(1/2)$, it is not guaranteed that $P_{d+1}(error) < P_d(error)$.

36. First note the definition of $k(\beta)$ given by Eq. 75 in the text:

$$\begin{aligned} k(\beta) &= \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t(\beta\boldsymbol{\Sigma}_2 + (1-\beta)\boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \\ &\quad + \frac{1}{2}\ln\left[\frac{|\beta\boldsymbol{\Sigma}_2 + (1-\beta)\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|^\beta|\boldsymbol{\Sigma}_1|^{1-\beta}}\right]. \end{aligned}$$

(a) Recall from Eq. 74 we have

$$\begin{aligned} e^{-k(\beta)} &= \int p^\beta(\mathbf{x}|\omega_1)p^{1-\beta}(\mathbf{x}|\omega_2)\, d\mathbf{x} \\ &= \int \left[\frac{\exp\left[-\frac{\beta}{2}\boldsymbol{\mu}^t\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \frac{(1-\beta)}{2}\boldsymbol{\mu}^t\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right]}{|\boldsymbol{\Sigma}_1|^{\beta/2}|\boldsymbol{\Sigma}_2|^{\beta/2}}\right. \\ &\quad \left.\times\frac{\exp\left[-\frac{1}{2}\{\mathbf{x}^t(\beta\boldsymbol{\Sigma}_1^{-1} + (1-\beta)\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - 2\mathbf{x}^t(\beta\boldsymbol{\Sigma}_1^{-1} + (1-\beta)\boldsymbol{\Sigma}_2^{-1})\}\right]}{(2\pi)^{d/2}}\right] d\mathbf{x} \end{aligned}$$

(b) Again from Eq. 74 we have

$$\begin{aligned} e^{-k(\beta)} &= \frac{\exp\left[-\beta/2\ \boldsymbol{\mu}_1^t\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - (1-\beta)/2\ \boldsymbol{\mu}_2^t\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right]}{|\boldsymbol{\Sigma}_1|^{\beta/2}|\boldsymbol{\Sigma}_2|^{(1-\beta)/2}} \\ &\quad \times\int\frac{\exp\left[-\frac{1}{2}\{\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x} - 2\mathbf{x}\mathbf{A}^{-1}\boldsymbol{\theta}\}\right]}{(2\pi)^{d/2}}d\mathbf{x} \end{aligned}$$

where

$$\mathbf{A} = \left(\beta\boldsymbol{\Sigma}_1^{-1} + (1-\beta)\boldsymbol{\Sigma}_2^{-1}\right)^{-1}$$

and

$$\mathbf{A}^{-1}\boldsymbol{\theta} = \beta\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + (1-\beta)\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2.$$

Thus we conclude that the vector $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta} = \mathbf{A}(\beta\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + (1-\beta)\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2).$$

(c) For the conditions given, we have

$$\begin{aligned} \int\exp\left[\frac{1}{2}(\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x} - 2\mathbf{x}^t\mathbf{A}^{-1}\boldsymbol{\theta})\right]d\mathbf{x} &= e^{\frac{1}{2}\boldsymbol{\theta}^t\mathbf{A}^{-1}\boldsymbol{\theta}}\int\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\theta})^t\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\theta})\right]d\mathbf{x} \\ &= (2\pi)^{d/2}e^{1/2\boldsymbol{\theta}^t\mathbf{A}^{-1}\boldsymbol{\theta}}|\mathbf{A}|^{1/2} \end{aligned}$$

since

$$g(\mathbf{x}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\theta})^t\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\theta})}}{(2\pi)^{d/2}|\mathbf{A}|^{1/2}},$$

where $g(\mathbf{x})$ has the form of a d-dimensional Gaussian density. So it follows that

$$e^{-k(\beta)} = \exp\left[-\frac{1}{2}\{-\boldsymbol{\theta}\mathbf{A}^{-1}\boldsymbol{\theta} + \beta\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu} + (1-\beta)\boldsymbol{\mu}_2^t\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\}\right] \times \frac{|\mathbf{A}|^{1/2}}{|\boldsymbol{\Sigma}_1|^{\beta/2}|\boldsymbol{\Sigma}_2|^{(1-\beta)/2}} \cdot []$$

PROBLEM NOT YET SOLVED

37. We are given that $P(\omega_1) = P(\omega_2) = 0.5$ and

$$\begin{aligned} p(\mathbf{x}|\omega_1) &\sim N(\mathbf{0}, \mathbf{I}) \\ p(\mathbf{x}|\omega_2) &\sim N(\mathbf{1}, \mathbf{I}) \end{aligned}$$

where $\mathbf{1}$ is a two-component vector of 1s.

(a) The inverse matrices are simple in this case:

$$\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We substitute these into Eqs. 53–55 in the text and find

$$\begin{aligned} g_1(\mathbf{x}) &= \mathbf{w}_1^t\mathbf{x} + w_{10} \\ &= \mathbf{0}^t\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + 0 + \ln(1/2) \\ &= \ln(1/2) \end{aligned}$$

and

$$\begin{aligned} g_2(\mathbf{x}) &= \mathbf{w}_2^t\mathbf{x} + w_{20} \\ &= (1,1)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \frac{1}{2}(1,1)\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \ln(1/2) \\ &= x_1 + x_2 - 1 + \ln(1/2). \end{aligned}$$

We set $g_1(\mathbf{x}) = g_2(\mathbf{x})$ and find the decision boundardy is $x_1 + x_2 = 1$, which passes through the midpoint of the two means, that is, at

$$(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}.$$

This result makes sense because these two categories have the same prior and conditional distributions except for their means.

(b) We use Eqs. 76 in the text and substitute the values given to find

$$\begin{aligned} k(1/2) &= \frac{1}{8}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right)^t \left[\frac{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}{2}\right]^{-1} \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) + \frac{1}{2}\ln\frac{\left|\frac{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}{2}\right|}{\sqrt{\left|\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right|\left|\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right|}} \\ &= \frac{1}{8}\begin{pmatrix} 1 \\ 1 \end{pmatrix}^t\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2}\ln\frac{1}{1} \\ &= 1/4. \end{aligned}$$

Equation 77 in the text gives the Bhatacharyya bound as

$$P(error) \leq \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)} = \sqrt{0.5 \cdot 0.5}e^{-1/4} = 0.3894.$$

(c) Here we have $P(\omega_1) = P(\omega_2) = 0.5$ and

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad , \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$
$$\boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad , \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}.$$

The inverse matrices are

$$\begin{aligned} \boldsymbol{\Sigma}_1^{-1} &= \begin{pmatrix} 8/5 & -2/15 \\ -2/15 & 8/15 \end{pmatrix} \\ \boldsymbol{\Sigma}_2^{-1} &= \begin{pmatrix} 5/9 & -4/9 \\ -4/9 & 5/9 \end{pmatrix}. \end{aligned}$$

We use Eqs. 66–69 and find

$$\begin{aligned} g_1(\mathbf{x}) &= -\frac{1}{2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^t \begin{pmatrix} 8/5 & -2/15 \\ -2/15 & 8/15 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \left(\begin{pmatrix} 8/5 & -2/15 \\ -2/15 & 8/15 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right)^t \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &\quad -\frac{1}{2}\begin{pmatrix} 0 \\ 0 \end{pmatrix}^t \begin{pmatrix} 8/5 & -2/15 \\ -2/15 & 8/15 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix} - \frac{1}{2}\ln\left|\begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right| + \ln\frac{1}{2} \\ &= -\frac{4}{15}x_1^2 + \frac{2}{15}x_1x_2 - \frac{4}{15}x_2^2 - 0.66 + \ln\frac{1}{2}, \end{aligned}$$

and

$$\begin{aligned} g_2(\mathbf{x}) &= -\frac{1}{2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^t \begin{pmatrix} 5/9 & -4/9 \\ -4/9 & 5/9 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \left(\begin{pmatrix} 5/9 & -4/9 \\ -4/9 & 5/9 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)^t \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &\quad -\frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}^t \begin{pmatrix} 5/9 & -4/9 \\ -4/9 & 5/9 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2}\ln\left|\begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}\right| + \ln\frac{1}{2} \\ &= -\frac{5}{18}x_1^2 + \frac{8}{18}x_1x_2 - \frac{5}{18}x_2^2 + \frac{1}{9}x_1 + \frac{1}{9}x_2 - \frac{1}{9} - 1.1 + \ln\frac{1}{2}. \end{aligned}$$

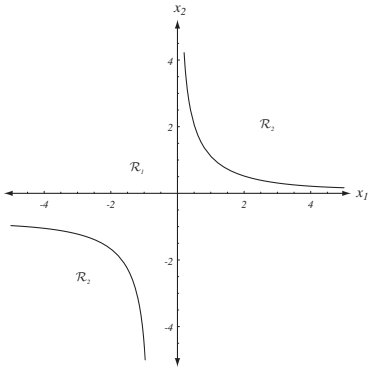
The Bayes decision boundary is the solution to $g_1(\mathbf{x}) = g_2(\mathbf{x})$ or

$$x_1^2 + x_2^2 - 28x_1x_2 - 10x_1 - 10x_2 + 50 = 0,$$

which consists of two hyperbolas, as shown in the figure.

We use Eqs. 76 and 77 in the text and find

$$\begin{aligned} k(1/2) &= \frac{1}{8}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right)^t \left[\frac{\begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix} + \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}}{2}\right]^{-1} \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) + \ln\frac{\left|\frac{\begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix} + \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}}{2}\right|}{\sqrt{\left|\begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right|\left|\begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}\right|}} \\ &= \frac{1}{8}\begin{pmatrix} 1 \\ 1 \end{pmatrix}^t \begin{pmatrix} 3.5 & 2.25 \\ 2.25 & 3.5 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2}\ln\frac{7.1875}{5.8095} \\ &= 0.1499. \end{aligned}$$



Equation 77 in the text gives the Bhatacharyya bound as

$$P(error) \leq \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)} = \sqrt{0.5 \cdot 0.5}e^{-1.5439} = 0.4304.$$

38. We derive the Bhattacharyya error bound without first examining the Chernoff bound as follows.

(a) We wish to show that $\min[a, b] \leq \sqrt{ab}$. We suppose without loss of generality that $a \leq b$, or equivalently $b \leq a + \delta$ for $\delta > 0$. Thus $\sqrt{ab} = \sqrt{a(a+\delta)} \geq \sqrt{a^2} = a = \min[a, b]$.

(b) Using the above result and the formula for the probability of error given by Eq. 7 in the text, we have:

$$\begin{aligned} P(error) &= \int \min\left[P(\omega_1)p(\mathbf{x}|\omega_1), P(\omega_2)p(\mathbf{x}|\omega_2)\right]\, d\mathbf{x} \\ &\leq \underbrace{\sqrt{P(\omega_1)P(\omega_2)}}_{\leq 1/2} \underbrace{\int \sqrt{p(\mathbf{x}|\omega_1)p(\mathbf{x}|\omega_2)}\, d\mathbf{x}}_{=\rho} \\ &\leq \rho/2, \end{aligned}$$

where for the last step we have used the fact that $\min[P(\omega_1), P(\omega_2)] \leq 1/2$, which follows from the normalization condition $P(\omega_1) + P(\omega_2) = 1$.

39. We assume the underlying distributions are Gaussian.

(a) Based on the Gaussian assumption, we can calculate $(x^* - \mu_2)/\sigma_2$ from the hit rate $P_{hit} = P(x > x^*|x \in \omega_2)$. We can also calculate $(x^* - \mu_1)/\sigma_1$ from the false alarm rate $P_{false} = P(x > x^*|x \in \omega_1)$. Since $\sigma_1 = \sigma_2 = \sigma$, the discriminability is simply

$$d' = \left|\frac{\mu_2 - \mu_1}{\sigma}\right| = \left|\frac{x^* - \mu_1}{\sigma_1} - \frac{x^* - \mu_2}{\sigma_2}\right|.$$

(b) Recall the error function from Eq. 96 in the Appendix of the text:

$$\text{erf}[u] = \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2} dx.$$

We can express the probability of a Gaussian distribution in terms of the error function as:

$$P(x > x^*) = 1/2 - \text{erf}\left[\frac{x^* - \mu}{\sigma}\right]$$

and thus

$$\frac{x^* - \mu}{\sigma} = \text{erf}^{-1}\left[1/2 - P(x > x^*)\right].$$

We let $P_{hit} = P(x > x^*|x \in \omega_2)$ and $P_{false} = P(x > x^*|x \in \omega_1)$. The discriminability can be written as

$$d' = \frac{\mu_2 - \mu_1}{\sigma} = \frac{x^* - \mu_1}{\sigma} - \frac{x^* - \mu_2}{\sigma} = \text{erf}^{-1}[1/2 - P_{false}] - \text{erf}^{-1}[1/2 - P_{hit}].$$

We substitute the values for this problem and find

$$\begin{aligned} d'_1 &= \text{erf}^{-1}[0.2] - \text{erf}^{-1}[-0.3] = 0.52 + 0.84 = 1.36 \\ d'_2 &= \text{erf}^{-1}[0.1] - \text{erf}^{-1}[-0.2] = 0.26 + 0.52 = 0.78. \end{aligned}$$

(c) According to Eq. 70 in the text, we have

$$\begin{aligned} \text{Case 1}: \quad & P(error) = \frac{1}{2}[0.3 + (1 - 0.8)] = 0.25 \\ \text{Case 2}: \quad & P(error) = \frac{1}{2}[0.4 + (1 - 0.7)] = 0.35. \end{aligned}$$

(d) Because of the symmetry property of the ROC curve, the point (P_{hit}, P_{false}) and the point $(1 - P_{hit}, 1 - P_{false})$ will go through the same curve corresponding to some fixed d'. For case B, (0.1, 0.3) is also a point on ROC curve that (0.9, 0.7) lies. We can compare this point with case A, going through (0.8, 0.3) and the help of Fig. 2.20 in the text, we can see that case A has a higher discriminability d'.

40. We are to assume that the two Gaussians underlying the ROC curve have different variances.

(a) From the hit rate $P_{hit} = P(x > x^*|x \in \omega_2)$ we can calculate $(x^* - \mu_2)/\sigma_2$. From the false alarm rate $P_{false} = P(x > x^*|x \in \omega_1)$ we can calculate $(x^* - \mu_1)/\sigma_1$. Let us denote the ratio of the standard deviations as $\sigma_1/\sigma_2 = K$. Then we can write the discriminability in this case as

$$d'_a = \left|\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1\sigma_2}}\right| = \left|\frac{\mu_2 - x^*}{\sqrt{\sigma_1\sigma_2}} - \frac{x^* - \mu_1}{\sqrt{\sigma_1\sigma_2}}\right| = \left|\frac{x^* - \mu_2}{\sigma_2/K} - \frac{x^* - \mu_1}{K\sigma_1}\right|.$$

Because we cannot determine K from $(\mu_2 - x^*)/\sigma_2$ and $(x^* - \mu_1)/\sigma_1$, we cannot determine d' uniquely with only $P_{hit} = P(x > x^*|x \in \omega_2)$ and $P_{false} = P(x > x^*|x \in \omega_1)$.

(b) Suppose we are given the following four experimental rates:

$$\begin{aligned} P_{hit1} &= P(x > x_1^*|\omega_2) & P_{false1} &= P(x > x_1^*|\omega_1) \\ P_{hit2} &= P(x > x_2^*|\omega_2) & P_{false2} &= P(x > x_2^*|\omega_1). \end{aligned}$$

Then we can calculate the four quantities

$$a_1 = \frac{x_1^* - \mu_2}{\sigma_2} = \text{erf}^{-1}[1/2 - P_{hit1}] \qquad b_1 = \frac{x_1^* - \mu_1}{\sigma_1} = \text{erf}^{-1}[1/2 - P_{false1}] \text{ for } x_1^*$$
$$a_2 = \frac{x_2^* - \mu_2}{\sigma_2} = \text{erf}^{-1}[1/2 - P_{hit2}] \qquad b_2 = \frac{x_2^* - \mu_1}{\sigma_1} = \text{erf}^{-1}[1/2 - P_{false2}] \text{ for } x_2^*.$$

Then we have the following relations:

$$\begin{aligned} a_1 - a_2 &= \frac{x_1^* - x_2^*}{\sigma_2} \\ b_1 - b_2 &= \frac{x_1^* - x_2^*}{\sigma_1} \\ K &= \frac{a_1 - a_2}{b_1 - b_2} = \frac{\sigma_1}{\sigma_2}. \end{aligned}$$

Thus, with K we can calculate d_a' as

$$\begin{aligned} d_a' &= \left|\frac{\mu_2 - \mu_1}{\sqrt{\sigma_1\sigma_2}}\right| = \left|\frac{\mu_2 - x_1^*}{\sigma_2/K} - \frac{x_1^* - \mu_1}{K\sigma_1}\right| \\ &= \left| - \frac{(a_1 - a_2)a_1}{b_1 - b_2} - \frac{(b_1 - b_2)b_1}{a_1 - a_2}\right|. \end{aligned}$$

(c) For all those x_1^* and x_2^* that satisfy

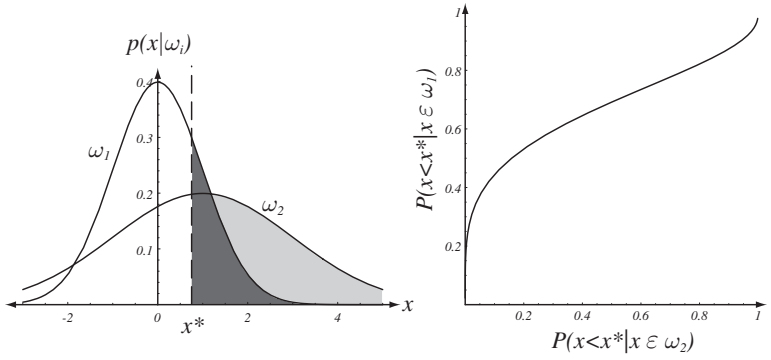
$$\frac{x_1^* - \mu_2}{\sigma_2} = -\frac{x_2^* - \mu_2}{\sigma_2}$$

or

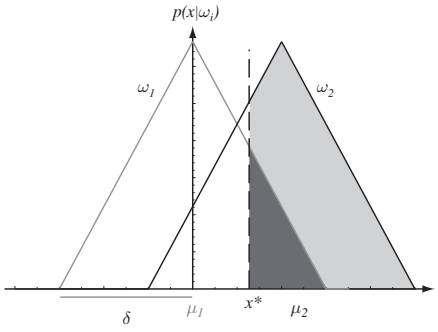
$$\frac{x_1^* - \mu_1}{\sigma_1} = -\frac{x_2^* - \mu_1}{\sigma_1}.$$

That is, the two different thresholds do not provide any additional information and conveys the same information as only one observation. As explained in part (a), this kind of result would not allow us to determine d_a'.

(d) See figure.



41. We use the notation shown in the figure.



(a) Here our triangle distribution is

$$T(\mu,\delta) = \begin{cases} \frac{\delta - |x-\mu|}{\delta^2} & |x-\mu| \leq \delta \\ 0 & \text{otherwise.} \end{cases}$$

We assume without loss of generality that $\mu_2 \geq \mu_1$ and then define $d'_T = (\mu_2 - \mu_1)/\delta$. We then limit our considerations to overlapping distributions, that is, $0 \leq d_T \leq 2$. There are two cases to consider:

- $1 \leq d'_T \leq 2$, where the decision boundary x^* is confined to the region of overlap, as shown in the figure.
- $0 \leq d'_T \leq 1$ where the decision boundary x^* can be in one of three regions, as shown in the figure.

We denote the hit rate $H = \int P(x > x^*|\omega_2)dx$, false alarm rate $F = \int P(x > x^*|\omega_1)dx$, and miss rate $M = 1 - F$. The hit rate is

$$\begin{aligned} H &= 1 - \int_{\mu_2-\delta}^{x^*} T(\mu_2,\delta)dx = 1 - \frac{(x^* - \mu_2 + \delta)^2}{2\delta^2} \\ F &= \int_{x^*}^{\mu_1+\delta} T(\mu_1,\delta)dx = \frac{(\mu_1 + \delta - x^*)^2}{2\delta^2} \\ d'_T &= 1 - \sqrt{2F} - \sqrt{2(1-H)} \\ H &= 1 - \frac{(2 - d'_T - \sqrt{2F})^2}{2} \end{aligned}$$

which is valid for $0 \leq F \leq (2 - d'_T)^2/2$.

For the second case $\mu_2 - \delta \leq x^* < \mu_1$, that is, $0 \leq d^*_T \leq 1$, we have

$$\begin{aligned} H &= 1 - \frac{(x^* - \mu_1 + \delta)^2}{2\delta^2} \\ F &= 1 - \frac{(x^* - \mu_1 + \delta)^2}{2\delta^2} \\ d'_T &= \sqrt{2(1-F)} - \sqrt{2(1-H)} \\ H &= 1 - \frac{(\sqrt{2(1-F)} - d'_T)^2}{2} \end{aligned}$$

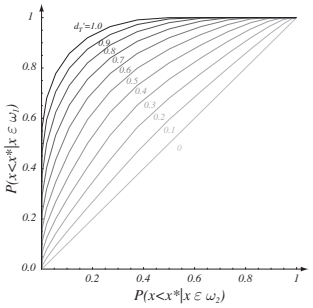
which is valid for $1/2 \leq F \leq 1 - d_T'^2/2$.

For the third case, $\mu_2 \leq x^* \leq \mu_1 + \delta$ and $0 \leq d_T' \leq 1$ we have

$$\begin{aligned} F &= \frac{(\mu_1 + \delta - x^*)^2}{2\delta^2} \\ H &= \frac{(\mu_2 + \delta - x^*)^2}{2\delta^2} \\ d_T' &= \sqrt{2H} - \sqrt{2F}, \end{aligned}$$

valid for $0 \leq F \leq (1 - d_T')^2/2$.

(b) If $d_T' = 2.0$, then the two densities do not overlap. If x^* is closer to the mean of ω_1 than to the mean of ω_2, then the modified ROC curve goes through $P(x > x^*|x \in \omega_1) = 1, P(x > x^*|x \in \omega_2) = 0$. Alternatively, if x^* is closer to the mean of ω_2 than to the mean of ω_1, then we get the converse. In short, there is no meaningful value of d_T' greater than 2.0.



(c) For $H = 0.4$ and $F = 0.2$, we can rule out case I because $H < 0.5$. We can rule out Case II-i because $F < 0.5$. Can rule out Case II-ii because $H < 0.5$. Thus $d_T' = \sqrt{2H} - \sqrt{2F} = 0.262$. We assume $P(\omega_1) = P(\omega_2) = 0.5$, then by symmetry, $x^* = \mu_1 + 0.131\delta = \mu_2 - 0.131\delta$. The Bayes optimal is then $P_E^* = 1/2(0.378 + 0.378) = 0.378$.

(d) Here $x^* = \mu_1 + (d_T'/2)\delta = \mu_2 - (d_T'/2)\delta$.

(e) Here $H = 0.9$ and $F = 0.3$. Case I: $d_T' = 2 - \sqrt{2F} - \sqrt{2(1-H)} = 0.778$. This is not valide because $d_T' < 1$. Cannot be case II-i since $F < 0.5$. Cannot be case II-iii since $H > 0.5$. Thus it is case II-ii and $d_T' = 0.778$. Thus $x^* = \mu_1 + 0.389\delta = \mu_2 - 0.389\delta$. Also, $F = M = 0.187$, and the Bayes error is $P_E^* = 0.187$.

42. We consider bounds on $\min[p, 1-p]$, where $p = p(x|\omega_1)$.

(a) Here the candidate lower bound is

$$b_L(p) = \frac{1}{\beta} \ln \left[\frac{1 + e^{-\beta}}{e^{-\beta p} + e^{-\beta(1-p)}} \right]$$

with $\beta > 0$. First, note that $b_L(p)$ is symmetry with respect to the interchange $p \leftrightarrow (1-p)$, and thus it is symmetric with respect to the value $p = 0.5$. We need consider only the range $0 \leq p \leq 1/2$ as the limit properties we prove there will also hold for $1/2 \leq p \leq 1$. In the range $0 \leq p \leq 1/2$, we have $\min[p, 1-p] = p$.

The simplest way to prove that $b_L(p)$ is a lower bound for p in the range $0 \leq p \leq 1/2$ is to note that $b_L(0) = 0$, and the derivative is always less than the derivative of $\min[p, 1-p]$. Indeed, at $p = 0$ we have

$$b_L(0) = \frac{1}{\beta}\ln\left[\frac{1+e^{-\beta}}{1+e^{-\beta}}\right] = \frac{1}{\beta}\ln[1] = 0.$$

Moreover the derivative is

$$\begin{aligned}\frac{\partial}{\partial p}b_L(p) &= \frac{e^{\beta}-e^{2\beta p}}{e^{\beta}+e^{2\beta p}} \\ &< 1 = \frac{\partial}{\partial p}\min[p, 1-p]\end{aligned}$$

in the range $0 \leq p \leq 1/2$ for $\beta < \infty$.

(b) To show that $b_L(p)$ is an arbitrarily tight bound, we need show only that in the limit $\beta \to \infty$, the derivative, $\partial b_L(p)/\partial p$ approaches 1, the same as

$$\frac{\partial}{\partial p}\min[p, 1-p]$$

in this range. Using the results from part (a) we find

$$\lim_{\beta\to\infty}\frac{\partial}{\partial p}b_L(p) = \lim_{\beta\to\infty}\frac{e^{\beta}-e^{2\beta p}}{e^{\beta}+e^{2\beta p}} = 1$$

in the range $0 \leq p < 1/2$.

(c) Our candidate upper bound is specified by

$$b_U(p) = b_L(p) + [1 - 2b_L(0.5)]b_G(p),$$

where $g_U(p)$ obeys several simple conditions, restated in part (d) below. We let $b_L(p) = p - \theta(p)$, where from part (a) we know that $\theta(p)$ is non-negative and in fact is at least linear in p. By the conditions given, we can write $b_G(p) = p+\phi(p)$, where $\phi(p)$ is non-negative and $\phi(0) = \phi(1/2) = 0$. Then our candidate upper limit obeys

$$\begin{aligned}b_U(p) &= p - \theta(p) + [1 - 2(1/2 - \theta(1/2))](p + \phi(p)) \\ &= p - \theta(p) + \theta(1/2)(p + \phi(p)).\end{aligned}$$

We show that this is an upper bound by calculating the difference between this bound and the Bayes limit (which is $\min[p, 1-p] = p$ in the range $0 \leq p \leq 1/2$). Thus we have

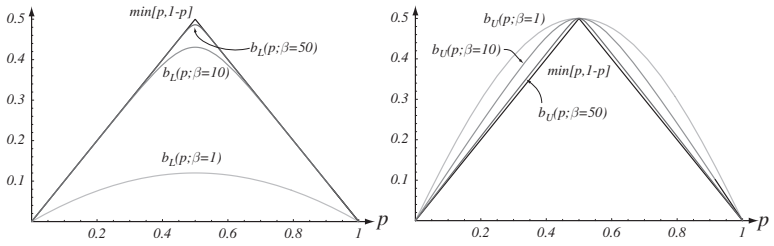
$$\begin{aligned}b_U(p) - p &= -\theta(p) + p\theta(1/2) + \theta(1/2)\phi(p) \\ &> 0.\end{aligned}$$

(d) We seek to confirm that $b_G(p) = 1/2\sin[\pi p]$ has the following four properties:

- $b_G(p) \geq \min[p, 1-p]$: Indeed, $1/2\sin[\pi p] \geq p$ for $0 \leq p \leq 1/2$, with equality holding at the extremes of the interval (that is, at $p = 0$ and $p = 1/2$). By symmetry (see below), the relation holds for the interval $1/2 \leq p \leq 1$.

- $b_G(p) = b_G(1-p)$: Indeed, the sine function is symmetric about the point $\pi/2$, that is, $1/2\ \sin[\pi/2 + \theta] = 1/2\ \sin[\pi/2 - \theta]$. Hence by a simple substitution we see that $1/2\ \sin[\pi p] = 1/2\ \sin[\pi(1-p)]$.
- $b_G(0) = b_G(1) = 0$: Indeed, $1/2\ \sin[\pi \cdot 0] = 1/2\ \sin[\pi \cdot 1] = 0$ — a special case of the fact $b_G(p) = b_G(1-p)$, as shown immediately above.
- $b_G(0.5) = 0.5$: Indeed, $1/2\ \sin[\pi 0.5] = 1/2 \cdot 1 = 0.5$.

(e) See figure.



Section 2.9

43. Here the components of the vector $\mathbf{x} = (x_1, \ldots, x_d)^t$ are binary-valued (0 or 1), and

$$p_{ij} = \Pr[x_i = 1|\omega_j] \qquad \begin{matrix} i = 1, \ldots, d \\ j = 1, \ldots, c. \end{matrix}$$

(a) Thus p_{ij} is simply the probability we get a 1 in feature x_i given that the category is ω_j. This is the kind of probability structure we find when each category has a set of independent binary features (or even real-valued features, thresholded in the form "$y_i > y_{i0}$?").

(b) The discriminant functions are then

$$g_j(\mathbf{x}) = \ln\ p(\mathbf{x}|\omega_j) + \ln\ P(\omega_j).$$

The components of $\mathbf{x}$ are statistically independent for all $\mathbf{x}$ in ω_j, then we can write the density as a product:

$$\begin{aligned} p(\mathbf{x}|\omega_j) &= p((x_1, \ldots, x_d)^t|\omega_j) \\ &= \prod_{i=1}^{d} p(x_i|\omega_j) = \prod_{i=1}^{d} p_{ij}^{x_i}(1-p_{ij})^{1-x_i}. \end{aligned}$$

Thus, we have the discriminant function

$$\begin{aligned} g_j(\mathbf{x}) &= \sum_{i=1}^{d} [x_i\ \ln\ p_{ij} + (1-x_i)\ \ln\ (1-p_{ij})] + \ln\ P(\omega_j) \\ &= \sum_{i=1}^{d} x_i\ \ln\ \frac{p_{ij}}{1-p_{ij}} + \sum_{i=1}^{d} \ln\ (1-p_{ij}) + \ln\ P(\omega_j). \end{aligned}$$

44. The minimum probability of error is achieved by the following decision rule:

$$\text{Choose } \omega_k \text{ if } g_k(\mathbf{x}) \geq g_j(\mathbf{x}) \text{ for all } j \neq k,$$

where here we will use the discriminant function

$$g_j(\mathbf{x}) = \ln\, p(\mathbf{x}|\omega_j) + \ln\, P(\omega_j).$$

The components of $\mathbf{x}$ are statistically independent for all $\mathbf{x}$ in ω_j, and therefore,

$$p(\mathbf{x}|\omega_j) = p((x_1, \ldots, x_d)^t|\omega_j) = \prod_{i=1}^{d} p(x_i|\omega_j),$$

where

$$\begin{aligned} p_{ij} &= \Pr[x_i = \ \ 1\ |\omega_j], \\ q_{ij} &= \Pr[x_i = \ \ 0\ |\omega_j], \\ r_{ij} &= \Pr[x_i = -1\ |\omega_j]. \end{aligned}$$

As in Sect. 2.9.1 in the text, we use exponents to "select" the proper probability, that is, exponents that have value 1.0 when x_i has the value corresponding to the particular probability and value 0.0 for the other values of x_i. For instance, for the p_{ij} term, we seek an exponent that has value 1.0 when $x_i = +1$ but is 0.0 when $x_i = 0$ and when $x_i = -1$. The simplest such exponent is $\frac{1}{2}x_i + \frac{1}{2}x_i^2$. For the q_{ij} term, the simplest exponent is $1 - x_i^2$, and so on. Thus we write the class-conditional probability for a single component x_i as:

$$p(x_i|\omega_j) = p_{ij}^{\frac{1}{2}x_i + \frac{1}{2}x_i^2} q_{ij}^{1-x_i^2} r_{ij}^{-\frac{1}{2}x_i + \frac{1}{2}x_i^2} \quad \begin{matrix} i = 1, \ldots, d \\ j = 1, \ldots, c \end{matrix}$$

and thus for the full vector $\mathbf{x}$ the conditional probability is

$$p(\mathbf{x}|\omega_j) = \prod_{i=1}^{d} p_{ij}^{\frac{1}{2}x_i + \frac{1}{2}x_i^2} q_{ij}^{1-x_i^2} r_{ij}^{-\frac{1}{2}x_i + \frac{1}{2}x_i^2}.$$

Thus the discriminant functions can be written as

$$\begin{aligned} g_j(\mathbf{x}) &= \ln\, p(\mathbf{x}|\omega_j) + \ln\, P(\omega_j) \\ &= \sum_{i=1}^{d} \left[\left(\frac{1}{2}x_i + \frac{1}{2}x_i^2 \right) \ln\, p_{ij} + (1 - x_i^2) \ln\, q_{ij} + \left(-\frac{1}{2}x_i + \frac{1}{2}x_i^2 \ln\, r_{ij} \right) \right] + \ln\, P(\omega_j) \\ &= \sum_{i=1}^{d} x_i^2 \ln\, \frac{\sqrt{p_{ij} r_{ij}}}{q_{ij}} + \frac{1}{2} \sum_{i=1}^{d} x_i \ln\, \frac{p_{ij}}{r_{ij}} + \sum_{i+1}^{d} \ln\, q_{ij} + \ln\, P(\omega_j), \end{aligned}$$

which are quadratic functions of the components x_i.

45. We are given that $P(\omega_1) = P(\omega_2) = 1/2$ and

$$\begin{aligned} p_{i1} &= p > 1/2 \\ p_{i2} &= 1 - p \qquad i = 1, \ldots, d, \end{aligned}$$

where d is the dimension, or number of features.

(a) The minimum error decision rule is

$$\begin{aligned}\text{Choose } \omega_1 \text{ if } \quad & \sum_{i=1}^{d} x_i \ln \tfrac{p}{1-p} + \sum_{i=1}^{d} \ln\,(1-p) + \ln\,\tfrac{1}{2} \\ & > \sum_{i=1}^{d} x_i \ln \tfrac{1-p}{p} + \sum_{i=1}^{d} \ln\, p + \ln\,\tfrac{1}{2}.\end{aligned}$$

This rule can be expressed as

$$\begin{aligned}\left(\sum_{i=1}^{d} x_i\right)\left[\ln\,\frac{p}{1-p} - \ln\,\frac{1-p}{p}\right] &> d\,\ln\,p - d\,\ln\,(1-p) \\ \left(\sum_{i=1}^{d} x_i\right)\left(\ln\,\frac{p}{1-p}\right)\times 2 &> d\,\ln\,\frac{p}{1-p}\end{aligned}$$

or simply

$$\sum_{i=1}^{d} x_i > d/2.$$

(b) We denote the minimum probability of error as $P_e(d,p)$. Then we have:

$$\begin{aligned}P_e(d,p) &= P(error|\omega_1)P(\omega_1) + P(error|\omega_2)P(\omega_2) \\ &= P\left(\sum_{i=1}^{d} x_i \le d/2\middle|\omega_1\right)\times 1/2 + P\left(\sum_{i=1}^{d} x_i > d/2\middle|\omega_1\right)\times 1/2.\end{aligned}$$

As d is odd, and $\sum_{i=1}^{d}$ is an integer, we have

$$\begin{aligned}P_e(d,p) &= P\left(\sum_{i=1}^{d} x_i \le \frac{d-1}{2}\middle|\omega_1\right)\times 1/2 + P\left(\sum_{i=1}^{d} x_i \ge \frac{d+1}{2}\middle|\omega_2\right)\times 1/2 \\ &= 1/2\sum_{k=0}^{(d-1)/2}\binom{d}{k}p^k(1-p)^{d-k} + 1/2\sum_{k=(d+1)/2}^{d}\binom{d}{k}(1-p)^k p^k.\end{aligned}$$

We substitute $k' = d-k$ in the second summation, use the fact that $\binom{d}{k'} = \binom{d}{d-k'}$, and find

$$\begin{aligned}P_e(d,p) &= 1/2\sum_{k=0}^{(d-1)/2}\binom{d}{k}p^k(1-p)^{d-k} + 1/2\sum_{k'=0}^{(d-1)/2}\binom{d}{k'}p^{k'}(1-p)^{d-k'} \\ &= \sum_{k=0}^{(d-1)/2}\binom{d}{k}p^k(1-p)^{d-k}.\end{aligned}$$

(c) We seek the limit for $p \to 1/2$ of $P_e(d,p)$. Formally, we write

$$\lim_{p\to 1/2} P_e(d,p) = \sum_{k=0}^{(d-1)/2}\binom{d}{k}\lim_{p\to 1/2} p^k(1-p)^{d-k}$$

$$= \sum_{k=0}^{(d-1)/2} \binom{d}{k} \left(\frac{1}{2}\right)^d = \left(\frac{1}{2}\right)^d \sum_{k=0}^{(d-1)/2} \binom{d}{k}$$

$$= \left(\frac{1}{2}\right)^d \frac{1}{2} \left[\sum_{k=0}^{d} \binom{d}{k}\right] = \left(\frac{1}{2}\right)^d \times \frac{2^d}{2} = \frac{1}{2}.$$

Indeed, in the case $p \to 1/2$, the probability of error will be 1/2.

(d) Note that as $d \to \infty$, the binomial probability $\binom{d}{k} p^k (1-p)^{d-k}$ can be approximated by the normal density with mean dp and variance $dp\ q$. Thus we have

$$\begin{aligned}
\lim_{d\to\infty} P_e(d,p) &= \lim_{d\to\infty} \sum_{k=0}^{(d-1)/2} \binom{d}{k} p^k (1-p)^{d-k} \\
&= \lim_{d\to\infty} P(0 \le X \le (d-1)/2) \text{ where } X \sim N(dp, dp\ q) \\
&= \lim_{d\to\infty} P\left(\frac{-dp}{\sqrt{dp\ q}} \le Z \le \frac{(d-1)/2 - dp}{\sqrt{dp\ q}}\right) \text{ where } Z \sim N(0,1) \\
&= \lim_{d\to\infty} P\left(-\sqrt{dp\ q} \le Z \le \frac{\sqrt{z}(1/2-p) - d/2}{\sqrt{dp\ q}}\right).
\end{aligned}$$

As $P > 1/2$, $\lim_{d\to\infty} -\sqrt{dp\ q} = -\infty$ and $\lim_{d\to\infty} \sqrt{d}\ (1/2 - p) = -\infty$. Thus in the limit of very large dimension, we have

$$\lim_{d\to\infty} P_e(d,p) = \Pr(-\infty \le Z \le -\infty) = 0.$$

46. The general minimum-risk discriminant rule is given by Eq. 16 in the text:

Choose ω_1 if $(\lambda_{11} - \lambda_{21}) p(\mathbf{x}|\omega_1) P(\omega_1) < (\lambda_{22} - \lambda_{12}) p(\mathbf{x}|\omega_2) P(\omega_2)$; otherwise choose ω_2.

Under the assumption $\lambda_{21} > \lambda_{11}$ and $\lambda_{12} > \lambda_{22}$, we thus have

$$\text{Choose } \omega_1 \text{ if } \frac{(\lambda_{21} - \lambda_{11}) p(\mathbf{x}|\omega_1) P(\omega_1)}{(\lambda_{12} - \lambda_{22}) p(\mathbf{x}|\omega_2) P(\omega_2)} > 1,$$

or

$$\text{Choose } \omega_1 \text{ if } \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} + \ln \frac{\lambda_{21} - \lambda_{11}}{\lambda_{12} - \lambda_{22}} > 0.$$

Thus the discriminant function for minimum risk, derived by taking logarithms in Eq. 17 in the text, is:

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} + \ln \frac{\lambda_{21} - \lambda_{11}}{\lambda_{12} - \lambda_{22}}.$$

For the case of independent features given in this problem, we have

$$\ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} = \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i},$$

where

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)}, \qquad i = 1, \ldots, d.$$

Therefore, the discriminant function can be written as:

$$\begin{aligned} g(\mathbf{x}) &= \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)} + \ln \frac{\lambda_{21}-\lambda_{11}}{\lambda_{12}-\lambda_{22}} \\ &= \mathbf{w}^t\mathbf{x} + w_0. \end{aligned}$$

47. Recall the Poisson distribution for a discrete variable $x = 0, 1, 2, \ldots$, is given by

$$P(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

(a) The mean, or expectation of this distribution is defined as

$$\mathcal{E}[x] = \sum_{x=0}^{\infty} xP(x|\lambda) = \sum_{x=0}^{\infty} xe^{-\lambda}\frac{\lambda^x}{x!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

(b) The variance of this distribution is

$$\mathrm{Var}[x] = \mathcal{E}[(x-\mathcal{E}[x])^2] = \mathcal{E}[x^2] - \underbrace{(\mathcal{E}[x])^2}_{\lambda^2}.$$

To evaluate this variance, we need $\mathcal{E}[x^2]$:

$$\begin{aligned} \mathcal{E}[x^2] &= \mathcal{E}[x(x-1)+x] = \mathcal{E}[x(x-1)] + \mathcal{E}[x] \\ &= \sum_{x=0}^{\infty}(x(x-1)e^{-\lambda}\frac{\lambda^x}{x!} + \lambda \\ &= \lambda^2 e^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} + \lambda \\ &= \lambda^2 e^{-\lambda}\sum_{x'=0}^{\infty}\frac{\lambda^{x'}}{x'!} + \lambda \\ &= \lambda^2 e^{-\lambda}e^{\lambda} + \lambda \\ &= \lambda^2 + \lambda, \end{aligned}$$

where we made the substitution $x' \leftarrow (x-2)$ in the fourth step. We put the above results together and find

$$\mathrm{Var}[x] = \mathcal{E}[(x-\mathcal{E}[x])^2] = \mathcal{E}[x^2] - (\mathcal{E}[x])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

(c) First recall the notation $\lfloor\lambda\rfloor$, read "floor of λ," indicates the greatest integer less than λ. For all $x < \lfloor\lambda\rfloor \leq \lambda$, then, we have

$$\frac{\lambda^x}{x!} = \frac{\lambda^{x-1}}{(x-1)!}\underbrace{\frac{\lambda}{x}}_{>1} > \frac{\lambda^{x-1}}{(x-1)!}.$$

That is, for $x < \lfloor\lambda\rfloor \leq \lambda$, the probability increases as x increases. Conversely, for $x > \lambda \geq \lfloor\lambda\rfloor$ we have

$$\frac{\lambda^x}{x!} = \frac{\lambda^{x-1}}{(x-1)!}\underbrace{\frac{\lambda}{x}}_{<1} > \frac{\lambda^{x-1}}{(x-1)!}.$$

That is, for $x > \lambda \geq \lfloor \lambda \rfloor$, the probability decreases as x increases. Thus, the probability is maximized for integers $x \in [\lfloor \lambda \rfloor, \lambda]$. In short, if λ is not an integer, then $x = \lfloor \lambda \rfloor$ is the mode; if λ is an integer, then both $\lfloor \lambda \rfloor$ and λ are modes.

(d) We denote the two distributions

$$P(x|\lambda_i) = e^{-\lambda_i} \frac{\lambda_i^x}{x!}$$

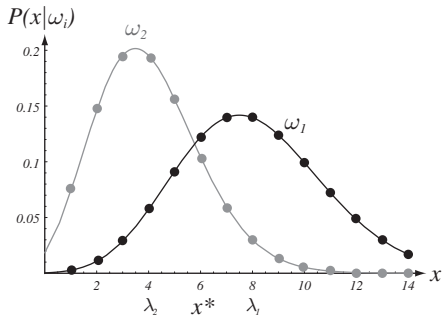
for $i = 1, 2$, and by assumption $\lambda_1 > \lambda_2$. The likelihood ratio is

$$\frac{P(x|\lambda_1)}{P(x|\lambda_2)} = \frac{e^{-\lambda_1}\lambda_1^x}{e^{-\lambda_2}\lambda_2^x} = e^{\lambda_2 - \lambda_1} \left(\frac{\lambda_1}{\lambda_2} \right)^x .$$

Thus the Bayes decision rule is

$$\begin{aligned} \text{Choose } \omega_2 \quad &\text{if} \quad e^{\lambda_2 - \lambda_1} \left(\frac{\lambda_1}{\lambda_2} \right)^x > 1, \\ \text{or equivalently} \quad &\text{if} \quad x < \frac{(\lambda_2 - \lambda_1)}{\ln[\lambda_1] - \ln[\lambda_2]}; \\ \text{otherwise choose } \omega_1, & \end{aligned}$$

as illustrated in the figure (where the actual x values are discrete).



(e) The conditional Bayes error rate is

$$P(error|x) = \min \left[e^{-\lambda_1} \frac{\lambda_1^x}{x!}, e^{-\lambda_2} \frac{\lambda_2^x}{x!} \right].$$

The Bayes error, given the decision rule in part (d) is

$$P_B(error) = \sum_{x=0}^{x^*} e^{\lambda_2} \frac{\lambda_2^x}{x!} + \sum_{x=x^*}^{\infty} e^{-\lambda_1} \frac{\lambda_1^x}{x!},$$

where $x^* = \lfloor (\lambda_2 - \lambda_1)/(\ln[\lambda_1] - \ln[\lambda_2]) \rfloor$.

Section 2.10

48. In two dimensions, the Gaussian distribution is

$$p(\mathbf{x}|\omega_i) = \frac{1}{2\pi |\mathbf{\Sigma}_i|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^t \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right].$$

(a) By direct calculation using the densities stated in the problem, we find that for $\mathbf{x} = \binom{.3}{.3}$ that $p(\mathbf{x}|\omega_1)P(\omega_1) = 0.04849$, $p(\mathbf{x}|\omega_2)P(\omega_2) = 0.03250$ and $p(\mathbf{x}|\omega_3)P(\omega_3) = 0.04437$, and thus the pattern should be classified as category ω_1.

(b) To classify $\binom{*}{.3}$, that is, a vector whose first component is missing and its second component is 0.3, we need to marginalize over the unknown feature. Thus we compute numerically

$$P(\omega_i)p\left(\binom{*}{.3}\Big|\omega_i\right) = P(\omega_i)\int_{-\infty}^{\infty} p\left(\binom{x}{.3}\Big|\omega_i\right)\,dx$$

and find that $P(\omega_1)p((*,.3)^t|\omega_1) = 0.12713$, $P(\omega_1)p((*,.3)^t|\omega_2) = 0.10409$, and $P(\omega_1)p((*,.3)^t|\omega_3) = 0.13035$. Thus the pattern should be categorized as ω_3.

(c) As in part (a), we calculate numerically

$$P(\omega_i)\tilde{p}\left(\binom{.3}{*}\Big|\omega_i\right) = P(\omega_i)\int_{-\infty}^{\infty} p\left(\binom{.3}{y}\Big|\omega_i\right)\,dy$$

and find that $P(\omega_1)p((.3,*)^t|\omega_1) = 0.12713$, $P(\omega_1)p((.3,*)^t|\omega_2) = 0.10409$, and $P(\omega_1)p((.3,*)^t|\omega_3) = 0.11346$. Thus the pattern should be categorized as ω_1.

(d) We follow the procedure in part (c) above:

$\mathbf{x} = (.2,.6)^t$

- $P(\omega_1)p(\mathbf{x}|\omega_1) = 0.04344$.
- $P(\omega_2)p(\mathbf{x}|\omega_2) = 0.03556$.
- $P(\omega_3)p(\mathbf{x}|\omega_3) = 0.04589$.

Thus $\mathbf{x} = (.2,.6)^t$ should be categorized as ω_3.

$\mathbf{x} = (*,.6)^t$

- $P(\omega_1)p(\mathbf{x}|\omega_1) = 0.11108$.
- $P(\omega_2)p(\mathbf{x}|\omega_2) = 0.12276$.
- $P(\omega_3)p(\mathbf{x}|\omega_3) = 0.13232$.

Thus $\mathbf{x} = (*,.6)^t$ should be categorized as ω_3.

$\mathbf{x} = (.2,*)^t$

- $P(\omega_1)p(\mathbf{x}|\omega_1) = 0.11108$.
- $P(\omega_2)p(\mathbf{x}|\omega_2) = 0.12276$.
- $P(\omega_3)p(\mathbf{x}|\omega_3) = 0.10247$.

Thus $\mathbf{x} = (*,.6)^t$ should be categorized as ω_2.

49. Equation 95 in the text states

$$P(\omega_i|\mathbf{x}_g,\mathbf{x}_b) = \frac{\int g_i(\mathbf{x})p(\mathbf{x})p(\mathbf{x}_b|\mathbf{x}_t)d\mathbf{x}_t}{\int p(\mathbf{x})p(\mathbf{x}_b|\mathbf{x}_t)d\mathbf{x}_t}.$$

We are given that the "true" features, $\mathbf{x}_t$, are corrupted by Gaussian noise to give us the measured "bad" data, that is,

$$p(\mathbf{x}_b|\mathbf{x}_t) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|}\exp\left[-\frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu}_i)^t\mathbf{\Sigma}^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_i)\right].$$

We drop the needless subscript i indicating category membership and substitute this Gaussian form into the above. The constant terms, $(2\pi)^{d/2}|\mathbf{\Sigma}|$, cancel, and we then find that the probability of the category given the good and the measured bad data is

$$P(\omega|\mathbf{x}_g,\mathbf{x}_b)=\frac{\int g(\mathbf{x})p(\mathbf{x})\exp[-\frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_t-\boldsymbol{\mu})]d\mathbf{x}_t}{\int p(\mathbf{x})\exp[-\frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_t-\boldsymbol{\mu})]d\mathbf{x}_t]d\mathbf{x}_t}.$$

After integration, this gives us the final answer,

$$P(\omega|\mathbf{x}_g,\mathbf{x}_b)=\frac{P(\omega)p(\mathbf{x}_g,\mathbf{x}_b|\omega)}{p(\mathbf{x}_g,\mathbf{x}_b)},$$

which is the result from standard Bayesian considerations.

Section 2.11

50. We use the values from Example 4 in the text.

(a) For this case, the probabilities are:

$$\begin{aligned}P(a_1) &= P(a_4)=0.5\\ P(a_2) &= P(a_3)=0\\ P(b_1) &= 1\\ P(b_2) &= 0\\ P(d_1) &= 0\\ P(d_2) &= 1.\end{aligned}$$

Then using Eq. 99 in the text we have

$$\begin{aligned}P_{\mathcal{P}}(x_1) &\sim P(x_1|a_1,b_1)P(a_1)P(b_1)+0+0+0+0+0+P(x_1|a_4,b_1)P(a_4)P(b_1)+0\\ &= \frac{0.9\cdot 0.65}{0.9\cdot 0.65+0.1\cdot 0.35}\cdot 0.5\cdot 1+\frac{0.8\cdot 0.65}{0.8\cdot 0.65+0.2\cdot 0.35}0.5\cdot 1\\ &= 0.472+0.441\\ &= 0.913.\end{aligned}$$

A similar calculation gives

$$\begin{aligned}P_{\mathcal{P}}(x_2) &\sim P(x_2|a_1,b_1)P(a_1)P(b_1)+0+0+0+0+0+P(x_2|a_4,b_1)P(a_4)P(b_1)+0\\ &= \frac{0.35\cdot 0.1}{0.9\cdot 0.65+0.1\cdot 0.35}0.5\cdot 1+\frac{0.2\cdot 0.35}{0.8\cdot 0.65+0.2\cdot 0.35}0.5\cdot 1\\ &= 0.87.\end{aligned}$$

Since the lightness is not measured, we can consider only thickness

$$\begin{aligned}P_{\mathcal{C}}(x_1) &\sim P(e_{\mathcal{D}}|x_2)\\ &= P(e_{\mathcal{D}}|d_1)P(d_1|x_1)+P(e_{\mathcal{D}}|d_2)P(d_2|x_1)\\ &= 0\cdot 0.4+1\cdot 0.6\\ &= 0.6.\end{aligned}$$

We normalize these and find

$$\begin{aligned} P(x_1|e) &= \frac{0.913 \cdot 0.6}{0.913 \cdot 0.6 + 0.087 \cdot 0.05} = 0.992 \\ P(x_2|e) &= \frac{0.087 \cdot 0.05}{0.913 \cdot 0.6 + 0.087 \cdot 0.05} = 0.008. \end{aligned}$$

Thus, given all the evidence e throughout the belief net, the most probable outcome is x_1, that is, salmon. The expected error is the probability of finding a sea bass, that is, 0.008.

(b) Here we are given that the fish is thin and medium lightness, which implies

$$\begin{aligned} P(e_{\mathcal{D}}|d_1) = 0, &\qquad P(e_{\mathcal{D}}|d_2) = 1 \\ P(e_{\mathcal{C}}|c_1) = 0, &\quad P(e_{\mathcal{C}}|c_2) = 1, P(e_{\mathcal{C}}|c_3) = 0. \end{aligned}$$

and as in part (a) we have

$$\begin{aligned} P_{\mathcal{C}}(x_1) &\sim P(e_{\mathcal{C}}|x_1)P(e_{\mathcal{D}}|x_1) \\ &= [P(e_{\mathcal{C}}|c_1)P(c_1|x_1) + P(e_{\mathcal{C}}|c_2)P(c_2|x_1) + P(e_{\mathcal{C}}|c_3)P(c_3|x_1)] \\ &\quad \times [P(e_{\mathcal{D}}|d_1)P(d_1|x_1) + P(e_{\mathcal{D}}|d_2)P(d_2|x_1)] \\ &= [0 + 1 \cdot 0.33 + 0][0 + 1 \cdot 0.6] \\ &= 0.198. \end{aligned}$$

Likewise we have

$$\begin{aligned} P_{\mathcal{C}}(x_2) &\sim P(e_{\mathcal{C}}|x_2)P(e_{\mathcal{D}}|x_2) \\ &= [P(e_{\mathcal{C}}|c_1)P(c_2|x_2) + P(e_{\mathcal{C}}|c_2)P(c_2|x_2) + P(e_{\mathcal{C}}|c_3)P(c_3|x_2)] \\ &\quad \times [P(e_{\mathcal{D}}|d_1)P(d_1|x_2) + P(e_{\mathcal{D}}|d_2)P(d_2|x_2)] \\ &= [0 + 1 \cdot 0.1 + 0][0 + 1 \cdot 0.05] \\ &= 0.005. \end{aligned}$$

We normalize and find

$$\begin{aligned} P(x_1|e_{\mathcal{C},\mathcal{D}}) &= \frac{0.198}{0.198 + 0.005} = 0.975 \\ P(x_2|e_{\mathcal{C},\mathcal{D}}) &= \frac{0.005}{0.198 + 0.005} = 0.025. \end{aligned}$$

Thus, from the evidence of the children nodes, we classify the fish as x_1, that is, salmon.

Now we infer the probability $P(a_i|x_1)$. We have $P(a_1) = P(a_2) = P(a_3) = P(a_4) = 1/4$. Then we normalize and find

$$\begin{aligned} P(a_1|x_1) &= \frac{P(x_1|a_1)P(a_1)}{P(x_1|a_1) \cdot P(a_1) + P(x_1|a_2) \cdot P(a_2) + P(x_1|a_3)P(a_3) + P(x_1|a_4)P(a_4)} \\ &= \frac{0.9 \cdot 0.25}{0.9 \cdot 0.25 + 0.3 \cdot 0.25 + 0.4 \cdot 0.25 + 0.8 \cdot 0.25} \\ &= 0.375. \end{aligned}$$

We also have

$$\begin{aligned} P(a_2|x_1) &= 0.125 \\ P(a_3|x_1) &= 0.167 \\ P(a_4|x_1) &= 0.333. \end{aligned}$$

Thus the most probable season is a_1, winter. The probability of being correct is

$$P(x_1|e_{\mathcal{C},\mathcal{D}})P(a_1|x_1) = 0.975 \cdot 0.375 = 0.367.$$

(c) Fish is thin and medium lightness, so from part (b) we have

$$P(x_1|e_{\mathcal{C},\mathcal{D}}) = 0.975, P(x_2|e_{\mathcal{C},\mathcal{D}}) = 0.025,$$

and we classify the fish as salmon.

For the fish caught in the north Atlantic, we have $P(b_1|e_{\mathcal{B}}) = 1$ and $P(b_2|e_{\mathcal{B}}) = 0$.

$$\begin{aligned} P(a_1|x_1,b_1) &\sim P(x_1|a_1)P(x_1|b_1)P(a_1)P(b_1) \\ &= 0.9 \cdot 0.65 \cdot 0.25 \cdot = 0.146 \\ P(a_2|x_1,b_1) &\sim P(x_1|a_2)P(x_1|b_1)P(a_2)P(b_1) \\ &= 0.3 \cdot 0.65 \cdot 0.25 \cdot = 0.049 \\ P(a_3|x_1,b_1) &\sim P(x_1|a_3)P(x_1|b_1)P(a_3)P(b_1) \\ &= 0.4 \cdot 0.65 \cdot 0.25 \cdot = 0.065 \\ P(a_4|x_1,b_1) &\sim P(x_1|a_4)P(x_1|b_1)P(a_4)P(b_1) \\ &= 0.8 \cdot 0.65 \cdot 0.25 \cdot = 0.13. \end{aligned}$$

So the most probable season is a_1, winter. The probability of being correct is

$$P(x_1|e_{\mathcal{C},\mathcal{D}})P(a_1|x_1) = 0.975 \cdot 0.375 = 0.367.$$

51. Problem not yet solved

Section 2.12

52. We have the priors $P(\omega_1) = 1/2$ and $P(\omega_2) = P(\omega_3) = 1/4$, and the Gaussian densities $p(x|\omega_1) \sim N(0,1)$, $p(x|\omega_2) \sim N(0.5,1)$, and $p(x|\omega_3) \sim N(1,1)$. We use the general form of a Gaussian, and by straightforward calculation find:

x	$p(x\|\omega_1)$	$p(x\|\omega_2)$	$p(x\|\omega_3)$
0.6	0.333225	0.396953	0.368270
0.1	0.396953	0.368270	0.266085
0.9	0.266085	0.368270	0.396953
1.1	0.217852	0.333225	0.396953

We denote $\mathbf{X} = (x_1, x_2, x_3, x_4)$ and $\boldsymbol{\omega} = (\omega(1), \omega(2), \omega(3), \omega(4))$. Using the notation in Section 2.12, we have $n = 4$ and $c = 3$, and thus there are $c^n = 3^4 = 81$ possible

values of $\boldsymbol{\omega}$, such as

$$\begin{array}{ccc}(\omega_1,\omega_1,\omega_1,\omega_1),(\omega_1,\omega_1,\omega_1,\omega_2),(\omega_1,\omega_1,\omega_1,\omega_3),\\ (\omega_1,\omega_1,\omega_2,\omega_1),(\omega_1,\omega_1,\omega_2,\omega_2),(\omega_1,\omega_1,\omega_2,\omega_3),\\ (\omega_1,\omega_1,\omega_3,\omega_1),(\omega_1,\omega_1,\omega_3,\omega_2),(\omega_1,\omega_1,\omega_3,\omega_3),\\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots\\ (\omega_3,\omega_3,\omega_3,\omega_1),(\omega_3,\omega_3,\omega_3,\omega_2),(\omega_3,\omega_3,\omega_3,\omega_3)\end{array}$$

For each possible value of $\boldsymbol{\omega}$, we calculate $P(\boldsymbol{\omega})$ and $p(\mathbf{X}|\boldsymbol{\omega})$ using the following, which assume the independences of x_i and $\omega(i)$:

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\omega}) &= \prod_{i=1}^{4} p(x_i|\omega(i)) \\ P(\boldsymbol{\omega}) &= \prod_{i=1}^{4} P(\omega(i)). \end{aligned}$$

For example, if $\boldsymbol{\omega} = (\omega_1, \omega_3, \omega_3, \omega_2)$ and $\mathbf{X} = (0.6, 0.1, 0.9, 1.1)$, then we have

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\omega}) &= p((0.6,0.1,0.9,1.1)|(\omega_1,\omega_3,\omega_3,\omega_2)) \\ &= p(0.6|\omega_1)p(0.1|\omega_3)p(0.9|\omega_3)p(1.1|\omega_2) \\ &= 0.333225 \times 0.266085 \times 0.396953 \times 0.333225 \\ &= 0.01173 \end{aligned}$$

and

$$\begin{aligned} P(\boldsymbol{\omega}) &= P(\omega_1)P(\omega_3)P(\omega_3)P(\omega_2) \\ &= \frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \\ &= \frac{1}{128} = 0.0078125. \end{aligned}$$

(a) Here we have $\mathbf{X} = (0.6, 0.1, 0.9, 1.1)$ and $\boldsymbol{\omega} = (\omega_1, \omega_3, \omega_3, \omega_2)$. Thus, we have

$$\begin{aligned} p(\mathbf{X}) &= p(x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1) \\ &= \sum_{\boldsymbol{\omega}} p(x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1|\boldsymbol{\omega})P(\boldsymbol{\omega}) \\ &= p((x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1)|(\omega_1,\omega_1,\omega_1,\omega_1))P(\omega_1,\omega_1,\omega_1,\omega_1) \\ &\quad + p((x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1)|(\omega_1,\omega_1,\omega_1,\omega_2))P(\omega_1,\omega_1,\omega_1,\omega_2) \\ &\quad \vdots \\ &\quad + p((x_1 = 0.6, x_2 = 0.1, x_3 = 0.9, x_4 = 1.1)|(\omega_3,\omega_3,\omega_3,\omega_3))P(\omega_3,\omega_3,\omega_3,\omega_3) \\ &= p(0.6|\omega_1)p(0.1|\omega_1)p(0.9|\omega_1)p(1.1|\omega_1)P(\omega_1)P(\omega_1)P(\omega_1)P(\omega_1) \\ &\quad + p(0.6|\omega_1)p(0.1|\omega_1)p(0.9|\omega_1)p(1.1|\omega_2)P(\omega_1)P(\omega_1)P(\omega_1)P(\omega_2) \\ &\quad \vdots \\ &\quad + p(0.6|\omega_3)p(0.1|\omega_3)p(0.9|\omega_3)p(1.1|\omega_3)P(\omega_3)P(\omega_3)P(\omega_3)P(\omega_3) \\ &= 0.012083, \end{aligned}$$

where the sum of the 81 terms can be computed by a simple program, or somewhat tediously by hand. We also have

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\omega}) &= p(0.6, 0.1, 0.9, 1.1|\omega_1, \omega_3, \omega_3, \omega_2) = p(0.6|\omega_1)p(0.1|\omega_3)p(0.9|\omega_3)p(1.1|\omega_2) \\ &= 0.33325 \times 0.266085 \times 0.396953 \times 0.333225 \\ &= 0.01173. \end{aligned}$$

Also we have

$$\begin{aligned} P(\boldsymbol{\omega}) &= P(\omega_1)P(\omega_3)P(\omega_3)P(\omega_2) \\ &= \frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \\ &= \frac{1}{128} = 0.0078125. \end{aligned}$$

According to Eq. 103 in the text, we have

$$\begin{aligned} P(\boldsymbol{\omega}|\mathbf{X}) &= P(\omega_1, \omega_3, \omega_3, \omega_2|0.6, 0.1, 0.9, 1.1) \\ &= \frac{p(0.6, 0.1, 0.8, 1.1|\omega_1, \omega_3, \omega_3, \omega_2)P(\omega_1, \omega_3, \omega_3, \omega_2)}{p(\mathbf{X})} \\ &= \frac{0.01173 \cdot 0.0078125}{0.012083} = 0.007584. \end{aligned}$$

(b) We follow the procedure in part (a), with the values $\mathbf{X} = (0.6, 0.1, 0.9, 1.1)$ and $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_2, \omega_3)$. We have

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\omega}) &= p(0.6, 0.1, 0.9, 1.1|\omega_1, \omega_2, \omega_2, \omega_3) \\ &= p(0.6|\omega_1)p(0.1|\omega_2)p(0.9|\omega_2)p(1.1|\omega_3) \\ &= 0.33225 \times 0.368270 \times 0.368270 \times 0.396953 = 0.01794. \end{aligned}$$

Likewise, we have

$$\begin{aligned} P(\boldsymbol{\omega}) &= P(\omega_1)P(\omega_2)P(\omega_2)P(\omega_3) \\ &= \frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \\ &= \frac{1}{128} = 0.0078125. \end{aligned}$$

Thus

$$\begin{aligned} P(\boldsymbol{\omega}|\mathbf{X}) &= P(\omega_1, \omega_2, \omega_2, \omega_3|0.6, 0.1, 0.9, 1.1) \\ &= \frac{p(0.6, 0.1, 0.9, 1.1|\omega_1, \omega_2, \omega_2, \omega_3)P(\omega_1, \omega_2, \omega_2, \omega_3)}{P(\mathbf{X})} \\ &= \frac{0.01794 \cdot 0.0078125}{0.012083} = 0.01160. \end{aligned}$$

(c) Here we have $\mathbf{X} = (0.6, 0.1, 0.9, 1.1)$ and $\boldsymbol{\omega} = (\omega(1), \omega(2), \omega(3), \omega(4))$. According to Eq. 103 in the text, the sequence $\boldsymbol{\omega}$ that maximizes $p(\mathbf{X}|\boldsymbol{\omega})P(\boldsymbol{\omega})$ has the maximum probability, since $p(\mathbf{X})$ is fixed for a given observed $\mathbf{X}$. With the simplifications above, we have

$$p(\mathbf{X}|\boldsymbol{\omega})P(\boldsymbol{\omega}) = p(x_1|\omega(1))p(x_2|\omega(2))p(x_3|\omega(3))p(x_4|\omega(4))P(\omega(1))P(\omega(2))P(\omega(3))P(\omega(4)).$$

For $\mathbf{X} = (0.6, 0.1, 0.9, 1.1)$, we have

$$\begin{aligned} & \max\left[p(0.6|\omega_1)P(\omega_1), p(0.6|\omega_2)P(\omega_2), p(0.6|\omega_3)P(\omega_3)\right] \\ = & \max\left[0.333225 \times 0.5, 0.396953 \times 0.25, 0.368270 \times 0.25\right] \\ = & 0.333225 \times 0.5 \\ = & 0.1666125. \end{aligned}$$

Likewise, for the second step we have

$$\begin{aligned} & \max\left[p(0.1|\omega_1)P(\omega_1), p(0.1|\omega_2)P(\omega_2), p(0.1|\omega_3)P(\omega_3)\right] \\ = & \max\left[0.396953 \times 0.5, 0.368270 \times 0.25, 0.266085 \times 0.25\right] \\ = & 0.396953 \times 0.5 \\ = & 0.1984765. \end{aligned}$$

For the third step we have

$$\begin{aligned} & \max\left[p(0.9|\omega_1)P(\omega_1), p(0.9|\omega_2)P(\omega_2), p(0.9|\omega_3)P(\omega_3)\right] \\ = & \max\left[0.266085 \times 0.5, 0.368270 \times 0.25, 0.396953 \times 0.25\right] \\ = & 0.266085 \times 0.5 \\ = & 0.133042. \end{aligned}$$

For the final step we have

$$\begin{aligned} & \max\left[p(1.1|\omega_1)P(\omega_1), p(1.1|\omega_2)P(\omega_2), p(1.1|\omega_3)P(\omega_3)\right] \\ = & \max\left[0.217852 \times 0.5, 0.333225 \times 0.25, 0.396953 \times 0.25\right] \\ = & 0.217852 \times 0.5 \\ = & 0.108926. \end{aligned}$$

Thus the sequence $\boldsymbol{\omega} = (\omega_1, \omega_1, \omega_1, \omega_1)$ has the maximum probability to observe $\mathbf{X} = (0.6, 0.1, 0.9, 1.1)$. This maximum probability is

$$0.166625 \times 0.1984765 \times 0.133042 \times 0.108926 \times \frac{1}{0.012083} = 0.03966.$$

Computer Exercises

Section 2.5

1. Computer exercise not yet solved

2.

```
1                                                          MATLAB program
2  load samples.mat;                                       the data
3  [n,m] = size(samples);
4  for i=1:3
5     mu{i} = mean(samples(:, (i-1)*3+1;i*3))';
6     sigma{i} = zeros(3);
7     for j=1:n
8        sigma{i} = sigma{i} + ...                         The ... continues the line
9           (samples(j,(i-1)*3+1:i*3)' - mu{i}) ...
10          * (samples(j,(i-1)*3+1;i*3)' - mu{i})';
11    end
12    sigma{i} = sigma{i}./n;
13 end
14 s = [1 2 1; 5 3 2; 0 0 0; 1 0 0]'
15 for j=1:size(s,2)
16    for i=1;3
17       d = sqrt((s:,j)-mu{i})'*inv(sigma{i})*(s(:,j)-mu{i}));
18       fprintf('Mahal. dist. for class %d and point %d: %: %f\n', i, j, d);
19    end
20 end
21 pw(1:)  =[1/3 0.8];
22 pw(2:)  =[1/3 0.1];
23 pw(3:)  =[1/3 0.1];
24 for p=1:2
25    fprintf('\n\n\n\n');
26    for j=1:size(s,2)
27       class = 0; max_gi = -1000000;
28       for i=1:3
29          d_i = (s(:,j)-mu{i})'*inv(sigma{i})*(s(:,j)-mu{i});
30          g_i = -0.5*d_i - 1.5*log(2*pi) - 0.5*log(det(sigma{i})) + log(pw(i,p));
31          if gi > max_gi,
32             max_gi = gi;
33             class = i;
34          end
35       end
36       fprintf('Point %d classified in category %d\n', j, class);
37    end
38 end
```

Output

```
Mahal. dist. for class 1 and point 1: 1.069873
Mahal. dist. for class 2 and point 1: 0.904465
Mahal. dist. for class 3 and point 1: 2.819441
```

```
Mahal. dist. for class 1 and point 2: 1.641368
Mahal. dist. for class 2 and point 2: 1.850650
Mahal. dist. for class 3 and point 2: 0.682007
Mahal. dist. for class 1 and point 3: 0.516465
Mahal. dist. for class 2 and point 3: 0.282953
Mahal. dist. for class 3 and point 3: 2.362750
Mahal. dist. for class 1 and point 4: 0.513593
Mahal. dist. for class 2 and point 4: 0.476275
Mahal. dist. for class 3 and point 4: 1.541438


Point 1 classified in category 2
Point 2 classified in category 3
Point 3 classified in category 1
Point 4 classified in category 1


Point 1 classified in category 1
Point 2 classified in category 1
Point 3 classified in category 1
Point 4 classified in category 1
```

3. COMPUTER EXERCISE NOT YET SOLVED
4. COMPUTER EXERCISE NOT YET SOLVED
5. COMPUTER EXERCISE NOT YET SOLVED

Section 2.8

6. COMPUTER EXERCISE NOT YET SOLVED
7. COMPUTER EXERCISE NOT YET SOLVED
8. COMPUTER EXERCISE NOT YET SOLVED

Section 2.11

9. COMPUTER EXERCISE NOT YET SOLVED

Chapter 3

Maximum likelihood and Bayesian parameter estimation

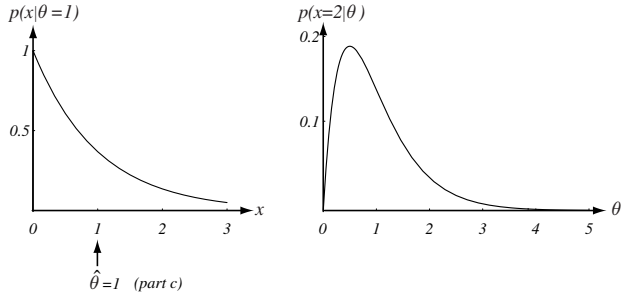
Problem Solutions

Section 3.2

1. Our exponential function is:

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) See Figure. Note that $p(x = 2|\theta)$ is not maximized when $\theta = 2$ but instead for a value less than 1.0.



(b) The log-likelihood function is

$$l(\theta) = \sum_{k=1}^{n} \ln\, p(x_k|\theta) = \sum_{k=1}^{n} [\ln\, \theta - \theta x_k] = n \ln\, \theta - \theta \sum_{k=1}^{n} x_k.$$

We solve $\nabla_\theta l(\theta) = 0$ to find $\hat{\theta}$ as

$$\begin{aligned}\nabla_\theta l(\theta) &= \frac{\partial}{\partial\theta}\left[n \ln \theta - \theta \sum_{k=1}^{n} x_k\right] \\ &= \frac{n}{\theta} - \sum_{k=1}^{n} x_k = 0.\end{aligned}$$

Thus the maximum-likelihood solution is

$$\hat{\theta} = \frac{1}{\frac{1}{n}\sum\limits_{k=1}^{n} x_k}.$$

(c) Here we approximate the mean

$$\frac{1}{n}\sum_{k=1}^{n} x_n$$

by the integral

$$\int_0^\infty xp(x)dx,$$

which is valid in the large n limit. Noting that

$$\int_0^\infty xe^{-x}dx = 1,$$

we put these results together and see that $\hat{\theta} = 1$, as shown on the figure in part (a).

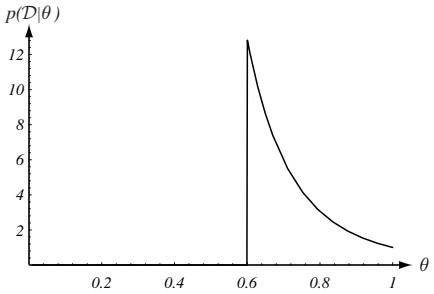
2. Our (normalized) distribution function is

$$p(x|\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

(a) We will use the notation of an *indicator function* $I(\cdot)$, whose value is equal to 1.0 if the logical value of its argument is TRUE, and 0.0 otherwise. We can write the likelihood function using $I(\cdot)$ as

$$\begin{aligned}p(\mathcal{D}|\theta) &= \prod_{k=1}^{n} p(x_k|\theta) \\ &= \prod_{k=1}^{n} \frac{1}{\theta} I\left(0 \le x_k \le \theta\right) \\ &= \frac{1}{\theta^n} I\left(\theta \ge \max_k x_k\right)\ I\left(\min_k x_k \ge 0\right).\end{aligned}$$

We note that $1/\theta^n$ decreases monotonically as θ increases but also that $I(\theta \ge \max\limits_k x_k)$ is 0.0 if θ is less than the maximum value of x_k. Therefore, our likelihood function is maximized at $\hat{\theta} = \max\limits_k x_k$.



(b) See Figure.

3. We are given that

$$z_{ik} = \begin{cases} 1 & \text{if the state of nature for the } k^{th} \text{ sample is } \omega_i \\ 0 & \text{otherwise.} \end{cases}$$

(a) The samples are drawn by successive independent selection of a state of nature ω_i with probability $P(\omega_i)$. We have then

$$\Pr[z_{ik} = 1|P(\omega_i)] = P(\omega_i)$$

and

$$\Pr[z_{ik} = 0|P(\omega_i)] = 1 - P(\omega_i).$$

These two equations can be unified as

$$P(z_{ik}|P(\omega_i)) \quad = \quad [P(\omega_i)]^{z_{ik}}[1 - P(\omega_i)]^{1-z_{ik}}.$$

By the independence of the successive selections, we have

$$\begin{aligned} P(z_{i1}, \cdots, z_{in}|P(\omega_i)) &= \prod_{k=1}^{n} P(z_{ik}|P(\omega_i)) \\ &= \prod_{k=1}^{n} [P(\omega_i)]^{z_{ik}}[1 - P(\omega_i)]^{1-z_{ik}}. \end{aligned}$$

(b) The log-likelihood as a function of $P(\omega_i)$ is

$$\begin{aligned} l(P(\omega_i)) &= \ln\, P(z_{i1}, \cdots, z_{in}|P(\omega_i)) \\ &= \ln\, \left[\prod_{k=1}^{n} [P(\omega_i)]^{z_{ik}}[1 - P(\omega_i)]^{1-z_{ik}}\right] \\ &= \sum_{k=1}^{n} \left[z_{ik}\, \ln\, P(\omega_i) + (1 - z_{ik})\, \ln\, (1 - P(\omega_i))\right]. \end{aligned}$$

Therefore, the maximum-likelihood values for the $P(\omega_i)$ must satisfy

$$\nabla_{P(\omega_i)} l(P(\omega_i)) = \frac{1}{P(\omega_i)} \sum_{k=1}^{n} z_{ik} - \frac{1}{1 - P(\omega_i)} \sum_{k=1}^{n} (1 - z_{ik}) = 0.$$

We solve this equation and find

$$(1-\hat{P}(\omega_i))\sum_{k=1}^{n} z_{ik} \quad = \quad \hat{P}(\omega_i)\sum_{k=1}^{n}(1-z_{ik}),$$

which can be rewritten as

$$\sum_{k=1}^{n} z_{ik} \quad = \quad \hat{P}(\omega_i)\sum_{k=1}^{n} z_{ik} + n\hat{P}(\omega_i) - \hat{P}(\omega_i)\sum_{k=1}^{n} z_{ik}.$$

The final solution is then

$$\hat{P}(\omega_i) \quad = \quad \frac{1}{n}\sum_{k=1}^{n} z_{ik}.$$

That is, the estimate of the probability of category ω_i is merely the probability of obtaining its indicatory value in the training data, just as we would expect.

4. We have n samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from the discrete distribution

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}.$$

The likelihood for a particular sequence of n samples is

$$P(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\theta}) = \prod_{k=1}^{n}\prod_{i=1}^{d} \theta_i^{x_{ki}}(1-\theta_i)^{1-x_{ki}},$$

and the log-likelihood function is then

$$l(\boldsymbol{\theta}) \quad = \quad \sum_{k=1}^{n}\sum_{i=1}^{d} x_{ki} \ln \theta_i + (1-x_{ki}) \ln (1-\theta_i).$$

To find the maximum of $l(\boldsymbol{\theta})$, we set $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \mathbf{0}$ and evaluate component by component ($i = 1, \ldots, d$) and get

$$\begin{aligned}
\left[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})\right]_i &= \nabla_{\theta_i} l(\theta) \\
&= \frac{1}{\theta_i}\sum_{k=1}^{n} - \frac{1}{1-\theta_i}\sum_{k=1}^{n}(1-x_{ki}) \\
&= 0.
\end{aligned}$$

This implies that for any i

$$\frac{1}{\hat{\theta}_i}\sum_{k=1}^{n} x_{ki} = \frac{1}{1-\hat{\theta}_i}\sum_{k=1}^{n}(1-x_{ki}),$$

which can be rewritten as

$$(1-\hat{\theta}_i)\sum_{k=1}^{n} x_{ki} \quad = \quad \hat{\theta}_i\left(n - \sum_{k=1}^{n} x_{ki}\right).$$

The final solution is then

$$\hat{\theta}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}.$$

Since this result is valid for all $i = 1, \ldots, d$, we can write this last equation in vector form as

$$\hat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k.$$

Thus the maximum-likelihood value of $\boldsymbol{\theta}$ is merely the sample mean, just as we would expect.

5. The probability of finding feature x_i to be 1.0 in category ω_1 is denoted p:

$$p(x_i = 1|\omega_1) = 1 - p(x_i = 0|\omega_1) = p_{i1} = p > \frac{1}{2},$$

for $i = 1, \ldots, d$. Moreover, the normalization condition gives $p_{i2} = p(x_i|\omega_2) = 1 - p_{i1}$.

(a) A single observation $\mathbf{x} = (x_1, \ldots, x_d)$ is drawn from class ω_1, and thus have

$$p(\mathbf{x}|\omega_1) = \prod_{i=1}^{d} p(x_i|\omega_1) = \prod_{i=1}^{d} p^{x_i}(1-p)^{1-x_i},$$

and the log-likelihood function for p is

$$l(p) = \ln\, p(\mathbf{x}|\omega_1) = \sum_{i=1}^{d}[x_i \ln\, p + (1 - x_i)\, \ln\, (1-p)].$$

Thus the derivative is

$$\nabla_p l(p) \quad = \quad \frac{1}{p}\sum_{i=1}^{d} x_i - \frac{1}{(1-p)}\sum_{i=1}^{d}(1 - x_i).$$

We set this derivative to zero, which gives

$$\frac{1}{\hat{p}}\sum_{i=1}^{d} x_i = \frac{1}{1-\hat{p}}\sum_{i=1}^{d}(1 - x_i),$$

which after simple rearrangement gives

$$(1-\hat{p})\sum_{i=1}^{d} x_i \quad = \quad \hat{p}\left(d - \sum_{i=1}^{d} x_i\right).$$

Thus our final solution is

$$\hat{p} \quad = \quad \frac{1}{d}\sum_{i=1}^{d} x_i.$$

That is, the maximum-likelihood estimate of the probability of obtaining a 1 in any position is simply the ratio of the number of 1's in a single sample divided by the total number of features, given that the number of features is large.

(b) We define $T = (1/d)\sum_{i=1}^{d} x_i$ to be the proportion of 1's in a single observation $\mathbf{x}$. As the number of dimensions d approaches infinity, we have

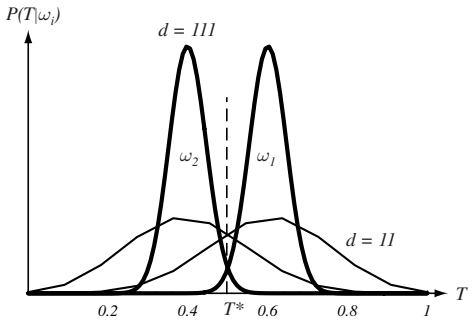
$$T = \frac{1}{d}\sum_{i=1}^{d} \mathcal{E}(x_i|\omega_1) = [1 \times p + 0 \times (1-p)] = p.$$

Likewise, the variance of T, given that we're considering just one class, ω_1, is

$$\begin{aligned} \mathrm{Var}(T|\omega_1) &= \frac{1}{d}\sum_{i=1}^{d} \mathrm{Var}(x_i|\omega_1) \\ &= \frac{1}{d^2}\sum_{i=1}^{d}[1^2 \times p + 0^2 \times (1-p) - p \times p] \\ &= \frac{p(1-p)}{d}, \end{aligned}$$

which vanishes as $d \to \infty$. Clearly, for minimum error, we choose ω_1 if $T > T^* = 1/2$ for $p > 1/2$. Since the variance vanishes for large d, the probability of error is zero for a single sample having a sufficiently large d.

(c) See figure.



6. The d-dimensional multivariate normal density is given by

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

We choose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ independent observations from $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The joint density is

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})\right].$$

The log-likelihood function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{n}{2}\ln\,(2\pi) - \frac{n}{2}\ln\,|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}) \\ &= -\frac{n}{2}\ln\,(2\pi) - \frac{n}{2}\ln\,|\boldsymbol{\Sigma}| - \frac{1}{2}\left[\sum_{k=1}^{n}\mathbf{x}_k^t\mathbf{x}_k - 2\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k + n\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]. \end{aligned}$$

We set the derivative of the log-likelihood to zero, that is,

$$\frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = -\frac{1}{2}\left[-2\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}\mathbf{x}_k + n2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] = \mathbf{0},$$

and find that

$$\hat{\boldsymbol{\Sigma}}^{-1}\sum_{k=1}^{n}\mathbf{x}_k = n\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}.$$

This gives the maximum-likelihood solution,

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k,$$

as expected. In order to simplify our calculation of $\hat{\boldsymbol{\Sigma}}$, we temporarily substitute $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$, and thus have

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\ln\,(2\pi) + \frac{n}{2}\ln\,|\mathbf{A}| - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})^t\mathbf{A}(\mathbf{x}_k - \boldsymbol{\mu}).$$

We use the above results and seek the solution to

$$\frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \mathbf{A}} = \frac{n}{2}\mathbf{A}^{-1} - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t = \mathbf{0}.$$

We now replace $\mathbf{A}$ by $\boldsymbol{\Sigma}^{-1}$ and find that

$$\frac{n}{2}\hat{\boldsymbol{\Sigma}} = \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t,$$

and then multiply both sides by $2/n$ and find our solution:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t.$$

As we would expect, the maximum-likelihood estimate of the covariance matrix is merely the covariance of the samples actually found.

7. The figure shows the model for ω_1, the incorrect model for ω_2, that is, $p(x|\omega_2) \sim N(0, 1)$ and the true model for ω_2 (dashed, and scaled by 30000 for visibility).

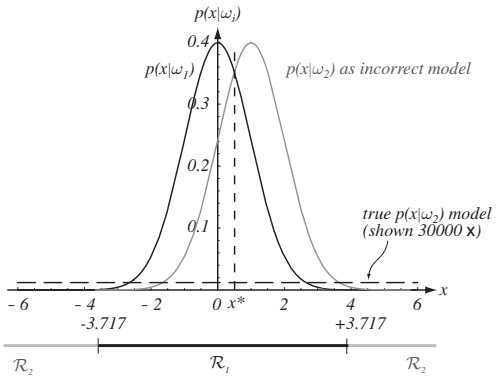
(a) According to Eq. 18 in the text, the maximum-likelihood estimate of the mean (for the incorrect model) is merely the sample mean of the data from the true model. In this case, then, we have $\hat{\mu} = 1.0$.

(b) Given two equal-variance Gaussians and equal priors, the decision boundary point is midway between the two means, here, $x^* = 0.5$.

(c) For the incorrect case, according to part (b), we have $\mathcal{R}_1$ is the line segment $x < 0.5$ and for $\mathcal{R}_2$ the line segment $x > 0.5$. For the correct case, we must solve numerically

$$\frac{1}{2\pi\sqrt{1}}\exp[-x^2/2] = \frac{1}{2\pi\sqrt{10^6}}\exp[-1(x-1)^2/(2\cdot 10^6)],$$

which gives the values $x = \pm 3.717$. The Bayesian decision boundaries and regions are shown along the bottom of the figure above.



(d) For the poor model, which has equal variances for the two Gaussians, the only decision boundary possible is a single point, as was illustrated in part (a). The best we can do with the incorrect model is to adjust the mean μ so as to match the rightmost of the decision boundaries given in part (c), that is, $x^* = 3.717$. This decision point is midway between the means of the two Gaussians — that is, at 0 and at μ. As such, we need to choose μ such that $(0 + \mu)/2 = 3.717$. This leads directly to the solution, $\mu = 7.43$.

(e) The maximum-likelihood solution in the incorrect model — here $p(x|\omega_2) \sim N(\mu, 1)$ — does not yield the minimum classification error. A different value of the parameter (here, $\mu = 7.43$) approximates the Bayes decision boundary better and gives a lower error than the maximum-likelihood solution in the incorrect model. As such, the faulty prior information about the model, which did not permit the true model to be expressed, could not be overcome with even an infinite amount of training data.

8. Consider a case in which the maximum-likelihood solution gives the *worst* possible classifier.

(a) In this case, the symmetry operation $x \leftrightarrow -x$ takes $p(x|\omega_1) \leftrightarrow p(x|\omega_2)$, and thus we are assured that the estimated distributions have this same symmetry property. For that reason, we are guaranteed that these distribution have the same value at $x = 0$, and thus $x^* = 0$ is a decision boundary. Since the Gaussian estimates must have the same variance, we are assured that there will be only a single intersection, and hence a single decision point, at $x^* = 0$.

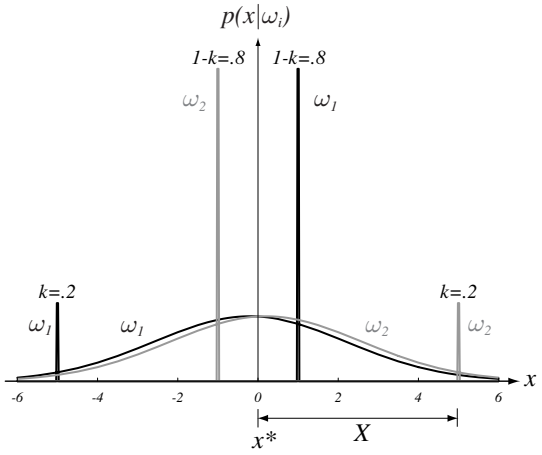
(b) See Figure.

(c) We have the estimate of the mean as

$$\hat{\mu}_1 = \int p(x|\omega_1)dx = (1-k)1 + k(-X) = 1 - k(1+X).$$

We are asked to "switch" the mean, that is, have $\hat{\mu}_1 < 0$ and $\hat{\mu}_2 > 0$. This can be assured if $X > (1-k)/k$. (We get the symmetric answer for ω_2.)

(d) Since the decision boundary is at $x^* = 0$, the error is $1 - k$, that is, the value of the distribution spikes on the "wrong" side of the decision boundary.



(e) Thus for the error to approach 1, we must have $k \to 0$, and this, in turn, requires X to become arbitrarily large, that is, $X \to (1-k)/k \to \infty$. In this way, a tiny "spike" in density sufficiently far away from the origin is sufficient to "switch" the estimated mean and give error equal 100%.

(f) There is no substantive difference in the above discussion if we constrain the variances to have a fixed value, since they will always be equal to each other, that is, $\hat{\sigma}_1^2 = \hat{\sigma}_2^2$. The equality of variances preserves the property that the decision point remains at $x^* = 0$; it consequently also ensures that for sufficiently small k and large X the error will be 100%.

(g) All the maximum-likelihood methods demand that the optimal solution exists in the model space. If the optimal solution does *not* lie in the solution space, then the proofs do not hold. This is actually a very strong restriction. Note that obtaining the error equal 100% solution is *not* dependent upon limited data, or getting caught in a local minimum — it arises because of the error in assuming that the optimal solution lies in the solution set. This is model error, as described on page 101 of the text.

9. The standard maximum-likelihood solution is

$$\hat{\theta} = \arg\max_{\theta} p(x|\theta).$$

Now consider a mapping $x \to \tau(x)$ where $\tau(\cdot)$ is continuous. Then we can write $p(\tau|\theta)d\tau = p(x|\theta)dx$, and

$$p(\tau|\theta) = \frac{p(x|\theta)}{d\tau/dx}.$$

Then we find the value of θ maximizing $p(\tau(x)|\theta)$ as

$$\begin{aligned} \arg\max_{\theta} p(\tau(x)|\theta) &= \arg\max_{\theta} \frac{p(x|\theta)}{d\tau/dx} \\ &= \arg\max_{\theta} p(x|\theta) \\ &= \hat{\theta}, \end{aligned}$$

where we have assumed $d\tau/dx \neq 0$ at $\theta = \hat{\theta}$. In short, then, the maximum-likelihood value of $\tau(\theta)$ is indeed $\hat{\theta}$. In practice, however, we must check whether the value of $\hat{\theta}$ derived this way gives a maximum or a minimum (or possibly inflection point) for $p(\tau|\theta)$.

10. Consider the novel method of estimating the mean of a set of points as taking its first value, which we denote $\mathbf{M} = \mathbf{x}_1$.

(a) Clearly, this unusual estimator of the mean is unbiased, that is, the expected value of this statistic is equal to the true value. In other words, if we repeat the selection of the first point of a data set we have

$$bias = \mathcal{E}[\mathbf{M}] - \boldsymbol{\mu} = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbf{M}(k) - \boldsymbol{\mu} = \mathbf{0},$$

where $\mathbf{M}(k)$ is the first point in data set k drawn from the given distribution.

(b) While the unusual method for estimating the mean may indeed be unbiased, it will generally have large variance, and this is an undesirable property. Note that $\mathcal{E}[(x_i - \mu)^2] = \sigma^2$, and the RMS error, σ, is independent of n. This undesirable behavior is quite different from that of the measurement of

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where we see

$$\begin{aligned}
\mathcal{E}[(\bar{x} - \mu)^2] &= \mathcal{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\mathcal{E}[(x_i - \mu)^2]\right] \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Thus the RMS error, $\sigma/\sqrt{n}$, approches 0 as $1/\sqrt{n}$. Note that there are many superior methods for estimating the mean, for instance the sample mean. (In Chapter 9 we shall see other techniques — ones based on resampling — such as the so-called "bootstrap" and "jackknife" methods.)

11. We assume $p_2(\mathbf{x}) \equiv p(\mathbf{x}|\omega_2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ but that $p_1(\mathbf{x}) \equiv p(\mathbf{x}|\omega_1)$ is arbitrary. The Kullback-Leibler divergence from $p_1(\mathbf{x})$ to $p_2(\mathbf{x})$ is

$$D_{KL}(p_1, p_2) = \int p_1(\mathbf{x}) \ln p_1(\mathbf{x}) d\mathbf{x} + \frac{1}{2}\int p_1(\mathbf{x})\left[d\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right] d\mathbf{x},$$

where we used the fact that p_2 is a Gaussian, that is,

$$p_2(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{2}\right].$$

We now seek $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to minimize this "distance." We set the derivative to zero and find

$$\frac{\partial}{\partial \boldsymbol{\mu}} D_{KL}(p_1, p_2) = -\int \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) p_1(\mathbf{x}) d\mathbf{x} = \mathbf{0},$$

and this implies

$$\boldsymbol{\Sigma}^{-1} \int p_1(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})d\mathbf{x} = \mathbf{0}.$$

We assume $\boldsymbol{\Sigma}$ is non-singular, and hence this equation implies

$$\int p_1(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})d\mathbf{x} = \mathcal{E}_1[\mathbf{x}-\boldsymbol{\mu}] = \mathbf{0},$$

or simply, $\mathcal{E}_1[\mathbf{x}] = \boldsymbol{\mu}$. In short, the mean of the second distribution should be the same as that of the Gaussian.

Now we turn to the covariance of the second distribution. Here for notational convenience we denote $\mathbf{A} = \boldsymbol{\Sigma}$. Again, we take a derivative of the Kullback-Leibler divergence and find:

$$\frac{\partial}{\partial \mathbf{A}} D_{KL}(p_1, p_2) = \mathbf{0} = \int p_1(\mathbf{x}) \left[-\mathbf{A}^{-1} + (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right] d\mathbf{x},$$

and thus

$$\mathcal{E}_1 \left[\boldsymbol{\Sigma} - (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right],$$

or

$$\mathcal{E}_1 \left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right] = \boldsymbol{\Sigma}.$$

In short, the covariance of the second distribution should indeed match that of the Gaussian.

Note that above, in taking the derivative above,

$$\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}|\mathbf{A}^{-1}$$

we relied on the fact that $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$ is symmetric since $\boldsymbol{\Sigma}$ is a covariance matrix. More generally, for an arbitrary non-singular matrix we would use

$$\frac{\partial |\mathbf{M}|}{\partial \mathbf{M}} = |\mathbf{M}|(\mathbf{M}^{-1})^t.$$

Section 3.3

12. In the text we saw the following results:

1. The posterior density can be computed as
$$p(\mathbf{x}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}.$$
2. $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) = p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta}|\mathcal{D}))$.
3. $p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) = p(\mathbf{x}|\boldsymbol{\theta})$, that is, the distribution of $\mathbf{x}$ is known completely once we know the value of the parameter vector, regardless of the data $\mathcal{D}$.
4. $p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}$.

These are justified as follows:

1. This statement reflects the conceptual difference between the maximum-likelihood estimator and Bayesian estimator. The Bayesian learning method considers the parameter vector $\boldsymbol{\theta}$ to be a random variable rather than a fixed value, as in maximum-likelihood estimation. The posterior density $p(\mathbf{x}|\mathcal{D})$ also depends upon the probability density $p(\boldsymbol{\theta})$ distributed over the entire $\boldsymbol{\theta}$ space instead of a single value. Therefore, the $p(\mathbf{x}|\mathcal{D})$ is the integration of $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ over the entire parameter space. The maximum-likelihood estimator can be regarded as a special case of Bayesian estimator, where $p(\boldsymbol{\theta})$ is uniformly distributed so that its effect disappears after the integration.

2. The $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ implies two steps in computation. One is the computation of the probability density $\boldsymbol{\theta}$ given the data set $\mathcal{D}$, that is, $p(\boldsymbol{\theta}|\mathcal{D})$. The other is the computation of the probability density of $\mathbf{x}$ given $\boldsymbol{\theta}$, that is, $p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D})$. The above two steps are independent of each other, and thus $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ is the product of the results of the two steps.

3. As mentioned in the text, the selection of $\mathbf{x}$ and that of the training samples $\mathcal{D}$ is done independently, that is, the selection of $\mathbf{x}$ does not depend upont $\mathcal{D}$. Therefore we have $p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) = p(\mathbf{x}|\boldsymbol{\theta})$.

4. We substitute the above relations into Eq. 24 in the text and get Eq. 25.

Section 3.4

13. We seek a novel approach for finding the maximum-likelihood estimate for $\boldsymbol{\Sigma}$.

(a) We first inspect the forms of a general vector $\mathbf{a}$ and matrix $\mathbf{A}$:

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \text{ and } \mathbf{A} = \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \dots & A_{nn} \end{pmatrix}.$$

Consider the scalar

$$\mathbf{a}^t\mathbf{A}\mathbf{a} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_j A_{ij} a_j.$$

The (i,i)th element of this scalar is $\sum_{j=1}^{n} A_{ij}a_j a_i$, and the trace of $\mathbf{A}\mathbf{a}\mathbf{a}^t$ is the sum of these diagonal elements, that is,

$$\text{tr } (\mathbf{A}\mathbf{a}\mathbf{a}^t) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}a_j a_i = \mathbf{a}^t\mathbf{A}\mathbf{a}.$$

(b) We seek to show that the likelihood function can be written as

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n|\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}^{-1}|^{n/2}} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right)\right].$$

We note that $p(\mathbf{x}|\boldsymbol{\Sigma}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is known and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent observations from $p(\mathbf{x}|\boldsymbol{\Sigma})$. Therefore the likelihood is

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) &= \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}^{-1}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})\right] \\ &= \frac{|\boldsymbol{\Sigma}|^{-n/2}}{(2\pi)^{nd/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})\right]. \end{aligned}$$

From the results in part (a), with $\mathbf{a} = \mathbf{x}_k - \boldsymbol{\mu}$ and $|\mathbf{A}| = |\boldsymbol{\Sigma}^{-1}|$, we have

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) &= \frac{|\boldsymbol{\Sigma}|^{-n/2}}{(2\pi)^{nd/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n} \text{tr}\ \left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right)\right] \\ &= \frac{|\boldsymbol{\Sigma}^{-1}|^{-n/2}}{(2\pi)^{nd/2}} \exp\left[-\frac{1}{2}\text{tr}\ \left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right)\right], \end{aligned}$$

where we used $\sum_{k=1}^{n} \text{tr}\ (A_k) = \text{tr}\ \left(\sum_{k=1}^{n} A_k\right)$ and $|\boldsymbol{\Sigma}^{-1}| = |\boldsymbol{\Sigma}|^{-1}$.

(c) Recall our definition of the sample covariance matrix:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t.$$

Here we let $\mathbf{A} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}$, which easily leads to the following equalities

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} &= \mathbf{A}\hat{\boldsymbol{\Sigma}}^{-1}, \\ |\boldsymbol{\Sigma}^{-1}| &= |\mathbf{A}\hat{\boldsymbol{\Sigma}}^{-1}| = |\mathbf{A}|\hat{|\boldsymbol{\Sigma}^{-1}|} \\ &= |\mathbf{A}||\hat{\boldsymbol{\Sigma}}|^{-1} = \frac{|\mathbf{A}|}{|\hat{\boldsymbol{\Sigma}}|} \\ &= \frac{\lambda_1\lambda_2\cdots\lambda_d}{|\hat{\boldsymbol{\Sigma}}|}, \end{aligned}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{A}$. We substitute these into our result in part (b) to get

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) &= \frac{|\boldsymbol{\Sigma}^{-1}|^{n/2}}{(2\pi)^{nd/2}} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right)\right] \\ &= \frac{(\lambda_1 \ldots \lambda_n)^{n/2}}{(2\pi)^{nd/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} \exp\left[-\frac{1}{2}\text{tr}\left(n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\right)\right]. \end{aligned}$$

Note, however, that $\text{tr}[n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}] = n[\text{tr}\ (\mathbf{A})] = n(\lambda_1 + \cdots + \lambda_d)$, and thus we have

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) = \frac{(\lambda_1 \ldots \lambda_n)^{n/2}}{(2\pi)^{nd/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} \exp\left[-\frac{n}{2}(\lambda_1 + \cdots + \lambda_d)\right].$$

(d) The expression for $p(\mathbf{x}_1, \cdots, \mathbf{x}_n|\boldsymbol{\Sigma})$ in part (c) depends on $\boldsymbol{\Sigma}$ only through $\lambda_1, \cdots, \lambda_d$, the eigenvalues of $\mathbf{A} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}$. We can write our likelihood, then, as

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{nd/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} \left[\prod_{i=1}^{d} \lambda_i e^{-\lambda_i}\right]^{n/2}.$$

Maximizing $p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ is equivalent to maximizing $\lambda_i e^{-\lambda_i}$ with respect to λ_i. We do this by setting the derivative to zero, that is,

$$\frac{\partial[\lambda_i e^{-\lambda_i}]}{\partial \lambda_i} = e^{-\lambda_i} + \lambda_i(-e^{-\lambda_i}) = 0,$$

which has solution $\lambda_i = 1$. In short, $p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma})$ is maximized by choosing $\lambda_1 = \lambda_2 = \cdots = \lambda_n = 1$. This means that $\mathbf{A} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}$, or $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$, as expected.

14. First we note that $p(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. We have also $l_k = i$ if the state of nature for $\mathbf{x}_k$ was ω_i.

(a) From Bayes' Rule we can write

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, l_1, \ldots, l_n|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, l_1, \ldots, l_n, \boldsymbol{\Sigma})p(l_1, \ldots, l_n).$$

Because the distribution of $l_1, \ldots, l_n$ does not depend on $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c$ or $\boldsymbol{\Sigma}$, we can write

$$\begin{aligned} &p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}, l_1, \ldots, l_n) \\ &= \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}, l_k) \\ &= \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})\right]. \end{aligned}$$

The l_i are independent, and thus the probability density of the ls is a product,

$$p(l_1, \ldots, l_n) = \prod_{k=1}^{n} p(l_k) = \prod_{k=1}^{n} p(\omega_{l_k}).$$

We combine the above equations and get

$$\begin{aligned} &p(\mathbf{x}_1, \ldots, \mathbf{x}_n, l_1, \ldots, l_n|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}) \\ = &\frac{\prod_{k=1}^{n} P(\omega_{l_k})}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})\right]. \end{aligned}$$

(b) We sum the result of part (a) over n samples to find

$$\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k}) = \sum_{i=1}^{l}\sum_{l_k=1}(\mathbf{x}_k - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_i).$$

The likelihood function for $\boldsymbol{\mu}_i$ is

$$\begin{aligned} l(\boldsymbol{\mu}_i) &= \frac{\prod_{k=1}^{n} P(\omega_i)}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{j=1}^{c}\sum_{k:l_k=j}(\mathbf{x}_k-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_i)\right] \\ &\propto \exp\left[-\frac{1}{2}\sum_{l_k=i}(\mathbf{x}_k-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_i)\right]. \end{aligned}$$

Thus, the maximum-likelihood estimate of $\boldsymbol{\mu}_i$ is a function of $\mathbf{x}_k$ restricted to $l_k = i$. But these $\mathbf{x}_k$'s are from an independent identically distributed (i.i.d.) sample from $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Thus, by the result for samples drawn from a single normal population, it follows that

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i}\sum_{l_k=i}\mathbf{x}_k,$$

where $n_i = \{\mathbf{x}_k : l_k = i\} = \sum_{l_k=i} 1$. Therefore, our solution is

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{l_k=i}\mathbf{x}_k}{\sum_{l_k=i}1}.$$

We also know that if we have a maximum-likelihood estimate for $\boldsymbol{\mu}$, then the maximum-likelihood estimate for $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^t.$$

But the maximum-likelihood estimate $\hat{\boldsymbol{\mu}}$ corresponds to the distribution of $\mathbf{x}_k$ is $\hat{\boldsymbol{\mu}}_{l_k}$. Thus, we have the estimate

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_{l_k})(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_{l_k})^t.$$

15. Consider the problem of learning the mean of a univariate normal distribution.

(a) From Eqs. 34 and 35 in the text, we have

$$\mu_n = \frac{n\sigma_o^2}{n\sigma_o^2+\sigma^2}m_n + \frac{\sigma^2}{n\sigma_o^2+\sigma^2}\mu_o,$$

and

$$\sigma_n^2 = \frac{\sigma_o^2\sigma^2}{n\sigma_o^2+\sigma^2},$$

where the sample mean is

$$m_n = \frac{1}{n}\sum_{k=1}^{n}x_k.$$

Here μ_o is formed by averaging n_o fictitious samples x_k for $k = -n_o+1, -n_o+2, \ldots, 0$. Thus we can write

$$\mu_o = \frac{1}{n_o}\sum_{k=-n_o+1}^{0} x_k,$$

and

$$\begin{aligned}\mu_n &= \frac{\sum_{k=1}^{n} x_k}{n+\sigma^2/\sigma_o^2} + \frac{\sigma^2/\sigma_o^2}{\sigma^2/\sigma_o^2+n}\frac{1}{n_o}\sum_{k=-n_o+1}^{0} x_k \\ &= \frac{\sum_{k=1}^{n} x_k}{n+n_o} + \frac{n_o}{n+n_o}\frac{1}{n_o}\sum_{k=1-n_o}^{0} x_k.\end{aligned}$$

We can use the fact that $n_o = \sigma^2/\sigma_o^2$ to write

$$\mu_n = \frac{1}{n+n_o}\sum_{k=-n_o+1}^{n} x_k.$$

Likewise, we have

$$\begin{aligned}\sigma_n^2 &= \frac{\sigma^2\sigma_o^2}{n\sigma_o^2+\sigma^2} \\ &= \frac{\sigma^2}{n+\sigma^2/\sigma_o^2} = \frac{\sigma^2}{n+n_o}.\end{aligned}$$

(b) The result of part (a) can be interpreted as follows: For a suitable choice of the prior density $p(\mu) \sim N(\mu_o, \sigma_o^2)$, maximum-likelihood inference on the "full" sample on $n+n_o$ observations coincides with Bayesian inference on the "second sample" of n observations. Thus, by suitable choice of prior, Bayesian learning can be interpreted as maximum-likelihood learning and here the suitable choice of prior in Bayesian learning is

$$\begin{aligned}\mu_o &= \frac{1}{n_o}\sum_{k=-n_o+1}^{0} x_k, \\ \sigma_o^2 &= \frac{\sigma^2}{n_o}.\end{aligned}$$

Here μ_o is the sample mean of the first n_o observations and σ_o^2 is the variance based on those n_o observations.

16. We assume that $\mathbf{A}$ and $\mathbf{B}$ are non-singular matrices of the same order.

(a) Consider Eq. 44 in the text. We write

$$\begin{aligned}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} &= \mathbf{A}[(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{B}^{-1})^{-1}] = \mathbf{A}[\mathbf{B}^{-1}(\mathbf{A}+\mathbf{B})]^{-1} \\ &= \mathbf{A}[\mathbf{B}^{-1}\mathbf{A}+\mathbf{I}]^{-1} = (\mathbf{A}^{-1})^{-1}(\mathbf{B}^{-1}\mathbf{A}+\mathbf{I})^{-1} \\ &= [(\mathbf{B}^{-1}\mathbf{A}+\mathbf{I})\mathbf{A}^{-1}]^{-1} = (\mathbf{B}^{-1}+\mathbf{A}^{-1})^{-1}.\end{aligned}$$

We interchange the roles of $\mathbf{A}$ and $\mathbf{B}$ in this equation to get our desired answer:

$$\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A} = (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}.$$

(b) Recall Eqs. 41 and 42 in the text:

$$\begin{aligned}\boldsymbol{\Sigma}_n^{-1} &= n\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_o^{-1}\\ \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n &= n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_n+\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o.\end{aligned}$$

We have solutions

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)\boldsymbol{\mu}_n+\frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_o,$$

and

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{n}\boldsymbol{\Sigma}.$$

Taking the inverse on both sides of Eq. 41 in the text gives

$$\boldsymbol{\Sigma}_n = \left(n\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_o^{-1}\right)^{-1}.$$

We use the result from part (a), letting $\mathbf{A}=\frac{1}{n}\boldsymbol{\Sigma}$ and $\mathbf{B}=\boldsymbol{\Sigma}_o$ to get

$$\begin{aligned}\boldsymbol{\Sigma}_n &= \frac{1}{n}\boldsymbol{\Sigma}\left(\frac{1}{n}\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_o\right)^{-1}\\ \boldsymbol{\Sigma}_o &= \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma},\end{aligned}$$

which proves Eqs. 41 and 42 in the text. We also compute the mean as

$$\begin{aligned}\boldsymbol{\mu}_n &= \boldsymbol{\Sigma}_n(n\boldsymbol{\Sigma}^{-1}\mathbf{m}_n+\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o)\\ &= \boldsymbol{\Sigma}_n n\boldsymbol{\Sigma}^{-1}\mathbf{m}_n+\boldsymbol{\Sigma}_n\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o\\ &= \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{n}\boldsymbol{\Sigma}n\boldsymbol{\Sigma}^{-1}\mathbf{m}_n+\frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}_o\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o\\ &= \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\mathbf{m}_n+\frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_o.\end{aligned}$$

Section 3.5

17. The Bernoulli distribution is written

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d}\theta_i^{x_i}(1-\theta_i)^{1-x_i}.$$

Let $\mathcal{D}$ be a set of n samples $\mathbf{x}_1,\ldots,\mathbf{x}_n$ independently drawn according to $p(\mathbf{x}|\boldsymbol{\theta})$.

(a) We denote $\mathbf{s} = (s_1, \cdots, s_d)^t$ as the sum of the n samples. If we denote $\mathbf{x}_k = (x_{k1}, \cdots, x_{kd})^t$ for $k = 1, \ldots, n$, then $s_i = \sum_{k=1}^{n} x_{ki}, i = 1, \ldots, d$, and the likelihood is

$$\begin{aligned} P(\mathcal{D}|\boldsymbol{\theta}) &= P(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\theta}) = \underbrace{\prod_{k=1}^{n} P(\mathbf{x}_k|\boldsymbol{\theta})}_{\mathbf{x}_k \text{ are indep.}} \\ &= \prod_{k=1}^{n}\prod_{i=1}^{d} \theta_i^{x_{ki}}(1-\theta_i)^{1-x_{ki}} \\ &= \prod_{i=1}^{d} \theta_i^{\sum_{k=1}^{n} x_{ki}}(1-\theta_i)^{\sum_{k=1}^{n}(1-x_{ki})} \\ &= \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i}. \end{aligned}$$

(b) We assume an (unnormalized) uniform prior for $\boldsymbol{\theta}$, that is, $p(\boldsymbol{\theta}) = 1$ for $0 \leq \theta_i \leq 1$ for $i = 1, \cdots, d$, and have by Bayes' Theorem

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}.$$

From part (a), we know that $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta)^{n-s_i}$, and therefore the probability density of obtaining data set $\mathcal{D}$ is

$$\begin{aligned} p(\mathcal{D}) &= \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\boldsymbol{\theta} \\ &= \int_0^1 \cdots \int_0^1 \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\theta_1 d\theta_2 \cdots d\theta_d \\ &= \prod_{i=1}^{d} \int_0^1 \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\theta_i. \end{aligned}$$

Now $s_i = \sum_{k=1}^{n} x_{ki}$ takes values in the set $\{0, 1, \ldots, n\}$ for $i = 1, \ldots, d$, and if we use the identity

$$\int_0^1 \theta^m(1-\theta)^n d\theta = \frac{m!n!}{(m+n+1)!},$$

and substitute into the above equation, we get

$$p(\mathcal{D}) = \prod_{i=1}^{d} \int_0^1 \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\theta_i = \prod_{i=1}^{d} \frac{s_i!(n-s_i)!}{(n+1)!}.$$

We consolidate these partial results and find

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \\ &= \frac{\prod_{i=1}^{d}\theta_i^{s_i}(1-\theta_i)^{n-s_i}}{\prod_{i=1}^{d}s_i!(n-s_i)!/(n+1)!} \\ &= \prod_{i=1}^{d}\frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{n-s_i}. \end{aligned}$$

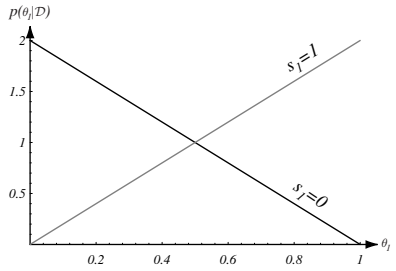
(c) We have $d=1, n=1$, and thus

$$p(\theta_1|\mathcal{D}) = \frac{2!}{s_1!(n-s_1)!}\theta_1^{s_1}(1-\theta_1)^{n-s_1} = \frac{2}{s_1!(1-s_1)!}\theta_1^{s_1}(1-\theta_1)^{1-s_1}.$$

Note that s_1 takes the discrete values 0 and 1. Thus the densities are of the form

$$\begin{aligned} s_1 = 0 \quad &: \quad p(\theta_1|\mathcal{D}) = 2(1-\theta_1) \\ s_1 = 1 \quad &: \quad p(\theta_1|\mathcal{D}) = 2\theta_1, \end{aligned}$$

for $0 \le \theta_1 \le 1$, as shown in the figure.



18. Consider how knowledge of an invariance can guide our choice of priors.

(a) We are given that s is actually the number of times that $x=1$ in the first n tests. Consider the $(n+1)$st test. If again $x=1$, then there are $\binom{n+1}{s+1}$ permutations of 0s and 1s in the $(n+1)$ tests, in which the number of 1s is $(s+1)$. Given the assumption of invariance of exchangeablility (that is, all permutations have the same chance to appear), the probability of each permutation is

$$P_{instance} = \frac{1}{\binom{n+1}{s+1}}.$$

Therefore, the probability of $x=1$ after n tests is the product of two probabilities: one is the probability of having $(s+1)$ number of 1s, and the other is the probability for a particular instance with $(s+1)$ number of 1s, that is,

$$\Pr[x_{n+1}=1|\mathcal{D}^n] = \Pr[x_1+\cdots+x_n = s+1]\cdot P_{instance} = \frac{p(s+1)}{\binom{n+1}{s+1}}.$$

The same analysis yields

$$P(x_{n+1} = 0|\mathcal{D}^n) = \frac{p(s)}{\binom{n+1}{s}}.$$

Therefore, the ratio can be written as

$$\frac{\Pr[x_{n+1} = 1|\mathcal{D}^n]}{\Pr[x_{n+1} = 0|\mathcal{D}^n]} = \frac{p(s+1)/\binom{n+1}{s+1}}{p(s)/\binom{n+1}{s}}.$$

(b) Given $p(s) \simeq p(s+1)$ for large s, we have

$$\begin{aligned}
\frac{p(s+1)/\binom{n+1}{s+1}}{p(s)/\binom{n+1}{s}} &\simeq \frac{\binom{n+1}{s}}{\binom{n+1}{s+1}} = \frac{(n+1)!}{s!(n+1-s)!}\frac{(s+1)!(n+1-s-1)!}{(n+1)!} \\
&= \frac{(s+1)!(n+1-s-1)!}{s!(n+1-s)!} = \frac{s!(s+1)(n+1-s-1)!}{s!(n+1-s-1)!(n+1-s)} \\
&= \frac{s+1}{n+1-s}.
\end{aligned}$$

We can see that for some n, the ratio will be small if s is small, and that the ratio will be large if s is large. This implies that with n increasing, if x is unlikely to be 1 in the first n tests (i.e., small s), it would be unlikely to be 1 in the $(n+1)$st test, and thus s remains small. On the contrary, if there are more 1s than 0s in the first n tests (i.e., large s), it is very likely that in the $(n+1)$st test the x remains 1, and thus retains s large.

(c) If $p(\theta) \sim U(0,1)$ for some n, then $p(\theta)$ is a constant, which we call c. Thus

$$\begin{aligned}
p(s) &= \int_0^1 \theta^s(1-\theta)^{n-s}p(\theta)d\theta \\
&= c\binom{n}{s}\int_0^1 \theta^s(1-\theta)^{n-s}d\theta \\
&= c\binom{n}{s}\left(-\frac{\theta^{s+1}}{s+1}\frac{(1-\theta)^{n-s+1}}{n-s+1}\right)\Bigg|_0^1 \\
&= 0,
\end{aligned}$$

which of course does not depend upon s.

19. Consider MAP estimators, that is, ones that maximize $l(\boldsymbol{\theta})p(\boldsymbol{\theta})$.

(a) In this problem, the parameter needed to be estimated is $\boldsymbol{\mu}$. Given the training data, we have

$$l(\boldsymbol{\mu})p(\boldsymbol{\mu}) = \ln[p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu})]$$

where for the Gaussian

$$\ln[p(\mathcal{D}|\boldsymbol{\mu})] = \ln\left(\prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\mu})\right)$$

$$\begin{aligned} &= \sum_{k=1}^{n} \ln[p(\mathbf{x}_k|\boldsymbol{\mu})] \\ &= -\frac{n}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \sum_{k=1}^{n}\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}) \end{aligned}$$

and

$$p(\boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_0)^t\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\mu}-\mathbf{m}_0)\right].$$

The MAP estimator for the mean is then

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \arg\max_{\boldsymbol{\mu}}\left\{\left[-\frac{n}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \sum_{k=1}^{n}\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})\right]\right. \\ &\left.\times\left[\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_0)^t\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{\mu}-\mathbf{m}_0)\right]\right]\right\}. \end{aligned}$$

(b) After the linear transform governed by the matrix $\mathbf{A}$, we have

$$\boldsymbol{\mu}' = \mathcal{E}[\mathbf{x}'] = \mathcal{E}[\mathbf{A}\mathbf{x}] = \mathbf{A}\mathcal{E}[\mathbf{x}] = \mathbf{A}\boldsymbol{\mu},$$

and

$$\begin{aligned} \boldsymbol{\Sigma}' &= \mathcal{E}[(\mathbf{x}'-\boldsymbol{\mu}')(\mathbf{x}'-\boldsymbol{\mu}')^t] \\ &= \mathcal{E}[(\mathbf{A}\mathbf{x}'-\mathbf{A}\boldsymbol{\mu}')(\mathbf{A}\mathbf{x}'-\mathbf{A}\boldsymbol{\mu}')^t] \\ &= \mathcal{E}[\mathbf{A}(\mathbf{x}'-\boldsymbol{\mu}')(\mathbf{x}'-\boldsymbol{\mu}')^t\mathbf{A}^t] \\ &= \mathbf{A}\mathcal{E}[(\mathbf{x}'-\boldsymbol{\mu}')(\mathbf{x}'-\boldsymbol{\mu}')^t]\mathbf{A}^t \\ &= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t. \end{aligned}$$

Thus we have the log-likelihood

$$\begin{aligned} \ln[p(\mathcal{D}'|\boldsymbol{\mu}')] &= \ln\left(\prod_{k=1}^{n} p(\mathbf{x}'_k|\boldsymbol{\mu}')\right) \\ &= \ln\left(\prod_{k=1}^{n} p(\mathbf{A}\mathbf{x}_k|\mathbf{A}\boldsymbol{\mu})\right) \\ &= \sum_{k=1}^{n}\ln[p(\mathbf{A}\mathbf{x}_k|\mathbf{A}\boldsymbol{\mu})] \\ &= -\frac{n}{2}\ln\left[(2\pi)^d|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t|\right] - \sum_{k=1}^{n}\frac{1}{2}(\mathbf{A}\mathbf{x}_k-\mathbf{A}\boldsymbol{\mu})^t(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{x}_k-\mathbf{A}\boldsymbol{\mu}) \\ &= -\frac{n}{2}\ln\left[(2\pi)^d|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t|\right] - \sum_{k=1}^{n}\frac{1}{2}((\mathbf{x}-\boldsymbol{\mu})^t\mathbf{A}^t)((\mathbf{A}^{-1})^t\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1})(\mathbf{A}(\mathbf{x}_k-\boldsymbol{\mu})) \\ &= -\frac{n}{2}\ln\left[(2\pi)^d|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t|\right] - \sum_{k=1}^{n}\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^t(\mathbf{A}^t(\mathbf{A}^{-1})^t)\boldsymbol{\Sigma}^{-1}(\mathbf{A}^{-1}\mathbf{A})(\mathbf{x}_k-\boldsymbol{\mu}) \\ &= -\frac{n}{2}\ln\left[(2\pi)^d|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t|\right] - \sum_{k=1}^{n}\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}). \end{aligned}$$

Likewise we have that the density of $\boldsymbol{\mu}'$ is a Gaussian of the form

$$\begin{aligned}
p(\boldsymbol{\mu}') &= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0'|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\mu}'-\mathbf{m}_0)^t\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}'-\mathbf{m}_0)\right] \\
&= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0'|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{A}\boldsymbol{\mu}-\mathbf{A}\mathbf{m}_0)^t(\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^t)^{-1}(\mathbf{A}\boldsymbol{\mu}-\mathbf{A}\mathbf{m}_0)\right] \\
&= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0'|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_0)^t\mathbf{A}^t(\mathbf{A}^{-1})^t\boldsymbol{\Sigma}_0^{-1}\mathbf{A}^{-1}\mathbf{A}(\boldsymbol{\mu}-\mathbf{m}_0)\right] \\
&= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0'|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_0)^t\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\mathbf{m}_0)\right].
\end{aligned}$$

Thus the new MAP estimator is

$$\begin{aligned}
\hat{\boldsymbol{\mu}}' = \arg\max_{\boldsymbol{\mu}}\Bigg\{ &-\frac{n}{2}\ln\left[(2\pi)^d|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t|\right] \\
&-\sum_{k=1}^{n}\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})\left[\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0'|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_0)^t\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\mathbf{m}_0)\right]\right]\Bigg\}.
\end{aligned}$$

We compare $\hat{\boldsymbol{\mu}}$ and see that the two equations are the same, up to a constant. Therefore the estimator gives the appropriate estimate for the transformed mean $\hat{\boldsymbol{\mu}}'$.

20. Consider the problem of noninformative priors.

(a) Assume that $\tilde{\sigma}$ has a uniform distribution, that is, $\tilde{p}(\tilde{\sigma}) = c$, where c is a constant. We make the correspondence $\tilde{\sigma} = \ln\sigma$ and thus $\sigma = \exp[\tilde{\sigma}] = f(\tilde{\sigma})$, for some $f(\cdot)$. Since $\tilde{p}(\tilde{\sigma}) = c$, we have

$$\begin{aligned}
p(\sigma) &= \tilde{p}(f^{-1}(\sigma))\frac{df^{-1}(\sigma)}{d\sigma} \\
&= \tilde{p}(\tilde{\sigma})\frac{d\ln(\sigma)}{d\sigma} \\
&= \frac{c}{\sigma}.
\end{aligned}$$

For the case $c = 1$, we have $p(\sigma) = 1/\sigma$.

(b) The noninformative priors here are very simple,

$$\begin{aligned}
p(\theta_0) &= \frac{1}{2\pi-0} = \frac{1}{2\pi} \\
p(\sigma_\theta) &= 1/\sigma_\theta.
\end{aligned}$$

21. Note that $0 \le p(\boldsymbol{\theta}|\mathcal{D}) \le 1$. In order to converge while $n \to \infty$, it must be true that

$$\lim_{n\to\infty} p(\boldsymbol{\theta}|\mathcal{D}) = \lim_{n\to\infty}\frac{p(\mathbf{x}_k|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}^{n-1})}{\int p(\mathbf{x}_k|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}^{n-1})d\boldsymbol{\theta}} = \lim_{n\to\infty} p(\boldsymbol{\theta}|\mathcal{D}^{n-1}).$$

Note that $\lim_{n\to\infty} p(\boldsymbol{\theta}|\mathcal{D}^n) \to p(\boldsymbol{\theta})$. We assume $\lim_{n\to\infty} p(\boldsymbol{\theta}|\mathcal{D}) \to p(\boldsymbol{\theta}) \neq 0$ and $\lim_{n\to\infty}\mathbf{x}_n \to \mathbf{x}^*$. Then the above equation implies

$$p(\mathbf{x}^*|\boldsymbol{\theta}) = \lim_{n\to\infty} p(\mathbf{x}_n|\boldsymbol{\theta})$$

$$\begin{aligned}
&= \lim_{n\to\infty} \int p(\mathbf{x}_n|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \int \lim_{n\to\infty} p(\mathbf{x}_n|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \int p(\mathbf{x}^*|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= p(\mathbf{x}^*).
\end{aligned}$$

In summary, we have the conditions:

- $\lim_{n\to\infty} p(\boldsymbol{\theta}|\mathcal{D}^n) \to p(\boldsymbol{\theta}) \neq 0$
- $\lim_{n\to\infty} \mathbf{x}_n \to \mathbf{x}^*$
- $p(\mathbf{x}^*|\boldsymbol{\theta}) = p(\mathbf{x}^*)$, that is, $p(\mathbf{x}^*|\boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$.

22. Consider the Gibbs algorithm.

(a) Note that $p(x|\omega_2, \mu) \neq 0$ for $|x - \mu| < 1$ and this implies $x - 1 < \mu < x + 1$ and $p(\mu) \neq 0$ for $0 \leq \mu \leq 2$. We have the following cases:

- $x - 1 > 2$ and thus $x > 3$ or $x + 1 < 0$ and thus $x < -1$.
- $0 \leq x - 1 \leq 2$ and thus $1 \leq x \leq 3$. In that case,

$$p(x|\omega_2) = \int_{x-1}^{2} p(x|\omega_2, \mu)p(\mu)d\mu = \int_{x-1}^{2} \frac{1}{2}\frac{1}{2}d\mu = \frac{3-x}{4}.$$

- $0 \leq x + 1 \leq 2$ and thus $-1 \leq x \leq 1$. In that case,

$$p(x|\omega_2) = \int_{0}^{x+1} p(x|\omega_2, \mu)p(\mu)d\mu = \int_{0}^{x+1} \frac{1}{2}\frac{1}{2}d\mu = \frac{x+1}{4}.$$

Therefore, the class-conditional density is:

$$p(x|\omega_2) = \begin{cases} 0 & x < -1 \\ (x+1)/4 & -1 \leq x \leq 1 \\ (3-x)/4 & 1 \leq x \leq 3 \\ 0 & x > 3. \end{cases}$$

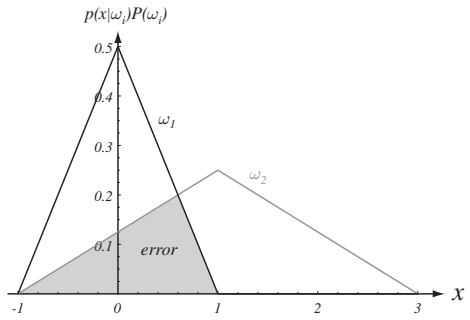
(b) The decision point is at

$$x^* = \arg_x [p(x|\omega_1)p(\omega_1) = p(x|\omega_2)p(\omega_2)].$$

(c) We let $P \equiv P(\omega_1)$ and note the normalization condition $P(\omega_1) + P(\omega_2) = 1$. There are two cases:

- $(1 + x)P \geq (x + 1)/4 \cdot (1 - P)$ and thus $1/5 \leq P \leq 1$, as shown in the figure.
  In this case, $(1 - x^*)P = (x^* + 1)/4 \cdot (1 - P)$ and this implies $x^* = (5P - 1)/(3P + 1)$, and the error is

$$P(error) = \int_{-1}^{x^*} p(x|\omega_2)P(\omega_2)dx + \int_{x^*}^{1} p(x|\omega_1)P(\omega_1)dx = \frac{24P^3 + 8P^2}{(3P+1)^2}.$$



- $0 \leq P \leq 1/5$, as shown in the figure.
  The error is

$$P(error) = \int_{-1}^{1} p(x|\omega_1)P(\omega_1)dx = P.$$

(d) Again, we are considering $p(x|\omega_1)P(\omega_1)$ and $p(x|\omega_2)P(\omega_2)$, but this time, we first use a single value of μ as the true one, that is, $p(\mu) = 1$. There are two cases for different relationships between P and μ.

- $p(x|\omega_1)P(\omega_1) \leq p(x|\omega_2, \mu)P(\omega_2)$ which implies $P \leq 1/2 \cdot (1 - P)$ or $0 \leq P \leq 1/3$. This case can be further divided into two sub-cases:

  subcase 1: $0 \leq \mu \leq 1$
  The error is

$$P(error) = \int_{0}^{1} \frac{1}{2}(1 - (-1)) \cdot \left(P - \frac{\mu^2}{2}\right) d\mu = P - P/6 = 5P/6.$$

  subcase 2: $1 < \mu \leq 2$
  The error is

$$P_2(error) = \int_{1}^{2} \frac{(2-\mu)^2}{2} P d\mu = P/6$$

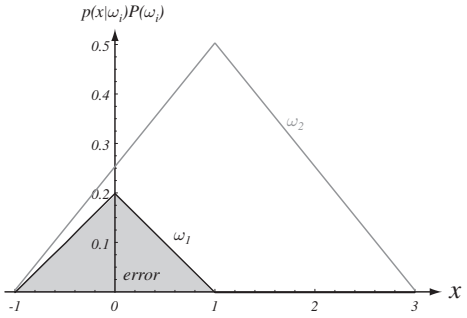
  and the total error for this case is

$$P_{Gibbs}(error) = P_1(error) + P_2(error) = \frac{5P}{6} + \frac{P}{6} = P.$$

- Other case: $1/3 \leq P \leq 1$. This case can be further subdivided into three subcases.

  subcase 1: $0 \leq \mu \leq (1 - P)/(2P)$ as shown in the figure.
  The error is

$$\begin{aligned} P_1(error) &= \int_{0}^{(1-P)/(2P)} \left[\frac{1-P}{4} \left(\frac{9}{2} - 2\mu - \frac{1}{2P} \right) \right. \\ &\qquad \left. - \frac{1}{2} \left(\left(\frac{1-P}{2} - P\mu \right) \left(\frac{1-3P}{2P} - \mu + 1 \right) \right) \right] d\mu \\ &= \frac{(1-P)^2(31P-7)}{48P^2}. \end{aligned}$$



subcase 2: $(1-P)/(2P) \leq \mu \leq (5P-1)/(2P)$, as shown in the figure.
The error is

$$\begin{aligned} P_2(error) &= \int_{(1-P)/(2P)}^{(5P-1)/(2P)} \frac{1-P}{4}\left(\frac{9}{2} - 2\mu - \frac{1}{2P}\right) d\mu \\ &= \frac{(5P-1)(1-P)(3P-1)}{8P^2}. \end{aligned}$$

subcase 3: $(5P-1)/(2P) \leq \mu \leq 2$, as shown in the figure.
The error is

$$\begin{aligned} P_3(error) &= \int_{(5P-1)/(2P)}^{2} \frac{P}{2}(2-\mu)^2 d\mu \\ &= \frac{P}{6}\left(\frac{1-P}{2P}\right)^3. \end{aligned}$$

Thus the probability of error under Gibbs sampling is

$$\begin{aligned} P_{Gibbs}(error) &= P_1(error) + P_2(error) + P_3(error) \\ &= -\frac{5P^2 - 6P + 1}{4P} \end{aligned}$$

(e) It can be easily confirmed that $P_{Gibbs}(error) < 2P_{Bayes}(error)$.

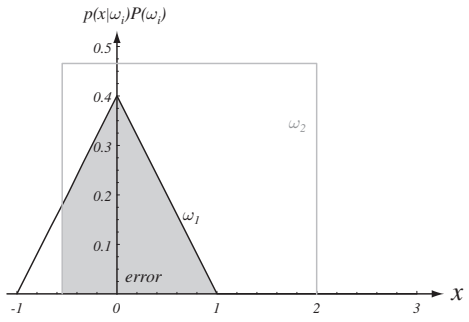
Section 3.6

23. Let $\mathbf{s}$ be a sufficient statistic for which $p(\theta|\mathbf{s}, \mathcal{D}) = p(\boldsymbol{\theta}|\mathbf{s})$; we assume $p(\boldsymbol{\theta}|\mathbf{s}) \neq 0$. In that case, we can write Bayes' law

$$p(\mathcal{D}|\mathbf{s}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}|\mathbf{s}, \mathcal{D})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})}$$

as

$$\begin{aligned} p(\mathcal{D}|\mathbf{s}, \boldsymbol{\theta}) &= \frac{p(\mathcal{D}, \mathbf{s}, \boldsymbol{\theta})}{p(\mathbf{s}, \boldsymbol{\theta})} \\ &= \frac{p(\boldsymbol{\theta}|\mathbf{s}, \mathcal{D})P(\mathcal{D}, \mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})p(\mathbf{s})} \end{aligned}$$



$$= \frac{p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D})p(\mathcal{D}|\mathbf{s})p(\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})p(\mathbf{s})}$$
$$= \frac{p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})}.$$

Note that the probability density of the parameter $\boldsymbol{\theta}$ is fully specified by the sufficient statistic; the data gives no further information, and this implies

$$p(\theta|\mathbf{s},\mathcal{D}) = p(\theta|\mathbf{s}).$$

Since $p(\boldsymbol{\theta}|\mathbf{s}) \neq 0$, we can write

$$\begin{aligned} p(\mathcal{D}|\mathbf{s},\boldsymbol{\theta}) &= \frac{p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})} \\ &= \frac{p(\boldsymbol{\theta}|\mathbf{s})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})} \\ &= p(\mathcal{D}|\mathbf{s}), \end{aligned}$$

which does not involve $\boldsymbol{\theta}$. Thus, $p(\mathcal{D}|\mathbf{s},\boldsymbol{\theta})$ is indeed independent of $\boldsymbol{\theta}$.

24. To obtain the maximum-likelihood estimate, we must maximize the likelihood function $p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{x}_1,\ldots,\mathbf{x}_n|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. However, by the Factorization Theorem (Theorem 3.1) in the text, we have

$$p(\mathcal{D}|\boldsymbol{\theta}) = g(\mathbf{s},\boldsymbol{\theta})h(\mathcal{D}),$$

where $\mathbf{s}$ is a sufficient statistic for $\boldsymbol{\theta}$. Thus, if we maximize $g(\mathbf{s},\boldsymbol{\theta})$ or equivalently $[g(\mathbf{s},\boldsymbol{\theta})]^{1/n}$, we will have the maximum-likelihoood solution we seek.

For the Rayleigh distribution, we have from Table 3.1 in the text,

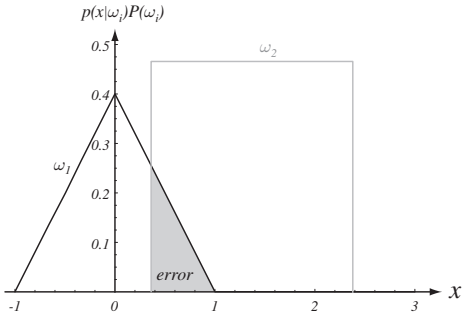
$$[g(s,\theta)]^{1/n} = \theta e^{-\theta s}$$

for $\theta > 0$, where

$$s = \frac{1}{n}\sum_{k=1}^{n} x_k^2.$$

Then, we take the derivative with respect to θ and find

$$\nabla_\theta [g(s,\theta)]^{1/n} = e^{-\theta s} - s\theta e^{-\theta s}.$$



We set this to 0 and solve to get

$$e^{-\hat{\theta}s} = s\hat{\theta}e^{-\hat{\theta}s},$$

which gives the maximum-likelihood solution,

$$\hat{\theta} = \frac{1}{s} = \left(\frac{1}{n}\sum_{k=1}^{n} x_k^2\right)^{-1}.$$

We next evaluate the second derivative at this value of $\hat{\theta}$ to see if the solution represents a maximum, a minimum, or possibly an inflection point:

$$\begin{aligned}\nabla_\theta^2[g(s,\theta)]^{1/n}\Big|_{\theta=\hat{\theta}} &= -se^{-\theta s} - se^{-\theta s} + s^2\theta e^{-\theta s}\Big|_{\theta=\hat{\theta}} \\ &= e^{-\hat{\theta}s}(s^2\hat{\theta} - 2s) = -se^{-1} < 0.\end{aligned}$$

Thus $\hat{\theta}$ indeed gives a maximum (and not a minimum or an inflection point).

25. The maximum-likelihood solution is obtained by maximizing $[g(\mathbf{s},\theta)]^{1/n}$. From Table 3.1 in the text, we have for a Maxwell distribution

$$[g(s,\theta)]^{1/n} = \theta^{3/2}e^{-\theta s}$$

where $s = \frac{1}{n}\sum_{k=1}^{n} x_k^2$. The derivative is

$$\nabla_\theta[g(s,\theta)]^{1/n} = \frac{3}{2}\theta^{1/2}e^{-\theta s} - s\theta^{3/2}e^{-\theta s}.$$

We set this to zero to obtain

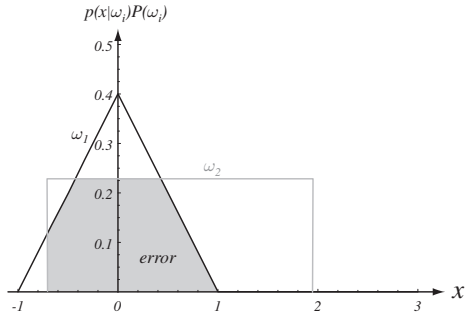
$$\frac{3}{2}\theta^{1/2}e^{-\theta s} = s\theta^{3/2}e^{-\theta s},$$

and thus the maximum-likelihood solution is

$$\hat{\theta} = \frac{3/2}{s} = \frac{3}{2}\left(\frac{1}{n}\sum_{k=1}^{n} x_k^2\right)^{-1}.$$

We next evaluate the second derivative at this value of $\hat{\theta}$ to see if the solution represents a maximum, a minimum, or possibly an inflection point:

$$\nabla_\theta^2[g(s,\theta)]^{1/n}\Big|_{\theta=\hat{\theta}} = \frac{3}{2}\frac{1}{2}\theta^{1/2}e^{-\theta s} - \frac{3}{2}\theta^{1/2}se^{-\theta s} - \frac{3}{2}\theta^{1/2}se^{-\theta s} + s^2\theta^{3/2}e^{-\theta s}\Big|_{\theta=\hat{\theta}}$$



$$
\begin{aligned}
&= \frac{3}{4}\hat{\theta}^{-1/2}e^{-\hat{\theta}s} - 3s\hat{\theta}^{1/2}e^{-\hat{\theta}s} + s^2\hat{\theta}^{3/2}e^{-\hat{\theta}s} \\
&= e^{-3/2}\left(\frac{3}{4} - 3\frac{3}{2} + \frac{9}{4}\right)\hat{\theta}^{-1/2} = -\frac{3}{2}\hat{\theta}^{-1/2}e^{-3/2} < 0.
\end{aligned}
$$

Thus $\hat{\theta}$ indeed gives a maximum (and not a minimum or an inflection point).

26. We fine the maximum-likelihood solution by maximizing $[g(\mathbf{s},\theta)]^{1/n}$. In this case, we have

$$[g(\mathbf{s},\theta)]^{1/n} = \prod_{i=1}^{d}\theta_i^{s_i}$$

and

$$\mathbf{s} = (s_1,\ldots,s_d)^t = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$$

Our goal is to maximize $[g(\mathbf{s},\theta)]^{1/n}$ with respect to $\boldsymbol{\theta}$ over the set $0 < \theta_i < 1, i = 1,\ldots,d$ subject to the constraint $\sum_{i=1}^{d}\theta_i = 1$. We set up the objective function

$$
\begin{aligned}
l(\theta,\lambda|\mathbf{s}) &= \ln\,[g(\mathbf{s},\theta)]^{1/n} + \lambda\left(\sum_{i=1}^{d}\theta_i - 1\right) \\
&= \sum_{i=1}^{d} s_i \ln\,\theta_i + \lambda\left(\sum_{i=1}^{d}\theta_i - 1\right).
\end{aligned}
$$

The derivative is thus

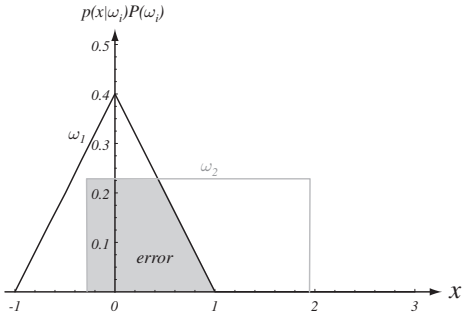
$$(\nabla_\theta l)_i = \frac{\delta l}{\delta\theta_i} = \frac{s_i}{\theta_i} + \lambda$$

for $i = 1,\ldots,d$. We set this derivative to zero and get

$$\frac{s_i}{\hat{\theta}_i} = -\lambda,$$

which yields our intermediate solution:

$$\hat{\theta}_i = -\frac{s_i}{\lambda}.$$



We impose the normalization condition, $\sum_{i=1}^{d} \hat{\theta}_i = 1$, to find λ, which leads to our final solution

$$\hat{\theta}_i = \frac{s_j}{\sum_{i=1}^{d} s_i}$$

for $j = 1, \ldots, d$.

27. Consider the notion that sufficiency is an integral concept.

(a) Using Eq. 51 in the text, we have

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{k=1}^{n} p(x|\boldsymbol{\theta}) \\ &= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_k - \mu)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[\frac{1}{2\sigma^2}\sum_{k=1}^{n}\left(x_k^2 - 2\mu x_k + \mu^2\right)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n} x_k^2 - 2\mu\sum_{k=1}^{n} x_k + \sum_{k=1}^{n}\mu^2\right)\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left(ns_2 - 2ns_1 + n\mu^2\right)\right]. \end{aligned}$$

Now let $g(\mathbf{s}, \boldsymbol{\theta})$ be a Gaussian of the form

$$g(\mathbf{s}, \boldsymbol{\theta}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}(ns_2 - 2\mu n s_1 + n\mu^2)\right]$$

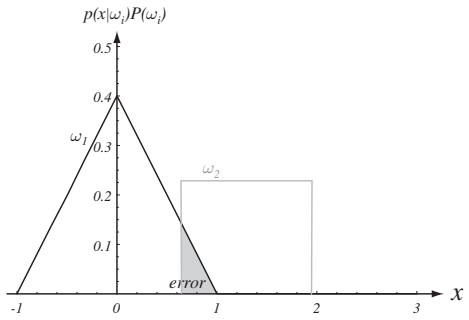
and $h(\mathcal{D}) = 1$. Then we can write

$$p(\mathcal{D}|\boldsymbol{\theta}) = g(\mathbf{s}, \boldsymbol{\theta})h(\mathcal{D}).$$

According to Theorem 3.1 in the text, the statistic $\mathbf{s}$ is indeed sufficient for $\boldsymbol{\theta}$.

(b) We apply the result from part (a) to the case of $g(s_1, \mu, \sigma^2)$, that is,

$$g(s_1, \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[\frac{1}{2\sigma^2}(-2\mu n s_1 + n\mu^2)\right]$$



and

$$h(\mathcal{D},\sigma^2)=\exp\left[\frac{1}{2\sigma^2}ns_2\right].$$

In the general case, the $g(s_1,\mu,\sigma^2)$ and $h(\mathcal{D},\sigma^2)$ are dependent on σ^2, that is, there is o way to have $p(\mathcal{D}|\mu,\sigma^2)=g(s_1,\mu)h(\mathcal{D})$. Thus according to Theorem 3.1 in the text, s_1 is not sufficient for μ. However, if σ^2 is known, then $g(s_1,\mu,\sigma^2)=g(s_1,\mu)$ and $h(\mathcal{D},\sigma^2)=h(\mathcal{D})$. Then we indeed have $p(\mathcal{D}|\mu)=g(s_1,\mu)h(\mathcal{D})$, and thus s_1 is sufficient for μ.

(c) As in part (b), we let

$$g(s_2,\mu,\sigma^2)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left[\frac{1}{2\sigma^2}(ns_2-2\mu ns_1+n\mu^2)\right]$$

and $h(\mathcal{D})=1$. In the general case, the $g(s_2,\mu,\sigma^2)$ is dependent upon μ, that is, there is no way to have $p(\mathcal{D}|\mu,\sigma^2)=g(s_2,\sigma^2)h(\mathcal{D})$. Thus according to Theorem 3.1 in the text, s_2 is not sufficient for σ^2. However, if μ is known, then $g(s_2,\mu,\sigma^2)=g(s_2,\sigma^2)$. Then we have $p(\mathcal{D}|\sigma^2)=g(s_2,\sigma^2)h(\mathcal{D})$, and thus s_2 is sufficient for σ^2.

28. We are to suppose that $\mathbf{s}$ is a statistic for which $p(\boldsymbol{\theta}|\mathbf{x},\mathcal{D})=p(\boldsymbol{\theta}|\mathbf{s})$, that is, that $\mathbf{s}$ does not depend upon the data set $\mathcal{D}$.

(a) Let $\mathbf{s}$ be a sufficient statistic for which $p(\theta|\mathbf{s},\mathcal{D})=p(\boldsymbol{\theta}|\mathbf{s})$; we assume $p(\boldsymbol{\theta}|\mathbf{s})\neq 0$. In that case, we can write Bayes' law

$$p(\mathcal{D}|\mathbf{s},\boldsymbol{\theta})=\frac{p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})}$$

as

$$\begin{aligned}p(\mathcal{D}|\mathbf{s},\boldsymbol{\theta}) &= \frac{p(\mathcal{D},\mathbf{s},\boldsymbol{\theta})}{p(\mathbf{s},\boldsymbol{\theta})}\\ &= \frac{p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D})P(\mathcal{D},\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})p(\mathbf{s})}\\ &= \frac{p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D})p(\mathcal{D}|\mathbf{s})p(\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})p(\mathbf{s})}\\ &= \frac{p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})}.\end{aligned}$$

Note that the probability density of the parameter $\boldsymbol{\theta}$ is fully specified by the sufficient statistic; the data gives no further information, and this implies

$$p(\theta|\mathbf{s},\mathcal{D}) = p(\theta|\mathbf{s}).$$

Since $p(\boldsymbol{\theta}|\mathbf{s}) \neq 0$, we can write

$$\begin{aligned} p(\mathcal{D}|\mathbf{s},\boldsymbol{\theta}) &= \frac{p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})} \\ &= \frac{p(\boldsymbol{\theta}|\mathbf{s})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})} \\ &= p(\mathcal{D}|\mathbf{s}), \end{aligned}$$

which does not involve $\boldsymbol{\theta}$. Thus, $p(\mathcal{D}|\mathbf{s},\boldsymbol{\theta})$ is indeed independent of $\boldsymbol{\theta}$.

(b) Assume the variable x comes from a uniform distribution, $p(x) \sim U(\mu, 1)$. Let $s = \frac{1}{n}\sum_{k=1}^{n} x_k$ be a statistic of a sample data set $\mathcal{D}$ of x, that is, s is the sample mean of $\mathcal{D}$. Now assume that we estimate $\mu = \mu_0 = 0$, but from a particular sample data set $\mathcal{D}_0$ we have $s_0 = 5$. Note that we estimate $p(x) \sim U(0,1)$, we have $p(\mu_0|s_0,\mathcal{D}) = p(\mu_0|s_0) = 0$, that is, it is impossible that given the data set $\mathcal{D}_0$ whose mean is $s_0 = 5$ that the distribution is $p(x) \sim U(0,1)$. Therefore, if given $\mu = \mu_0 = 0$ and $s_0 = 5$, there does not exist a data set $\mathcal{D}_0$ whose mean is $s_0 = 5$ but whose elements are distributed according to $p(x) \sim U(0,1)$. Said another way, we have $p(\mathcal{D}_0|s_0,\mu_0) = 0$ for these particular values of s_0 and μ_0. However, $p(\mathcal{D}_0|s_0)$ is not necessarily zero. It is easy to set up a data set whose mean is s_0, giving $p(\mathcal{D}_0|s_0,\mu_0) \neq p(\mathcal{D}_0|s_0)$. Therefore, $p(\theta|s) \neq 0$, as required for the proof.

29. Recall the Cauchy distribution,

$$p(x) = \frac{1}{\pi b}\cdot\frac{1}{1+\left(\frac{x-a}{b}\right)^2}.$$

(a) We integrate $p(x)$ over all x, and have

$$\int_{-\infty}^{\infty} p(x)dx = \frac{1}{\pi b}\int_{-\infty}^{\infty}\frac{1}{1+\left(\frac{x-a}{b}\right)^2}dx.$$

We substitute $y = (x-a)/b$ and $dx = b\,dy$ and find

$$\frac{1}{\pi b}\int_{-\infty}^{\infty}\frac{1}{1+\left(\frac{x-a}{b}\right)^2}dx = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{1+y^2}dy = \frac{1}{\pi}\tan^{-1}[y]\Big|_{-\infty}^{\infty} = \frac{1}{\pi}\left(\frac{\pi}{2}-\left(-\frac{\pi}{2}\right)\right) = 1.$$

We find, then, that indeed $p(x)$ is normalized.

(b) According to the definition of the mean, we have

$$\mu = \int_{-\infty}^{\infty} xp(x)dx = \frac{1}{\pi b}\frac{x}{1+\left(\frac{x-a}{b}\right)^2}dx.$$

As in part (a), we let $y = (x - a)/b$ and $dx = b\ dy$. Then the mean is

$$\begin{aligned}\mu &= \frac{1}{\pi b}\int_{-\infty}^{\infty}\frac{by+a}{1+y^2}b\ dy \\ &= \frac{1}{\pi}\left(\underbrace{b\int_{-\infty}^{\infty}\frac{y}{1+y^2}dy}_{0} + a\int_{-\infty}^{\infty}\frac{1}{1+y^2}b\ dy\right).\end{aligned}$$

Note that $f(y) = y/(1 + y^2)$ is an odd function, and thus the first integral vanishes. We substitute our result from part (a) and find

$$\mu = \frac{a}{\pi}\int_{-\infty}^{\infty}\frac{1}{1+y^2}b\ dy = a.$$

This conclusion is satisfying, since when $x = a$, the distribution has its maximum; moreover, the distribution dies symmetrically around this peak.

According to the definition of the standard deviation,

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty}x^2p(x)dx - \mu^2 \\ &= \frac{1}{\pi b}\int_{-\infty}^{\infty}x^2\frac{1}{1+\left(\frac{x-a}{b}\right)^2}dx - \mu^2.\end{aligned}$$

We substitute $y = (x - a)/b$ and $dx = bdy$, and have

$$\begin{aligned}\sigma^2 &= \frac{1}{\pi b}\int_{-\infty}^{\infty}\frac{b^2y^2+2aby+a^2}{1+y^2}b\ dy - \mu^2 \\ &= \frac{1}{\pi}\left[\int_{-\infty}^{\infty}b^2dy + 2ab\int_{-\infty}^{\infty}\frac{y}{1+y^2}dy + (a^2-b^2)\int_{-\infty}^{\infty}\frac{1}{1+y^2}dy\right] - \mu^2 \\ &= \frac{1}{\pi}\left[\infty + 2ab\cdot 0 + (a^2-b^2)\pi\right] - \mu^2 \\ &= \infty.\end{aligned}$$

This result is again as we would expect, since according to the above calculation, $\mathcal{E}[(x - \mu)^2] = \infty$.

(c) Since we know $\boldsymbol{\theta} = (\mu\ \infty)^t = (a\ \infty)^t$, for any defined $\mathbf{s} = (s_1, \ldots, s_n), s_i < \infty$, it is impossible to have $p(\mathcal{D}|\mathbf{s}, (a\ \infty)^t) = p(\mathcal{D}|\mathbf{s})$. Thus the Cauchy distribution has no sufficient statistics for the mean and standard deviation.

30. We consider the univariate case. Let μ and σ^2 denote the mean and variance of the Gaussian, Cauchy and binomial distributions and $\hat{\mu} = 1/n\sum_{i=1}^{n}x_i$ be the sample mean of the data points x_i, for $i = 1, \ldots, n$.

(a) According to Eq. 21 in the text, we have

$$\begin{aligned}
\sigma_n^2 &= \mathcal{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\right] \\
&= \frac{1}{n-1}\mathcal{E}\left[\sum_{i=1}^{n}((x_i-\mu)-(\hat{\mu}-\mu))^2\right] \\
&= \frac{1}{n-1}\mathcal{E}\left[\sum_{i=1}^{n}\left((x_i-\mu)^2-2(x_i-\mu)(\hat{\mu}-\mu)+(\hat{\mu}-\mu)^2\right)\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[\mathcal{E}[(x_i-\mu)^2]-2\mathcal{E}[(x_i-\mu)(\hat{\mu}-\mu)]+\mathcal{E}[(\hat{\mu}-\mu)^2]\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\mathcal{E}[(x_i-\mu)(\hat{\mu}-\mu)] &= \mathcal{E}\left[(x_i-\mu)\left(\frac{1}{n}\sum_{j=1}^{n}x_j-\mu\right)\right] \\
&= \mathcal{E}\left[(x_i-\mu)\left(\frac{x_i-\mu}{n}+\frac{1}{n}\sum_{k=1;\,k\neq i}^{n}x_k-\mu\right)\right] \\
&= \mathcal{E}\left[\frac{1}{n}(x_i-\mu)^2\right]+\mathcal{E}\left[\frac{1}{n}(x_i-\mu)\left(\sum_{k=1;\,k\neq i}^{n}x_k-(n-1)\mu\right)\right] \\
&= \frac{1}{n}\sigma^2+0=\sigma^2/n.
\end{aligned}$$

Similarly, we have the variance,

$$\mathcal{E}[(\hat{\mu}-\mu)^2]=\sigma^2/n,$$

and furthermore

$$\begin{aligned}
\sigma_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sigma^2-\frac{2}{n}\sigma^2+\frac{1}{n}\sigma^2\right) \\
&= \frac{n-1}{n-1}\sigma^2=\sigma^2.
\end{aligned}$$

Thus indeed the estimator in Eq. 21 in the text is unbiased.

(b) The result in part (a) applies to a Cauchy distribution.

(c) The result in part (a) applies to a Binomial distribution.

(d) From Eq. 20 in the text, we have

$$\lim_{n\to\infty}\frac{n-1}{n}\sigma^2=\sigma^2,$$

that is, asymptotically unbiased.

Section 3.7

31. We assume that a and b are positive constants and n a variable. Recall the definition on page 633 in the text that

$$O(g(x)) = \{f(x) : \text{There exist positive constants } c \text{ and } x_0 \text{ such that } 0 \le f(x) \le cg(x) \text{ for all } x \ge x_0\}.$$

(a) Is $a^{n+1} = O(a^n)$? Yes. If let $c = a$ and $x_0 = 1$, then $c \cdot a^n = a^{n+1} \ge a^{n+1}$ for $x \ge 1$.

(b) Is $a^{bn} = O(a^n)$? No. No matter what constant c we choose, we can always find a value n for which $a^{bn} > c \cdot a^n$.

(c) Is $a^{n+b} = O(a^n)$? Yes. If choose $c = a^b$ and $x_0 = 1$, then $c \cdot a^n = a^{n+b} \le a^{n+b}$ for $x > 1$.

(d) Clearly $f(n) = O(f(n))$. To prove this, we let $c = 1$ and $x_0 = 0$. Then of course $f(x) \ge f(x)$ for $x \ge 0$. (In fact, we have $f(n) = f(n)$.)

32. We seek to evaluate $f(x) = \sum_{i=0}^{n-1} a_i x^i$ at a point x where the n coefficients a_i are given.

(a) A straightforward $\mathbf{\Theta}(n^2)$ algorithm is:

Algorithm 0 (Basic polynomial evaluation)

```
1   begin initialize x, a_i for i = 1, ..., n-1
2         f ← a_0
3         i ← 0
4         for i ← i + 1
5               x_0 ← 1
6               j ← 0
7               for j ← j + 1
8                     x_j ← x_{j-1} x
9               until j = i
10              f ← f + a_i x_i
11        until i = n - 1
12  end
```

(b) A $\mathbf{\Theta}(n)$ algorithm that takes advantage of Horner's rule is:

Algorithm 0 (Horner evaluation)

```
1   begin initialize x, a_i for i = 1, ..., n-1
2         f ← a_{n-1}
3         i ← n
4         for i ← i - 1
5               f ← fx + a_{i-1}
6         until i = 1
7   end
```

33. The complexities are as follows:

(a) $O(N)$

(b) $O(N)$

(c) $O(J(I+K))$

34. We assume that the uniprocessor can perform one operation per nanosecond (10^{-9} second). There are 3600 *seconds* in an hour, $24 \times 3600 = 86,400$ *seconds* in a day, and $31,557,600$ *seconds* in a year. Given one operation per nanosecond, the total number of basic operations in the periods are: 10^9 in one second; 3.6×10^{12} in one hour; 8.64×10^{13} in one day; 3.156×10^{16} in one year. Thus, for the table below, we find n such that $f(n)$ is equal to these total number of operations. For instance, for the "1 day" entry in the $n\log_2 n$ row, we solve for n such that $n\log_2 n = 8.64 \times 10^{13}$.

$f(n)$	1 sec	1 hour	1 day	1 year
operations:	10^9	3.6×10^{12}	8.64×10^{13}	3.156×10^{16}
$\log_2 n$	$2^{10^9} \simeq 4.6\times10^{301029995}$	$2^{3.6\times10^{12}} \simeq 10^{10^{12}}$	$2^{8.64\times10^{13}} \simeq 10^{10^{13}}$	$2^{3.156\times10^{16}} \simeq 10^{10^{16}}$
$\sqrt{n}$	$(10^9)^2 = 10^{18}$	$(3.6\times10^{12})^2 \simeq 1.3\times10^{25}$	$(8.64\times10^{13})^2 \simeq 7.46\times10^{27}$	$(3.156\times10^{16})^2 \simeq 9.96\times10^{32}$
n	10^9	3.6×10^{12}	8.64×10^{13}	3.156×10^{16}
$n\log_2 n$ *	3.9×10^7	9.86×10^{10}	2.11×10^{12}	6.41×10^{14}
n^2	$\sqrt{10^9} \simeq 3.16 \times 10^4$	$\sqrt{3.6\times10^{12}} \simeq 1.9\times10^6$	$\sqrt{8.64\times10^{13}} \simeq 9.3\times10^6$	$\sqrt{3.156\times10^{16}} \simeq 1.78\times10^8$
n^3	$\sqrt[3]{10^9} = 10^3$	$\sqrt[3]{3.6\times10^{12}} \simeq 1.53\times10^4$	$\sqrt[3]{8.64\times10^{13}} \simeq 4.42\times10^4$	$\sqrt[3]{3.156\times10^{16}} \simeq 3.16\times10^5$
2^n	$\log_2 10^9 = 29.90$	$\log_2(3.6\times10^{12}) \sim 41.71$	$\log_2(8.64\times10^{13}) \sim 46.30$	$\log_2(3.156\times10^{16}) \sim 54.81$
e^n	$\ln 10^9 = 10.72$	$\ln(3.6\times10^{12}) \sim 28.91$	$\ln(8.64\times10^{13}) \sim 32.09$	$\ln(3.156\times10^{16}) \sim 38.0$
$n!$ **	13.1	16.1	17.8	19.3

* For entries in this row, we solve numerically for n such that, for instance, $n\log_2 n = 3.9 \times 10^7$, and so on.

** For entries in this row, we use Stirling's approximation — $n! \simeq (n/e)^n$ for large n — and solve numerically.

35. Recall first the sample mean and sample covariance based on n samples:

$$\begin{aligned}\mathbf{m}_n &= \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \\ \mathbf{C}_n &= \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m}_n)(\mathbf{x}_k - \mathbf{m}_n)^t.\end{aligned}$$

(a) The computational complexity for $\mathbf{m}_n$ is $O(dn)$, and for $\mathbf{C}_n$ is $O(dn^2)$.

(b) We can express the sample mean for $n+1$ observations as

$$\begin{aligned}\mathbf{m}_{n+1} = \frac{1}{n+1}\sum_{k=1}^{n+1}\mathbf{x}_k &= \frac{1}{n+1}\left[\sum_{k=1}^{n}\mathbf{x}_k + \mathbf{x}_{n+1}\right] \\ &= \frac{1}{n+1}\left[n\mathbf{m}_n + \mathbf{x}_{n+1}\right] \\ &= \mathbf{m}_n - \frac{1}{n+1}\mathbf{m}_n + \frac{1}{n+1}\mathbf{x}_{n+1} \\ &= \mathbf{m}_n + \frac{1}{n+1}(\mathbf{x}_{n+1} - \mathbf{m}_n).\end{aligned}$$

For the sample covariance, we have

$$\begin{aligned}
\mathbf{C}_{n+1} &= \frac{1}{n}\sum_{k=1}^{n+1}(\mathbf{x}_k-\mathbf{m}_{n+1})(\mathbf{x}_k-\mathbf{m}_{n+1})^t \\
&= \frac{1}{n}\left[\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}_{n+1})(\mathbf{x}_k-\mathbf{m}_{n+1})^t+(\mathbf{x}_{n+1}-\mathbf{m}_{n+1})(\mathbf{x}_{n+1}-\mathbf{m}_{n+1})^t\right] \\
&= \frac{1}{n}\Big[\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}_n)(\mathbf{x}_k-\mathbf{m}_n)^t-\frac{1}{(n+1)}(\mathbf{x}_{n+1}-\mathbf{m}_n)\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}_n)^t \\
&\quad -\frac{1}{n+1}\left(\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}_n)\right)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t+\frac{1}{(n+1)^2}\sum_{k=1}^{n}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t\Big] \\
&\quad +\frac{1}{n}\left((\mathbf{x}_{n+1}-\mathbf{m}_n)-\frac{1}{n+1}(\mathbf{x}_{n+1}-\mathbf{m}_n)\right)\left((\mathbf{x}_{n+1}-\mathbf{m}_n)-\frac{1}{n+1}(\mathbf{x}_{n+1}-\mathbf{m}_n)\right)^t \\
&= \frac{1}{n}\left[(n-1)\mathbf{C}_n+\frac{n}{(n+1)^2}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t\right] \\
&\qquad +\frac{1}{n}\left(\left(\frac{n}{n+1}\right)^2(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t\right) \\
&= \frac{n-1}{n}\mathbf{C}_n+\left(\frac{1}{(n+1)^2}+\frac{n}{(n+1)^2}\right)(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-bfm_n)^t \\
&= \frac{n-1}{n}\mathbf{C}_n+\frac{1}{n+1}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t.
\end{aligned}$$

(c) To compute $\mathbf{m}_n$ from scratch, requires $O(dn)$ operations. To update $\mathbf{m}_n$ according to

$$\mathbf{m}_{n+1}=\mathbf{m}_n+\frac{1}{n+1}(\mathbf{x}_{n+1}-\mathbf{m}_n)$$

requires $O(d)$ operations, one for each component.

To update $\mathbf{C}_n$ according to

$$\mathbf{C}_{n+1}=\frac{n-1}{n}\mathbf{C}_n+\frac{1}{(n+1)}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t$$

given $\mathbf{C}_n, \mathbf{m}_n$ and $\mathbf{x}_n$ requires $O(d)$ operations for computing $\mathbf{x}_{n+1}-\mathbf{m}_n$; and $O(d^2)$ operations for computing $(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t$. Thus, $\mathbf{C}_{n+1}$ is computed in $O(d^2)$ operations, given $\mathbf{C}_n, \mathbf{m}_n$ and $\mathbf{x}_n$. If we must compute $\mathbf{C}_{n+1}$ from scratch, then we need $O(dn^2)$ operations.

(d) The recursive method is on-line, and the classifier can be used as the data is coming in. One possible advantage of the non-recursive method is that it might avoid compounding roundoff errors in the successive additions of new information from samples, and hence be more accurate.

36. We seek a recursive method for computing $\mathbf{C}_{n+1}^{-1}$.

(a) We first seek to show

$$(\mathbf{A}+\mathbf{x}\mathbf{x}^t)^{-1}=\mathbf{A}^{-1}\frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}}{1+\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}.$$

We prove this most simply by verifying the inverse as:

$$\begin{aligned}
&(\mathbf{A}+\mathbf{x}\mathbf{x}^t)\left(\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}}{1+\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}\right)\\
&\quad=\mathbf{A}\mathbf{A}^{-1}-\frac{\mathbf{A}\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}}{1+\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}+\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}-\frac{\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}}{1+\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}\\
&\quad=\mathbf{I}-\frac{\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}}{1+\underbrace{\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}_{a\ scalar}}+\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}-\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}\frac{\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}}{1+\underbrace{\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}_{a\ scalar}}\\
&\quad=\mathbf{I}+\mathbf{x}^t\mathbf{x}\mathbf{A}^{-1}\left(1-\frac{1}{1+\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}-\frac{\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}{1+\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}\right)\\
&\quad=\mathbf{I}+\mathbf{x}^t\mathbf{x}\mathbf{A}^{-1}\cdot 0\\
&\quad=\mathbf{I}.
\end{aligned}$$

Note the left-hand-side and the right-hand-side of this equation and see

$$(\mathbf{A}+\mathbf{x}^t\mathbf{x})^{-1} \quad = \quad \mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}}{1+\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}.$$

(b) We apply the result in part (a) to the result discussed in Problem 35 and see that

$$\begin{aligned}
\mathbf{C}_{n+1} &= \frac{n-1}{n}\mathbf{C}_n+\frac{1}{n+1}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t\\
&= \frac{n-1}{n}\left[\mathbf{C}_n+\frac{n}{n^2-1}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t\right]\\
&= \frac{n-1}{n}(\mathbf{A}+\mathbf{x}\mathbf{x}^t).
\end{aligned}$$

We set $\mathbf{A}=\mathbf{C}_n$ and $\mathbf{x}=\sqrt{\frac{n}{n^2-1}}(\mathbf{x}_{n+1}-\mathbf{m}_n)$ and substitute to get

$$\begin{aligned}
\mathbf{C}_{n+1}^{-1} &= \left[\frac{n-1}{n}(\mathbf{A}+\mathbf{x}\mathbf{x}^t)\right]^{-1}=\frac{n}{n-1}(\mathbf{A}+\mathbf{x}\mathbf{x}^t)^{-1}\\
&= \frac{n}{n+1}\left[\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^t\mathbf{A}^{-1}}{1+\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}\right]\\
&= \frac{n}{n+1}\left[\mathbf{C}_n^{-1}-\frac{\mathbf{C}_n^{-1}\sqrt{\frac{n}{n^2-1}}(\mathbf{x}_{n+1}-\mathbf{m}_n)\sqrt{\frac{n}{n^2-1}}(\mathbf{x}_{n+1}-\mathbf{m}_n)^t\mathbf{C}_n^{-1}}{1+\sqrt{\frac{n}{n^2-1}}(\mathbf{x}_{n+1}-\mathbf{m}_n)^t\mathbf{C}_n^{-1}\sqrt{\frac{n}{n^2-1}}(\mathbf{x}_{n+1}-\mathbf{m}_n)}\right]\\
&= \frac{n}{n-1}\left[\mathbf{C}_n^{-1}-\frac{\mathbf{C}_n^{-1}(\mathbf{x}_{n+1}-\mathbf{m}_n)(\mathbf{x}_{n+1}-\mathbf{m}_n)^t\mathbf{C}_n^{-1}}{\frac{n^2-1}{n}+(\mathbf{x}_{n+1}-\mathbf{m}_n)^t\mathbf{C}_n^{-1}(\mathbf{x}_{n+1}-\mathbf{m}_n)}\right],
\end{aligned}$$

where we used our result from part (a).

(c) We determine the computational complexity as follows. Consider the formula for $\mathbf{C}_{n+1}^{-1}$:

$$\mathbf{C}_{n+1}^{-1}=\frac{n}{n-1}\left[\mathbf{C}_n^{-1}-\frac{\mathbf{u}\mathbf{u}^t}{\frac{n^2-1}{n}+\mathbf{u}^t(\mathbf{x}_{n+1}-\mathbf{m}_n)}\right],$$

where $\mathbf{u} = \mathbf{C}_n^{-1}(\mathbf{x}_{n+1} - \mathbf{m}_n)$ is of $O(d^2)$ complexity, given that $\mathbf{C}_n^{-1}, \mathbf{x}_{n+1}$ and $\mathbf{m}_n$ are known. Hence, clearly $\mathbf{C}_n^{-1}$ can be computed from $\mathbf{C}_{n-1}^{-1}$ in $O(d^2)$ operations, as $\mathbf{u}\mathbf{u}^t, \mathbf{u}^t(\mathbf{x}_{n+1} - \mathbf{m}_n)$ is computed in $O(d^2)$ operations. The complexity associated with determining $\mathbf{C}_n^{-1}$ is $O(nd^2)$.

37. We assume the symmetric non-negative covariance matrix is of otherwise general form:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}.$$

To employ shrinkage of an assumed common covariance toward the identity matrix, then Eq. 77 requires

$$\mathbf{\Sigma}(\beta) = (1-\beta)\mathbf{\Sigma} + \beta\mathbf{I} = \mathbf{I},$$

and this implies $(1-\beta)\sigma_{ii} + \beta \cdot 1 = 1$, and thus

$$\sigma_{ii} = \frac{1-\beta}{1-\beta} = 1$$

for all $0 < \beta < 1$. Therefore, we must first normalize the data to have unit variance.

Section 3.8

38. Note that in this problem our densities need not be normal.

(a) Here we have the criterion function

$$J_1(\mathbf{w}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}.$$

We make use of the following facts for $i = 1, 2$:

$$\begin{aligned} y &= \mathbf{w}^t\mathbf{x} \\ \mu_i &= \frac{1}{n_i}\sum_{y\in\mathcal{Y}_i} y = \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i} \mathbf{w}^t\mathbf{x} = \mathbf{w}^t\boldsymbol{\mu}_i \\ \sigma_i^2 &= \sum_{y\in\mathcal{Y}_i}(y-\mu_i)^2 = \mathbf{w}^t\left[\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^t\right]\mathbf{w} \\ \mathbf{\Sigma}_i &= \sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^t. \end{aligned}$$

We define the within- and between-scatter matrices to be

$$\begin{aligned} \mathbf{S}_W &= \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 \\ \mathbf{S}_B &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t. \end{aligned}$$

Then we can write

$$\begin{aligned} \sigma_1^2 + \sigma_2^2 &= \mathbf{w}^t\mathbf{S}_W\mathbf{w} \\ (\mu_1 - \mu_2)^2 &= \mathbf{w}^t\mathbf{S}_B\mathbf{w}. \end{aligned}$$

The criterion function can be written as

$$J_1(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}.$$

For the same reason Eq. 103 in the text is maximized, we have that $J_1(\mathbf{w})$ is maximized at $\mathbf{w}\mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. In sum, that $J_1(\mathbf{w})$ is maximized at $\mathbf{w} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

(b) Consider the criterion function

$$J_2(\mathbf{w}) = \frac{(\mu_1 - \mu_2)^2}{P(\omega_1)\sigma_1^2 + P(\omega_2)\sigma_2^2}.$$

Except for letting $\mathbf{S}_W = P(\omega_1)\boldsymbol{\Sigma}_1 + P(\omega_2)\boldsymbol{\Sigma}_2$, we retain all the notations in part (a). Then we write the criterion function as a Rayleigh quotient

$$J_2(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}.$$

For the same reason Eq. 103 is maximized, we have that $J_2(\mathbf{w})$ is maximized at

$$\mathbf{w} = \left(P(\omega_1)\boldsymbol{\Sigma}_1 + P(\omega_2)\boldsymbol{\Sigma}_2\right)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

(c) Equation 96 of the text is more closely related to the criterion function in part (a) above. In Eq. 96 in the text, we let $\tilde{m}_i = \mu_i$, and $\tilde{s}_i^2 = \sigma_i^2$ and the statistical meanings are unchanged. Then we see the exact correspondence between $J(\mathbf{w})$ and $J_1(\mathbf{w})$.

39. The expression for the criterion function

$$J_1 = \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}(y_i - y_j)^2$$

clearly measures the total within-group scatter.

(a) We can rewrite J_1 by expanding

$$\begin{aligned}
J_1 &= \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}[(y_i - m_1) - (y_j - m_2) + (m_1 - m_2)]^2 \\
&= \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}\big[(y_i - m_1)^2 + (y_j - m_2)^2 + (m_1 - m_2)^2 \\
&\qquad +2(y_i - m_1)(y_j - m_2) + 2(y_i - m_1)(m_1 - m_2) + 2(y_j - m_2)(m_1 - m_2)\big] \\
&= \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}(y_i - m_1)^2 + \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}(y_j - m_2)^2 + (m_1 - m_2)^2 \\
&\qquad +\frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}2(y_i - m_1)(y_j - m_2) + \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}2(y_i - m_1)(m_1 - m_2) \\
&\quad +\frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}2(y_j - m_2)(m_1 - m_2) \\
&= \frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2 + (m_1 - m_2)^2,
\end{aligned}$$

where

$$\begin{aligned} m_1 &= \sum_{y_i \in \mathcal{Y}_1} y_i \\ m_2 &= \sum_{y_j \in \mathcal{Y}_2} y_j \\ s_1^2 &= \sum_{y_i \in \mathcal{Y}_1} (y_i - m_1)^2 \\ s_2^2 &= \sum_{y_j \in \mathcal{Y}_2} (y_j - m_2)^2. \end{aligned}$$

(b) The prior probability of class one is $P(\omega_1) = 1/n_1$, and likewise the prior probability of class two is $P(\omega_2) = 1/n_2$. Thus the total within-class scatter is

$$J_2 = P(\omega_1)s_1^2 + P(\omega_2)s_2^2 = \frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2.$$

(c) We write the criterion function as

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2 + \frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2}{\frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2} = \frac{(m_1 - m_2)^2}{\frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2} + 1,$$

then $J(\mathbf{w}) = J_1/J_2$. Thus optimizing J_1 subject to the constraint $J_2 = 1$ is equivalent to optimizing $J(\mathbf{w})$, except that the optimal solution of the latter is not unique. We let $y = \mathbf{w}^t\mathbf{x}$, and this implies

$$\begin{aligned} (m_1 - m_2)^2 &= \mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} \\ J_2 &= \frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2 = \mathbf{w}^t\left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]\mathbf{w} \end{aligned}$$

where the mean and scatter matrices are

$$\begin{aligned} \mathbf{m}_i &= \frac{1}{n_i}\sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x} \\ \mathbf{S}_i &= \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t, \end{aligned}$$

respectively. Thus the criterion function is

$$J(\mathbf{w}) = \frac{\mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w}}{\mathbf{w}^t\left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]\mathbf{w}} + 1.$$

For the same reason Eq. 103 in the text is optimized, we have that $J(\mathbf{w})$ is maximized at

$$\mathbf{w} = \lambda\left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1}(\mathbf{m}_1 - \mathbf{m}_2).$$

We need to guarantee $J_2 = 1$, and this requires

$$\mathbf{w}^t\left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]\mathbf{w} = 1,$$

which in turn requires

$$\lambda^2(\mathbf{m}_1-\mathbf{m}_2)^t\left[\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right]^{-1}(\mathbf{m}_1-\mathbf{m}_2)=1.$$

This equation implies that the parameter λ has the value

$$\lambda=\left[(\mathbf{m}_1-\mathbf{m}_2)^t\left[\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right]^{-1}(\mathbf{m}_1-\mathbf{m}_2)\right]^{1/2}.$$

40. In the below, $\mathbf{W}$ is a d-by-n matrix whose columns correspond to n distinct eigenvectors.

(a) Suppose that the set $\{\mathbf{e}_i\}$ are normalized eigenvectors, then we have

$$\begin{aligned}\mathbf{e}_i^t\mathbf{S}_B\mathbf{e}_j &= \lambda_i\delta_{ij}\\ \mathbf{e}_i^t\mathbf{S}_W\mathbf{e}_j &= \lambda_i\delta_{ij},\end{aligned}$$

where δ_{ij} is the Kronecker symbol. We denote the matrix consisting of eigenvectors as $\mathbf{W}=[\mathbf{e}_1\ \ \mathbf{e}_2\ \ \ldots\mathbf{e}_n]$. Then we can write the within scatter matrix in the new representation as

$$\begin{aligned}\tilde{\mathbf{S}}_W &= \mathbf{W}^t\mathbf{S}_W\mathbf{W}=\begin{pmatrix}\mathbf{e}_1^t\\ \mathbf{e}_2^t\\ \vdots\\ \mathbf{e}_n^t\end{pmatrix}\mathbf{S}_W(\mathbf{e}_1\ \ \mathbf{e}_2\cdots\mathbf{e}_n)\\ &= \begin{pmatrix}\mathbf{e}_1^t\mathbf{S}_W\mathbf{e}_1 & \cdots & \mathbf{e}_1^t\mathbf{S}_W\mathbf{e}_n\\ \vdots & \ddots & \vdots\\ \mathbf{e}_n^t\mathbf{S}_W\mathbf{e}_1 & \cdots & \mathbf{e}_n^t\mathbf{S}_W\mathbf{e}_n\end{pmatrix}=\mathbf{I}.\end{aligned}$$

Likewise we can write the between scatter matrix in the new representation as

$$\begin{aligned}\tilde{\mathbf{S}}_B &= \mathbf{W}^t\mathbf{S}_B\mathbf{W}=\begin{pmatrix}\mathbf{e}_1^t\\ \mathbf{e}_2^t\\ \vdots\\ \mathbf{e}_n^t\end{pmatrix}\mathbf{S}_B(\mathbf{e}_1\ \ \mathbf{e}_2\cdots\mathbf{e}_n)\\ &= \begin{pmatrix}\mathbf{e}_1^t\mathbf{S}_B\mathbf{e}_1 & \cdots & \mathbf{e}_1^t\mathbf{S}_B\mathbf{e}_n\\ \vdots & \ddots & \vdots\\ \mathbf{e}_n^t\mathbf{S}_B\mathbf{e}_1 & \cdots & \mathbf{e}_n^t\mathbf{S}_B\mathbf{e}_n\end{pmatrix}=\begin{pmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_n\end{pmatrix}.\end{aligned}$$

Thus $\tilde{\mathbf{S}}_W$ is an n-by-n matrix and $\tilde{\mathbf{S}}_B$ is a diagonal matrix whose elements are the corresponding eigenvalues.

(b) Here we have the determinant of the transformed between-scatter matrix is just the product of the eigenvalues, that is,

$$|\tilde{\mathbf{S}}_B|=\lambda_1\lambda_2\cdots\lambda_n,$$

and $|\tilde{\mathbf{S}}_W|=1$. Then the criterion function is simply $J=\lambda_1\lambda_2\cdots\lambda_n$.

(c) We make the following definitions:

$$\begin{aligned}\tilde{\mathbf{W}}^t &= \mathbf{QDW}^t \\ \widetilde{\tilde{\mathbf{S}}}_W &= \tilde{\mathbf{W}}^t\mathbf{S}_W\tilde{\mathbf{W}} = \mathbf{QDW}^t\mathbf{S}_W\mathbf{WDQ}^t.\end{aligned}$$

Then we have $\left|\widetilde{\tilde{\mathbf{S}}}_W\right| = |\mathbf{D}|^2$ and

$$\widetilde{\tilde{\mathbf{S}}}_B = \tilde{\mathbf{W}}^t\mathbf{S}_B\tilde{\mathbf{W}} = \mathbf{QDW}^t\mathbf{S}_B\mathbf{WDQ}^t = \mathbf{QD}\tilde{\mathbf{S}}_B\mathbf{DQ}^t,$$

then $\left|\widetilde{\tilde{\mathbf{S}}}_B\right| = |\mathbf{D}|^2\lambda_1\lambda_2\cdots\lambda_n$. This implies that the criterion function obeys

$$J = \frac{\left|\widetilde{\tilde{\mathbf{S}}}_B\right|}{\left|\widetilde{\tilde{\mathbf{S}}}_W\right|},$$

and thus J is invariant to this transformation.

41. Our two Gaussian distributions are $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for $i = 1, 2$. We denote the samples after projection as $\tilde{\mathcal{D}}_i$ and the distributions

$$p(y|\tilde{\boldsymbol{\theta}}_i) = \frac{1}{\sqrt{2\pi}\tilde{s}}\exp[-(y-\tilde{\mu})^2/(2\tilde{s}^2)],$$

and $\tilde{\boldsymbol{\theta}}_i = \binom{\tilde{\mu}_i}{\tilde{s}}$ for $i = 1, 2$. The log-likelihood ratio is

$$\begin{aligned}
r &= \frac{\ln p(\tilde{\mathcal{D}}|\tilde{\boldsymbol{\theta}}_1)}{\ln p(\tilde{\mathcal{D}}|\tilde{\boldsymbol{\theta}}_2)} = \frac{\ln\left[\prod_{k=1}^{n} p(y_k|\tilde{\boldsymbol{\theta}}_1)\right]}{\ln\left[\prod_{k=1}^{n} p(y_k|\tilde{\boldsymbol{\theta}}_2)\right]} \\
&= \frac{\sum_{k=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\tilde{s}}\exp\left[\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2}\right]\right]}{\sum_{k=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\tilde{s}}\exp\left[\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2}\right]\right]} = \frac{\sum_{k=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\tilde{s}}\right] + \sum_{k=1}^{n}\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2}}{\sum_{k=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\tilde{s}}\right] + \sum_{k=1}^{n}\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2}} \\
&= \frac{c_1 + \sum_{y_k\in\mathcal{D}_1}\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2} + \sum_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2}}{c_1 + \sum_{y_k\in\mathcal{D}_1}\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2} + \sum_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2}} \\
&= \frac{c_1 + \frac{1}{2} + \sum_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2}}{c_1 + \frac{1}{2} + \sum_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2}} = \frac{c_1 + \frac{1}{2} + \sum_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_2)+(\tilde{\mu}_2-\tilde{\mu}_1))^2}{2\tilde{s}^2}}{c_1 + \frac{1}{2} + \sum_{y_k\in\mathcal{D}_1}\frac{(y_k-\tilde{\mu}_2)+(\tilde{\mu}_2-\tilde{\mu}_1))^2}{2\tilde{s}^2}} \\
&= \frac{c_1 + \frac{1}{2} + \frac{1}{2\tilde{s}^2}\sum_{y_k\in\tilde{\mathcal{D}}_2}\left((y_k-\tilde{\mu}_2)^2 + (\tilde{\mu}_2-\tilde{\mu}_1)^2 + 2(y_k-\tilde{\mu}_2)(\tilde{\mu}_2-\tilde{\mu}_1)\right)}{c_1 + \frac{1}{2} + \frac{1}{2\tilde{s}^2}\sum_{y_k\in\tilde{\mathcal{D}}_1}((y_k-\tilde{\mu}_1)^2 + (\tilde{\mu}_1-\tilde{\mu}_2)^2 + 2(y_k-\tilde{\mu}_1)(\tilde{\mu}_1-\tilde{\mu}_2))} \\
&= \frac{c_1 + 1 + \frac{1}{2\tilde{s}^2}n_2(\tilde{\mu}_2-\tilde{\mu}_1)^2}{c_1 + 1 + \frac{1}{2\tilde{s}^2}n_1(\tilde{\mu}_1-\tilde{\mu}_2)^2} = \frac{c + n_2J(\mathbf{w})}{c + n_1J(\mathbf{w})}.
\end{aligned}$$

Thus we can write the criterion function as

$$J(\mathbf{w}) = \frac{rc - c}{n_2 - rn_1}.$$

This implies that the Fisher linear discriminant can be derived from the negative of the log-likelihood ratio.

42. Consider the criterion function $J(\mathbf{w})$ required for the Fisher linear discriminant.

(a) We are given Eqs. 96, 97, and 98 in the text:

$$\begin{aligned} J_1(\mathbf{w}) &= \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \quad (96) \\ \mathbf{S}_i &= \sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \quad (97) \\ \mathbf{S}_W &= \mathbf{S}_1 + \mathbf{S}_2 \quad (98) \end{aligned}$$

where $y = \mathbf{w}^t\mathbf{x}$, $\tilde{m}_i = 1/n_i \sum_{y\in\mathcal{Y}_i} y = \mathbf{w}^t\mathbf{m}_i$. From these we can write Eq. 99 in the text, that is,

$$\begin{aligned} \tilde{s}_i^2 &= \sum_{y\in\mathcal{Y}_i} (y - \tilde{m}_i)^2 \\ &= \sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{w}^t\mathbf{x} - \mathbf{w}^t\mathbf{m}_i)^2 \\ &= \sum_{\mathbf{x}\in\mathcal{D}} \mathbf{w}^t(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t\mathbf{w} \\ &= \mathbf{w}^t\mathbf{S}_i\mathbf{w}. \end{aligned}$$

Therefore, the sum of the scatter matrixes can be written as

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t\mathbf{S}_W\mathbf{w} \quad (100)$$

$$\begin{aligned} (\tilde{m}_1 - \tilde{m}_2)^2 &= (\mathbf{w}^t\mathbf{m}_1 - \mathbf{w}^t\mathbf{m}_2)^2 \quad (101) \\ &= \mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} \\ &= \mathbf{w}^t\mathbf{S}_B\mathbf{w}, \end{aligned}$$

where $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$, as given by Eq. 102 in the text. Putting these together we get Eq. 103 in the text,

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}. \quad (103)$$

(b) Part (a) gave us Eq. 103. It is easy to see that the $\mathbf{w}$ that optimizes Eq. 103 is not unique. Here we optimize $J_1(\mathbf{w}) = \mathbf{w}^t\mathbf{S}_B\mathbf{w}$ subject to the constraint that $J_2(\mathbf{w}) = \mathbf{w}^t\mathbf{S}_W\mathbf{w} = 1$. We use the method of Lagrange undetermined multipliers and form the functional

$$g(\mathbf{w}, \lambda) = J_1(\mathbf{w}) - \lambda(J_2(\mathbf{w}) - 1).$$

We set its derivative to zero, that is,

$$\begin{aligned} \frac{\partial g(\mathbf{w}, \lambda)}{\partial w_i} &= \left(\mathbf{u}_i^t\mathbf{S}_B\mathbf{w} + \mathbf{w}^t\mathbf{S}_B\mathbf{u}_i\right) - \lambda\left(\mathbf{u}_i^t\mathbf{S}_W\mathbf{w} + \mathbf{w}^t\mathbf{S}_w\mathbf{u}_i\right) \\ &= 2\mathbf{u}_i^t(\mathbf{S}_B\mathbf{w} - \lambda\mathbf{S}_W\mathbf{w}) = 0, \end{aligned}$$

where $\mathbf{u}_i = (0 \;\; 0 \;\; \cdots \;\; 1 \;\; \cdots \;\; 0 \;\; 0)^t$ is the n-dimensional unit vector in the ith direction. This equation implies

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}.$$

(c) We assume that $J(\mathbf{w})$ reaches the extremum at $\mathbf{w}$ and that $\Delta\mathbf{w}$ is a small deviation. Then we can expand the criterion function around $\mathbf{w}$ as

$$J(\mathbf{w}+\Delta\mathbf{w}) = \frac{(\mathbf{w}+\Delta\mathbf{w})^t\mathbf{S}_B(\mathbf{w}+\Delta\mathbf{w})}{(\mathbf{w}+\Delta\mathbf{w})^t\mathbf{S}_W(\mathbf{w}+\Delta\mathbf{w})}.$$

From $J(\mathbf{w}) = J(\mathbf{w}+\Delta\mathbf{w}) = \lambda$, to first order we have

$$\lambda = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}} = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}+2\Delta\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}+2\Delta\mathbf{w}^t\mathbf{S}_W\mathbf{w}},$$

that is,

$$\lambda = \frac{\Delta\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\Delta\mathbf{w}^t\mathbf{S}_W\mathbf{w}}.$$

Therefore the following equation holds:

$$\Delta\mathbf{w}^t(\mathbf{S}_B\mathbf{w}-\lambda\mathbf{S}_W\mathbf{w}) = 0.$$

Because $\Delta\mathbf{w}$ is arbitrary, we conclude that $\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}$.

43. Here we have the between-scatter matrix is

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t,$$

where the group and full means are

$$\begin{aligned}\mathbf{m}_i &= \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i}\mathbf{x}\\ \mathbf{m} &= \sum_{\mathbf{x}\in\mathcal{D}}\mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i\mathbf{m}_i,\end{aligned}$$

and for this case of equal covariances,

$$\mathbf{S}_W = \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t = \sum_{i=1}^{c}\boldsymbol{\Sigma} = c\boldsymbol{\Sigma}.$$

In order to maximize

$$J(\mathbf{W}) = \frac{|\mathbf{W}^t\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^t\mathbf{S}_W\mathbf{W}|},$$

the columns of an optimal $\mathbf{W}$ must be generalized eigenvectors that correspond to the largest eigenvalues in

$$\mathbf{S}_B\mathbf{w}_i = \lambda_i\mathbf{S}_W\mathbf{w}_i = c\lambda_i\boldsymbol{\Sigma}\mathbf{w}_i,$$

with $\mathbf{W} = [\mathbf{w}_1 \;\; \mathbf{w}_2 \;\; \cdots \;\; \mathbf{w}_{c-1}]$ in terms of $\boldsymbol{\Sigma}$ and d-dimensional mean vectors.

Section 3.9

44. Consider the convergence of the expectation-maximization algorithm, where as usual $\mathcal{D}_b$ and $\mathcal{D}_g$ are the "bad"and the "good" data, as described in the text.

(a) The expected value of $\boldsymbol{\theta}$ as determined in the primed coordinates is

$$\begin{aligned}
\mathcal{E}'(\boldsymbol{\theta};\ \mathcal{D}_g) &= \int l(\boldsymbol{\theta};\ \mathcal{D}_g)p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}')d\mathcal{D}_b \\
&= \int \left(\ln p(\mathcal{D}_g,\mathcal{D}_g;\ \boldsymbol{\theta}) - \ln p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta})\right) p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}')d\mathcal{D}_b \\
&= \int \left(\ln p(\mathcal{D}_g,\mathcal{D}_b;\ \boldsymbol{\theta})p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}')\right) d\mathcal{D}_b - \int \ln p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta})p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}')d\mathcal{D}_b \\
&= \mathcal{E}_{\mathcal{D}_b}[\ln p(\mathcal{D}_g,\mathcal{D}_b;\ \boldsymbol{\theta})|\mathcal{D}_g;\ \boldsymbol{\theta}'] - \mathcal{E}'_{\mathcal{D}_b}[\ln p(\mathcal{D}_b : \mathcal{D}_g;\ \boldsymbol{\theta})] \\
&= Q(\boldsymbol{\theta};\ \boldsymbol{\theta}') - \mathcal{E}'_{\mathcal{D}_b}[\ln p(\mathcal{D}_b : \mathcal{D}_g;\ \boldsymbol{\theta})].
\end{aligned}$$

(b) If we define $\phi(\mathcal{D}_b) = p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta})/p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}')$, then

$$\begin{aligned}
\mathcal{E}'[\phi(\mathcal{D}_b)] - 1 &= \int \frac{p(\mathcal{D}_b)|\mathcal{D}_g;\ \boldsymbol{\theta})}{p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}'}p(\mathcal{D}_b)|\mathcal{D}_g;\ \boldsymbol{\theta})d\mathcal{D}_b - 1 \\
&= \int p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta})d\mathcal{D}_b - 1 \\
&= 1 - 1 = 0.
\end{aligned}$$

According to Jensen's inequality for this case — $\mathcal{E}'[\ln(\cdot)] \leq \mathcal{E}'[\cdot] - 1$. Thus we have in our particular case $\mathcal{E}'[\ln\phi(\mathcal{D}_b)] \leq \mathcal{E}'[\phi(\mathcal{D}_b)] - 1 = 0$.

(c) The expected value of the log-likelihood at step $t+1$ and at step t are

$$\begin{aligned}
\mathcal{E}[l(\boldsymbol{\theta}^{t+1};\ \mathcal{D}_g)] &= Q(\boldsymbol{\theta}^{t+1};\ \boldsymbol{\theta}^t) - \mathcal{E}^t_{\mathcal{D}_b}[\ln p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}^{t+1})] \\
\mathcal{E}[l(\boldsymbol{\theta}^{t};\ \mathcal{D}_g)] &= Q(\boldsymbol{\theta}^{t};\ \boldsymbol{\theta}^t) - \mathcal{E}^t_{\mathcal{D}_b}[\ln p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}^{t})],
\end{aligned}$$

respectively. We take the difference and find

$$\begin{aligned}
\mathcal{E}[l(\boldsymbol{\theta}^{t+1};\ \mathcal{D}_g)] - \mathcal{E}[l(\boldsymbol{\theta}^{t};\ \mathcal{D}_g)] &= \left[Q(\boldsymbol{\theta}^{t+1};\ \boldsymbol{\theta}^t) - Q(\boldsymbol{\theta}^{t};\ \boldsymbol{\theta}^t)\right] \\
&\quad - \left[\mathcal{E}^t_{\mathcal{D}_b}[\ln p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}^{t+1})] - \mathcal{E}^t_{\mathcal{D}_b}[\ln p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}^{t})]\right] \\
&= \left[Q(\boldsymbol{\theta}^{t+1};\ \boldsymbol{\theta}^t) - Q(\boldsymbol{\theta}^{t};\ \boldsymbol{\theta}^t)\right] - \mathcal{E}^t_{\mathcal{D}_b}\left[\ln\frac{p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}^{t+1})}{p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}^{t})}\right].
\end{aligned}$$

From part (b), we have

$$\mathcal{E}^t_{\mathcal{D}_b}\left[\ln\frac{p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}^{t+1})}{p(\mathcal{D}_b|\mathcal{D}_g;\ \boldsymbol{\theta}^{t}}\right] = \mathcal{E}^t_{\mathcal{D}_b}[\ln\phi(\mathcal{D}_b)] < 0,$$

and thus $-\mathcal{E}^t_{\mathcal{D}_b}[\ln\phi(\mathcal{D}_b)] > 0$. Given the fact that $Q(\boldsymbol{\theta}^{t+1};\ \boldsymbol{\theta}^t) > Q(\boldsymbol{\theta}^{t};\ \boldsymbol{\theta}^t)$ and that $\mathcal{E}[l(\boldsymbol{\theta}^{t+1};\ \mathcal{D}_g)] - \mathcal{E}[l(\boldsymbol{\theta}^{t};\ \mathcal{D}_g)] > 0$, we conclude

$$\mathcal{E}[l(\boldsymbol{\theta}^{t+1};\ \mathcal{D}_g)] > \mathcal{E}[l(\boldsymbol{\theta}^{t};\ \mathcal{D}_g)],$$

and thus

$$l(\boldsymbol{\theta}^{t+1};\ \mathcal{D}_g) > l(\boldsymbol{\theta}^{t};\ \mathcal{D}_g)$$

for increasing t. In short, the log-likelihood increases through the application of the expectation-maximization algorithm.

45. We consider an iterative algorithm in which the maximum-likelihood value of missing values is calculated, then assumed to be correct for the purposes of re-estimating $\boldsymbol{\theta}$ and the process iterated.

(a) This is a generalized expectation-maximization algorithm, but not necessarily a full expectation-maximization algorithm because there is no guarantee that the *optimal* likelihood value is achieved on each step.

(b) We have $Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i) = \mathcal{E}_{\mathcal{D}_b}[\ln p(\mathcal{D}_b;\ \boldsymbol{\theta})|\mathcal{D}_g;\ \boldsymbol{\theta}^i)]$.

46. Recall first a special case Jensen's inequality, that is

$$\mathcal{E}[\ln x] < \ln[\mathcal{E}[x]].$$

Our data set is $\mathcal{D} = \{\binom{2}{3}, \binom{3}{1}, \binom{5}{4}, \binom{4}{*}, \binom{8}{6}\}$ sampled from a two-dimensional uniform distribution bounded by $x_{l1} \le x_1 \le x_{u1}$ and $x_{l2} \le x_2 \le x_{l2}$.

(a) According to the definition of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^0)$ of Eq. 129 in the text, we have

$$\begin{aligned}
Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^0) &= \mathcal{E}_{x_{42}x_{51}}[\ln p(\mathbf{x}_g, \mathbf{x}_b;\ \boldsymbol{\theta}|\boldsymbol{\theta}^0;\ \mathcal{D}_g)] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\left(\sum_{k=1}^{3}\ln p(\mathbf{x}_k|\boldsymbol{\theta})\right) + \ln p(\mathbf{x}_4|\boldsymbol{\theta}) + \ln p(\mathbf{x}_5|\boldsymbol{\theta})\right] \\
&\qquad \times p(x_{42}|\boldsymbol{\theta}^0;\ x_{41}=4)p(x_{51}|\boldsymbol{\theta}^0;\ x_{52}=6)dx_{42}dx_{51} \\
&= \sum_{k=1}^{3}\ln p(\mathbf{x}_k|\boldsymbol{\theta}) + \underbrace{\int_{-\infty}^{\infty}\ln p(\mathbf{x}_4|\boldsymbol{\theta})p(x_{42}|\boldsymbol{\theta}^0;\ x_{41}=4)dx_{42}}_{\equiv K} \\
&\quad + \underbrace{\int_{-\infty}^{\infty}p(\mathbf{x}_5|\boldsymbol{\theta})p(x_{51}|\boldsymbol{\theta}^0;\ x_{52}=6)dx_{51}}_{\equiv L}.
\end{aligned}$$

We now calculate the second term, K:

$$\begin{aligned}
K &= \int_{-\infty}^{\infty}\ln p(\mathbf{x}_4|\boldsymbol{\theta})p(x_{42}|\boldsymbol{\theta}^0;\ x_{41}=4)dx_{42} \\
&= \int_{-\infty}^{\infty}\ln p\left(\binom{4}{x_{42}}\middle|\boldsymbol{\theta}\right)\frac{p\left(\binom{4}{x_{42}}\middle|\boldsymbol{\theta}^0\right)}{\int_{-\infty}^{\infty}p\left(\binom{4}{x'_{42}}\middle|\boldsymbol{\theta}^0\right)dx'_{42}}dx_{42} \\
&= \int_{-\infty}^{\infty}\ln\left(\binom{4}{x_{42}}\middle|\boldsymbol{\theta}\right)\frac{p\left(\binom{4}{x_{42}}\middle|\boldsymbol{\theta}^0\right)}{\int_{0}^{10}\frac{1}{10\times 10}dx'_{42}}dx_{42} \\
&= 10\int_{-\infty}^{\infty}\ln p\left(\binom{4}{x_{42}}\middle|\boldsymbol{\theta}\right)p\left(\binom{4}{x_{42}}\middle|\boldsymbol{\theta}^0\right)dx_{42}.
\end{aligned}$$

There are four cases we must consider when calculating K:

1. $0 \le x_{l2} < x_{u2} \le 10$ and $x_{l1} \le 4 \le x_{u1}$, which gives:

$$\begin{aligned} K &= 10 \int_{x_{l2}}^{x_{u2}} \ln p\left(\binom{4}{x_{42}} \middle| \boldsymbol{\theta}\right) \frac{1}{10 \times 10} dx_{42} \\ &= \frac{1}{10} \int_{-\infty}^{\infty} (x_{u2} - x_{l2}) \ln \frac{1}{|x_{u1} - x_{l1}| \cdot |x_{u2} - x_{l2}|}. \end{aligned}$$

2. $x_{l2} \le 0 < x_{u2} \le 10$ and $x_{l1} \le 4 \le x_{u1}$:

$$\begin{aligned} K &= 10 \int_{0}^{x_{u2}} \ln p\left(\binom{4}{x_{42}} \middle| \boldsymbol{\theta}\right) \frac{1}{10 \times 10} dx_{42} \\ &= \frac{1}{10} x_{u2} \ln \frac{1}{|x_{u1} - x_{l1}| \cdot |x_{u2} - x_{l2}|}. \end{aligned}$$

3. $0 \le x_{l2} < 10 \le x_{u2}$ and $x_{l1} \le 4 \le x_{u1}$:

$$\begin{aligned} K &= 10 \int_{x_{u1}}^{10} \ln p\left(\binom{4}{x_{42}} \middle| \boldsymbol{\theta}\right) \frac{1}{10 \times 10} dx_{42} \\ &= \frac{1}{10} (10 - x_{u1}) \ln \frac{1}{|x_{u1} - x_{l1}| \cdot |x_{u2} - x_{l2}|}. \end{aligned}$$

4. Otherwise, $K = 0$.

Similarly, there are four cases we must consider when calculating L:

1. $x_{l1} < x_{u1} \le 0 < 10$ or $0 < 10 \le x_{l1} < x_{u1}$ or $6 \le x_{l2}$ or $x_{u2} \le 6$: then $L = 0$.

2. $0 \le x_{l1} < x_{u1} \le 10$ and $x_{l2} \le 6 \le x_{u2}$, then

$$\begin{aligned} L &= 10 \int_{x_{l1}}^{x_{u1}} \ln p\left(\binom{x_{41}}{6} \middle| \boldsymbol{\theta}\right) \frac{1}{10 \times 10} dx_{41} \\ &= \frac{1}{10} (x_{u1} - x_{l1}) \ln \frac{1}{|x_{u1} - x_{l1}| \cdot |x_{u2} - x_{l2}|}. \end{aligned}$$

3. $x_{l1} \le 0 < x_{u1} \le 10$ and $x_{l2} \le 6 \le x_{u2}$, then

$$\begin{aligned} L &= 10 \int_{0}^{x_{u1}} \ln p\left(\binom{x_{41}}{6} \middle| \boldsymbol{\theta}\right) \frac{1}{10 \times 10} dx_{41} \\ &= \frac{1}{10} x_{u1} \ln \frac{1}{|x_{u1} - x_{l1}| \cdot |x_{u2} - x_{l2}|}. \end{aligned}$$

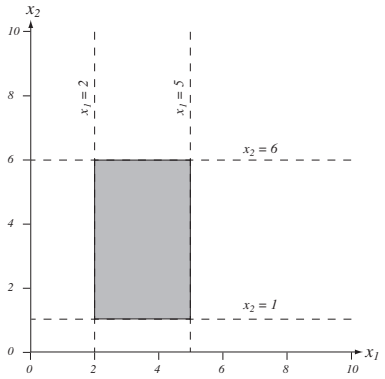
4. $0 \le x_{l1} < 10 \le x_{u1}$ and $x_{l2} \le 6 \le x_{u2}$:

$$\begin{aligned} L &= 10 \int_{x_{l1}}^{10} \ln p\left(\binom{x_{41}}{6} \middle| \boldsymbol{\theta}\right) \frac{1}{10 \times 10} dx_{41} \\ &= \frac{1}{10} (10 - x_{l1}) \ln \frac{1}{|x_{u1} - x_{l1}| \cdot |x_{u2} - x_{l2}|}. \end{aligned}$$

Therefore $Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^0) = \sum_{k=1}^{3} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) + K + L$ has different forms depending upon the different values of $\boldsymbol{\theta}$, as shown above.

(b) Here we have $\boldsymbol{\theta} = (2\ \ 1\ \ 5\ \ 6)^t$.

(c) See figure.



(d) Here we have $\boldsymbol{\theta} = (2\ \ 1\ \ 5\ \ 6)^t$.

47. Our data set is $\mathcal{D} = \left\{\binom{1}{1}, \binom{3}{3}, \binom{2}{*}\right\}$, where $*$ represents an unknown value for the x_2 component of the third data point.

(a) For the **E** step we have:

$$\begin{aligned}
Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^0) &= \mathcal{E}_{x_{32}}\left[\ln p(\mathbf{x}_g, \mathbf{x}_b;\ \boldsymbol{\theta})|\boldsymbol{\theta}^0, \mathcal{D}_g\right] \\
&= \int_{-\infty}^{\infty} \left(\ln p(\mathbf{x}_1|\boldsymbol{\theta}) + \ln p(\mathbf{x}_2|\boldsymbol{\theta}) + \ln p(\mathbf{x}_3|\boldsymbol{\theta})\right) p(x_{32}|\boldsymbol{\theta}^0, x_{31} = 2) dx_{32} \\
&= \ln p(\mathbf{x}_1|\boldsymbol{\theta}) + \ln p(\mathbf{x}_2|\boldsymbol{\theta}) + \int_{-\infty}^{\infty} \ln p(\mathbf{x}_3|\boldsymbol{\theta}) \cdot p(x_{32}|\boldsymbol{\theta}^0, x_{31} = 2) dx_{32} \\
&= \ln p(\mathbf{x}_1|\boldsymbol{\theta}) + \ln p(\mathbf{x}_2|\boldsymbol{\theta}) + \int_{-\infty}^{\infty} p\left(\binom{2}{x_{32}}\Big|\boldsymbol{\theta}\right) \cdot \underbrace{\frac{p\left(\binom{2}{x_{32}}\Big|\boldsymbol{\theta}^0\right)}{\int_{-\infty}^{\infty} p\left(\binom{2}{x'_{32}}\Big|\boldsymbol{\theta}^0\right) d'_{32}}}_{1/(2e)} dx_{32} \\
&= \ln p(\mathbf{x}_1|\boldsymbol{\theta}) + \ln p(\mathbf{x}_2|\boldsymbol{\theta}) + 2e \int_{-\infty}^{\infty} \ln p\left(\binom{2}{x_{32}}\Big|\boldsymbol{\theta}\right) \cdot p\left(\binom{2}{x_{32}}\Big|\boldsymbol{\theta}^0\right) dx_{32} \\
&= \ln p(\mathbf{x}_1|\boldsymbol{\theta}) + \ln p(\mathbf{x}_2|\boldsymbol{\theta}) + K.
\end{aligned}$$

There are three cases for K:

1. $3 \le \theta_2 \le 4$:

$$K = \frac{1}{4}\int_0^{\theta_2} \ln\left(\frac{1}{\theta_1} e^{-2\theta_1} \frac{1}{\theta_2}\right) dx_{32}$$

$$= \frac{1}{4}\theta_2 \ln\left(\frac{1}{\theta_1}e^{-2\theta_1}\frac{1}{\theta_2}\right).$$

2. $\theta_2 \geq 4$:

$$\begin{aligned} K &= \frac{1}{4}\int_0^4 \ln\left(\frac{1}{\theta_1}e^{-2\theta_1}\frac{1}{\theta_2}\right)dx_{32} \\ &= \frac{1}{4}4\ln\left(\frac{1}{\theta_1}e^{-2\theta_1}\frac{1}{\theta_2}\right) \\ &= \ln\left(\frac{1}{\theta_1}e^{-2\theta_1}\frac{1}{\theta_2}\right). \end{aligned}$$

3. Otherwise $K = 0$.

Thus we have

$$\begin{aligned} Q(\boldsymbol{\theta};\, \boldsymbol{\theta}^0) &= \ln p(\mathbf{x}_1|\boldsymbol{\theta}) + \ln p(\mathbf{x}_2|\boldsymbol{\theta}) + K \\ &= \ln\left(\frac{1}{\theta_1}e^{-\theta_1}\frac{1}{\theta_2}\right) + \ln\left(\frac{1}{\theta_1}e^{-3\theta_1}\frac{1}{\theta_2}\right) + K \\ &= -\theta_1 - \ln(\theta_1\theta_2) - 3\theta_1 - \ln(\theta_1\theta_2) + K \\ &= -4\theta_1 - 2\ln(\theta_1\theta_2) + K, \end{aligned}$$

according to the different cases of θ_2. Note the normalization condition $\int_{-\infty}^{\infty} p(\mathbf{x}_1)dx_1 = 1$, or

$$\int_{-\infty}^{\infty} \frac{1}{\theta_1}e^{-\theta_1 x_1}dx_1 = 1.$$

This equation yields the solution $\theta_1 = 1$.

(b) There are two cases:

1. $3 \leq \theta_2 \leq 4$:

$$Q(\boldsymbol{\theta};\, \boldsymbol{\theta}^0) = -4 - \left(2\ln\theta_2 + \frac{1}{4}\theta_2(2 + \ln\theta_2)\right).$$

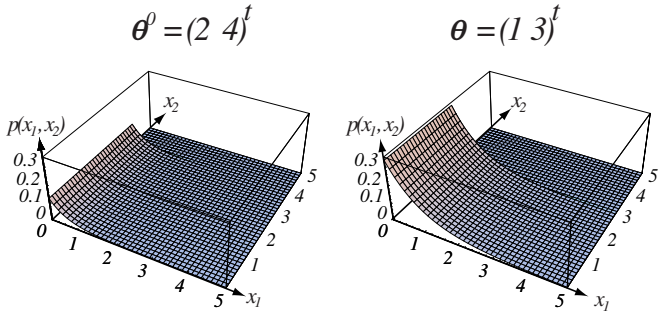
Note that this is a monotonic function and that $\arg\max_{\theta_2} Q(\boldsymbol{\theta};\, \boldsymbol{\theta}^0) = 3$, which leads to $\max Q(\boldsymbol{\theta};\, \boldsymbol{\theta}^0) \simeq -8.5212$.

2. $\theta_2 \geq 4$: In this case $Q(\boldsymbol{\theta};\, \boldsymbol{\theta}^0) = -6 - 3\ln\theta_2$. Note that this is a monotomic function and that $\arg\max_{\theta_2} Q(\boldsymbol{\theta};\, \boldsymbol{\theta}^0) = 4$, which leads to $\max Q(\boldsymbol{\theta};\, \boldsymbol{\theta}^0) \simeq -10.1589$.

In short, then, $\boldsymbol{\theta} = (1 \quad 3)^t$.

(c) See figure.

48. Our data set is $\mathcal{D} = \left\{\binom{1}{1}.\binom{3}{3}, \binom{*}{2}\right\}$.



(a) The **E** step is:

$$\begin{aligned}
Q(\boldsymbol{\theta};\,\boldsymbol{\theta}^0) &= \mathcal{E}_{x_{31}}\left[\ln p(\mathbf{x}_g,\mathbf{x}_b;\,\boldsymbol{\theta})|\boldsymbol{\theta}^0,\mathcal{D}_g\right] \\
&= \int_{-\infty}^{\infty}\left(\ln p(\mathbf{x}_1|\boldsymbol{\theta})+\ln p(\mathbf{x}_2|\boldsymbol{\theta})+\ln p(\mathbf{x}_3|\boldsymbol{\theta})\right)p(x_{31}|\boldsymbol{\theta}^0,x_{32}=2)dx_{31} \\
&= \ln p(\mathbf{x}_1|\boldsymbol{\theta})+\ln p(\mathbf{x}_2|\boldsymbol{\theta})+\int_{-\infty}^{\infty}\ln p\left(\binom{x_{31}}{2}\Big|\boldsymbol{\theta}\right)\cdot\underbrace{\frac{p\left(\binom{x_{31}}{2}\Big|\boldsymbol{\theta}^0\right)}{\int_{-\infty}^{\infty}p\left(\binom{x'_{31}}{2}\Big|\boldsymbol{\theta}^0\right)dx'_{31}}}_{1/16}dx_{31} \\
&= \ln p(\mathbf{x}_1|\boldsymbol{\theta})+\ln p(\mathbf{x}_2|\boldsymbol{\theta})+16\int_{-\infty}^{\infty}\ln p\left(\binom{x_{31}}{2}\Big|\boldsymbol{\theta}\right)\cdot p\left(\binom{x_{31}}{2}\Big|\boldsymbol{\theta}^0\right)dx_{31} \\
&= \ln p(\mathbf{x}_1|\boldsymbol{\theta})+\ln p(\mathbf{x}_2|\boldsymbol{\theta})+K.
\end{aligned}$$

There are two cases for K:

1. $\theta_2 \geq 2$:

$$\begin{aligned}
K &= \int_0^{\infty}2e^{-2x_{31}}\ln\left(\frac{1}{\theta_1}e^{-\theta_1 x_{31}}\frac{1}{\theta_2}\right)dx_{31} \\
&= \int_0^{\infty}2e^{-2x_{31}}(-\theta_1 x_{31}-\ln(\theta_1\theta_2))dx_{31} \\
&= -\ln(\theta_1\theta_2)\int_0^{\infty}2e^{-2x_{31}}dx_{31}-2\theta_1\int_0^{\infty}x_{31}e^{-2x_{31}}dx_{31} \\
&= -\ln(\theta_1\theta_2)-2\theta_1\frac{1}{4} \\
&= \frac{1}{2}\theta_1-\ln(\theta_1\theta_2).
\end{aligned}$$

2. Otherwise $K=0$.

Therefore we have

$$Q(\boldsymbol{\theta};\,\boldsymbol{\theta}^0) = \ln p(\mathbf{x}_1|\boldsymbol{\theta})+\ln p(\mathbf{x}_2|\boldsymbol{\theta})+K$$

$$\begin{aligned} &= \ln\left(\frac{1}{\theta_1}e^{-\theta_1}\frac{1}{\theta_2}\right) + \ln\left(\frac{1}{\theta_1}e^{-3\theta_1}\frac{1}{\theta_2}\right) + K \\ &= -\theta_1 - \ln(\theta_1\theta_2) - 3\theta_1 - \ln(\theta_1\theta_2) + K \\ &= -4\theta_1 - 2\ln(\theta_1\theta_2) + K, \end{aligned}$$

according to the cases for K, directly above. Note that $\int_{-\infty}^{\infty} p(\mathbf{x}_1)dx_1 = 1$, and thus

$$\int_{-\infty}^{\infty} \frac{1}{\theta_1}e^{-\theta_1 x}dx_1 = 1,$$

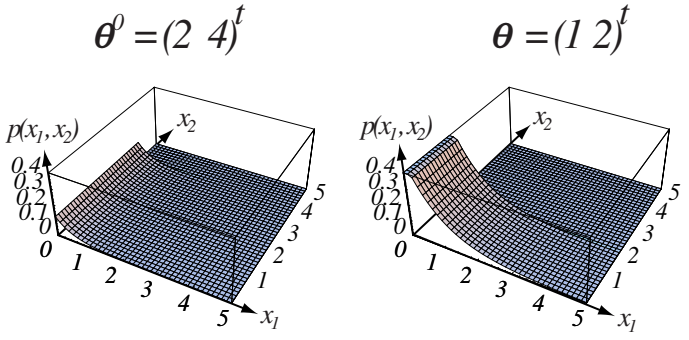
and thus $\theta_1 = 1$. The above results can be simplified.

(b) When $\theta_2 \geq 2$, we have

$$\begin{aligned} Q(\boldsymbol{\theta};\, \boldsymbol{\theta}^0) &= -4 - 2\ln\theta_2 + \frac{1}{2} - \ln\theta_2 \\ &= -\frac{7}{2} - 3\ln\theta_2. \end{aligned}$$

Note that this is a monotonic function, and thus $\arg\max_{\theta_2} Q(\boldsymbol{\theta};\, \boldsymbol{\theta}^0) = 2$. Thus the parameter vector is $\boldsymbol{\theta} = (1 \;\; 2)^t$.

(c) See figure.



Section 3.10

49. A single revision of $\hat{a}_{ij}$'s and $\hat{b}_{ij}$'s involves computing (via Eqs. **??** – **??**)

$$\begin{aligned} \hat{a}_{ij} &= \frac{\sum_{t=1}^{T}\gamma_{ij}(t)}{\sum_{t=1}^{T}\sum_{k}\gamma_{ik}(t)} \\ \hat{b}_{ij} &= \frac{\sum_{??}^{??}\gamma_{ij}}{\sum_{t=1}^{T}\gamma_{ij}(t)} \\ \gamma_{ij}(t) &= \frac{\alpha_i(t-1)a_{ij}b_{ij}\beta_i(t)}{P(V^T|M)}. \end{aligned}$$

$\alpha_i(t)$'s and $P(V^T|M)$ are computed by the `Forward Algorithm`, which requires $O(c^2T)$ operations. The $\beta_i(t)$'s can be computed recursively as follows:

For t=T to 1 (by -1)
For i=1 to c
$\beta_i(t) = \sum_j a_{ij}b_{jk}v(t+1)\beta_j(t+1)$
End

This requires $O(c^2T)$ operations.

Similarly, γij's can be computed by $O(c^2T)$ operations given $\alpha_i(t)$'s, a_{ij}'s, b_{ij}'s, $\beta_i(t)$'s and $P(V^T|M)$. So, $\gamma_{ij}(t)$'s are computed by

$$\underbrace{O(c^2T)}_{\alpha_i(t)\text{'s}} + \underbrace{O(c^2T)}_{\beta_i(t)\text{'s}} + \underbrace{O(c^2T)}_{\gamma_{ij}(t)\text{'s}} = O(c^2T)\text{operations.}$$

Then, given $\hat{\gamma}_{ij}(t)$'s, the $\hat{a}_{ij}$'s can be computed by $O(c^2T)$ operations and $\hat{b}_{ij}$'s by $O(c^2T)$ operations. Therefore, a single revision requires $O(c^2T)$ operations.

50. The standard method for calculating the probability of a sequence in a given HMM is to use the forward probabilities $\alpha_i(t)$.

(a) In the forward algorithm, for $t = 0, 1, \ldots, T$, we have

$$\alpha_j(t) = \begin{cases} 0 & t = 0 \text{ and } j \neq \text{initial status} \\ 1 & t = 0 \text{ and } j = \text{initial status} \\ \sum_{i=1}^{c} \alpha_i(t-1)a_{ij}b_{jk}v(t) & \text{otherwise.} \end{cases}$$

In the backward algorithm, we use for $t = T, T-1, \ldots, 0$,

$$\beta_j(t) = \begin{cases} 0 & t = T \text{ and } j \neq \text{final status} \\ 1 & t = T \text{ and } j = \text{final status} \\ \sum_{i=1}^{c} \beta_i(t+1)a_{ij}b_{jk}v(t+1) & \text{otherwise.} \end{cases}$$

Thus in the forward algorithm, if we first reverse the observed sequence $\mathbf{V}^T$ (that is, set $b_{jk}v(t) = b_{jk}(T+1-t)$ and then set $\beta_j(t) = \alpha_j(T-t)$, we can obtain the backward algorithm.

(b) Consider splitting the sequence $\mathbf{V}^T$ into two parts — $\mathbf{V}_1$ and $\mathbf{V}_2$ — before, during, and after each time step T' where $T' < T$. We know that $\alpha_i(T')$ represents the probability that the HMM is in hidden state ω_i at step T', having generated the firt T' elements of $\mathbf{V}^T$, that is $\mathbf{V}_1$. Likewise, $\beta_i(T')$ represents the probability that the HMM given that it is in ω_i at step T' generates the remaining elements of $\mathbf{V}^T$, that is, $\mathbf{V}_2$. Hence, for the complete sequence we have

$$\begin{aligned} P(\mathbf{V}^T) &= P(\mathbf{V}_1, \mathbf{V}_2) = \sum_{i=1}^{c} P(\mathbf{V}_1, \mathbf{V}_2, \text{hidden state } \omega_i \text{ at step } T') \\ &= \sum_{i=1}^{c} P(\mathbf{V}_1, \text{hidden state } \omega_i \text{ at step } T')P(\mathbf{V}_2|\text{hidden state } \omega_i \text{ at step } T') \\ &= \sum_{i=1}^{c} \alpha_i(T')\beta_i(T'). \end{aligned}$$

(c) At $T' = 0$, the above reduces to $P(\mathbf{V}^T) = \sum_{i=1}^{c} \alpha_i(0)\beta_i(0) = \beta_j(0)$, where j is the known initial state. This is the same as line 5 in Algorithm 3. Likewise, at $T' = T$, the above reduces to $P(\mathbf{V}^T) = \sum_{i=1}^{c} \alpha_i(T)\beta_i(T) = \alpha_j(T)$, where j is the known final state. This is the same as line 5 in Algorithm 2.

51. From the learning algorithm in the text, we have for a giveen HMM with model parameters $\boldsymbol{\theta}$:

$$\gamma_{ij}(t) = \frac{\alpha_i(t-1)a_{ij}b_{jk}v(t)\beta_j(t)}{P(\mathbf{V}^T|\boldsymbol{\theta})} \qquad (*)$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T}\gamma_{ij}(t)}{\sum_{t=1}^{T}\sum_{k=1}^{c}\gamma_{ik}(t)}. \qquad (**)$$

For a new HMM with $a_{i'j'} = 0$, from $(*)$ we have $\gamma_{i'j'} = 0$ for all t. Substituting $\gamma_{i'j'}(t)$ into $(**)$, we have $\hat{a}_{i'j'} = 0$. Therefore, keeping this substitution throughout the iterations in the learning algorithm, we see that $\hat{a}_{i'j'} = 0$ remains unchanged.

52. Consider the decoding algorithm (Algorithm 4).

(a) the algorithm is:

Algorithm 0 (Modified decoding)

```
1   begin initialize Path ← {}, t ← 0
2       for t ← t + 1
3           j ← 0; δ_0 ← 0
4           for j ← j + 1
5               δ_j(t) ← min_{1≤i≤c} [δ_i(t − 1) − ln(a_ij)] − ln[b_jk v(t)]
6           until j = c
7           j' ← arg min_j [δ_j(t)]
8           Append ω_j' to Path
9       until t = T
10      return Path
11  end
```

(b) Taking the logarithm is an $O(c^2)$ computation since we only need to calculate $\ln a_{ij}$ for all $i, j = 1, 2, \ldots, c$, and $\ln[b_{jk}v(t)]$ for $j = 1, 2, \ldots, c$. Then, the whole complexity of this algorithm is $O(c^2T)$.

Computer Exercises

Section 3.2

1. Computer exercise not yet solved

Section 3.3

2. Computer exercise not yet solved

Section 3.4

3. Computer exercise not yet solved

Section 3.5

4. Computer exercise not yet solved

Section 3.6

5. Computer exercise not yet solved

Section 3.7

6. Computer exercise not yet solved
7. Computer exercise not yet solved
8. Computer exercise not yet solved

Section 3.8

9. Computer exercise not yet solved
10. Computer exercise not yet solved

Section 3.9

11. Computer exercise not yet solved
12. Computer exercise not yet solved

Section 3.10

13. Computer exercise not yet solved

Chapter 4

Nonparametric techniques

Problem Solutions

Section 4.3

1. Our goal is to show that Eqs. 19–22 are sufficient to assure convergence of Eqs. 17 and 18. We first need to show that $\lim_{n\to\infty} \bar{p}_n(\mathbf{x}) = p(\mathbf{x})$, and this requires us to show that

$$\lim_{n\to\infty} \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{V}}{h_n}\right) = \delta_n(\mathbf{x}-\mathbf{V}) \to \delta(\mathbf{x}-\mathbf{V}).$$

Without loss of generality, we set $\mathbf{V} = \mathbf{0}$ and write

$$\frac{1}{V_n}\varphi\left(\frac{\mathbf{x}}{h_n}\right) = \frac{1}{\prod_{i=1}^{d} x_i}\prod_{i=1}^{d}\frac{x_i}{h_n}\varphi\left(\frac{x}{h_n}\right) = \delta_n(\mathbf{x}),$$

where we used the fact that $V_n = h_n^d$. We also see that $\delta_n(\mathbf{x}) \to 0$ if $\mathbf{x} \neq 0$, with normalization

$$\int \delta_n(\mathbf{x})d\mathbf{x} = 1.$$

Thus we have in the limit $n \to \infty$, $\delta_n(\mathbf{x}) \to \delta(\mathbf{x})$ as required.

We substitute our definition of $\bar{p}_n(\mathbf{x})$ and get

$$\lim_{n\to\infty} \bar{p}_n(\mathbf{x}) = \mathcal{E}[\bar{p}(\mathbf{x})] = \int \delta(\mathbf{x}-\mathbf{V})p(\mathbf{V})d\mathbf{V} = p(\mathbf{x}).$$

Furthermore, the variance is

$$\sigma_n^2(x) = \frac{1}{nV_n}\int \frac{1}{V_n}\varphi^2\left(\frac{\mathbf{x}-\mathbf{V}}{h_n}\right)p(\mathbf{V})d\mathbf{v} - \frac{1}{n}[\bar{p}_n(x)]^2$$

$$\begin{aligned}
&\leq \frac{1}{nV_n}\int \frac{1}{V_n}\varphi^2\left(\frac{\mathbf{x}-\mathbf{V}}{h_n}\right)p(\mathbf{V})d\mathbf{V}\\
&\leq \frac{\mathrm{Sup}(\varphi)}{nV_n}\int \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{V}}{h_n}\right)p(\mathbf{V})d\mathbf{V}\\
&\leq \frac{\mathrm{Sup}(\varphi)\bar{p}_n(\mathbf{x})}{nV_n}.
\end{aligned}$$

We note that $\mathrm{Sup}(\varphi) < \infty$ and that in the limit $n \to \infty$ we have $\bar{p}_n(\mathbf{x}) \to p(\mathbf{x})$ and $nV_n \to \infty$. We put these results together to conclude that

$$\lim_{n\to\infty} \sigma_n^2(\mathbf{x}) = 0,$$

for all $\mathbf{x}$.

2. Our normal distribution is $p(x) \sim N(\mu, \sigma^2)$ and our Parzen window is $\varphi(x) \sim N(0,1)$, or more explicitly,

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2},$$

and our estimate is

$$p_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{h_n}\right).$$

(a) The expected value of the probability at x, based on a window width parameter h_n, is

$$\begin{aligned}
\bar{p}_n(x) &= \mathcal{E}[p_n(x)] = \frac{1}{nh_n}\sum_{i=1}^{n}\mathcal{E}\left[\varphi\left(\frac{x-x_i}{h_n}\right)\right]\\
&= \frac{1}{h_n}\int_{-\infty}^{\infty}\varphi\left(\frac{x-v}{h_n}\right)p(v)dv\\
&= \frac{1}{h_n}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-v}{h_n}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right]dv\\
&= \frac{1}{2\pi h_n\sigma}\exp\left[-\frac{1}{2}\left(\frac{x^2}{h_n^2}+\frac{\mu^2}{\sigma^2}\right)\right]\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}v^2\left(\frac{1}{h_n^2}+\frac{1}{\sigma^2}\right)-2v\left(\frac{x}{h_n^2}+\frac{\mu}{\sigma^2}\right)\right]dv\\
&= \frac{1}{2\pi h_n\sigma}\exp\left[-\frac{1}{2}\left(\frac{x^2}{h_n^2}+\frac{\mu^2}{\sigma^2}\right)+\frac{1}{2}\frac{\alpha^2}{\theta^2}\right]\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{v-\alpha}{\theta}\right)^2\right]dv,
\end{aligned}$$

where we have defined

$$\theta^2 = \frac{1}{1/h_n^2+1/\sigma^2} = \frac{h_n^2\sigma^2}{h_n^2+\sigma^2}$$

and

$$\alpha = \theta^2\left(\frac{x}{h_n^2}+\frac{\mu}{\sigma^2}\right).$$

We perform the integration and find

$$\begin{aligned}\bar{p}_n(x) &= \frac{\sqrt{2\pi}\theta}{2\pi h_n\sigma}\exp\left[-\frac{1}{2}\left(\frac{x^2}{h_n^2}+\frac{\mu^2}{\sigma^2}\right)+\frac{1}{2}\frac{\alpha^2}{\theta^2}\right]\\ &= \frac{1}{\sqrt{2\pi}h_n\sigma}\frac{h_n\sigma}{\sqrt{h_n^2+\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{x^2}{h_n^2}+\frac{\mu^2}{\sigma^2}-\frac{\alpha^2}{\theta^2}\right)\right].\end{aligned}$$

The argument of the exponentiation can be expressed as follows

$$\begin{aligned}\frac{x^2}{h_n^2}+\frac{\mu^2}{\sigma^2}-\frac{\alpha^2}{\theta^2} &= \frac{x^2}{h_n^2}+\frac{\mu^2}{\sigma^2}-\frac{\theta^4}{\theta^2}\left(\frac{x}{h_n^2}+\frac{\mu}{\sigma^2}\right)^2\\ &= \frac{x^2h_n^2}{(h_n^2+\sigma^2)h_n^2}+\frac{\mu^2\sigma^2}{(h_n^2+\sigma^2)\sigma^2}-\frac{2x\mu}{h_n^2+\sigma^2}\\ &= \frac{(x-\mu)^2}{h_n^2+\sigma^2}.\end{aligned}$$

We substitute this back to find

$$\bar{p}_n(x) = \frac{1}{\sqrt{2\pi}\sqrt{h_n^2+\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}\right],$$

which is the form of a Gaussian, denoted

$$\bar{p}_n(x)\sim N(\mu, h_n^2+\sigma^2).$$

(b) We calculate the variance as follows:

$$\begin{aligned}\mathrm{Var}[p_n(x)] &= \mathrm{Var}\left[\frac{1}{nh_n}\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{h_n}\right)\right]\\ &= \frac{1}{n^2h_n^2}\sum_{i=1}^{n}\mathrm{Var}\left[\varphi\left(\frac{x-x_i}{h_n}\right)\right]\\ &= \frac{1}{nh_n^2}\mathrm{Var}\left[\varphi\left(\frac{x-v}{h_n}\right)\right]\\ &= \frac{1}{nh_n^2}\left\{\mathcal{E}\left[\varphi^2\left(\frac{x-v}{h_n}\right)\right]-\left(\mathcal{E}\left[\varphi\left(\frac{x-v}{h_n}\right)\right]\right)^2\right\},\end{aligned}$$

where in the first step we used the fact that $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are independent samples drawn according to $p(\mathbf{x})$. We thus now need to calculate the expected value of the square of the kernel function

$$\begin{aligned}\mathcal{E}\left[\varphi^2\left(\frac{x-v}{h_n}\right)\right] &= \int\varphi^2\left(\frac{x-v}{h_n}\right)p(v)dv\\ &= \int_{-\infty}^{\infty}\frac{1}{2\pi}\exp\left[-\left(\frac{x-v}{h_n}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma}dv\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-v}{h_n/\sqrt{2}}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma}dv.\end{aligned}$$

From part (a) by a similar argument with $h_n/\sqrt{2}$ replacing h_n, we have

$$\frac{1}{h_n/\sqrt{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-v}{h_n/\sqrt{2}}\right)^2\right]\exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma}dv$$
$$=\frac{1}{\sqrt{2\pi}\sqrt{h_n^2/2+\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{h_n^2/2+\sigma^2}\right].$$

We make the substitution and find

$$\mathcal{E}\left[\varphi^2\left(\frac{x-v}{h_n}\right)\right]=\frac{h_n/\sqrt{2}}{2\pi\sqrt{h_n^2/2+\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{h_n^2/2+\sigma^2}\right],$$

and thus conclude

$$\frac{1}{nh_n^2}\mathcal{E}\left[\varphi^2\left(\frac{x-v}{h_n}\right)\right]=\frac{1}{nh_n}\frac{1}{2\sqrt{\pi}}\frac{1}{\sqrt{2\pi}\sqrt{h_n^2/2+\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{h_n^2/2+\sigma^2}\right].$$

For small h_n, $\sqrt{h_n^2/2+\sigma^2}\simeq\sigma$, and thus the above equation can be approximated as

$$\begin{aligned}\frac{1}{nh_n^2}\mathcal{E}\left[\varphi^2\left(\frac{x-v}{h_n}\right)\right] &\simeq \frac{1}{nh_n 2\sqrt{\pi}}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\\ &= \frac{1}{2nh_n\sqrt{\pi}}p(x). \qquad (*)\end{aligned}$$

Similarly, we have

$$\begin{aligned}\frac{1}{nh_n^2}\left\{\mathcal{E}\left[\varphi\left(\frac{x-v}{h_n}\right)\right]\right\}^2 &= \frac{1}{nh_n^2}h_n^2\frac{1}{\sqrt{2\pi}\sqrt{h_n^2+\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}\right]\\ &= \frac{h_n}{nh_n}\frac{1}{\sqrt{2\pi}\sqrt{h_n^2+\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}\right]\\ &\simeq \frac{h_n}{nh_n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]\simeq 0, (**)\end{aligned}$$

valid for small h_n. From $(*)$ and $(**)$ we have, (still for small h_n)

$$\mathrm{Var}[P_n(x)]\simeq\frac{p(x)}{2nh_n\sqrt{\pi}}.$$

(c) We write the bias as

$$\begin{aligned}p(x)-\bar{p}_n(x) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]-\frac{1}{\sqrt{2\pi}\sqrt{h_n^2+\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}\right]\\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}\right]\left\{1-\frac{\sigma}{\sqrt{h_n^2+\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}+\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]\right\}\\ &= p(x)\left\{1-\frac{1}{\sqrt{1+(h_n/\sigma)^2}}\exp\left[-\frac{(x-\mu)^2}{2}\left\{\frac{1}{h_n^2+\sigma^2}-\frac{1}{\sigma^2}\right\}\right]\right\}\\ &= p(x)\left\{1-\frac{1}{\sqrt{1+(h_n/\sigma)^2}}\exp\left[\frac{1}{2}\frac{h_n^2(x-\mu)^2}{h_n^2+\sigma^2}\right]\right\}.\end{aligned}$$

For small h_n we expand to second order:

$$\frac{1}{\sqrt{1+(\frac{h_n}{\sigma})^2}} \simeq 1 - \frac{1}{2}\left(h_n/\sigma\right)^2$$

and

$$\exp\left[\frac{h_n^2}{2\sigma^2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}\right] \simeq 1 + \frac{h_n^2}{2\sigma^2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}.$$

We ignore terms of order higher than h_n^2 and find

$$\begin{aligned}
p(x) - \bar{p}_n(x) &\simeq p(x)\left\{1 - \left(1 - \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2\right)\left(1 + \frac{h_n^2}{2\sigma^2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}\right)\right\} \\
&\simeq p(x)\left\{1 - 1 + \frac{1}{2}\frac{h_n^2}{\sigma^2} - \frac{h_n^2}{2\sigma^2}\frac{(x-\mu)^2}{h_n^2+\sigma^2}\right\} \\
&\simeq \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2\left[1 - \frac{(x-\mu)^2}{h_n^2+\sigma^2}\right]p(x) \\
&\simeq \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2\left[1 - \left(\frac{x-\mu}{\sigma}\right)^2\right]p(x).
\end{aligned}$$

3. Our (normalized) distribution is

$$p(x) = \begin{cases} 1/a & 0 \le x \le a \\ 0 & \text{otherwise,} \end{cases}$$

and our Parzen window is

$$\varphi(x) = \begin{cases} e^{-x} & x \ge 0 \\ 0 & x < 0. \end{cases}$$

(a) The expected value of the Parzen window estimate is

$$\begin{aligned}
\bar{p}_n(x) &= \mathcal{E}\left[\frac{1}{nh_n}\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{h_n}\right)\right] = \frac{1}{h_n}\int\varphi\left(\frac{x-v}{h_n}\right)p(v)dv \\
&= \frac{1}{h_n}\int_{x\ge v} e^{-\frac{(x-v)}{h_n}}p(v)dv \\
&= \frac{\exp[-x/h_n]}{h_n}\int_{\substack{x\ge v,\\ 0<v<a}}\frac{1}{a}\exp[v/h_n]dv \\
&= \begin{cases} 0 & \text{if} \quad x < 0 \\ \frac{e^{-x/h_n}}{ah_n}\int_0^x e^{v/h_n}dv & \text{if} \quad 0 \le x \le a \\ \frac{e^{-x/h_n}}{ah_n}\int_0^a e^{v/h_n}dv & \text{if} \quad x \ge a \end{cases}
\end{aligned}$$

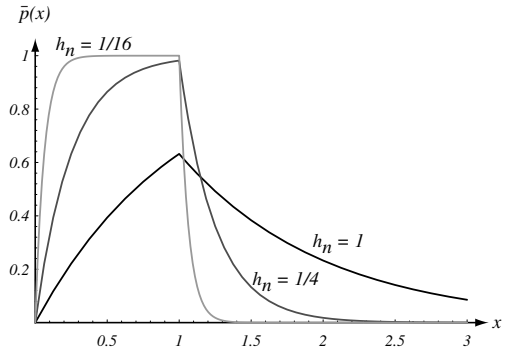
$$= \begin{cases} 0 & \text{if} \quad x < 0 \\ \frac{1}{a}(1 - e^{-x/h_n}) & \text{if} \quad 0 \le x \le a \\ \frac{1}{a}(e^{a/h_n} - 1)e^{-x/h_n} & \text{if} \quad x \ge a, \end{cases}$$

where in the first step we used the fact that $x_1, \ldots, x_n$ are independent samples drawn according to $p(v)$.

(b) For the case $a = 1$, we have

$$\bar{p}_n(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x/h_n} & 0 \le x \le 1 \\ (e^{1/h_n} - 1)e^{-x/h_n} & x > 1, \end{cases}$$

as shown in the figure.



(c) The bias is

$$p(x) - \bar{p}_n(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ \frac{1}{a} - \frac{1}{a}(1 - e^{-x/h_n}) & \text{if} \quad 0 \le x \le a \\ 0 - \frac{1}{a}(e^{a/h_n} - 1)e^{-x/h_n} & \text{if} \quad x \ge a \end{cases}$$

$$= \begin{cases} 0 & \text{if} \quad x < 0 \\ \frac{1}{a}e^{-x/h_n} & \text{if} \quad 0 \le x \le a \\ -\frac{1}{a}(e^{a/h_n} - 1)e^{-x/h_n} & \text{if} \quad x \ge a. \end{cases}$$

Formally, a bias lower than 1% over 99% of the range $0 < x < a$, means that

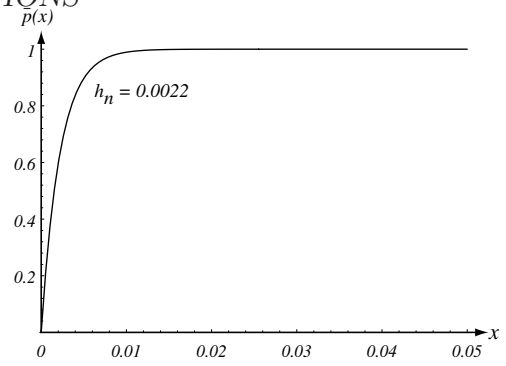
$$\frac{p(x) - \bar{p}(x)}{p(x)} \le 0.01$$

over 99% of the range $0 < x < a$. This, in turn, implies

$$\frac{1/ae^{-x/h_n}}{1/a} \le 0.01 \quad \text{over 99\% of } 0 < x < a \text{ or}$$

$$h_n \le \frac{0.01a}{\ln\,(100)}.$$

For the case $a = 1$, we have that $h_n \le 0.01/(\ln 100) = 0.0022$, as shown in the figure. Notice that the estimate is within 1% of $p(x) = 1/a = 1.0$ above $x \sim 0.01$, fulfilling the conditions of the problem.



4. We have from Algorithm 2 in the text that the discriminant functions of the PNN classifier are given by

$$g_i(\mathbf{x}) = \sum_{k=1}^{n_i} \exp\left[\frac{\mathbf{w}_k^t\mathbf{x} - 1}{\sigma_2}\right] a_{k_i} \qquad i = 1, \ldots, c$$

where $\|\mathbf{w}_k\| = \|\mathbf{x}\| = 1$, n_i is the number of training patterns belonging to ω_i and

$$a_{k_i} = \begin{cases} 1 & \text{if } \mathbf{w}_k \in \omega_i \\ 0 & \text{otherwise.} \end{cases}$$

(a) Since $\|\mathbf{w}_k\| = \|\mathbf{x}\| = 1$, $g_i(\mathbf{x})$ can be written as

$$g_i(\mathbf{x}) = \sum_{i=1}^{n_i} \exp\left[-\frac{\|\mathbf{x} - \mathbf{w}_k\|^2}{2\sigma^2}\right] a_{k_i}.$$

Note that $g_i(\mathbf{x})/n_i$ is a radial Gaussian based kernel estimate, $p_n(\mathbf{x}|\omega_i)$, of $p(\mathbf{x}|\omega_i)$. If we use n_i/n as the estimate of the prior class probability $P(\omega_i)$, then $g_i(\mathbf{x})$ can be rewritten as

$$g_i(\mathbf{x}) = nP_n(\omega_i)p(\mathbf{x}|\omega_i).$$

Thus $g_i(\mathbf{x})$ properly accounts for the class priors.

(b) The optimal classification rule for unequal costs is given by

$$\text{Choose } \omega_k \text{ if } g_k^* = \min_{i \le c} g_i^*,$$

where the λ_{ij} represent the costs and

$$g_i^*(\mathbf{x}) = \sum_{j=1}^{c} \lambda_{ij} P(\omega_j|\mathbf{x}) = \sum_{j=1}^{c} \frac{P(\omega_j)p(\mathbf{x}|\omega_j}{p(\mathbf{x})}.$$

This discriminant function can be written simply as

$$g_i^*(\mathbf{x}) = \sum_{j=1}^{c} \lambda_{ij} P(\omega_j)p(\mathbf{x}|\omega_j).$$

Consequently, the PNN classifier must estimate $g_i^*(\mathbf{x})$. From part (a) we have that $g_i = nP_n(\omega_j)p(\mathbf{x}|\omega_j)$. Thus the new discriminant functions are simply

$$\hat{g}_i(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^{c} \lambda_{ij} g_j(\mathbf{x}),$$

where $g_i(\mathbf{x})$ are the discriminant functions of the PNN classifier previously defined. An equivalent discriminant function is

$$\hat{g}_i(\mathbf{x}) = \sum_{j=1}^{c} \lambda_{ij} g_j(\mathbf{x}).$$

(c) The training algorithm as defined is unaffected since it only performs a normalization of the training patterns. However the PNN classification algorithm must be rewritten as:

Algorithm 0 (PNN with costs)

```
1   begin initialize x ← test pattern, g_i ← 0, g'_i ← 0, λ_ij
2         k ← 0
3     do k ← k + 1
4        net_k ← w_k^t x
5        if a_{k_i} = 1 then g'_i ← g'_i + exp[(net_k − 1)/σ^2]
6     until k = n
7     k ← 0
8     do k ← k + 1
9        g_i ← g_i + λ_ik g'_k
10    until k = n
11    return class ← arg max_i g_i(x)
12  end
```

Section 4.4

5. Our goal is to show that $p_n(\mathbf{x})$ converges in probability to $p(\mathbf{x})$ given the below conditions:

$$\lim_{n\to\infty} k_n \to \infty \quad \text{(condition 1)}$$
$$\lim_{n\to\infty} k_n/n \to 0 \quad \text{(condition 2)}$$

As we know from probability theory, when $p_n(\mathbf{x})$ converges in the rth mean ($r \geq 1$) to $p(\mathbf{x})$ — that is, $\mathcal{E}[|p_n(\mathbf{x}) - p(\mathbf{x})|^r] \to 0$ as $n \to 0$ — then $p(\mathbf{x})$ also converges in probability to $p(\mathbf{x})$.

It is simple to show for $r = 2$ that

$$\mathcal{E}[(p_n(\mathbf{x}) - p(\mathbf{x}))^2] = \mathrm{Var}[p_n(\mathbf{x})] + bias^2[p_n(\mathbf{x})],$$

where

$$\begin{aligned} \mathrm{Var}[p_n(\mathbf{x})] &= \mathcal{E}\left[p_n(\mathbf{x}) - \mathcal{E}[p_n(\mathbf{x})]\right] \\ bias^2[p_n(\mathbf{x})] &= \mathcal{E}[p_n(\mathbf{x}) - p(\mathbf{x})]. \end{aligned}$$

Consequently, $p_n(\mathbf{x})$ converges in probability to $p(\mathbf{x})$ if and only if

$$\mathrm{Var}[p_n(\mathbf{x})] \to 0 \quad \text{as } n \to \infty \quad \text{(condition 3)}$$
$$\mathcal{E}[p_n(\mathbf{x})] = p(\mathbf{x}) \quad \text{as } n \to \infty \quad \text{(condition 4)}.$$

Note that the k-nearest-neighbor estimate $p_n(\mathbf{x}) = \frac{k_n/n}{V_n}$ can be rewritten as a kernel estimate according to Eq. 11 in the text:

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right),$$

where h_n denotes here the radius of a ball which includes the k_n prototypes nearest to $\mathbf{x}$, V_n is the volume of the ball, and

$$\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) = \begin{cases} 1 & \text{if } \|\mathbf{x}-\mathbf{x}_i\| \le h_n \\ 0 & \text{otherwise.} \end{cases}$$

Hence we can apply the results of page 167 in the text. According to Eq. 23 in the text,

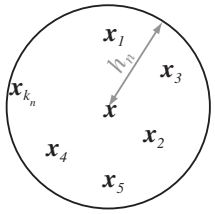
$$\mathcal{E}[p_n(\mathbf{x})] = p(\mathbf{x}) \quad \text{as } n \to \infty \text{ if and only if } \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}}{h_n}\right) = \delta(\mathbf{x}) \quad \text{as } n \to \infty,$$

where $\delta(\mathbf{x})$ is the delta function.

Given a test pattern $\mathbf{x}$ (see figure), we will have a very small ball (or more precisely, have $V_n \to 0$) as $n \to \infty$, since we can get k prototypes arbitrarily close to $\mathbf{x}$ due to condition 2. If $V_n \to 0$, $h_n \to 0$ and therefore

$$\varphi\left(\frac{\mathbf{x}}{h_n}\right) = \begin{cases} 1 & \text{if } \mathbf{x} = \mathbf{0} \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, $\frac{1}{V_n}\varphi(\mathbf{x}/h_n) = \delta(\mathbf{x})$ and hence condition 4 is fulfilled.



We can compute $\lim_{n\to\infty} \mathrm{Var}[p_n(\mathbf{x})]$ using Eq. 24 as follows:

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{Var}[p_n(\mathbf{x})] &= \lim_{n\to\infty} \frac{1}{nV_n}\int \frac{1}{V_n}\varphi^2\left(\frac{\mathbf{x}-\mathbf{u}}{h_n}\right)p(\mathbf{u})d\mathbf{u} \\
&= \frac{\int \lim_{n\to\infty}\left[\frac{1}{V_n}\varphi^2\left(\frac{\mathbf{x}-\mathbf{u}}{h_n}\right)\right]p(\mathbf{u})d\mathbf{u}}{\lim_{n\to\infty} nV_n} \\
&= \frac{\int \delta(\mathbf{x}-\mathbf{u})p(\mathbf{u})d\mathbf{u}}{\lim_{n\to\infty} nV_n} \\
&= \frac{p(\mathbf{x})}{\lim_{n\to\infty} nV_n}.
\end{aligned}$$

Since $\lim_{n\to\infty} \frac{k_n}{nV_n} = P(\mathbf{x})$ in probability and condition 1 is met, then $\lim_{n\to\infty} nV_n = \infty$. Thus $\mathrm{Var}[p(\mathbf{x})] \to 0$ as $n \to 0$.

6. We are given that $P(\omega_1) = P(\omega_2) = 1/2$, and that

$$\begin{aligned} P(\omega_1|\mathbf{x}) &= \begin{cases} 1 & \text{if } ||\mathbf{x}|| \leq 1 \\ 0 & \text{otherwise} \end{cases} \\ P(\omega_2|\mathbf{x}) &= \begin{cases} 1 & \text{if } ||\mathbf{x}-\mathbf{a}|| \leq 1 \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

where we have assumed, without loss of generality, that category ω_1 is centered on the origin $\mathbf{0}$, and category ω_2 is centered on $\mathbf{a}$, a point other than the origin.

(a) We denote by $P_n(e)$ the average probability of error based on n points.

$$\begin{aligned} P_n(e) &= \text{Pr[true category is } \omega_1 \text{ while } \omega_2 \text{ is most frequently labeled]} \\ &\quad + \text{Pr[true category is } \omega_2 \text{ while } \omega_1 \text{ is most frequently labeled]} \\ &= 2\text{Pr[true category is } \omega_1 \text{ while } \omega_2 \text{ is most frequently labeled]} \\ &= 2P(\omega_1)\text{Pr[label of } \omega_1 \text{ for fewer than } (k-1)/2 \text{points, and the rest labeled } \omega_2] \\ &= 2\frac{1}{2}\sum_{j=0}^{(k-1)/2} \text{Pr}[j \text{ of } n \text{ chosen points are labeled } \omega_1, \text{the rest } \omega_2] \\ &= \sum_{j=0}^{(k-1)/2} \binom{n}{j} \frac{1}{2^j}\frac{1}{2^{(n-j)}} \\ &= \frac{1}{2^n}\sum_{j=0}^{(k-1)/2} \binom{n}{j}. \end{aligned}$$

(b) We make explicit the k dependence on the probability by writing $P_n(e) = P_n(e;\ k)$ and have

$$P_n(e;\ 1) = \frac{1}{2^n} < P_n(e;\ k) = \frac{1}{2^n} \underbrace{\sum_{j=0}^{(k-1)/2} \binom{n}{j}}_{>0 \ for \ k>1}.$$

(c) We have in this case

$$\begin{aligned} P_n(e) &= \frac{1}{2^n}\sum_{j=0}^{(k-1)/2} \binom{n}{j} = \text{Pr}\left[B(n, 1/2) \leq (k-1)/2\right] \\ &= \text{Pr}\left[Y_1 + \cdots + Y_n \leq \frac{k-1}{2}\right], \end{aligned}$$

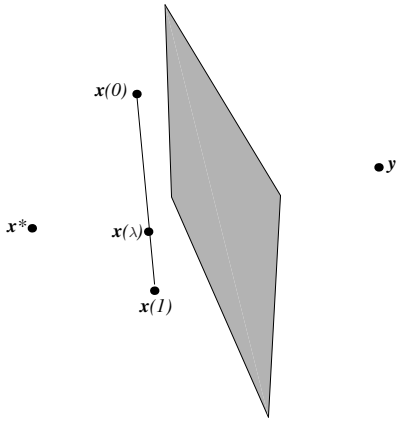
where $Y_1, \cdots, Y_n$ are independent, $B(\cdot,\cdot)$ is a binomial distribution and $\text{Pr}[Y_i = 1] = \text{Pr}[Y_i = 0] = 1/2$. If k is allowed to increase with n, but is restricted by $k < a/\sqrt{n}$, then we have

$$\begin{aligned} P_n(e) &\leq \text{Pr}\left(Y_1 + \cdots + Y_n \leq \frac{a/\sqrt{n} - 1}{2}\right) \\ &= Pr(Y_1 + \cdots + Y_n \leq 0) \text{ for } n \text{ sufficiently large} \\ &= 0, \end{aligned}$$

and this guarantees $P_n(e) \to 0$ as $n \to \infty$.

Section 4.5

7. Our goal is to show that Voronoi cells induced by the nearest-neighbor algorithm are convex, that is, given any two points in the cell, the line connecting them also lies in the cell. We let $\mathbf{x}^*$ be the labeled sample point in the Voronoi cell, and $\mathbf{y}$ be



any other labeled sample point. A unique hyperplane separates the space into those that are closer to $\mathbf{x}^*$ than to $\mathbf{y}$, as shown in the figure. Consider any two points $\mathbf{x}(0)$ and $\mathbf{x}(1)$ inside the Voronoi cell of $\mathbf{x}^*$; these points are surely on the side of the hyperplane nearest $\mathbf{x}^*$. Now consider the line connecting those points, parameterized as $\mathbf{x}(\lambda) = (1-\lambda)\mathbf{x}(0) + \lambda\mathbf{x}(1)$ for $0 \le \lambda \le 1$. Because the half-space defined by the hyperplane is convex, all the points $\mathbf{x}(\lambda)$ lie nearer to $\mathbf{x}^*$ than to $\mathbf{y}$. This is true for *every* pair of points $\mathbf{x}(0)$ and $\mathbf{x}(1)$ in the Voronoi cell. Furthermore, the result holds for *every* other sample point $\mathbf{y}_i$. Thus $\mathbf{x}(\lambda)$ remains closer to $\mathbf{x}^*$ than any other labeled point. By our definition of convexity, we have, then, that the Voronoi cell is convex.

8. It is indeed possible to have the nearest-neighbor error rate P equal to the Bayes error rate P^* for non-trivial distributions.

(a) Consider uniform priors over c categories, that is, $P(\omega_i) = 1/c$, and one-dimensional distributions

$$p(x|\omega_i) = \begin{cases} 1 & 0 \le x \le \frac{cr}{c-1} \\ 1 & i \le x \le i+1-\frac{cr}{c-1} \\ 0 & \text{elsewhere.} \end{cases}$$

The evidence is

$$p(x) = \sum_{i=1}^{c} p(x|\omega_i)P(\omega_i) = \begin{cases} 1 & 0 \le x \le \frac{cr}{c-1} \\ 1/c & i \le x \le (i+1)-\frac{cr}{c-1} \\ 0 & \text{elsewhere.} \end{cases}$$

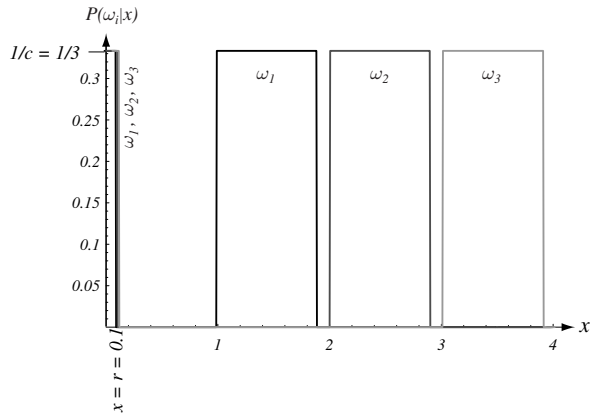
Note that this automatically imposes the restriction

$$0 \le \frac{cr}{c-1} \le 1.$$

Because the $P(\omega_i)$'s are constant, we have $P(\omega_i|x) \propto p(x|\omega_i)$ and thus

$$P(\omega_i|x) = \begin{cases} \frac{P(\omega_i)}{p(x)} = \frac{1/c}{p(x)} & 0 \le x \le \frac{cr}{c-1} \\ \left.\begin{array}{ll} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{array}\right\} & j \le x \le j + 1 - \frac{cr}{c-1} \\ 0 & \text{otherwise,} \end{cases}$$

as shown in the figure for $c = 3$ and $r = 0.1$.



The conditional Bayesian probability of error at a point x is

$$\begin{aligned} P^*(e|x) &= 1 - P(\omega_{max}|x) \\ &= \begin{cases} 1 - \frac{1/c}{p(x)} & \text{if } \quad 0 \le x \le \frac{cr}{c-1} \\ 1 - 1 = 0 & \text{if } \quad i \le x \le i + 1 - \frac{cr}{c-1}, \end{cases} \end{aligned}$$

and to calculate the full Bayes probability of error, we integrate as

$$\begin{aligned} P^* &= \int P^*(e|x)p(x)dx \\ &= \int_0^{cr/(c-1)} \left[1 - \frac{1/c}{p(x)}\right] p(x)dx \\ &= \left(1 - \frac{1}{c}\right)\frac{cr}{c-1} = r. \end{aligned}$$

(b) The nearest-neighbor error rate is

$$\begin{aligned} P &= \int \left[1 - \sum_{i=1}^{c} P^2(\omega_i|x)\right] p(x)dx \\ &= \int_0^{cr/(c-1)} \left[1 - \frac{c(\frac{1}{c})^2}{p^2(x)}\right] p(x)dx + \underbrace{\sum_{j=1}^{c} \int_j^{j+1-\frac{cr}{c-1}} [1-1]\, p(x)dx}_{0} \end{aligned}$$

$$= \int_0^{cr/(c-1)} \left(1 - \frac{1/c}{p^2(x)}\right) p(x)dx$$

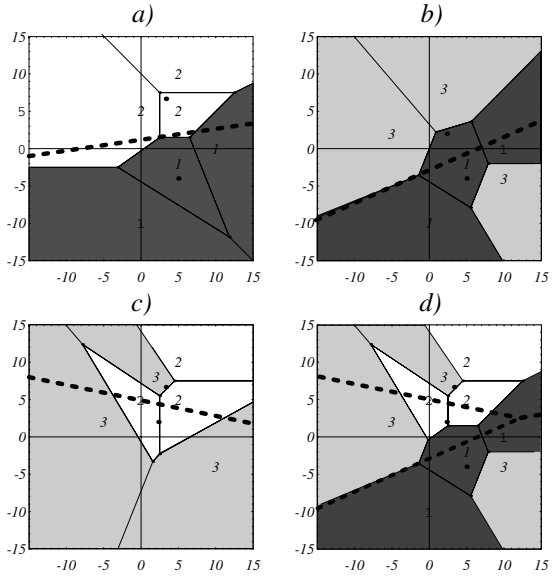
$$= \int_0^{cr/(c-1)} \left(1 - \frac{1}{c}\right) dx = \left(1 - \frac{1}{c}\right) \frac{cr}{c-1} = r.$$

Thus we have demonstrated that $P^* = P = r$ in this nontrivial case.

9. Our data are given in the table.

ω_1	ω_2	ω_3
(10,0)	(5,10)	(2,8)
(0,-10)	(0,5)	(-5,2)
(5,-2)	(5,5)	(10,-4)

Throughout the figures below, we let dark gray represent category ω_1, white represent category ω_2 and light gray represent category ω_3. The data points are labeled by their category numbers, the means are shown as small black dots, and the dashed straight lines separate the means.



10. The Voronoi diagram of n points in d-dimensional space is the same as the convex hull of those same points projected orthogonally to a hyperboloid in $(d+1)$-dimensional space. So the editing algorithm can be solved either with a Voronoi diagram algorithm in d-space or a convex hull algorithm in $(d+1)$-dimensional space. Now there are scores of algorithms available for both problems all with different complexities.

A theorem in the book by Preparata and Shamos refers to the complexity of the Voronoi diagram itself, which is of course a lower bound on the complexity of computing it. This complexity was solved by Victor Klee, "On the complexity of d-dimensional Voronoi diagrams," *Archiv. de Mathematik.*, vol. 34, 1980, pp. 75-80. The complexity formula given in this problem is the complexity of the convex hull algorithm of Raimund Seidel, "Constructing higher dimensional convex hulls at

logarithmic cost per face," *Proc. 18th ACM Conf. on the Theory of Computing*, 1986, pp. 404-413.

So here d is one bigger than for Voronoi diagrams. If we substitute d in the formula in this problem with $(d-1)$ we get the complexity of Seidel's algorithm for Voronoi diagrams, as discussed in A. Okabe, B. Boots and K. Sugihara, **Spatial Tessellations: Concepts and Applications of Voronoi Diagrams**, John Wiley, 1992.

11. Consider the "curse of dimensionality" and its relation to the separation of points randomly selected in a high-dimensional space.

(a) The sampling density goes as $n^{1/d}$, and thus if we need n_1 samples in $d = 1$ dimensions, an "equivalent" number samples in d dimensions is n_1^d. Thus if we needed 100 points in a line (i.e., $n_1 = 100$), then for $d = 20$, we would need $n_{20} = (100)^{20} = 10^{40}$ points — roughly the number of atoms in the universe.

(b) We assume roughly uniform distribution, and thus the typical inter-point Euclidean (i.e., L_2) distance δ goes as $\delta^d \sim$ volume, or $\delta \sim (\text{volume})^{1/d}$.

(c) Consider points uniformly distributed in the unit interval $0 \leq x \leq 1$. The length containing fraction p of all the points is of course p. In d dimensions, the width of a hypercube containing fraction p of points is $l_d(p) = p^{1/d}$. Thus we have

$$\begin{aligned} l_5(0.01) &= (0.01)^{1/5} &&= 0.3910 \\ l_5(0.1) &= (0.1)^{1/5} &&= 0.7248 \\ l_{20}(0.01) &= (0.01)^{1/20} &&= 0.8609 \\ l_{20}(0.1) &= (0.1)^{1/20} &&= 0.8609. \end{aligned}$$

(d) The L_∞ distance between two points in d-dimensional space is given by Eq. 57 in the text, with $k \to \infty$:

$$\begin{aligned} L_\infty(\mathbf{x}, \mathbf{y}) &= \lim_{k \to \infty} \sqrt[1/k]{\sum_{i=1}^{d} |x_i - y_i|^k} \\ &= \max[|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_d - y_d|] \\ &= \max_i |x_i - y_i|. \end{aligned}$$

In other words, consider each axis separately, $i = 1, \ldots, d$. There is a separation between two points $\mathbf{x}$ and $\mathbf{y}$ along each individual direction i, that is, $|x_i - y_i|$. One of these distances is the greatest. The L_∞ distance between two points is merely this maximum distance.

Informally we can see that for two points randomly selected in the unit d-dimensional hypercube $[0, 1]^d$, this L_∞ distance approaches 1.0 as we can nearly always find an axis i for which the separation is large. In contrast, the L_∞ distance to *the closest* of the faces of the hypercube approaches 0.0, because we can nearly always find an axis for which the distance to a face is small. Thus, nearly every point is closer to a face than to another randomly selected point. In short, nearly every point is on the "outside" (that is, on the "convex hull") of the set of points in a high-dimensional space — nearly every point is an "outlier."

We now demonstrate this result formally. Of course, $\mathbf{x}$ is always closer to a wall than 0.5 — even for $d = 1$ — and thus we consider distances l^* in the

range $0 \leq l^* < 0.5$. The figure at the left shows the x_i and y_i coordinates plotted in a unit square. There are d points, one corresponding to each axis in the hypercube. The probability that a particular single coordinate i will have $|x_i - y_i| \leq l^*$ is equal to 1.0 minus the area in the small triangles, that is,

$$\Pr[|x_i - y_i| \geq l^* \text{ for a particular } i] = (1 - l^*)^2.$$

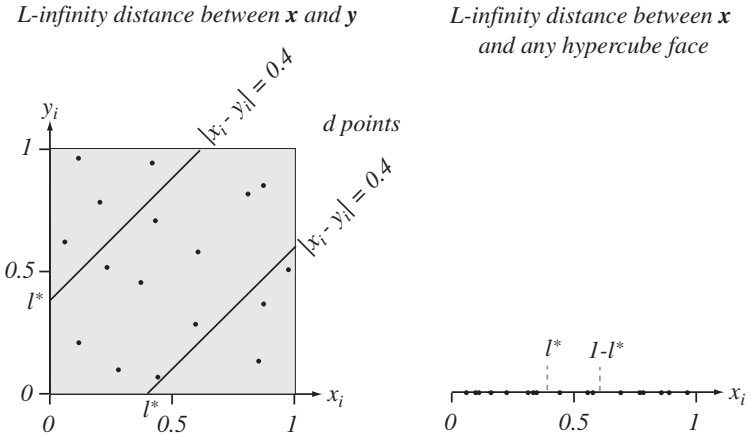
The probability that *all* of the d points will have $|x_i - y_i| \geq l^*$ is then

$$\Pr[|x_i - y_i| \geq l^* \text{ for all } i] = (1 - (1 - l^*)^2)^d.$$

The probability that at least one coordinate has separation less than l^* is just 1.0 minus the above, that is,

$$\Pr[\text{at least one coordinate is closer than } l^*] = 1 - (1 - (1 - l^*)^2)^d.$$

Now consider the distance of a point $\mathbf{x}$ to the faces of the hypercube. The figure



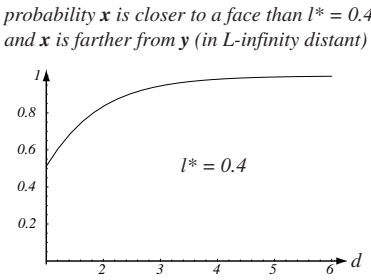
at the right shows d points corresponding to the coordinates of $\mathbf{x}$. The probability that a single particular coordinate value x_i is closer to a corresponding face (0 or 1) than a distance l^* is clearly $2l^*$ since the position of x_i is uniformly distributed in the unit interval. The probability that *any* of the coordinates is closer to a face than l^* is

$$\Pr[\text{at least one coordinate is closer to a face than } l^*] = 1 - (1 - 2l^*)^d.$$

We put these two results together to find that for a given l^* and d, the probability that $\mathbf{x}$ is closer to a wall than l^* *and* that $L_\infty(\mathbf{x}, \mathbf{y}) \leq l^*$ is the product of the two independent probabilities, that is,

$$[1 - (1 - 2l^*)^d][1 - (1 - l^*)^{2d}],$$

which approaches 1.0 quickly as a function of d, as shown in the graph (for the case $l^* = 0.4$). As shown in Fig. 4.19 in the text, the unit "hypersphere" in the L_∞ distance always encloses the unit hypersphere in the L_2 or Euclidean metric. Therefore, our discussion above is "pessimistic," that is, if $\mathbf{x}$ is closer to a face than to point $\mathbf{y}$ according to the L_∞ metric, than it is "even closer" to the wall than to a face in the L_2 metric. Thus our conclusion above holds for the Euclidean metric too. Consequently in high dimensions, we generally must *extrapolate* from sample points, rather than *interpolate* between them.



12. The "curse of dimensionality" can be "overcome" if we know the form of the target function.

(a) Our target function is linear,

$$f(\mathbf{x}) = \sum_{j=1}^{d} a_j x_j = \mathbf{a}^t \mathbf{x},$$

and $y = f(\mathbf{x}) + N(0, \sigma^2)$, that is, we have Gaussian noise. The approximation is $\hat{f}(\mathbf{x}) = \sum_{j=1}^{d} \hat{a}_j x_j$, where

$$\hat{a}_j = \arg\min_{a_j} \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{d} a_j x_{ij} \right]^2 ,$$

where x_{ij} is the jth component of point i, for $j = 1, \ldots, d$ and $i = 1, \ldots, n$. In short, these are the best fit coefficients in a sum-squared-error sense. The expected value (over the data set) of the difference between the target and fit functions (squared) is

$$\mathcal{E}[f(\mathbf{x}) - \hat{f}(\mathbf{x})]^2 \quad = \quad \mathcal{E}\left[\sum_{j=1}^{d} a_j x_j - \sum_{j=1}^{d} \left(\arg\min_{a_j} \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{d} a_j x_{ij} \right]^2 \right) x_j \right]^2 .$$

For some set of sample points $\mathbf{x}_i$, we let y_i be the corresponding sample values $y_i = f(\mathbf{x}_i) + N(0, \sigma^2)$. Consider the n-by-d matrix $\mathbf{X} = [x_{ij}]$ where x_{ij} is the jth component of x_i, that is

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}$$

where $\mathbf{x}_i$ is a d-dimensional vector. As stated above, we have $f(\mathbf{x}) = \sum_{j=1}^{d} a_j x_j$.

We define the weight vector

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

then we can write

$$\mathbf{Xa} = \begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix}$$

and hence $\mathbf{y} = \mathbf{Xa} + N(0, \sigma^2)^n$ where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

We approximate $f(\mathbf{x})$ with $\hat{f}(\mathbf{x}) = \sum_{j=1}^{d} \hat{a}_j x_j$ where the $\hat{a}_j$ are chosen to minimize

$$\sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{d} \hat{a}_j x_{ij} \right]^2 = \|\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}\|^2$$

where the definition of $\hat{\mathbf{a}}$ is self-evident. Thus we choose $\hat{\mathbf{a}}$ such that $\mathbf{X}\hat{\mathbf{a}}$ is as close to $\mathbf{y}$ as possible. However, $\mathbf{X}\hat{\mathbf{a}}$ is in $\mathbf{C_X}$, the column space of $\mathbf{X}$, that is, the d-dimensional subspace of $\mathbf{R}^n$ spanned by the columns of $\mathbf{X}$. (Of course, we assume $n > d$.)

It is clear that we wish to choose $\hat{\mathbf{a}}$ such that $\mathbf{X}\hat{\mathbf{a}}$ is the projection of $\mathbf{y}$ onto $\mathbf{C_X}$, which we denote $\text{Proj}_{\mathbf{C_X}}\mathbf{y}$. Now $\mathbf{y}$ is distributed according to an n-dimensional Gaussian distribution with mean $\mathbf{Xn}$. Projecting an n-dimensional Gaussian onto a d-dimensional subspace yields a d-dimensional Gaussian, that is,

$$\mathbf{X}\hat{\mathbf{a}} = \text{Proj}_{\mathbf{C_X}}\mathbf{y} = \text{Proj}_{\mathbf{C_X}}[\mathbf{X}_n + N(0, \sigma^2)^n] = \mathbf{Xa} + N(0, \sigma^2)^d,$$

where $N(0, \sigma^2)^d$ is rotated to lie in $\mathbf{C_X}$. Thus we have

$$\mathcal{E}\left[\|\mathbf{X}\hat{\mathbf{a}} - \mathbf{Xa}\|^2\right] = \text{Var}\left[N(0, \sigma^2)^d\right] = d\sigma^2.$$

But $\|\mathbf{X}\hat{\mathbf{a}} - \mathbf{Xa}\|^2 = \sum_{i=1}^{n} (\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2$. Since the terms $(\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2$ are independent for each i, we have

$$\mathcal{E}\left[\|\mathbf{X}\hat{\mathbf{a}} - \mathbf{Xa}\|^2\right] = n\mathcal{E}\left[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2\right] = d\sigma^2,$$

and thus

$$\mathcal{E}\left[(f(\mathbf{x}) - \hat{f}(\mathbf{x}))^2\right] = \frac{d\sigma^2}{n}.$$

In short, the squared error increases linearly with the dimension d, not exponentially as we might expect from the general curse of dimension.

(b) We follow the logic of part (a). Now our target function is $f(\mathbf{x}) = \sum_{i=1}^{d} a_i B_i(\mathbf{x})$ where each member in the basis set of M basis functions $B_i(\mathbf{x})$ is a function of the d-component vector $\mathbf{x}$. The approximation function is

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} \hat{a}_m B_m(\mathbf{x}),$$

and, as before, the coefficients are least-squares estimates

$$\hat{a}_i = \arg\min_{a_i} \sum_{i=1}^{n} \left[y_i - \sum_{m=1}^{q} a_m B_m(\mathbf{x}) \right]^2$$

and $y_i = f(\mathbf{x}_i) + N(0, \sigma^2)$. Now $\mathbf{y}$ will be approximated by $\mathbf{B}\hat{\mathbf{a}}$, the projection of $\mathbf{y}$ onto the column space of $\mathbf{B}_j$, that is, the subspace spanned by the M vectors

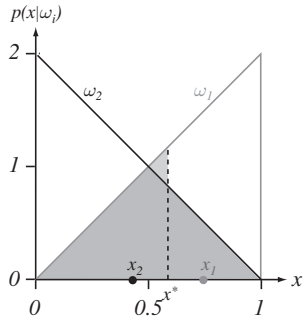
$$\begin{bmatrix} B_i(\mathbf{x}_1) \\ B_i(\mathbf{x}_2) \\ \vdots \\ B_i(\mathbf{x}_n) \end{bmatrix}.$$

As in part (a), we have

$$\mathcal{E}[(f(\mathbf{x}) - \hat{f}(\mathbf{x}))^2] = \frac{M\sigma^2}{n},$$

which is independent of d, the dimensionality of the original space.

13. We assume $P(\omega_1) = P(\omega_2) = 0.5$ and the distributions are as given in the figure.



(a) Clearly, by the symmetry of the problem, the Bayes decision boundary is $x^* = 0.5$. The error is then the area of the dark shading in the figure, divided by the total possible area, that is

$$P^* = \int_0^1 \min[P(\omega_1)p(x|\omega_1), P(\omega_2)p(x|\omega_2)]\, dx$$

$$\begin{aligned} &= P(\omega_1)\int_0^{0.5} 2x\ dx + P(\omega_2)\int_{0.5}^{1}(2-2x)\ dx \\ &= 0.5\frac{1}{4} + 0.5\frac{1}{4} = 0.25. \end{aligned}$$

(b) Suppose a point is randomly selected from ω_1 according to $p(x|\omega_1)$ and another point from ω_2 according to $p(x|\omega_2)$. We have that the error is

$$\int_0^1 dx_1\ p(x_1|\omega_1)\int_0^1 dx_2\ p(x_2|\omega_2)\int_{(x_1-x_2)/2}^{1} dx\ p(x|\omega_2)\int_0^{(x_1-x_2)/2} dx\ p(x|\omega_1).$$

(c) From part (d), below, we have for the special case $n = 2$,

$$P_2(e) = \frac{1}{3} + \frac{1}{(2+1)(2+3)} + \frac{1}{2(2+2)(2+3)} = \frac{51}{120} = 0.425.$$

(d) By symmetry, we may assume that the test point belongs to category ω_2. Then the chance of error is the chance that the nearest sample point ot the test point is in ω_1. Thus the probability of error is

$$\begin{aligned} P_n(e) &= \int_0^1 P(x|\omega_2)\Pr[\text{nearest } y_i \text{ to } x \text{ is in } \omega_1]dx \\ &= \int_0^1 P(x|\omega_2)\sum_{i=1}^{n}\Pr[y_i \in \omega_1 \text{ and } y_i \text{ is closer to } x \text{ than } y_j, \forall j \neq i]dx. \end{aligned}$$

By symmetry the summands are the same for all i, and thus we can write

$$\begin{aligned} P_n(e) &= \int_0^1 P(x|\omega_2)n\Pr[y_1 \in \omega_1 \text{ and } |y_1 - x| > |y_i - x|, \forall i > 1]dx \\ &= \int_0^1 P(x|\omega_2)n\int_0^1 P(\omega_1|y_1)\Pr[|y_i - x| > |y_1 - x|, \forall i > 1]dy_1\ dx \\ &= \int_0^1 P(x|\omega_2)n\int_0^1 P(\omega_1|y_1)\Pr[|y_2 - x| > |y_1 - x|]^{n-1}dy\ dx, \end{aligned}$$

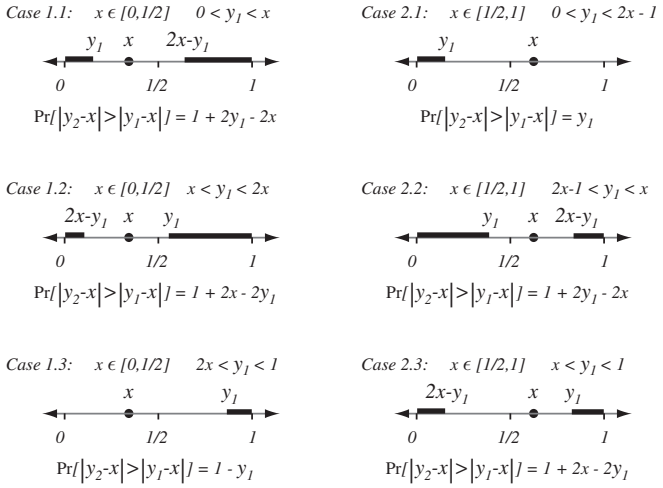
where the last step again relies on the symmetry of the problem.

To evalute $\Pr[|y_2 - x| > |y_1 - x|]$, we divide the integral into six cases, as shown in the figure.

We substitute these values into the above integral, and break the integral into the six cases as

$$P_n(e) = \int_0^{1/2} P(x|\omega_2)n\left[\int_0^x P(\omega_1|y_1)(1 + 2y_1 - 2x)^{n-1}dy_1\right.$$

denotes possible locations of y_2 with $|y_2-x|>|y_1-x|$



$$
\begin{aligned}
&+\int_x^{2x} P(\omega_1|y_1)(1+2x-2y_1)^{n-1}dy_1 \\
&+\int_{2x}^{1} P(\omega_1|y_1)(1-y_1)^{n-1}dy_1\Bigg]dx \\
+\int_{1/2}^{1} P(x|\omega_2)n\Bigg[&\int_0^{2x-1} P(\omega_1|y_1)y_1^{n-1}dy_1 \\
&+\int_{2x-1}^{x} P(\omega_1|y_1)(1+2y_1-2x)^{n-1}dy_1 \\
&+\int_x^{1} P(\omega_1|y_1)(1+2x-2y_1)^{n-1}dy_1\Bigg]dx.
\end{aligned}
$$

Our density and posterior are given as $p(x|\omega_2) = 2(1-x)$ and $P(\omega_1|y) = y$ for $x \in [0,1]$ and $y \in [0,1]$. We substitute these forms into the above large integral and find

$$
\begin{aligned}
P_n(e) \quad = \quad \int_0^{1/2} 2n(1-x)\Bigg[&\int_0^{x} y_1(1+2y_1-2x)^{n-1}dy_1 \\
&\int_x^{2x} y_1(1+2x-2y_1)^{n-1}dy_1 \\
&\int_{2x}^{1} y_1(1-y_1)^{n-1}dy_1\Bigg]dx
\end{aligned}
$$

$$+\int_{1/2}^{1} 2n(1-x)\left[\int_{0}^{2x-1} y_1^n dy_1 \right.$$
$$\int_{2x-1}^{x} y_1(1+2y_1-2x)^{n-1}dy_1$$
$$\left.\int_{x}^{1} y_1(1+2x-2y_1)^{n-1}dy_1\right]dx.$$

There are two integrals we must do twice with different bounds. The first is:

$$\int_{a}^{b} y_1(1+2y_1-2x)^{n-1}dy_1.$$

We define the function $u(y_1) = 1+2y_1-2x$, and thus $y_1 = (u+2x-1)/2$ and $dy_1 = du/2$. Then the integral is

$$\begin{aligned}\int_{a}^{b} y(1+2y_1-2x)^{n-1}dy_1 &= \frac{1}{4}\int_{u(a)}^{u(b)}(u+2x-1)u^{n-1}du \\ &= \frac{1}{4}\left[\frac{2x-1}{n}u^n + \frac{1}{n+1}u^{n+1}\right]_{u(a)}^{u(b)}.\end{aligned}$$

The second general integral is:

$$\int_{a}^{b} y_1(1+2x-2y_1)^{n-1}dy_1.$$

We define the function $u(y_1) = 1+2x-2y_1$, and thus $y_1 = (1+2x-u)/2$ and $dy_1 = -du/2$. Then the integral is

$$\begin{aligned}\int_{a}^{b} y_1(1+2x-2y_1)^{n-1}dy_1 &= -\frac{1}{4}\int_{u(a)}^{u(b)}(1+2x+u)u^{n-1}du \\ &= -\frac{1}{4}\left[\frac{2x+1}{n}u^n - \frac{1}{n+1}u^{n+1}\right]_{u(a)}^{u(b)}.\end{aligned}$$

We use these general forms to evaluate three of the six components of our full integral for $P_n(e)$:

$$\begin{aligned}\int_{0}^{x} y_1(1+2y_1-2x)^{n-1}dy_1 &= \frac{1}{4}\left[\frac{2x-1}{n}u^n - \frac{1}{n+1}u^{n+1}\right]_{1=u(x)}^{1-2x=u(0)} \\ &= \frac{1}{4}\left(\frac{2x+1}{n}+\frac{1}{n+1}\right) + \frac{1}{4}(1-2x)^{n+1}\left(\frac{1}{n}-\frac{1}{n+1}\right)\end{aligned}$$

$$\begin{aligned}
\int_x^{2x} y_1(1+2x-2y_1)^{n-1}dy_1 &= -\frac{1}{4}\left[\frac{2x+1}{n}u^n - \frac{1}{n+1}u^{n+1}\right]_{1=u(x)}^{1-2x=u(2x)} \\
&= \frac{1}{4}\left(\frac{2x+1}{n} - \frac{1}{n+1}\right) - \frac{1}{2n}(1-2x)^n + \frac{1}{4}(1-2x)^{n+1}\left(\frac{1}{n}+\frac{1}{n+1}\right) \\
\int_{2x}^{1} y_1(1-y_1)^{n-1}dy_1 &= \int_0^{1-2x}(1-u)u^{n-1}du = \left[\frac{1}{n}u^n - \frac{1}{n+1}u^{n+1}\right]_0^{1-2x} \\
&= \frac{1}{n}(1-2x)^n - \frac{1}{n+1}(1-2x)^{n+1}.
\end{aligned}$$

We add these three integrals together and find, after a straightforward calculation that the sum is

$$\frac{x}{n} + \frac{1}{2n}(1-2x)^n + \left(\frac{1}{2n} - \frac{1}{n+1}\right)(1-2x)^{n+1}. \qquad (*)$$

We next turn to the three remaining three of the six components of our full integral for $P_n(e)$:

$$\begin{aligned}
\int_0^{2x-1} y_1^n dy_1 &= \frac{1}{n+1}(2x-1)^{n+1} \\
\int_{2x-1}^{x} y_1(1+2y_1-2x)^{n-1}dy_1 &= \frac{1}{4}\left[\frac{2x-1}{n}u^n + \frac{1}{n+1}u^{n+1}\right]_{2x-1=u(2x-1)}^{1=u(x)} \\
&= \frac{1}{4}\left(\frac{2x-1}{n} + \frac{1}{n+1}\right) - \frac{1}{4}(2x-1)^{n+1}\left(\frac{1}{n}+\frac{1}{n+1}\right) \\
\int_x^{1} y_1(1+2x-2y_1)^{n-1}dy_1 &= -\frac{1}{4}\left[\frac{2x+1}{n}u^n - \frac{1}{n+1}u^{n+1}\right]_{1=u(x)}^{2x-1=u(1)} \\
&= \frac{1}{4}\left(\frac{2x+1}{n} - \frac{1}{n+1}\right) - \frac{1}{2n}(2x-1)^n - \frac{1}{4}(2x-1)^{n+1}\left(\frac{1}{n}-\frac{1}{n+1}\right.
\end{aligned}$$

The sum of these latter three terms is

$$\frac{x}{n} - \frac{1}{2n}(2x-1)^n - \left(\frac{1}{2n} - \frac{1}{n+1}\right)(2x-1)^{n+1}. \qquad (**)$$

Thus the sum of the three Case 1 terms is

$$\begin{aligned}
\int_0^{1/2} 2n(1-x)\Bigg[&\int_0^x y_1(1+2y_1-2x)^{n-1}dy_1 \\
&+ \int_x^{2x} y_1(1+2x-2y_1)^{n-1}dy_1 \\
&+ \int_{2x}^{1} y_1(1-y_1)^{n-1}dy_1\Bigg]dx
\end{aligned}$$

$$= \int_0^{1/2} 2n(1-x)\left[\frac{x}{n} + \frac{1}{2n}(1-2x)^n + (1-2x)^{n+1}\left(\frac{1}{2n} - \frac{1}{n+1}\right)\right] dx$$

$$= \int_0^1 n\left(\frac{1+u}{2}\right)\left[\frac{1-u}{2n} + \frac{1}{2n}u^n + \left(\frac{1}{2n} - \frac{1}{n+1}\right)u^{n+1}\right] du,$$

where we used the substitution $u = 1 - 2x$, and thus $x = (1-u)/2$, and $dx = -du/2$ and $1 - x = (1+u)/2$.

Likewise, we have for the three Case 2 terms

$$\int_{1/2}^1 2n(1-x)\left[\int_0^{2x-1} y_1^n dy_1 + \int_{2x-1}^x y_1(1+2y_1-2x)^{n-1} dy_1 + \int_x^1 y_1(1+2x-2y_1)^{n-1} dy_1\right] dx$$

$$= \int_{1/2}^1 2n(1-x)\left[\frac{x}{n} - \frac{1}{2n}(2x-1)^n - (2x-1)^{n+1}\left(\frac{1}{2n} - \frac{1}{n+1}\right)\right] dx$$

$$= \int_0^1 n\left(\frac{1-u}{2}\right)\left[\frac{1+u}{2n} - \frac{1}{2n}u^n - \left(\frac{1}{2n} - \frac{1}{n+1}\right)u^{n+1}\right] du,$$

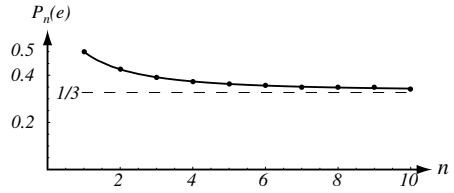
where we used the substitution $u = 2x-1$, and thus $x = (1+u)/2$ and $dx = du/2$ and $1 - x = (1-u)/2$.

Now we are ready to put all these results together:

$$\begin{aligned}
P_n(e) &= \int_0^1 n\left(\frac{1+u}{2}\right)\left[\frac{1-u}{2n} + \frac{1}{2n}u^n + \left(\frac{1}{2n} - \frac{1}{n+1}\right)u^{n+1}\right] du \\
&\quad + \int_0^1 n\left(\frac{1-u}{2}\right)\left[\frac{1+u}{2n} - \frac{1}{2n}u^n - \left(\frac{1}{2n} - \frac{1}{n+1}\right)u^{n+1}\right] du \\
&= n\int_0^1 \left[\frac{1}{2n}(1-u^2) + \frac{1}{2n}u^{n+1} + \left(\frac{1}{2n} - \frac{1}{n+1}\right)u^{n+2}\right] du \\
&= n\left[\frac{1}{2n}\left(u - \frac{u^3}{3}\right) \frac{1}{2n(n+2}u^{n+2} + \frac{1}{n+3}\left(\frac{1}{2n} - \frac{1}{n+1}\right)u^{n+3}\right]_0^1 \\
&= n\left[\frac{1}{2n}\frac{2}{3} + \frac{1}{2n(n+2)} + \frac{1}{n+3}\left(\frac{1-n}{2n(n+1)}\right)\right] \\
&= \frac{1}{3} + \frac{(n+1)(n+3) - (n-1)(n+2)}{2(n+1)(n+2)(n+3)}
\end{aligned}$$

$$\begin{aligned} &= \frac{1}{3} + \frac{3n+5}{2(n+1)(n+2)(n+3)} \\ &= \frac{1}{3} + \frac{1}{(n+1)(n+3)} + \frac{1}{2(n+2)(n+3)}, \end{aligned}$$

which decays as $1/n^2$, as shown in the figure.



We may check this result for the case $n = 1$ where there is only one sample. Of course, the error in that case is $P_1(e) = 1/2$, since the true label on the test point may either match or mismatch that of the single sample point with equal probability. The above formula above confirms this

$$\begin{aligned} P_1(e) &= \frac{1}{3} + \frac{1}{(1+1)(1+3)} + \frac{1}{2(1+2)(1+3)} \\ &= \frac{1}{3} + \frac{1}{8} + \frac{1}{24} = \frac{1}{2}. \end{aligned}$$

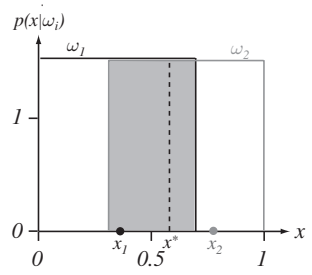
(e) The limit for infinite data is simply

$$\lim_{n\to\infty} P_n(e) = \frac{1}{3},$$

which is larger than the Bayes error, as indeed it must be. In fact, this solution also illustrates the bounds of Eq. 52 in the text:

$$\begin{aligned} P^* &\leq P \leq P^*(2 - 2P^*) \\ \frac{1}{4} &\leq \frac{1}{3} \leq \frac{3}{8}. \end{aligned}$$

14. We assume $P(\omega_1) = P(\omega_2) = 0.5$ and the distributions are as given in the figure.

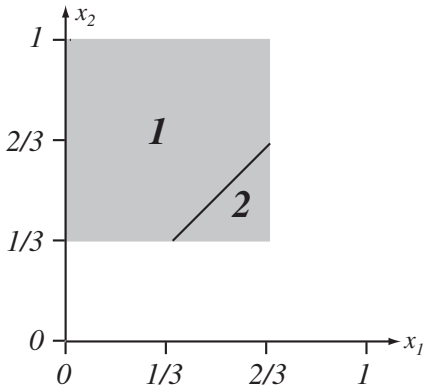


(a) This is a somewhat unusual problem in that the Bayes decision can be any point $1/3 \leq x^* \leq 2/3$. For simplicity, we can take $x^* = 1/3$. Then the Bayes error is

then simply

$$\begin{aligned} P^* &= \int_0^1 \min[P(\omega_1)p(x|\omega_1), P(\omega_2)p(x|\omega_2)]dx \\ &= \int_{1/3}^{2/3} P(\omega_1)p(x|\omega_1)dx \\ &= 0.5(1/3)(3/2) = 0.25. \end{aligned}$$

(b) The shaded area in the figure shows the possible (and equally likely) values of a point x_1 chosen from $p(x|\omega_1)$ and a point x_2 chosen from $p(x|\omega_2)$.



There are two functionally separate cases, as numbered, corresponding to the position of the decision boundary $x^* = (x_1 + x_2)/2$. (Note that we will also have to consider which is larger, x_1 or x_2. We now turn to the decision rule and probability of error in the single nearest-neighbor classifier in these two cases:

case 1 : $x_2 \geq x_1$ and $1/3 \leq (x_1 + x_2)/2 \leq 2/3$: Here the decision point x^* is between 1/3 and 2/3, with $\mathcal{R}_2$ at large values of x. This is just the Bayes case described in part (a) and the error rate is thus 0.25, as we saw. The relative probability of **case 2** occuring is the relative area of the gray region, that is, 7/8.

case 2 : $x_1 \geq x_2$ and $1/3 \leq (x_1 + x_2)/2 \leq 2/3$: Here the decision boundary is between 1/3 and 2/3 (in the Bayes region) but note especially that $\mathcal{R}_1$ is for large values of x, that is, the decision is the opposite of the Bayes decision. Thus the error is 1.0 minus the Bayes error, or 0.75. The relative probability of **case 2** occuring is the relative area of the gray region, that is, 1/8.

We calculate the average error rate in the case of one point from each category by merely adding the probability of occurrence of each of the three cases (proportional to the area in the figure), times the expected error given that case, that is,

$$P_1 = \frac{7}{8}0.25 + \frac{1}{8}0.75 = \frac{5}{16} = 0.3125,$$

which is of course greater than the Bayes error.

(c) Problem not yet solved

(d) Problem not yet solved

(e) In the limit $n \to \infty$, every test point x in the range $0 \leq x \leq 1/3$ will be properly classified as ω_1 and every point in the range $2/3 \leq x \leq 1$ will be properly classified as ω_2. Test points in the range $1/3 \leq x \leq 2/3$ will be misclassified half of the time, of course. Thus the expected error in the $n \to \infty$ case is

$$\begin{aligned} P_\infty &= P(\omega_1)\Pr[0 \leq x \leq 1/3|\omega_1] \cdot 0 + P(\omega_1)\Pr[1/3 \leq x \leq 2/3|\omega_1] \cdot 0.5 \\ &\quad + P(\omega_2)\Pr[1/3 \leq x \leq 2/3|\omega_2] \cdot 0.5 + P(\omega_2)\Pr[2/3 \leq x \leq 1|\omega_2] \cdot 0 \\ &= 0.5\ 0.5\ 0.5 + 0.5\ 0.5\ 0.5 = 0.25. \end{aligned}$$

Note that this is the same as the Bayes rate. This problem is closely related to the "zero information" case, where the posterior probabilities of the two categories are equal over a range of x. If the problem specified that the distributions were equal throughout the full range of x, then the Bayes error and the P_∞ errors would equal 0.5.

15. An faster version of Algorithm 3 in the text deletes prototypes as follows:

Algorithm 0 (Faster nearest-neighbor)

1 **begin initialize** $j \leftarrow 0, \mathcal{D}, n =$ number of prototypes
2 Construct the full Voronoi diagram of $\mathcal{D}$
3 **do** $j \leftarrow j + 1$ (for each prototype $\mathbf{x}'_j$)
4 **if** $\mathbf{x}'_j$ is not marked **then** find the Voronoi neighbors of $\mathbf{x}'_j$
5 **if** any neighbor is not from the$\mathbf{x}'_j$class **then** mark $\mathbf{x}'_j$ and its neighbors in other classes
6 **until** $j = n$
7 Discard all unmarked prototypes
8 **return**Voronoi diagram of the remaining (marked) prototypes
9 **end**

If we have k Voronoi neighbors on average of any point $\mathbf{x}'_j$, then the probability that i out of these k neighbors are not from the same class as $\mathbf{x}'_j$ is given by the binomial law:

$$P(i) = \binom{k}{i}(1 - 1/c)^k(1/c)^{k-1},$$

where we have assumed that all the classes have the same prior probability. Then the expected number of neighbors of any point $\mathbf{x}'_j$ belonging to different class is

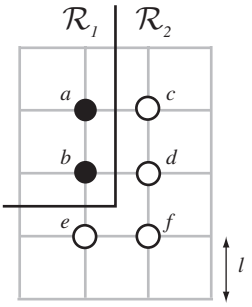
$$E(i) = k(1 - 1/c).$$

Since each time we find a prototype to delete we will remove $k(1-1/c)$ more prototypes on average, we will be able to speed up the search by a factor $k(1 - 1/c)$.

16. Consider Algorithm 3 in the text.

(a) In the figure, the training points (black for ω_1, white for ω_2) are constrained to the intersections of a two-dimensional grid. Note that prototype f does not contribute to the class boundary due to the existence of points e and d. Hence f should be removed from the set of prototypes by the editing algorithm (Algorithm 3 in the text). However, this algorithm detects that f has a prototype

from another class (prototype c) as a neighbor, and thus f is retained according to step 5 in the algorithm. Consequently, the editing algorithm does not only select the points a, b, c, d and e which are are the minimum set of points, but it also retains the "useless" prototype f.



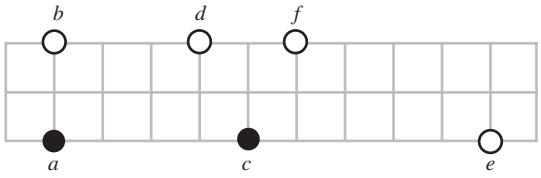
(b) A sequential editing algorithm in which each point is considered in turn and retained or rejected before the next point is considered is as follows:

Algorithm 0 (Sequential editing)

```
1   begin initialize j ← 0, D, n = number of prototypes
2         do j ← j + 1
3            Remove x'_j from D
4         if x'_j is not well classified by D, then restore x'_j to D
5           until J = n
6         return D
7     end
```

This sequential editing algorithm picks up one training sample and checks whether it can be correctly classified with the remaining prototypes. If an error is detected, the prototype under consideration must be kept, since it may contribute to the class boundaries. Otherwise it may be removed. This procedure is repeated for all the training patterns.

Given a set of training points, the solution computed by the sequential editing algorithm is not unique, since it clearly depends upon the order in which the data are presented. This can be seen in the following simple example in the figure, where black points are in ω_1, and white points in ω_2. If d is the first point



presented to the editing algorithm, then it will be removed, since its nearest-neighbor is f so d can be correctly classified. Then, the other points are kept, except e, which can be removed since f is also its nearest neighbor. Suppose now that the first point to be considered is f. Then, f will be removed since d is its nearest neighbor. However, point e will not be deleted once f has been

removed, due to c, which will be its nearest neighbor. According to the above considerations, the algorithm will return points a, b, c, f in the first ordering, and will return points a, b, c, d, e for the second ordering.

17. We have that $P(\omega_i) = 1/c$ and $p(\mathbf{x}|\omega_i) = p(\mathbf{x})$ for $i = 1, \ldots, c$. Thus the probability density of finding $\mathbf{x}$ is

$$p(\mathbf{x}) = \sum_{i=1}^{c} p(\mathbf{x}|\omega_i)P(\omega_i) = \sum_{i=1}^{c} p(\mathbf{x}|\omega)\frac{1}{c} = p(\mathbf{x}|\omega),$$

and accordingly

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\omega)1/c}{p(\mathbf{x}|\omega)} = \frac{1}{c}. \qquad (*)$$

We use $(*)$ in Eq. 45 in the text to find that the asymptotic nearest-neighbor error rate is

$$\begin{aligned} P &= \lim_{n\to\infty} P_n(e) \\ &= \int \left[1 - \sum_{i=1}^{c} P^2(\omega_i|\mathbf{x})\right] p(\mathbf{x})d\mathbf{x} \\ &= \int \left[1 - \sum_{i=1}^{c} \frac{1}{c^2}\right] p(\mathbf{x})d\mathbf{x} \\ &= \left(1 - \frac{1}{c}\right)\int p(\mathbf{x})d\mathbf{x} = 1 - \frac{1}{c}. \end{aligned}$$

On the other hand, the Bayes error is

$$\begin{aligned} P^* &= \int P^*(error|\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \sum_{i=1}^{c} \int_{\mathcal{R}_i} [1 - P(\omega_i|\mathbf{x})]p(\mathbf{x})d\mathbf{x}, \qquad (**) \end{aligned}$$

where $\mathcal{R}_i$ denotes the Bayes decision regions for class ω_i. We now substitute $(*)$ into $(**)$ and find

$$P^* = \sum_{i=1}^{c} \int_{\mathcal{R}_i} \left(1 - \frac{1}{c}\right) p(\mathbf{x})d\mathbf{x} = \left(1 - \frac{1}{c}\right)\int p(\mathbf{x})d\mathbf{x} = 1 - \frac{1}{c}.$$

Thus we have $P = P^*$, which agrees with the upper bound $P^*(2 - \frac{c}{c-1}P^*) = 1 - 1/c$ in this "no information" case.

18. The probability of error of a classifier, and the k-nearest-neighbor classifier in particular, is

$$P(e) = \int p(e|\mathbf{x})d\mathbf{x}.$$

In a two-class case, where $P(\omega_1|\mathbf{x}) + P(\omega_2|\mathbf{x}) = 1$, we have

$$P(e|\mathbf{x}) = P(e|\mathbf{x}, \omega_1)P(\omega_1|\mathbf{x}) + P(e|\mathbf{x}, \omega_2)P(\omega_2|\mathbf{x}).$$

The probability of error given a pattern $\mathbf{x}$ which belongs to class ω_1 can be computed as the probability that the number of nearest neighbors which belong to ω_1 (which we denote by a) is less than $k/2$ for k odd. Thus we have

$$P(e|\mathbf{x},\omega_1) = Pr[a \leq (k-1)/2].$$

The above probability is the sum of the probabilities of finding $a = 0$ to $a = (k-1)/2$, that is,

$$\Pr[a \leq (k-1)/2] = \sum_{i=0}^{(k-1)/2} \Pr[a = i].$$

The probability of finding i nearest prototopyes which belong to ω_i among k is $P(\omega_1|\mathbf{x})^i P(\omega_2|\mathbf{x})^{k-i}$ multiplied by the number of possible combinations, $\binom{k}{i}$; that is,

$$P(e|\mathbf{x},\omega_1) = \sum_{i=0}^{(k-1)/2} \binom{k}{i} P(\omega_1|\mathbf{x})^i P(\omega_2|\mathbf{x})^{k-i}.$$

By a simple interchange $\omega_1 \leftrightarrow \omega_2$, we find

$$P(e|\mathbf{x},\omega_2) = \sum_{i=0}^{(k-1)/2} \binom{k}{i} P(\omega_2|\mathbf{x})^i P(\omega_1|\mathbf{x})^{k-i}.$$

We put these results together and find

$$P(e|\mathbf{x}) = \sum_{i=0}^{(k-1)/2} \binom{k}{i} \left[P(\omega_1|\mathbf{x})^{i+1}[1-P(\omega_1|\mathbf{x})]^{k-i} + P(\omega_1|\mathbf{x})^{k-i}[1-P(\omega_1|\mathbf{x})]^{i+1}\right].$$

Recall that the conditional Bayes error rate is

$$P^*(e|\mathbf{x}) = \min[P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})].$$

Hence we can write $P(e|\mathbf{x})$ as

$$\begin{aligned} P(e|\mathbf{x}) &= \sum_{i=0}^{(k-1)/2} \binom{k}{i} \left[P^*(e|\mathbf{x})^{i+1}[1-P^*(e|\mathbf{x}))]^{k-i} + P^*(e|\mathbf{x})^{k-i}[1-P^*(e|\mathbf{x})]^{i+1}\right] \\ &= f_k[P(e|\mathbf{x})], \end{aligned}$$

where we have defined f_k to show the dependency upon k. This is Eq. 54 in the text.

Now we denote $C_k[P^*(e|\mathbf{x})]$ as the smallest concave function of $P^*(e|\mathbf{x})$ greater than $f_k[P^*(e|\mathbf{x})]$, that is,

$$f_k[P^*(e|\mathbf{x})] \leq C_k[P^*(e|\mathbf{x})].$$

Since this is true at any $\mathbf{x}$, we can integrate and find

$$P(e) = \int f_k[P^*(e|\mathbf{x})]p(\mathbf{x})d\mathbf{x} \leq \int C_k[P^*(e|\mathbf{x})]p(\mathbf{x})d\mathbf{x}.$$

Jensen's inequality here states that under very broad conditions, $\mathcal{E}_{\mathbf{x}}[u(\mathbf{x})] \leq u[\mathcal{E}[\mathbf{x}]]$. We apply Jensen's inequality to the expression for $P(e)$ and find

$$P(e) \leq \int C_k[P^*(e|\mathbf{x})]p(\mathbf{x})d\mathbf{x} \leq C_k\left[\int P^*(e|\mathbf{x})p(\mathbf{x})d\mathbf{x}\right] = C_k[P^*],$$

where again P^* is the Bayes error. In short, we see $P(e) \leq C_k[P^*]$, and from above $f_k[P^*] \leq C_k[P^*]$.

Section 4.6

19. We must show that the use of the distance measure

$$D(\mathbf{a},\mathbf{b}) = \sqrt{\sum_{k=1}^{d}\alpha_k(a_k - b_k)^2}$$

for $\alpha_k > 0$ generates a metric space. This happens if and only if $D(\mathbf{a},\mathbf{b})$ is a metric. Thus we first check whether D obeys the properties of a metric. First, D must be non-negative, that is, $D(\mathbf{a},\mathbf{b}) \geq 0$, which indeed holds because the sum of squares is non-negative. Next we consider reflexivity, that is, $D(\mathbf{a},\mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$. Clearly, $D = 0$ if and only if each term in the sum vanishes, and thus indeed $\mathbf{a} = \mathbf{b}$. Conversely, if $\mathbf{a} = \mathbf{b}$, then $D = 0$. Next we consider symmetry, that is, $D(\mathbf{a},\mathbf{b}) = D(\mathbf{b},\mathbf{a})$ for all $\mathbf{a}$ and $\mathbf{b}$. We can alter the order of the elements in the squared sum without affecting its value, that is, $(a_k - b_k)^2 = (b_k - a_k)^2$. Thus symmetry holds. The last property is the triangle inequality, that is,

$$D(\mathbf{a},\mathbf{b}) + D(\mathbf{b},\mathbf{c}) \geq D(\mathbf{a},\mathbf{c})$$

for all $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. Note that D can be written

$$D(\mathbf{a},\mathbf{b}) = \|\mathbf{A}(\mathbf{a} - \mathbf{b})\|_2$$

where $\mathbf{A} = \text{diag}[\alpha_1, \ldots, \alpha_d]$. We let $\mathbf{x} = \mathbf{a} - \mathbf{c}$ and $\mathbf{y} = \mathbf{c} - \mathbf{b}$, and then the triangle inequality test can be written

$$\|\mathbf{A}\mathbf{x}\|_2 + \|\mathbf{A}\mathbf{y}\|_2 \geq \|\mathbf{A}(\mathbf{x} + \mathbf{y})\|_2.$$

We denote $\mathbf{x}' = \mathbf{A}\mathbf{x}$ and $\mathbf{y}' = \mathbf{A}\mathbf{y}$, as the feature vectors computed using the matrix $\mathbf{A}$ from the input space. Then, the above inequality gives

$$\|\mathbf{x}'\|_2 + \|\mathbf{y}'\|_2 \geq \|\mathbf{x}' + \mathbf{y}'\|_2.$$

We can square both sides of this equation and see

$$\left(\|\mathbf{x}'\|_2 + \|\mathbf{y}'\|_2\right)^2 = \|\mathbf{x}'\|_2^2 + \|\mathbf{y}'\|_2^2 + 2\|\mathbf{x}'\|_2\|\mathbf{y}'\|_2 \geq \|\mathbf{x}'\|_2^2 + \|\mathbf{y}'\|_2^2 + 2\sum_{i=1}^{d} x_i' y_i' = \|\mathbf{x}' + \mathbf{y}'\|_2,$$

which is equivalent to the simple test

$$\|\mathbf{x}'\|_2\|\mathbf{y}'\|_2 \geq \sum_{i=1}^{d} x_i' y_i'.$$

Observe that the above inequality if fulfilled, since the Cauchy-Schwarz inequality states that

$$\|\mathbf{x}'\|_2\|\mathbf{y}'\|_2 \geq \left|\sum_{i=1}^{d} x_i' y_i'\right|.$$

If we work with metric spaces in the nearest-neighbor method, we can ensure the existence of best approximations in this space. In other words, there will always be a stored prototype $\mathbf{p}$ of minimum distance from an input pattern $\mathbf{x}^*$. Let

$$\delta = \min_{\mathbf{x}} D(\mathbf{p}, \mathbf{x}),$$

where $\mathbf{x}$ belongs to a set $\mathcal{S}$ of the metric space which is compact. (A subset $\mathcal{S}$ is said to be compact if every sequence of points in $\mathcal{S}$ has a sequence that converges to a point in $\mathcal{S}$.) Suppse we define a sequence of points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\}$ with the property

$$D(\mathbf{p}, \mathbf{x}_n) \to \delta \text{ as } n \to \infty,$$

where $\mathbf{x}_n \to \mathbf{x}^*$ using the compactness of $\mathcal{S}$. By the triangle inequality, we have that

$$D(\mathbf{p}, \mathbf{x}_n) + D(\mathbf{x}_n, \mathbf{x}^*) \geq D(\mathbf{p}, \mathbf{x}^*).$$

The right-hand side of the inequality does not depend on n, and the left side approches δ as $n \to \infty$. Thus we have $\delta \geq D(\mathbf{p}, \mathbf{x}^*)$. Nevertheless, we also have $D(\mathbf{p}, \mathbf{x}^*) \geq \delta$ because $\mathbf{x}^*$ belongs to $\mathcal{S}$. Hence we have that $D(\mathbf{p}, \mathbf{x}^*) = \delta$.

The import of this property for nearest-neighbor classifiers is that given enough prototypes, we can find a prototype $\mathbf{p}$ very close to an input pattern $\mathbf{x}^*$. Then, a good approximation of the posterior class probability of the input pattern can be achieved with that of the stored prototype, that is, $\mathbf{p} \simeq \mathbf{x}^*$ implies

$$P(\omega_i|\mathbf{p}) \simeq P(\omega_i|\mathbf{x}^*).$$

20. We must prove for

$$L_k(\mathbf{a}, \mathbf{b}) = \left(\sum_{i=1}^{d} |a_i - b_i|^k\right)^{1/k} = \|\mathbf{a} - \mathbf{b}\|_k$$

the four properties a metric are obeyed.

- The first states that L_k must be positive definite. Since $|a_i - b_i| \geq 0$ for all i, $L_k(\mathbf{a}, \mathbf{b}) \geq 0$.

- The second states that $L_k(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$. If $\mathbf{a} = \mathbf{b}$, then each term of the sum $|a_i - b_i|$ is 0, and $L_k(\mathbf{a}, \mathbf{b}) = 0$. Conversely, if $L_k(\mathbf{a}, \mathbf{b}) = 0$, then each term $|a_i - b_i|$ must be 0, and thus $a_i = b_i$ for all i, that is, $\mathbf{a} = \mathbf{b}$.

- The third condition is symmetry, that is $L_k(\mathbf{a}, \mathbf{b}) = L_k(\mathbf{b}, \mathbf{a})$. This follows directly from the fact

$$|a_i - b_i| = |-(b_i - a_i)| = |-1|\ |b_i - a_i| = |b_i - a_i| \quad \text{for } i = 1, \ldots, d.$$

- The last property is the triangle inequality, that is

$$L_k(\mathbf{a},\mathbf{b}) + L_k(\mathbf{b},\mathbf{c}) \geq L_k(\mathbf{a},\mathbf{c})$$

or

$$\|\mathbf{a}-\mathbf{b}\|_k + \|\mathbf{b}-\mathbf{c}\|_k \geq \|\mathbf{a}-\mathbf{c}\|_k \qquad (*)$$

for arbitrary $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. We define $\mathbf{a}-\mathbf{b}=\mathbf{x}$ and $\mathbf{b}-\mathbf{c}=\mathbf{y}$, then $(*)$ can be written

$$\|\mathbf{x}\|_k + \|\mathbf{y}\|_k \quad \geq \|\mathbf{x}+\mathbf{y}\|_k. \qquad (**)$$

We exponentiate both sides and find an equivalent condition

$$\left(\|\mathbf{x}\|_k + \|\mathbf{y}\|_k\right)^k \quad \geq \left(\|\mathbf{x}+\mathbf{y}\|_k\right)^k \qquad (***)$$

for $k>1$. We expand the left-hand side of $(***)$ to find

$$\left(\|\mathbf{x}\|_k + \|\mathbf{y}\|_k\right)^k = \sum_{i=1}^{d}|x_i|^k + \sum_{i=1}^{d}|y_i|^k + \sum_{j=1}^{k-1}\binom{k}{j}\|\mathbf{x}\|_k^{k-j}\cdot\|\mathbf{y}\|_k^j,$$

and the right-hand side to find

$$\|\mathbf{x}+\mathbf{y}\|_k^k = \sum_{i=1}^{d}|x_i|^k + \sum_{i=1}^{d}|b_i|^k + \sum_{j=1}^{k-1}\binom{k}{j}\sum_{i=1}^{d}|x_i|^{k-j}|b_i|^j.$$

In short, then $(***)$ can be written as

$$\|\mathbf{x}\|_k^{k-j}\cdot\|\mathbf{y}\|_k^j \geq \sum_{i=1}^{d}|x_i|^{k-j}|b_i|^j \qquad j=1,\ldots,k-1.$$

Note that

$$\|\mathbf{x}\|_k^{k-j} \geq \sum_{i=1}^{d}|x_i|^{k-j} \quad \text{and} \quad \|\mathbf{y}\|_k^j \geq \sum_{i=1}^{d}|y_i|^j$$

because

$$\left(\sum a\right)^q \geq \sum a^q \qquad \text{for } a>0.$$

Then we have for the above special case

$$\|\mathbf{x}\|_k^{k-j}\|\mathbf{y}\|_k^j \geq \left(\sum_{i=1}^{d}|x_i|^{k-j}\right)\left(\sum_{i=1}^{d}|y_i|^j\right).$$

Since

$$\left(\sum_i a_i\right)\left(\sum_i b_i\right) \geq \sum_i a_i b_i \quad \text{for } a_i, b_i > 0,$$

we have

$$\|\mathbf{x}\|_k^{k-j}\|\mathbf{y}\|_k^j \geq \left(\sum_{i=1}^{d}|x_i|^{k-j}\right)\left(\sum_{i=1}^{d}|y_i|^j\right) \geq \sum_{i=1}^{d}|x_i|^{k-j}|b_i|^j.$$

We replace $\mathbf{x} = \mathbf{a} - \mathbf{b}$ and $\mathbf{y} = \mathbf{b} - \mathbf{c}$ and conclude

$$L_k(\mathbf{a},\mathbf{b}) + L_k(\mathbf{b},\mathbf{c}) \geq L_k(\mathbf{a},\mathbf{c}),$$

and hence the triangle inequality is obeyed.

21. PROBLEM NOT YET SOLVED
22. PROBLEM NOT YET SOLVED
23. PROBLEM NOT YET SOLVED
24. PROBLEM NOT YET SOLVED
25. Consider the computational complexity of calculating distances under different metrics.

(a) The Euclidean distance between two prototypes is given by

$$\|\mathbf{x}'' - \mathbf{x}'\|^2 = \sum_{i=1}^{d}(x_i'' - x_i')^2,$$

where $\mathbf{x}''$ denotes the transformed image, $\mathbf{x}'$ is a stored prototype, and d is the number of pixels in the handwritten digit. As we can see, d subtractions, d multiplications and d additions are needed. Consequently, this is an $O(d)$ process.

(b) Given a text sample $\mathbf{x}$ and a prototype $\mathbf{x}'$, we must compute r non-linear transforms, which depend on a set of parameters subject to optimization. If an iterative algorithm is used for optimizing these parameters, we will have to compute the Euclidean distance of the transformed image and the stored prototype $\mathbf{x}'$ for each step. Since the computation of the transformed image invlves r non-linear transformations (with $a_i k^2$ operations per transform), the number of operations required is

$$a_i k^2 r + 3k^2 = (a_i r + 3)k^2.$$

(c) For each prototype, we will have to perform A searches, and this implies $(a_i r + 3)k^2 A$ operations. Accordingly, the total number of operations is

$$(a_i r + 3)k^2 An.$$

(d) If $n = 10^6$, $r = 6$, $a_i \simeq 10$, $A \simeq 5$, and basic operations on our computer require 10^{-9} *second*, then the classification of a single point is performed in

$$(10 \cdot 6 + 3)k^2 \cdot 5 \cdot 10^6 operations \cdot \frac{10^{-9}\ seconds}{operation} = 0.315k^2 seconds.$$

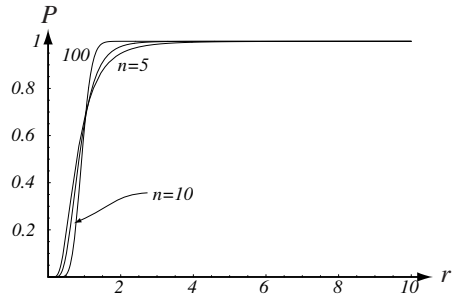
Suppose that $k = 64$. Then a test sample will be classified in 21 *minutes* and 30 *seconds*.

26. Explore the effect of r on the accuracy of nearest-neighbor search based on partial distance.

(a) We make some simplifications, specifically that points are uniformly distributed in a d-dimensional unit hypercube. The closest point to a given test point in, on average, enclosed in a small hypercube whose volume is $1/n$ times the volume of the full space. The length of a side of such a hypercube is thus $1/n^{1/d}$. Consider a single dimension. The probability that any point falls within the projection of the hypercube is thus $1/n^{1/d}$. Now in $r \leq d$ dimensions, the probability of falling inside the projection of the small hypercube onto r-dimensional subspace is $1/n^{r/d}$, i.e., the product of r indpendendent events, each having probability $1/n^{1/d}$. So, for each point the probability it will *not* fall inside the volume in the r-dimensional space is $1 - 1/n^{r/d}$. Since we have n indpendent such points, the probability that the partial distance in r dimensions will give us the true nearest neighbor is

$$Pr[\text{nearest neighbor in } d-\text{dim. is found in } r < d \text{ dim.}] = \left(1 - \frac{1}{n^{r/d}}\right)^n,$$

as shown in the figure for $0 < r \leq d = 10$ and $n = 5, 10$ and 100.



(b) The Bayes decision boundary in one dimenion is clearly $x_1^* = 0.5$. In two dimensions it is the line $x_2^* = 1 - x_1$. In three dimension the boundardy occurs when

$$x_1 x_2 x_3 = (1 - x_1)(1 - x_2)(1 - x_3)$$

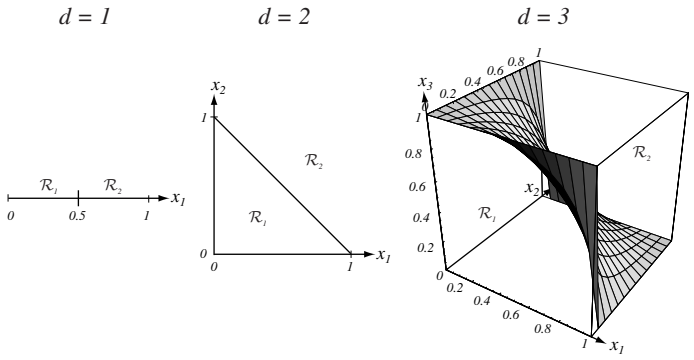
which implies

$$x_3^* = \frac{(x_1 - 1)(x_2 - 1)}{1 - x_1 - x_2 + 2x_1 x_2},$$

as shown in the figure.

(c) The Bayes error in one dimension is clearly

$$E_B = \int_0^1 \text{Min}[P(\omega_1)p(x|\omega_1), P(\omega_2)p(x|\omega_2)]p(x)\, dx$$



$$= 2 \cdot 0.5 \int_0^{x^*=0.5} (1-x)\, dx$$
$$= 0.1875.$$

The Bayes error in two dimensions is

$$E_B = 2 \cdot 0.5 \int_{x_1=0}^{1} dx_1 \int_0^{x_2^*=1-x_1} dx_2 (1-x_1)(1-x_2)$$
$$= 0.104167.$$

The Bayes error in three dimensions is

$$E_B = \int_0^1 dx_1 \int_0^1 dx_2 \int_0^{x_3^*(x_1,x_2)} dx_3\ (1-x_1)(1-x_2)(1-x_3)$$
$$= 0.0551969,$$

where $x_3^*(x_1, x_2)$ is the position of the decision boundary, found by solving

$$x_1 x_2 x_3 = (1-x_1)(1-x_2)(1-x_3)$$

for x_3, that is,

$$x_3^*(x_1, x_2) = \frac{(x_1 - 1)\,(x_2 - 1)}{1 - x_1 - x_2 + 2x_1 x_2}.$$

The general decision boundary in d dimensions comes from solving

$$\prod_{i=1}^{d} x_i = \prod_{i=1}^{d} (1 - x_i)$$

or

$$x_d \prod_{i=1}^{d-1} x_i = (1 - x_d) \prod_{i=1}^{d-1} (1 - x_i)$$

for x_d, which has solution

$$x_d^*(x_1, x_2, \ldots, x_{d-1}) = \frac{\prod_{i=1}^{d-1}(1-x_i)}{\prod_{j=1}^{d-1}(1-x_j) + \prod_{k=1}^{d-1} x_k}.$$

The Bayes error is

$$E_B = \underbrace{\int_0^1 dx_1 \int_0^1 dx_2 \cdots \int_0^1 dx_{d-1} \int_0^{x_d^*(x_1,\ldots,x_{d-1})} dx_d}_{d \text{ integrals}} \prod_{j=1}^{d}(1-x_i).$$

We first compute the integral(x[1] - 1)) over x_d, since it has the other variables implicit. This integral I is

$$\begin{aligned} I = \int_0^{x_d^*(x_1,\ldots,x_{d-1})} (1-x_d)\, dx_d &= x_d\|_0^{x_d^*} - \frac{1}{2}\left[x_d^2\right]_0^{x_d^*} \\ &= x_d^* - \frac{1}{2}(x_d^*)^2. \end{aligned}$$

Substituting x_d^* from above we have

$$\begin{aligned} I &= \frac{\left[\prod_{i=1}^{d-1}(1-x_i)\right]\left[\prod_{j=1}^{d-1}(1-x_j) + \prod_{k=1}^{d-1} x_k - 1/2\prod_{i=1}^{d-1}(1-x_i)\right]}{\left[\prod_{j=1}^{d-1}(1-x_j) + \prod_{k=1}^{d-1} x_k\right]^2} \\ &= \frac{1/2\prod_{i=1}^{d-1}(1-x_i)^2 + \prod_{i=1}^{d-1} x_i(1-x_i)}{\left[\prod_{j=1}^{d-1}(1-x_j) + \prod_{k=1}^{d-1} x_k\right]^2}. \end{aligned}$$

(d) Problem not yet solved

(e) Problem not yet solved

27. Consider the Tanimoto metric described by Eq. 58 in the text.

(a) The Tanimoto metric measures the distance between two sets $\mathcal{A}$ and $\mathcal{B}$, according to

$$D_{Tanimoto}(\mathcal{A}, \mathcal{B}) = \frac{n_a + n_b - 2n_{ab}}{n_a + n_b - n_{ab}}$$

where n_a and n_b are the sizes of the two sets, and n_{ab} is the number of elements common to both sets.

We must check whether the four properties of a metric are always fulfilled. The first property is non-negativity, that is,

$$D_{Tanimoto}(\mathcal{A},\mathcal{B}) = \frac{n_a + n_b - 2n_{ab}}{n_a + n_b - n_{ab}} \geq 0.$$

Since the sum of the elements in $\mathcal{A}$ and $\mathcal{B}$ is always greater than their common elements, we can write

$$n_a + n_b - n_{ab} > 0.$$

Furthermore, the term $n_a + n_b - 2n_{ab}$ gives account of the number of different elements in sets $\mathcal{A}$ and $\mathcal{B}$, and thus

$$n_a + n_b - 2n_{ab} \geq 0.$$

Consequently, $D_{Tanimoto}(\mathcal{A},\mathcal{B}) \geq 0$.

The second property, reflexivity, states that

$$D_{Tanimoto}(\mathcal{A},\mathcal{B}) = 0 \qquad \text{if and only if } \mathcal{A} = \mathcal{B}.$$

From the definition of the Tanimoto measure above, we see that $D_{Tanimoto}(\mathcal{A},\mathcal{B})$ will be 0 if and only if $n_a + n_b - 2n_{ab} = 0$. This numerator is the number of different elements in $\mathcal{A}$ and $\mathcal{B}$, so it will yield 0 only when $\mathcal{A}$ and $\mathcal{B}$ have no different elements, that is, when $\mathcal{A} = \mathcal{B}$. Conversely, if $\mathcal{A} = \mathcal{B}$, then $n_a + n_b - 2n_{ab} = 0$, and hence $D_{Tanimoto}(\mathcal{A},\mathcal{B}) = 0$.

The third property is symmetry, or

$$D_{Tanimoto}(\mathcal{A},\mathcal{B}) = D_{Tanimoto}(\mathcal{B},\mathcal{A})$$

for all $\mathcal{A}$ and $\mathcal{B}$. Since the terms n_a and n_b appear in the numerator and denominator in the same way, only the term n_{ab} can affect the fulfilment of this property. However, n_{ab} and n_{ba} give the same measure: the number of elements common to both $\mathcal{A}$ and $\mathcal{B}$. Thus the Tanimoto metric is indeed symmetric.

The final property is the triangle inequality:

$$D_{Tanimoto}(\mathcal{A},\mathcal{B}) + D_{Tanimoto}(\mathcal{B},\mathcal{C}) \geq D_{Tanimoto}(\mathcal{A},\mathcal{C}). \qquad (*)$$

We substitute $(*)$ into the definition of the Tanimoto metric and find

$$\frac{n_a + n_b - 2n_{ab}}{n_a + n_b - n_{ab}} + \frac{n_b + n_c - 2n_{bc}}{n_b + n_c - n_{bc}} \geq \frac{n_a + n_c - 2n_{ac}}{n_a + n_c - n_{ac}}.$$

After some simple manipulation, the inequality can be expressed as

$$3BDE - 2ADE - 2BCE + ACE - 2BDF + ADF + BCF \geq 0,$$

where

$$\begin{aligned} A &= n_a + n_b \\ B &= n_{ab} \\ C &= n_b + n_c \\ D &= n_{bc} \\ E &= n_a + n_c \\ F &= n_{ac}. \end{aligned}$$

We factor to find

$$E[C(A-2B)+D(3B-2A)]+F[D(A-2B)+BC]\geq 0.$$

Since $A-2B$ is the number of different elements in $\mathcal{A}$ and $\mathcal{B}$ and hence non-negative, the triangle inequality is fulfilled if some of the below conditions are met:

$$\begin{aligned} 3B-2A=3n_{ab}-2(n_a+n_b)\geq 0 \qquad (**)\\ D=n_{bc}=0 \qquad (***)\\ C(A-2B)+D(3B-2A)\geq 0. \end{aligned}$$

This last condition can be rewritten after some arrangment as

$$D_T\geq \frac{n_{bc}}{n_b+n_c-n_{bc}}.$$

On the other hand, if we take other common factors in the above equation, we have

$$B[E(3D-2C)+F(C-2D)]+A[E(C-2D)+DF]\geq 0.$$

The term $C-2D$ is the number of different elements in $\mathcal{C}$ and $\mathcal{D}$, so it is greater than zero. Accordingly, the triangle inequality can be also fulfilled if some of the below additional conditions are met:

$$\begin{aligned} 3D-2C=3n_{bc}-2(n_b+n_c) &\geq 0\\ E=n_b+n_a &= 0\\ E(3D-2C)+F(C-2D) &\geq 0. \end{aligned}$$

After some rearrangements, this last inequality can be written as

$$D_{Tanimoto}(\mathcal{B},\mathcal{C})\geq \frac{n_b+n_c}{n_{bc}-(n_b+n_c)}.$$

Since the denominator of the right-hand side is always negative, the inequality is met. Thus we can conclude that the Tanimoto metric also obeys the triangle inequality.

(b) Since the Tanimoto metric is symmetric, we consider here only 15 out of the 30 possible pairings.

$\mathcal{A}$	$\mathcal{B}$	n_a	n_b	n_{ab}	$D_{Tanimoto}(a,b)$	rank
pattern	pat	7	3	3	0.57	2
pattern	pots	7	4	2	0.77	7
pattern	stop	7	4	2	0.77	7
pattern	taxonomy	7	8	3	0.75	6
pattern	elementary	7	10	5	0.58	3
pat	pots	3	4	2	0.6	4
pat	stop	3	4	2	0.6	4
pat	taxonomy	3	8	2	0.77	7
pat	elementary	3	10	2	0.81	9
pots	stop	4	4	4	0	1
pots	taxonomy	4	8	2	0.8	8
pots	elementary	4	10	1	0.92	9
stop	taxonomy	4	8	2	0.8	8
stop	elementary	4	10	1	0.92	9
taxonomy	elementary	8	10	5	0.61	5

(c) Yes, the Tanimoto metric fulfills the triangle inequality for all possible sets given above.

Section 4.7

28. As shown in Fig. 23 in the text, there is a big difference between "categories" of a feature and the (true) categories or classes for the patterns. For this problem there is a fuzzy feature that we can call "temperature," and a class named "hot." Given a pattern, e.g., a cup, we can observe a degree of membership of this object onto the fuzzy feature "temperature," for instance, "temperature = warm." However "temperature = warm" does not necessarily imply that the membership to the "hot" class is less than 1.0 due to we could argue that the fuzzy feature "temperature" is not very hot but warm. We could reasonably suppose that if the "hot" class has something to do with temperature then it would have into account the fuzzy feature "temperature" for its membership function. Accordingly, values of "temperature" such as "warm" or "hot" might be active in some degree for the class "hot" though the exact dependence between them is unknown for us without more information aboute the problem.

29. Consider "fuzzy" classification.

(a) We first fit the designer's subjective feelings into fuzzy features, as suggested by the below table.

Feature	designer's feelings	fuzzy feature
lightness	medium-light	light
	medium-dark	dark
	dark	dark
length	short	short
	long	large

If we use $\text{Min}[a, b]$ as the conjunction rule between fuzzy sets a and b, then the discriminant functions can be written as:

$$\begin{aligned} d_1 &= \text{Min}[\hat{c}(length, 13, 5), \hat{c}(lightness, 70, 30)] \\ d_2 &= \text{Min}[\hat{c}'(length, 5, 5), \hat{c}'(lightness, 30, 30)] \\ d_1 &= \text{Min}[\hat{c}(length, 13, 5), \hat{c}'(lightness, 30, 30)] \end{aligned}$$

where

$$\hat{c}(x, \mu, \delta) = \begin{cases} 1 & x > \mu \\ 1 + (x - \mu)/\delta & \mu - \delta \leq x \leq \mu \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{c}'(x, \mu, \delta) = \begin{cases} 0 & x > \mu + \delta \\ 1 + (\mu - x)/\delta & \mu \leq x \leq \mu + \delta \\ 1 & \text{otherwise.} \end{cases}$$

(b) If every "category membership function" were rescaled by a constant α, the discriminant functions would be

$$d_1 = \text{Min}[\alpha\hat{c}(length, 13, 5), \alpha\hat{c}(lightness, 70, 30)]$$

$$\begin{aligned} d_2 &= \text{Min}[\alpha\hat{c}(length, 5, 5), \alpha\hat{c}'(lightness, 30, 30)] \\ d_3 &= \text{Min}[\alpha\hat{c}(length, 13, 5), \alpha\hat{c}'(lightness, 30, 30)] \end{aligned}$$

Since $\text{Min}[\alpha f_1, \alpha f_2] = \alpha\text{Min}[f_1, f_2]$, the above functions can be written

$$\begin{aligned} d_1 &= \alpha\text{Min}[\hat{c}(length, 13, 5), \hat{c}(lightness, 70, 30)] \\ d_2 &= \alpha\text{Min}[\hat{c}'(length, 13, 5), \hat{c}'(lightness, 70, 30)] \\ d_3 &= \alpha\text{Min}[\hat{c}(length, 13, 5), \hat{c}'(lightness, 70, 30)]. \end{aligned}$$

Hence, the classification borders would be unaffected, since all the discriminant functions would be rescaled by the same constant α, and we would have

$$\arg\max_{d_i}(\alpha d_1, \alpha d_2, \alpha d_3) = \alpha \arg\max_{d_i}(d_1, d_2, d_3) = \arg\max_{d_i}(d_1, d_2, d_3).$$

(c) Given the pattern $\mathbf{x} = (7.5, 60)^t$ and the above equations, the value of the discriminant functions can be computed as follows:

$$\begin{aligned} d_1 &= \text{Min}[\hat{c}(7.5, 13, 5), \hat{c}(60, 70, 30)] = \text{Min}[0, 0.66] = 0 \\ d_2 &= \text{Min}[\hat{c}'(7.5, 5, 5), \hat{c}'(60, 30, 30)] = \text{Min}[0.5, 0] = 0 \\ d_3 &= \text{Min}[\hat{c}(7.5, 13, 5), \hat{c}'(60, 30, 30)] = \text{Min}[0, 0] = 0. \end{aligned}$$

Since all the discriminant functions are equal to zero, the classifier cannot assign the input pattern to any of the existing classes.

(d) In this problem we deal with a handcrafted classifier. The designer has selected lightness and length as features that characterize the problem and has devined several membership functions for them without any theoretical or empirical justification. Moreover, the conjunction rule that fuses membership functions of the features to define the true discriminant functions is also imposed without any principle. Consequently, we cannot know where the sources of error come from.

30. PROBLEM NOT YET SOLVED

Section 4.8

31. If all the radii have been reduced to a value less than λ_m, Algorithm 4 in the text can be rewritten as:

Algorithm 0 (Modified RCE)

1 **begin** **initialize** $j \leftarrow 0, n \leftarrow$ number of patterns, $\epsilon \leftarrow$ small parameter
2 **do** $j \leftarrow j + 1$
3 $w_{ji} \leftarrow x_{ji}$ for $i = 1, \ldots, d$
4 $\hat{\mathbf{x}} \leftarrow \arg\min_{\mathbf{x} \in \omega_i} D(\mathbf{x}, \mathbf{x}_j)$
5 $\lambda_j \leftarrow D(\hat{\mathbf{x}}, \mathbf{x}_j)$
6 $\lambda_j \leftarrow D(\hat{\mathbf{x}}, \mathbf{x}_j) - \epsilon$
7 **if** $\mathbf{x}_j \in \omega_k$, **then** $a_{jk} \leftarrow 1$ for $k = 1, \ldots c$
8 **until** $j = n$
9 **end**

According to Algorithm 5 in the text, the decision boundaries of the RCE classifier depend on λ_j; an input pattern $\mathbf{x}$ is assigned to a class if and only if all its nearest prototypes belong to the same class, where these prototypes fulfill the condition $D(\mathbf{x}, \mathbf{w}_j), \lambda_j$. Since $\lambda_j = D(\hat{\mathbf{x}}, \mathbf{w}_j)$, where $\hat{\mathbf{x}}$ is the nearest training sample to $\mathbf{w}_j$ belonging to another class, we will have different class boundaries when we use different training points.

Section 4.9

32. We seek a Taylor series expansion for

$$p_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{h_n}\right).$$

(a) If the Parzen window is of the form $\varphi(x) \sim N(0,1)$, or

$$p(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2},$$

we can then write the estimate as

$$\begin{aligned} p_n(x) &= \frac{1}{nh_n}\sum_{i=1}^{n}\varphi\left(\frac{x-x_1}{h_n}\right) \\ &= \frac{1}{nh_n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-x_1}{h_n}\right)^2\right] \\ &= \frac{1}{\sqrt{2\pi}h_n}\frac{1}{n}\sum_{i=1}^{n}\exp\left[-\frac{1}{2}\left(\frac{x^2}{h_n^2}+\frac{x_i^2}{h_n^2}-\frac{2xx_i}{h_n^2}\right)\right] \\ &= \frac{\exp\left[-\frac{1}{2}\frac{x^2}{h_n^2}\right]}{\sqrt{2\pi}h_n}\frac{1}{n}\sum_{i=1}^{n}\exp\left[\frac{xx_i}{h_n^2}\right]\exp\left[-\frac{1}{2}\frac{x_i^2}{h_n^2}\right]. \end{aligned}$$

We express this result in terms of the normalized variables $u = x/h_n$ and $u_i = x_i/h_n$ as

$$p_n(x) = \frac{e^{-u^2/2}}{\sqrt{2\pi}h_n}\frac{1}{n}\sum_{i=1}^{n}e^{-u_i^2/2}e^{uu_i}.$$

We expand $e^{xx_i/h_n^2} = e^{uu_i}$ about the origin,

$$e^{uu_i} = \sum_{j=0}^{\infty}\frac{u^j u_i^j}{j!},$$

and therefore our density can be written as

$$\begin{aligned} p_n(x) &= \frac{e^{-u^2/2}}{\sqrt{2\pi}h_n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{\infty}\frac{u^j u_i^j}{j!}e^{-u_i^2/2} \\ &= \frac{e^{-u^2/2}}{\sqrt{2\pi}h_n}\sum_{j=0}^{\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{u_i^j e^{-u_i^2/2}}{j!}\right)u^j \end{aligned}$$

$$\begin{aligned} &= \frac{e^{-u^2/2}}{\sqrt{2\pi}h_n}\sum_{j=0}^{\infty} b_j u^j \\ &= \lim_{m\to\infty} p_{nm}(x), \end{aligned}$$

where we have defined

$$b_j = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{j!}u_i^j e^{-u_i^2/2}.$$

Thus we can express $p_{nm}(x)$ as the m-term approximation as

$$p_{nm}(x) = \frac{e^{-u^2/2}}{\sqrt{2\pi}h_n}\sum_{j=0}^{m-1} b_j u^j.$$

(b) If the n samples are extremely tightly clustered about $u = u_o$ (i.e., $u_i \simeq u_o$) for $i = 1, \ldots, n$, then the two-term approximation is

$$p_{n2}(x) = \frac{e^{-u^2/2}}{\sqrt{2\pi}h_n}(b_o + b_1 u),$$

where

$$b_o = \frac{1}{n}\sum_{i=1}^{n} e^{-u_i^2/2} \simeq e^{-u_o^2/2}$$

and

$$b_1 = \frac{1}{n}\sum_{i=1}^{n} u_i e^{-u_i^2/2} \simeq u_o e^{-u_o^2/2}.$$

The peak of $p_{n2}(x)$ is given by the solution of $\nabla_u p_{n2} = 0$, that is

$$\nabla_u p_{n2}(x) = \frac{-ue^{-u^2/2}(b_o + b_1 u)}{\sqrt{2\pi}h_n} + \frac{b_1 e^{-u^2/2}}{\sqrt{2\pi}h_n} = 0.$$

This equation then gives

$$\begin{aligned} 0 &= -u(b_o + b_1 u) + b_1 \\ &= u(e^{-u_o^2/2} + u_o e^{-u_o^2/2} u) - u_o e^{-u_o^2/2} \\ &= u^2 u_o + u - u_o = 0 \\ &= u^2 + \frac{u}{u_o} - 1. \end{aligned}$$

Thus the peaks of $p_{n2}(x)$ are the solutions of $u^2 + u/u_o - 1 = 0$.

(c) We have from part (b) that u is the solution to $u^2 + u/u_o - 1 = 0$, and thus the solution is

$$u = \frac{-1/u_o \pm \sqrt{1/u_o^2 + 4}}{2}.$$

In the case $u_o \ll 1$, the square root can be approximated as $\sqrt{1/u_0^2 + 4} \simeq 1/u_0$, and thus

$$\begin{aligned} u &= \frac{-1 \pm \sqrt{1 + 4u_o^2}}{2u_o} \\ &\simeq \frac{-1 \pm (1 + \frac{1}{2}4u_o^2)}{2u_o}, \end{aligned}$$

and this implies that one of the peaks of the probability is at

$$u \simeq \frac{-1 + (1 + 2u_o^2)}{2u_o} = u_o.$$

In the case $u_o \gg 1$, we have

$$\begin{aligned} u &= \frac{-1/u_o \pm \sqrt{1/u_o^2 + 4}}{2} \\ &= \frac{-1/u_o \pm 2\sqrt{1/(4u_o^2) + 1}}{2} \\ &\simeq \frac{-1/u_o \pm 2(1 + \frac{1}{2}1/(4u_o^2))}{2}. \end{aligned}$$

Thus, one of the peaks occurs at

$$u \simeq \frac{-1/u_o + 2(1 + 1/(8u_o^2))}{2} \simeq \frac{2}{2} = 1.$$

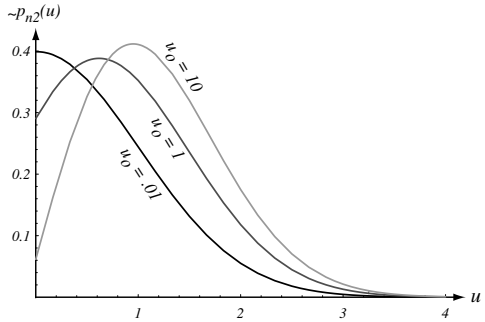
(d) In these conditions, we find

$$p_{n2}(u) = \frac{e^{-u^2/2}}{\sqrt{2\pi}h_n}(b_o + b_1 u)$$

where

$$b_o \simeq e^{-u_o^2/2}, \quad \text{and} \quad b_1 \simeq u_o e^{-u_o^2/2},$$

as is shown in the figure. (The curves have been rescaled so that their structure is apparent.) Indeed, as our calculation above showed, for small u_o, $p_{n2}(u)$ peaks near $u = 0$, and for large u_o, it peaks near $u = 1$.



Computer Exercises

Section 4.2

1. Computer exercise not yet solved

Section 4.3

2. Computer exercise not yet solved

Section 4.4

3. Computer exercise not yet solved

Section 4.5

4. xxx

MATLAB program

```
x = [0.58 0.27 0.0055 0.53 0.47 0.69 0.55 0.61 0.49 0.054]';
mu = mean(x);
delta = std(x);
mu_arr = linspace(min(x), max(x), 100);
delta_arr = linspace(2*(min(x)-mu), 8*(max(x)-mu), 100);
[M, D] = meshgrid(mu_arr, delta_arr);
p_D_theta = zeros(size(M));
for i=1:size(M,2)
    for j=1:size(M,2)
        x_minus_mu = abs(x - M(i,j));
        if max(x_minus_mu) > D(i,j)
            p_D_theta(i,j) = 0;
        else;
            a = (D(i,j) - x_minus_mu)/D(i,j)^2;
            p_D_theta(i,j) = prod(a);
        end
    end
end
p_D_theta = p_D_theta./(sum(p_D_theta(:))/...
    ((mu_arr(2) - mu_arr(1))*(delta_arr(2)-delta_arr(1))));
p_theta_D = p_D_theta;
ind = find(p_D_theta>0);
x_vect = linspace(min(x)-10*delta, max(x)+10*delta, 100);
post = zeros(size(x_vect));
for i=1:length(x_vect)
    for k=1:length(ind)
        if abs(x_vect(i) - M(ind(k))) < D(ind(k)),
            p_x_theta = (D(ind(k)) - abs(x_vect(i) - M(ind(k))))./D(ind(k)).^2;
        else
            p_x_theta = 0;
        end
        post(i) = post(i) + p_x_theta*p_theta_D(ind(k));
```

```
44     end
45 end
46 post = post./sum(post);
47 figure;
48 plot(x_vect, post, '-b');
49 hold on;
50 plot([mu mu], [min(post) max(post)], '-r');
51 plot(x, zeros(size(x)), '*m');
52 grid;
53 hold off;
```

[[FIGURE TO BE INSERTED]]

5. Computer exercise not yet solved

Section 4.6

6. Computer exercise not yet solved
7. Computer exercise not yet solved
8. Computer exercise not yet solved

Section 4.7

9. Computer exercise not yet solved

Section 4.8

10. Computer exercise not yet solved

Section 4.9

11. Computer exercise not yet solved

Chapter 5

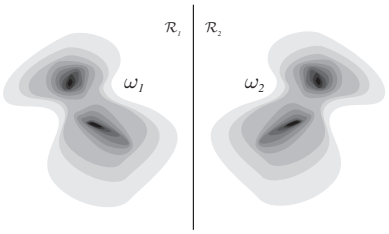
Linear discriminant functions

Problem Solutions

Section 5.2

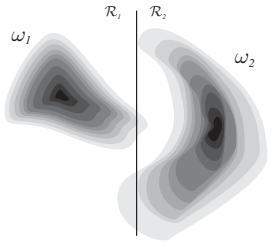
1. Consider the problem of linear discriminants with unimodal and multimodal distributions.

(a) The figure shows two bimodal distributions that obey reflective symmetry about a vertical axis. Thus their densities have the same value along this vertical line, and as a result this is the Bayesian decision boundary, giving minimum classification error.

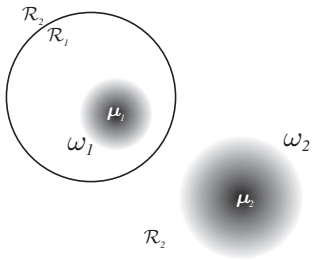


(b) Of course, if two unimodal distributions overlap significantly, the Bayes error will be high. As shown in Figs 2.14 and 2.5 in Chapter 2, the (unimodal) Gaussian case generally has quadratic (rather than linear) Bayes decision boundaries. Moreover, the two unimodal distributions shown in the figure below would be better classified using a curved boundary than the straight one shown.

(c) Suppose we have Gaussians of different variances, $\sigma_1^2 \neq \sigma_2^2$ and for definiteness (and without loss of generality) $\sigma_2 > \sigma_1$. Then at large distances from the means, the density $P(\omega_2|\mathbf{x}) > P(\omega_1|\mathbf{x})$. Moreover, the position of the mean of ω_1, that is, $\boldsymbol{\mu}_1$, will always be categorized as ω_1. No (single) straight line can separate the position of $\boldsymbol{\mu}_1$ from *all* distant points; thus, the optimal decision



boundary cannot be a straight line. In fact, for the case shown, the optimal boundary is a circle.



2. Consider a linear machine for c categories with discriminant functions $g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + \omega_{i0}$ for $i = 1, \ldots, c$. Our goal is to show that the decision regions are convex, that is, if $\mathbf{x}(1)$ and $\mathbf{x}(0)$ are in ω_i, then $\lambda\mathbf{x}(1) + (1-\lambda)\mathbf{x}(0)$ is also in ω_i, for all $0 \leq \lambda \leq 1$. The fact that $\mathbf{x}(1)$ and $\mathbf{x}(0)$ are in ω_i implies

$$\max_j g_j(\mathbf{x}(1)) = g_i(\mathbf{x}(1)) = \mathbf{w}_i^t\mathbf{x}(1) + w_{i0}$$

and

$$\max_j g_j(\mathbf{x}(0)) = g_i(\mathbf{x}(0)) = \mathbf{w}_i^t\mathbf{x}(0) + w_{i0}.$$

For any j and for $0 \leq \lambda \leq 1$, the above imply

$$\begin{aligned} \mathbf{w}_j^t[\lambda\mathbf{x}(1) + (1-\lambda)\mathbf{x}(0)] + w_{j0} &= \lambda[\mathbf{w}_j^t\mathbf{x}(1) + w_{j0}] + (1-\lambda)[\mathbf{w}_j^t\mathbf{x}(0) + w_{j0}] \\ &\leq \lambda[\mathbf{w}_i^t\mathbf{x}(1) + w_{i0}] + (1-\lambda)[\mathbf{w}_i^t\mathbf{x}(0) + w_{i0}]. \end{aligned}$$

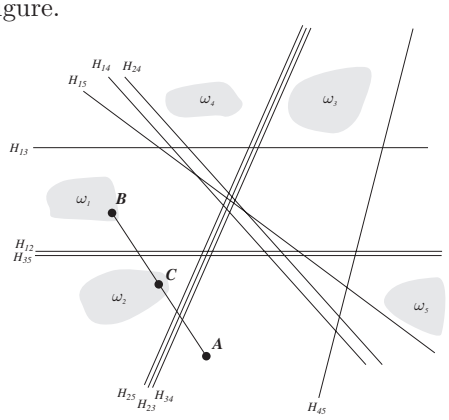
We next determine the category of a point between $\mathbf{x}(0)$ and $\mathbf{x}(1)$ by finding the maximum discriminant function:

$$\begin{aligned} \max_j \left[\mathbf{w}_j^t[\lambda\mathbf{x}(1) + (1-\lambda)\mathbf{x}(0)] + w_{j0}\right] &= \lambda \max_j[\mathbf{w}j^t\mathbf{x}(1) + w_{j0}] + (1-\lambda)\max_j[\mathbf{w}j^t\mathbf{x}(1) + w_{j0}] \\ &= \lambda[\mathbf{w}_i^t\mathbf{x}(1) + w_{i0}] + (1-\lambda)[\mathbf{w}_i^t\mathbf{x}(1) + w_{i0}] \\ &= \mathbf{w}_i^t[\lambda\mathbf{x}(1) + (1-\lambda)\mathbf{x}(0)] + w_{i0}. \end{aligned}$$

This shows that the point $\lambda\mathbf{x}(1) + (1-\lambda)\mathbf{x}(0)$ is in ω_i. Thus any point between $\mathbf{x}(0)$ and $\mathbf{x}(1)$ is also in category ω_i; and since this is true for any two points in ω_i, the region is convex. Furthermore, it is true for any *category* ω_j — all that is required in the proof is that $\mathbf{x}(1)$ and $\mathbf{x}(0)$ be in the category region of interest, ω_j. Thus every decision region is convex.

3. A counterexample will serve as proof that the decision regions need not be convex. In the example below, there are $c = 5$ classes and $\binom{c}{2} = 10$ hyperplanes separating

the pairs of classes; here H_{ij} means the hyperplane separating class ω_i from class ω_j. To show the non-convexity of the solution retion for class ω_1, we take points **A** and **B**, as shown in the figure.



The table below summarizes the voting results. Each row lists a hyperplane H_{ij}, used to disginguish category ω_i from ω_j. Each entry under a point states the category decision according to that hyperplane. For instance, hyperplane H_{12} would place **A** in ω_2, point **B** in ω_1, and point **C** in ω_2. The underlining indicates the winning votes in the full classifier system. Thus point **A** gets three ω_1 votes, only one ω_2 vote, two ω_3 votes, and so forth. The bottom row shows the final voting result of the full classifier system. Thus points **A** and **B** are classified as category ω_1, while point **C** (on the line between **A** and **B**) is category ω_2. Thus the solution region for ω_1 is not convex.

hyperplane	**A**	**B**	**C**
H_{12}	ω_2	$\underline{\omega_1}$	$\underline{\omega_2}$
H_{13}	$\underline{\omega_1}$	$\underline{\omega_1}$	ω_3
H_{14}	$\underline{\omega_1}$	$\underline{\omega_1}$	ω_1
H_{15}	$\underline{\omega_1}$	$\underline{\omega_1}$	ω_1
H_{23}	ω_3	ω_2	$\underline{\omega_2}$
H_{24}	ω_2	ω_2	$\underline{\omega_2}$
H_{25}	ω_5	ω_2	$\underline{\omega_2}$
H_{34}	ω_3	ω_4	ω_3
H_{35}	ω_5	ω_3	ω_5
H_{45}	ω_4	ω_4	ω_4
Vote result:	ω_1	ω_1	ω_2

4. Consider the hyperplane used for discriminant functions.

(a) In order to minimize $\|\mathbf{x}-\mathbf{x}_a\|^2$ subject to the constraint $g(\mathbf{x})=0$, we form the objective function

$$f(\mathbf{x},\lambda)=\|\mathbf{x}-\mathbf{x}_a\|^2+2\lambda[g(\mathbf{x})],$$

where λ is a Lagrange undetermined multiplier; the point $\mathbf{x}_a$ is fixed while $\mathbf{x}$ varies on the hyperplane. We expand $f(\cdot)$ and find

$$\begin{aligned} f(\mathbf{x},\lambda) &= \|\mathbf{x}-\mathbf{x}_a\|^2+2\lambda[\mathbf{w}^t\mathbf{x}+w_0] \\ &= (\mathbf{x}-\mathbf{x}_a)^t(\mathbf{x}-\mathbf{x}_a)+2\lambda(\mathbf{w}^t\mathbf{x}+w_0) \\ &= \mathbf{x}^t\mathbf{x}-2\mathbf{x}^t\mathbf{x}_a+\mathbf{x}_a^t\mathbf{x}_a+2\lambda(\mathbf{x}^t\mathbf{w}+w_0). \end{aligned}$$

We set the derivatives of $f(\mathbf{x}, \lambda)$ to zero and find

$$\begin{aligned}\frac{\partial f(\mathbf{x},\lambda)}{\partial \mathbf{x}} &= \mathbf{x} - \mathbf{x}_a + \lambda \mathbf{w} = 0\\ \frac{\partial f(\mathbf{x},\lambda)}{\partial \lambda} &= \mathbf{w}^t\mathbf{x} + w_0 = 0,\end{aligned}$$

and these equations imply

$$\mathbf{x} = \mathbf{x}_a - \lambda \mathbf{w},$$

and

$$\begin{aligned}\mathbf{w}^t\mathbf{x} + w_0 &= \mathbf{w}^t(\mathbf{x}_a - \lambda\mathbf{w}) + w_0\\ &= \mathbf{w}^t\mathbf{x}_a + w_0 - \lambda\mathbf{w}^t\mathbf{w}\\ &= 0.\end{aligned}$$

These can be solved for λ as

$$\lambda \mathbf{w}^t\mathbf{w} = \mathbf{w}^t\mathbf{x}_a + w_0$$

or (so long as $\mathbf{w} \neq \mathbf{0}$)

$$\lambda = \frac{\mathbf{w}^t\mathbf{x}_a + w_0}{\mathbf{w}^t\mathbf{w}}.$$

The value of $\mathbf{x}$ is then

$$\begin{aligned}\mathbf{x} &= \mathbf{x}_a - \lambda\mathbf{w}\\ &= \begin{cases}\mathbf{x}_a - \left[\frac{\mathbf{w}^t\mathbf{x}_a + w_0}{\mathbf{w}^t\mathbf{w}}\right]\mathbf{w} & \text{if } \mathbf{w} \neq \mathbf{0}\\ \mathbf{x}_a & \text{if } \mathbf{w} = \mathbf{0}.\end{cases}\end{aligned}$$

The minimum value of the distance is obtained by substituting the value of $\mathbf{x}$ into the distance:

$$\begin{aligned}\|\mathbf{x} - \mathbf{x}_a\| &= \left\|\mathbf{x}_a - \left[\frac{\mathbf{w}^t\mathbf{x}_a + w_0}{\mathbf{w}^t\mathbf{w}}\right]\mathbf{w} - \mathbf{x}_a\right\|\\ &= \left\|\left(\frac{\mathbf{w}^t\mathbf{x}_a + w_0}{\mathbf{w}^t\mathbf{w}}\right)\mathbf{w}\right\|\\ &= \frac{|g(\mathbf{x}_a)|\|\mathbf{w}\|}{\|\mathbf{w}\|^2} = \frac{|g(\mathbf{x}_a)|}{\|\mathbf{w}\|}.\end{aligned}$$

(b) From part (a) we have that the minimum distance from the hyperplane to the point $\mathbf{x}$ is attained uniquely at the point $\mathbf{x} = \mathbf{x}_a - \lambda\mathbf{w}$ on the hyperplane. Thus the projection of $\mathbf{x}_a$ onto the hyperplane is

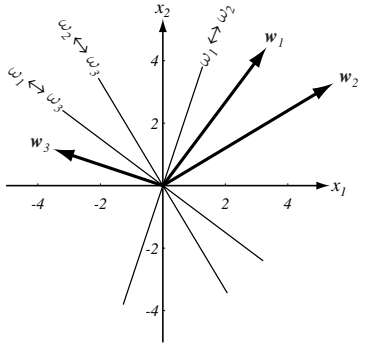
$$\begin{aligned}\mathbf{x}_o &= \mathbf{x}_a - \lambda\mathbf{w}\\ &= \mathbf{x}_a - \frac{g(\mathbf{x}_a)}{\|\mathbf{w}\|^2}\mathbf{w},\end{aligned}$$

where

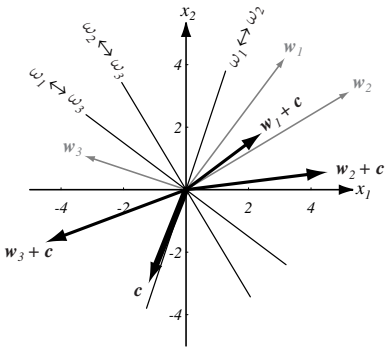
$$\lambda = \frac{\mathbf{w}^t\mathbf{x}_a + w_0}{\mathbf{w}^t\mathbf{w}} = \frac{g(\mathbf{x}_a)}{\|\mathbf{w}\|^2}.$$

5. Consider the three category linear machine with discriminant functions $g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} - w_{io}$, for $i = 1, 2, 3$.

(a) The decision boundaries are obtained when $\mathbf{w}_i^t\mathbf{x} = \mathbf{w}_j^t\mathbf{x}$ or equivalently $(\mathbf{w}_i - \mathbf{w}_j)^t\mathbf{x} = 0$. Thus, the relevant decision boundaries are the lines through the origin that are orthogonal to the vectors $\mathbf{w}_1 - \mathbf{w}_2, \mathbf{w}_1 - \mathbf{w}_3$ and $\mathbf{w}_2 - \mathbf{w}_3$, and illustrated by the lines marked $\omega_i \leftrightarrow \omega_j$ in the figure.



(b) If a constant vector $\mathbf{c}$ is added to the weight vectors, $\mathbf{w}_1, \mathbf{w}_2$ and $\mathbf{w}_3$, then the decision boundaries are given by $[(\mathbf{w}_i + \mathbf{c}) - (\mathbf{w}_j + \mathbf{c})]^t\mathbf{x} = (\mathbf{w}_i - \mathbf{w}_j)^t\mathbf{x} = 0$, just as in part (a); thus the decision boundaries do not change. Of course, the triangle formed by joining the heads of $\mathbf{w}_1, \mathbf{w}_2$, and $\mathbf{w}_3$ changes, as shown in the figure.



6. We first show that totally linearly separable samples must be linearly separable. For samples that are totally linearly separable, for each i there exists a hyperplane $g_i(\mathbf{x}) = 0$ that separates ω_i samples from the rest, that is,

$$g_i(\mathbf{x}) \geq 0 \quad \text{for } \mathbf{x} \in \omega_i$$
$$g_i(\mathbf{x}) < 0 \quad \text{for } \mathbf{x} \notin \omega_i.$$

The set of discriminant functions $g_i(\mathbf{x})$ form a linear machine that classifies correctly; the samples are linearly separable and each category can be separated from each of the others.

We shall show that by an example that, conversely, linearly separable samples need *not* be totally linearly separable. Consider three categories in two dimensions

specified by

$$\begin{aligned}\omega_1 &= \{\mathbf{x} : x_1 < -1\}\\ \omega_2 &= \{\mathbf{x} : -1 \leq x_1 < 1\}\\ \omega_3 &= \{\mathbf{x} : x_1 > 1\}.\end{aligned}$$

Clearly, these categories are linearly separable by vertical lines, but not totally linearly separable since there can be no linear function $g_2(\mathbf{x})$ that separates the category ω_2 samples from the rest.

7. Consider the four categories in two dimensions corresponding to the quadrants of the Euclidean plane:

$$\begin{aligned}\omega_1 &= \{\mathbf{x} : x_1 > 0, x_2 > 0\}\\ \omega_2 &= \{\mathbf{x} : x_1 > 0, x_2 < 0\}\\ \omega_3 &= \{\mathbf{x} : x_1 < 0, x_2 < 0\}\\ \omega_4 &= \{\mathbf{x} : x_1 < 0, x_2 > 0\}.\end{aligned}$$

Clearly, these are pairwise linearly separable but not totally linearly separable.

8. We employ proof by contradiction. Suppose there were two *distinct* minimum length solution vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ with $\mathbf{a}_1^t\mathbf{y} > b$ and $\mathbf{a}_2^t\mathbf{y} > b$. Then necessarily we would have $\|\mathbf{a}_1\| = \|\mathbf{a}_2\|$ (otherwise the longer of the two vectors would not be a *minimum* length solution). Next consider the average vector $\mathbf{a}_o = \frac{1}{2}(\mathbf{a}_1 + \mathbf{a}_2)$. We note that

$$\mathbf{a}_o^t\mathbf{y}_i = \frac{1}{2}(\mathbf{a}_1 + \mathbf{a}_2)^t\mathbf{y}_i = \frac{1}{2}\mathbf{a}_1^t\mathbf{y}_i + \frac{1}{2}\mathbf{a}_2^t\mathbf{y}_i \geq b,$$

and thus $\mathbf{a}_o$ is indeed a solution vector. Its length is

$$\|\mathbf{a}_o\| = \|1/2(\mathbf{a}_1 + \mathbf{a}_2)\| = 1/2\|\mathbf{a}_1 + \mathbf{a}_2\| \leq 1/2(\|\mathbf{a}_1\| + \|\mathbf{a}_2\|) = \|\mathbf{a}_1\| = \|\mathbf{a}_2\|,$$

where we used the triangle inequality for the Euclidean metric. Thus $\mathbf{a}_o$ is a solution vector such that $\|\mathbf{a}_o\| \leq \|\mathbf{a}_1\| = \|\mathbf{a}_2\|$. But by our hypothesis, $\mathbf{a}_1$ and $\mathbf{a}_2$ are minimum length solution vectors. Thus we must have $\|\mathbf{a}_o\| = \|\mathbf{a}_1\| = \|\mathbf{a}_2\|$, and thus

$$\frac{1}{2}\|\mathbf{a}_1 + \mathbf{a}_2\| = \|\mathbf{a}_1\| = \|\mathbf{a}_2\|.$$

We square both sides of this equation and find

$$\frac{1}{4}\|\mathbf{a}_1 + \mathbf{a}_2\|^2 = \|\mathbf{a}_1\|^2$$

or

$$\frac{1}{4}(\|\mathbf{a}_1\|^2 + \|\mathbf{a}_2\|^2 + 2\mathbf{a}_1^t\mathbf{a}_2) = \|\mathbf{a}_1\|^2.$$

We regroup and find

$$\begin{aligned}0 &= \|\mathbf{a}_1\|^2 + \|\mathbf{a}_2\|^2 - 2\mathbf{a}_1^t\mathbf{a}_2\\ &= \|\mathbf{a}_1 - \mathbf{a}_2\|^2,\end{aligned}$$

and thus $\mathbf{a}_1 = \mathbf{a}_2$, contradicting our hypothesis. Therefore, the minimum-length solution vector is unique.

9. Let the two sets of vectors be $\mathcal{S}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathcal{S}_2 = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$. We assume S_1 and S_2 are linearly seperable, that is, there exists a linear discriminant function $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$ such that

$$\begin{aligned} g(\mathbf{x}) &> 0 \text{ implies } \mathbf{x} \in \mathcal{S}_1 \quad \text{and} \\ g(\mathbf{x}) &< 0 \text{ implies } \mathbf{x} \in \mathcal{S}_2. \end{aligned}$$

Consider a point $\mathbf{x}$ in the convex hull of $\mathcal{S}_1$, or

$$\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i,$$

where the α_i's are non-negative and sum to 1. The discriminant function evaluated at $\mathbf{x}$ is

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{w}^t\mathbf{x} + w_0 \\ &= \mathbf{w}^t \left(\sum_{i=1}^{n} \alpha_i \mathbf{x}_i \right) + w_0 \\ &= \sum_{i=1}^{n} \alpha_i (\mathbf{w}^t\mathbf{x}_i + w_0) \\ &> 0, \end{aligned}$$

where we used the fact that $\mathbf{w}^t\mathbf{x}_i + w_0 > 0$ for $1 \le i \le n$ and $\sum_{i=1}^{n} \alpha_i = 1$.

Now let us assume that our point $\mathbf{x}$ is *also* in the convex hull of S_2, or

$$\mathbf{x} = \sum_{j=1}^{m} \beta_j \mathbf{y}_j,$$

where the β_j's are non-negative and sum to 1. We follow the approach immediately above and find

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{w}^t\mathbf{x} + w_0 \\ &= \mathbf{w}^t \left(\sum_{j=1}^{m} \beta_j \mathbf{y}_j \right) + w_0 \\ &= \sum_{j=1}^{m} \beta_j \underbrace{(\mathbf{w}^t\mathbf{y}_j + w_0)}_{g(\mathbf{y}_j)<0} \\ &< 0, \end{aligned}$$

where the last step comes from the realization that $g(\mathbf{y}_j) = \mathbf{w}^t\mathbf{y}_j + w_0 < 0$ for each $\mathbf{y}_i$, since they are each in $\mathcal{S}_2$. Thus, we have a contradiction: $g(\mathbf{x}) > 0$ and $g(\mathbf{x}) < 0$, and hence clearly the intersection is empty. In short, either two sets of vectors are either linearly separable or their convex hulls intersects.

10. Consider a piecewise linear machine.

(a) The discriminiant functions have the form

$$g_i(\mathbf{x}) = \max_{j=1,\ldots,n_i} g_{ij}(\mathbf{x}),$$

where the components are linear

$$g_{ij}(\mathbf{x}) = w_{ij}^t \mathbf{x} + w_{ij0},$$

and the decision rule is to classify $\mathbf{x}$ in category ω_i, if

$$\max_k g_k(\mathbf{x}) = g_i(\mathbf{x}),$$

for $j = 1, \ldots, c$. We can expand this discriminant function as

$$\max_k g_k(\mathbf{x}) = \max_k \max_{j=1,\ldots,n_k} g_{kj}(\mathbf{x}),$$

where $g_{kj}(\mathbf{x}) = \mathbf{w}_{kj}^t \mathbf{x} + w_{kjo}$, are linear functions. Thus our decision rule implies

$$\max_k \max_{j=1,\ldots,n_k} g_{kj}(\mathbf{x}) = \max_{1 \le j \le n_i} g_{ij}(\mathbf{x}) = g_{ij(i)}(\mathbf{x}).$$

Therefore, the piecewise linear machine can be viewed as a linear machine for classifying *subclasses* of patterns, as follows: Classify $\mathbf{x}$ in category ω_i if

$$\max_k \max_{j=1,\ldots,n_k} g_{kj}(\mathbf{x}) = g_{ij}(\mathbf{x}).$$

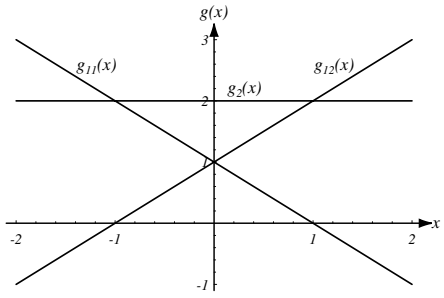
(b) Consider the following two categories in one dimension:

$$\begin{aligned} \omega_1 &= \{x : |x| > 1\}, \\ \omega_2 &= \{x : |x| < 1\} \end{aligned}$$

which clearly are not linearly separable. However, a classifier that is a piecewise linear machine with the following discriminant functions

$$\begin{aligned} g_{11}(x) &= 1 - x \\ g_{12}(x) &= 1 + x \\ g_1(x) &= \max_{j=1,2} g_{1j}(x) \\ g_2(x) &= 2 \end{aligned}$$

can indeed classify the patterns, as shown in the figure.



11. We denote the number of non-zero components in a vector by b. We let the d components of $\mathbf{x}$ be either 0 or 1 and the categories given by

$$\begin{aligned} \omega_1 &= \{\mathbf{x} : b \text{ is odd}\} \\ \omega_2 &= \{\mathbf{x} : b \text{ is even}\}. \end{aligned}$$

(a) Our goal is to show that ω_1 and ω_2 are not linearly separable if $d > 1$. We prove this by contradiction. Suppose that there exists a linear discriminant function

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$$

such that

$$\begin{aligned} g(\mathbf{x}) &\geq 0 \text{ for } \mathbf{x} \in \omega_1 \text{ and} \\ g(\mathbf{x}) &< 0 \text{ for } \mathbf{x} \in \omega_2. \end{aligned}$$

Consider the unique point having $b = 0$, i.e., $\mathbf{x} = \mathbf{0} = (0, 0, \ldots, 0)^t$. This pattern is in ω_2 and hence clearly $g(\mathbf{0}) = \mathbf{w}^t\mathbf{0} + w_0 = w_0 < 0$. Next consider patterns in ω_1 having $b = 1$. Such patterns are of the general form

$$\mathbf{x} = (0, \ldots, 0, 1, 0, \ldots, 0),$$

and thus

$$g(\mathbf{x}) = w_i + w_0 > 0$$

for any $i = 1, \ldots, d$. Next consider patterns with $b = 2$, for instance

$$\mathbf{x} = (0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0).$$

Because such patterns are in ω_2 we have

$$g(\mathbf{x}) = w_i + w_j + w_0 < 0$$

for $i \neq j$.

We summarize our results up to here as three equations:

$$\begin{aligned} w_0 &< 0 \\ w_i + w_0 &> 0 \\ w_i + w_j + w_0 &< 0, \end{aligned}$$

for $i \neq j$. The first two of these imply

$$\begin{aligned} \underbrace{(w_i + w_0)}_{>0} + \underbrace{(w_j + w_0)}_{>0} + \underbrace{(-w_0)}_{>0} &> 0, \\ \text{or} \quad w_i + w_j + w_0 &> 0, \end{aligned}$$

which contradicts the third equation. Thus our premise is false, and we have proven that ω_1, ω_2 are not linearly separable if $d > 1$.

(b) Consider the discriminant functions

$$\begin{aligned} g_1(\mathbf{x}) &= \max_j g_{1j}(\mathbf{x}) \quad \text{and} \\ g_2(\mathbf{x}) &= \max_j g_{2j}(\mathbf{x}), \end{aligned}$$

where

$$\begin{aligned} g_{ij}(\mathbf{x}) &= \alpha_{ij}(1, \ldots, 1)^t\mathbf{x} + w_{ijo} \\ &= \alpha_{ij}b + w_{ijo}, \end{aligned}$$

and, as in part (a), we let b denote the number of non-zero components of $\mathbf{x}$. We can set

$$\begin{aligned} \alpha_{1j} &= j + 1/2 \\ \alpha_{2j} &= j \\ w_{ijo} &= -j^2 - j - 1/4 \\ w_{2jo} &= -j^2, \end{aligned}$$

for $j = 0, 1, \ldots, d+1$ and hence write our discriminant function as

$$\begin{aligned} g_1(\mathbf{x}) &= (j + 1/2)b - j^2 - j - 1/4 \\ g_2(\mathbf{x}) &= jb - j^2. \end{aligned}$$

We now verify that these discriminant functions indeed solve the problem. Suppose $\mathbf{x}$ is in ω_1; then the number of non-zero components of $\mathbf{x}$ is $b = 2m+1, 0 \leq m \leq (d-1)/2$ for m an integer. The discriminant function is

$$\begin{aligned} g_{1j}(\mathbf{x}) &= (j + 1/2)\,(2m + 1) - j^2 - j - 1/4 \\ &= j(2m + 1) - j^2 - j + 1/2(2m + 1) - 1/4 \\ &= j(2m) - j^2 + 1/2(2m + 1) - 1/4 \\ &= j(2m - j) + 1/2(2m + 1) - 1/4. \end{aligned}$$

For patterns in ω_2, we have $b = 2m$, and thus

$$\begin{aligned} g_{2j}(\mathbf{x}) &= j(2m + 1) - j^2 \\ &= j(2m + 1 - j). \end{aligned}$$

It is a simple matter to verify that $j(2m - j)$ is maximized at $j = m$ and that $j(2m + 1 - j)$ is maximized at $j = m + 1/2$. But note that j is an integer and the maximum occurs at $j = m$ and $j = m + 1$. Thus we have

$$\begin{aligned} g_1(\mathbf{x}) &= \max_j g_{1j}(\mathbf{x}) \\ &= \max_j [j(2m - j)] + 1/2(2m + 1) - 1/4 \\ &= m(2m - m) + m + 1/2 - 1/4 \\ &= m^2 + m + 1/4 \\ &= m(m + 1) + 1/4, \quad \text{and} \\ g_2(\mathbf{x}) &= \max_j g_{2j}(\mathbf{x}) \\ &= \max_j j(2m + 1 - j) \\ &= m(m + 1). \end{aligned}$$

Thus if $\mathbf{x} \in \omega_1$ and the number of non-zero components of $\mathbf{x}$ is $b = 2m + 1$ (i.e., odd), we have

$$g_1(\mathbf{x}) = m(m + 1) + 1/4 > m(m + 1) = g_2(\mathbf{x}),$$

that is, $g_1(\mathbf{x}) > g_2(\mathbf{x})$, as desired. Conversely, if $\mathbf{x}$ is in ω_2, then the number of non-zero components of $\mathbf{x}$ is $b = 2m, 0 \leq m \leq d/2$ where m is an integer.

Following the above logic, we have

$$\begin{aligned} g_{1j}(\mathbf{x}) &= (j+1/2)(2m) - j^2 - j - 1/4 \\ &= j(2m-1) - j^2 + m - 1/4 \\ &= j(2m-j-1) + m - 1/4 \end{aligned}$$

which is maximized at $j = m$. Likewise, we have

$$\begin{aligned} g_{2j}(\mathbf{x}) &= j(2m) - j^2 \\ &= j(2m-j) \end{aligned}$$

which is maximized at $j = m$. Thus our discriminant functions are

$$\begin{aligned} g_1(\mathbf{x}) &= \max_j g_{1j}(\mathbf{x}) \\ &= \max_j j(2m-1-j) + m - 1/4 \\ &= m(m-1) + m - 1/4 = m^2 - 1/4, \\ g_2(\mathbf{x}) &= \max_j g_{2j}(\mathbf{x}) = \max_j j(2m-j) = m^2. \end{aligned}$$

Indeed, if $\mathbf{x}$ is in ω_2 and the number of non-zero components of $\mathbf{x}$ is $b = 2m$ (i.e., even), we have

$$g_1(\mathbf{x}) = m^2 - 1/4 < g_2(\mathbf{x}) = m^2,$$

or $g_1(\mathbf{x}) < g_2(\mathbf{x})$, and hence our piecewise linear machine solves the problem.

Section 5.3

12. Consider the quadratic discriminant function given by Eq. 4 in the text:

$$g_1(\mathbf{x}) - g_2(\mathbf{x}) = g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d}\sum_{i=1}^{d} w_{ij} x_i x_j,$$

which for convenience we write in vector and matrix form as

$$g(\mathbf{x}) = w_0 + \mathbf{w}^t\mathbf{x} + \mathbf{x}^t\mathbf{W}\mathbf{x}.$$

In this two-category case, the decision boundary is the set of $\mathbf{x}$ for which $g(\mathbf{x}) = 0$. If we did not have the $\mathbf{w}^t\mathbf{x}$ term, we could easily describe the boundary in the form $\mathbf{x}\mathbf{W}_0\mathbf{x} = k$ where k is some constant. We can, in fact, eliminate the $\mathbf{w}^t\mathbf{x}$ terms with a proper translation of axes. We seek a translation, described by a vector $\mathbf{m}$, that eliminates the linear term. Thus we have

$$\begin{aligned} g(\mathbf{x}-\mathbf{m}) &= w_0 + \mathbf{w}^t(\mathbf{x}-\mathbf{m}) + (\mathbf{x}-\mathbf{m})^t\mathbf{W}(\mathbf{x}-\mathbf{m}) \qquad (*) \\ &= (w_0 - \mathbf{w}^t\mathbf{m} + \mathbf{m}^t\mathbf{W}\mathbf{m}) + (\mathbf{w}^t - \mathbf{m}^t\mathbf{W}^t - \mathbf{m}\mathbf{W})\mathbf{x} + \mathbf{x}^t\mathbf{W}\mathbf{x}, \end{aligned}$$

where we used the fact that $\mathbf{W}$ is symmetric (i.e., $W_{ij} = W_{ji}$) and that taking the transpose of a scalar does not change its value. We can solve for $\mathbf{m}$ by setting the coefficients of $\mathbf{x}$ to zero, that is,

$$\mathbf{0} = \mathbf{w}^t - \mathbf{m}^t\mathbf{W} - \mathbf{m}\mathbf{W} = \mathbf{w}^t - 2\mathbf{m}^t\mathbf{W}.$$

This implies $\mathbf{m}^t\mathbf{W} = \mathbf{w}^t/2$ or $\mathbf{m} = \mathbf{W}^{-1}\mathbf{w}^t/2$.

Now we substitute this translation vector into the right-hand side of $(*)$ and find

$$\begin{aligned} g(\mathbf{x}-\mathbf{m}) &= w_0 - \mathbf{w}^t(\mathbf{W}^{-1}\mathbf{w}/2) + \mathbf{w}^t\mathbf{W}^{-1}/2 \cdot \mathbf{W} \cdot \mathbf{W}^{-1}\mathbf{w}/2 + \mathbf{x}^t\mathbf{W}\mathbf{x} \\ &= w_0 - \mathbf{w}^t\mathbf{W}^{-1}\mathbf{w}/4 + \mathbf{x}^t\mathbf{W}\mathbf{x}. \qquad (**) \end{aligned}$$

The decision boundary is given by $g(\mathbf{x}-\mathbf{m}) = 0$, and then $(**)$ implies

$$\mathbf{x}^t \frac{\mathbf{W}}{\mathbf{w}^t\mathbf{W}^{-1}\mathbf{w} - 4w_0}\mathbf{x} = \mathbf{x}^t\overline{\mathbf{W}}\mathbf{x} = 1/4,$$

and thus the matrix

$$\overline{\mathbf{W}} = \frac{\mathbf{W}}{\mathbf{w}^t\mathbf{W}^{-1}\mathbf{w} - 4w_0}$$

shows the basic character of the decision boundary.

(a) If $\overline{\mathbf{W}} = \mathbf{I}$, the identity matrix, then $\mathbf{x}^t\mathbf{I}\mathbf{x} = k/4$ is the equation of the decision boundary, where k is some constant. Then we have $\sum_{i=1}^{d} x_i^2 = k/4$, which defines a hypersphere of radius $\sqrt{k}/2$.

(b) Since $\overline{\mathbf{W}}$ is non-singular and symmetric, we can write $\overline{\mathbf{W}}$ in terms of $\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^t$, where $\mathbf{\Phi}$ is the matrix formed by the (linearly independent) eigenvectors of $\overline{\mathbf{W}}$ and $\mathbf{\Lambda}$ is the diagonal matrix formed by the eigenvalues of $\overline{\mathbf{W}}$. In this way, $\mathbf{x}^t\overline{\mathbf{W}}\mathbf{x} = 1/4$ can be written as $\mathbf{x}^t\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^t\mathbf{x} = 1/4$. We let $\mathbf{y}^t = \mathbf{x}^t\mathbf{\Phi}$; then we get $\mathbf{y}^t\mathbf{\Lambda}\mathbf{y} = 1/4$, or $\sum_i \lambda_i y_i^2 = 1/4$ where the eigenvalues $\lambda_i > 0$ when $\overline{\mathbf{W}}$ is positive definite, which it will be for non-pathological data distributions. Because the eigenvalues are positive and not necessarily all equal to each other, in this case the decision boundary is a hyperellipsoid.

(c) With an argument following that of part (b), we can represent the characteristics of the decision boundary with $\sum_{i=1}^{d} \lambda_i y_i^2 = 1/4$. Since, in this case we have at least one $\lambda_i < 0$, the geometric shape defined is a hyperhyperboloid.

(d) Our matrix and vector in this case are

$$\mathbf{W} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 5 & 1 \\ 0 & 1 & -3 \end{pmatrix}; \qquad \mathbf{w} = \begin{pmatrix} 5 \\ 2 \\ -3 \end{pmatrix}.$$

Thus we have

$$\mathbf{W}^{-1} = \begin{pmatrix} 16 & -6 & -2 \\ -6 & 3 & 1 \\ -2 & 1 & -1 \end{pmatrix} \cdot \frac{1}{4}; \qquad \mathbf{w}^t\mathbf{W}^{-1}\mathbf{w} = 82.75$$

and

$$\overline{\mathbf{W}} = \frac{\mathbf{W}}{82.75} = \begin{pmatrix} 0.0121 & 0.0242 & 0 \\ 0.0242 & 0.0604 & 0.0121 \\ 0 & 0.0121 & -0.0363 \end{pmatrix}$$

The eigenvalues are $\{\lambda_1, \lambda_2, \lambda_3\} = \{0.0026, 0.0716, -0.0379\}$. Because there is a negative eigenvalue, the decision boundary is a hyperhyperboloid.

(e) Our matrix and vector in this case are

$$\mathbf{W} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 0 & 4 \\ 3 & 4 & -5 \end{pmatrix}; \qquad \mathbf{w} = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix}.$$

Thus we have

$$\mathbf{W}^{-1} = \begin{pmatrix} -0.3077 & 0.4231 & 0.1538 \\ 0.4231 & -0.2692 & 0.0385 \\ 0.1538 & 0.0385 & -0.0769 \end{pmatrix}; \qquad \mathbf{w}^t\mathbf{W}^{-1}\mathbf{w} = -2.2692$$

and

$$\overline{\mathbf{W}} = \frac{\mathbf{W}}{-2.2692} = \begin{pmatrix} -0.4407 & -0.8814 & -1.3220 \\ -0.8814 & 0 & -1.7627 \\ -1.3220 & -1.7627 & 2.2034 \end{pmatrix}$$

The eigenvalues are $\{\lambda_1, \lambda_2, \lambda_3\} = \{0.6091, -2.1869, 3.3405\}$. Because of a negtative eigenvalue, the decision boundary is a hyperhyperboloid.

Section 5.4

13. We use a second-order Taylor series expansion of the criterion function at the point $\mathbf{a}(k)$:

$$J(\mathbf{a}) \simeq J(\mathbf{a}(k)) + \boldsymbol{\nabla} J^t(\mathbf{a}(k))(\mathbf{a} - \mathbf{a}(k)) + \frac{1}{2}(\mathbf{a} - \mathbf{a}(k))^t\mathbf{H}(\mathbf{a}(k))(\mathbf{a} - \mathbf{a}(k)),$$

where $\boldsymbol{\nabla} J(\mathbf{a}(k))$ is the derivative and $\mathbf{H}(\mathbf{a}(k))$ the Hessian matrix evaluated at the current point.

The update rule is

$$\mathbf{a}(k+1) = \mathbf{a}(k) - \eta(k)\boldsymbol{\nabla} J(\mathbf{a}(k)).$$

We use the expansion to write the criterion function as

$$J(\mathbf{a}(k+1) \simeq J(\mathbf{a}(k)) + \boldsymbol{\nabla} J^t(\mathbf{a}(k)[\eta(k)\boldsymbol{\nabla} J(\mathbf{a}(k)] + \frac{1}{2}[\eta(k)\boldsymbol{\nabla} J^t(\mathbf{a}(k))]\mathbf{H}(\mathbf{a}(k))[\eta(k)\boldsymbol{\nabla} J(\mathbf{a}(k))].$$

We minimize this with respect to $\eta(k)$ and find

$$0 = -\nabla J^t(\mathbf{a}(k))\boldsymbol{\nabla} J(\mathbf{a}(k)) + \eta(k)\boldsymbol{\nabla} J^t(\mathbf{a}(k))\mathbf{H}(\mathbf{a}(k))\boldsymbol{\nabla} J(\mathbf{a}(k))$$

which has solution

$$\eta(k) = \frac{\|J(\mathbf{a}(k))\|^2}{\boldsymbol{\nabla} J^t(\mathbf{a}(k))\mathbf{H}(\mathbf{a}(k))\boldsymbol{\nabla} J(\mathbf{a}(k))}.$$

We must, however, verify that this is indeed a minimum, and that $\eta(k) > 0$. We calculate the second derivative

$$\frac{\partial^2 J}{\partial \eta^2} = (\boldsymbol{\nabla} J)^t\mathbf{H}(\boldsymbol{\nabla} J).$$

If $\mathbf{a}$ is in a neighborhood of a local minimum of $J(\mathbf{a})$, the $\mathbf{H}$ will be positive definite and hence $(\boldsymbol{\nabla} J)^t \mathbf{H} (\boldsymbol{\nabla} J) > 0$. Thus J indeed attains minima at the optimal $\eta(k)$ and $\eta(k) > 0$ because $(\boldsymbol{\nabla} J)^t \mathbf{H} \boldsymbol{\nabla} J > 0$.

14. We are given samples from ω_1:

$$\begin{pmatrix} 1 \\ 5 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 9 \end{pmatrix}, \quad \begin{pmatrix} -5 \\ -3 \end{pmatrix}$$

and from ω_2:

$$\begin{pmatrix} 2 \\ -3 \end{pmatrix}, \quad \begin{pmatrix} -1 \\ -4 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

Augmenting the samples with an extra dimension, and inverting the sign of the samples from ω_2 define

$$\mathbf{Y} = \begin{pmatrix} 1 & 1 & 5 \\ 1 & 2 & 9 \\ 1 & -5 & -3 \\ -1 & -2 & 3 \\ -1 & 1 & 4 \\ -1 & 0 & -2 \end{pmatrix} \quad \mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b \\ b \\ b \\ b \\ b \\ b \end{pmatrix}.$$

Then the sum-of-squares error criterion is

$$\mathbf{J}_s(\mathbf{a}) = \frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{a}^t \mathbf{y}_i - b \right)^2 = \frac{(\mathbf{Ya} - \mathbf{b})^t (\mathbf{Ya} - \mathbf{b})}{2}.$$

(a) We differentiate the criterion function and find

$$\boldsymbol{\nabla} J_s(\mathbf{a}) = \boldsymbol{Y}^t (\mathbf{Ya} - \mathbf{b}) = \begin{pmatrix} 6a_1 & - & a_2 & + & 6a_3 & + & 0 \\ -a_1 & + & 35a_2 & + & 36a_3 & + & 3b \\ 6a_1 & + & 36a_2 & + & 144a_3 & - & 16b \end{pmatrix}$$

and

$$\mathbf{H} = \mathbf{Y}^t \mathbf{Y} = \begin{pmatrix} 6 & -1 & 6 \\ -1 & 35 & 36 \\ 6 & 36 & 144 \end{pmatrix}.$$

Notice that for this criterion, the Hessian is not a function of $\mathbf{a}$ since we assume a quadratic form throughout.

(b) The optimal step size varies with $\mathbf{a}$, that is,

$$\eta = \frac{[\boldsymbol{\nabla} J_s(\mathbf{a})]^t \boldsymbol{\nabla} J_s(\mathbf{a})}{[\boldsymbol{\nabla} J_s(\mathbf{a})]^t \mathbf{H} \boldsymbol{\nabla} J_s(\mathbf{a})}.$$

The range of optimal step sizes, however, will vary from the inverse of the largest eigenvalue of $\mathbf{H}$ to the inverse of the smallest eigenvalue of $\mathbf{H}$. By solving the characteristic equation or using singular value decomposition, the eigenvalues of $\mathbf{H}$ are found to be 5.417, 24.57, and 155.0, so $0.006452 \leq \eta \leq 0.1846$.

Section 5.5

15. We consider the Perceptron procedure (Algorithm 5.3 and Theorem 5.1 in the text).

(a) Equation 20 in the text gives the update rule

$$\begin{aligned} \mathbf{a}(1) &= \text{arbitrary} \\ \mathbf{a}(k+1) &= \mathbf{a}(k) + \mathbf{y}^k, \quad k \geq 1. \end{aligned}$$

It has been shown in the convergence proof in the text that

$$\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\|^2 \leq \|\mathbf{a}(k) - \alpha\hat{a}\|^2 - 2\alpha\nu + \beta^2,$$

where $\mathbf{a}(k)$ is the kth iterate, $\hat{\mathbf{a}}$ is a solution vector, α is a scale factor and β^2 and ν are specified by Eqs. 21 and 22 in the text:

$$\begin{aligned} \beta^2 &= \max_i \|y_i\|^2 \\ \nu &= \min_i \hat{\mathbf{a}}^t \mathbf{y}_i > 0. \end{aligned}$$

Thus we have

$$\|\mathbf{a}(k+1) - \alpha\hat{a}\|^2 \leq \|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\|^2 - (2\alpha\nu - \beta^2).$$

Because $\alpha > \beta^2/\nu$, we have the following inequality:

$$2\alpha\nu - \beta^2 2\beta^2 - \beta^2 = \beta^2 > 0.$$

Thus $\|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\|^2$ decreases (strictly) by an amount $(2\alpha\nu - \beta^2)$ at each iteration, and this implies

$$\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\|^2 \leq \|\mathbf{a}(1) - \alpha\hat{\mathbf{a}}\|^2 - k(2\alpha\nu - \beta)^2.$$

We denote the maximum number of correction convergence by k_o, we have

$$0 \leq \|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\|^2 \leq \|\mathbf{a}(1) - \alpha\hat{a}\|^2 - k(2\alpha\nu - \beta^2),$$

and thus

$$k \leq \frac{\|\mathbf{a}(1) - \alpha\hat{\mathbf{a}}\|^2}{2\alpha\nu - \beta^2},$$

and thus the *maximum* number of corrections is

$$k_o = \frac{\|\mathbf{a}(1) - \alpha\hat{\mathbf{a}}\|^2}{2\alpha\nu - \beta^2}.$$

(b) If we start out with a zero weight vector, that is, $\mathbf{a}(1) = \mathbf{0}$, we have

$$\begin{aligned} k_o &= \frac{\|\mathbf{a}(1) - \alpha\hat{\mathbf{a}}\|^2}{2\alpha\nu - \beta^2} \\ &= \frac{\|\alpha\hat{\mathbf{a}}\|^2}{2\alpha\nu - \beta^2} \\ &= \frac{\alpha^2\|\hat{\mathbf{a}}\|^2}{2\alpha\nu - \beta^2} \\ &= \frac{\alpha^2}{2\alpha\nu - \beta^2} \end{aligned}$$

In order to find the maximum of k_0 with respect to α, we set $\frac{\partial}{\partial\alpha}k_o(\alpha)$ to 0,

$$\begin{aligned}\frac{\partial}{\partial\alpha}k_o(\alpha) &= \frac{(2\alpha\nu-\beta^2)2\alpha\|\hat{\mathbf{a}}\|^2-\alpha^2\|\hat{\mathbf{a}}\|^2 2\nu}{(2\alpha\nu-\beta^2)^2}\\ &= \alpha^2(4\nu\|\hat{\mathbf{a}}\|^2-2\nu\|\hat{\mathbf{a}}\|^2)-2\alpha\beta^2\|\hat{\mathbf{a}}\|^2\\ &= 0,\end{aligned}$$

and this implies

$$\alpha\left[\alpha 2\nu\|\hat{\mathbf{a}}\|^2-2\beta^2\|\hat{\mathbf{a}}\|^2\right]=0.$$

We note that $\alpha>\beta^2/\nu>0$, which minimizes k_o if $\mathbf{a}(1)=\mathbf{0}$. We assume $b\geq 0$, and therefore

$$\nu=\min_i[\hat{\mathbf{a}}^t\mathbf{y}_i]>b\geq 0,$$

and $\nu>0$. We next use the fact that $0<\eta_a\leq\eta_k\leq\eta_b<\infty$, and $k\geq 1$ to find

$$\|\mathbf{a}(k+1)-\alpha\hat{\mathbf{a}}\|^2\leq\|\mathbf{a}(k)-\alpha\hat{\mathbf{a}}\|^2+\eta_b^2\beta^2+2\eta_b b-2\eta_a\alpha\nu.$$

We choose the scale factor α to be

$$\alpha=\frac{\eta_b^2\beta^2+2\eta_b b}{\eta_a\nu}$$

and find

$$\|\mathbf{a}(k+1)-\alpha\hat{\mathbf{a}}\|^2\leq\|\mathbf{a}(k)-\alpha\hat{\mathbf{a}}\|^2-(\eta_b^2\beta^2+2\eta_b b),$$

where $\eta_b^2\beta^2+2\eta_b b>0$. This means the difference between the weight vector at iteration $k+1$ and $\alpha\hat{\mathbf{a}}$ obeys:

$$\|\mathbf{a}(k+1)-\alpha\hat{\mathbf{a}}\|^2\leq\|\mathbf{a}(1)-\alpha\hat{\mathbf{a}}\|^2-k(\eta_b^2\beta^2+2\eta_b b).$$

Since a squared distance cannot become negative, it follows that the procedure must converge after almost k_o iterations where

$$k_o=\frac{\|\mathbf{a}(1)-\alpha\hat{\mathbf{a}}\|^2}{\eta_b^2\beta^2+2\eta_b b}.$$

In the case $b<0$ there is no guarantee that $\nu>0$ and thus there is no guarantee of convergence of the algorithm.

16. Our algorithm is given by Eq. 27 in the text:

$$\begin{aligned}\mathbf{a}(1) &= \text{arbitrary}\\ \mathbf{a}(k+1) &= \mathbf{a}(k)+\eta_k\mathbf{y}^k\end{aligned}$$

where in this case $\mathbf{a}^t(k)\mathbf{y}^k\leq b$, for all k. The η_k's are bounded by $0<\eta_a\leq\eta_k\leq\eta_b<\infty$, for $k\geq 1$. We shall modify the proof in the text to prove convergence given the above conditions.

Let $\hat{\mathbf{a}}$ be a solution vector. Then $\hat{\mathbf{a}}^t\mathbf{y}_i>b$, for all i. For a scale factor α we have

$$\begin{aligned}\mathbf{a}(k+1)-\alpha\hat{\mathbf{a}} &= \mathbf{a}(k)+\eta_k\mathbf{y}^k-\alpha\hat{\mathbf{a}}\\ &= (\mathbf{a}(k)-\alpha\hat{\mathbf{a}})+\eta_k\mathbf{y}^k,\end{aligned}$$

and thus

$$\begin{aligned}\|\mathbf{a}(k+1)-\alpha\hat{\mathbf{a}}\|^2 &= \|(\mathbf{a}(k)-\alpha\hat{\mathbf{a}})+\eta_k\mathbf{y}^k\|^2\\ &= \|\mathbf{a}(k)-\alpha\hat{\mathbf{a}}\|^2+\|\eta_k\mathbf{y}^k\|^2+2\eta_k(\mathbf{a}(k)-\alpha\hat{\mathbf{a}})^t\mathbf{y}^k\\ &= \|\mathbf{a}(k)-\alpha\hat{\mathbf{a}}\|^2+\eta_k^2\|\mathbf{y}^k\|^2+2\eta_k\mathbf{a}_k^t\mathbf{y}^k-2\eta_k\alpha\hat{\mathbf{a}}^t\mathbf{y}^k\\ &\leq \|\mathbf{a}(k)-\alpha\hat{\mathbf{a}}\|^2+\eta_k^2\|\mathbf{y}^k\|^2+2\eta_k b-2\eta_k\alpha\hat{\mathbf{a}}^t\mathbf{y}^k.\end{aligned}$$

As $\mathbf{y}^k$ was misclassified, $\mathbf{a}^t(k)\mathbf{y}^k \leq b$, for all k. Now we let

$$\begin{aligned}\beta^2 &= \max_i \|\mathbf{y}_i\|^2,\\ \nu &= \min_i[\hat{\mathbf{a}}^t\mathbf{y}_i] > b,\end{aligned}$$

since $\hat{\mathbf{a}}$ is a solution vector. Thus we substitute into the above

$$\|\mathbf{a}(k+1)-\alpha\hat{\mathbf{a}}\|^2 \leq \|\mathbf{a}(k)-\alpha\hat{\mathbf{a}}\|^2+\eta_k^2\beta^2+2\eta_k b-2\eta_k\alpha\nu,$$

and convergence is assured.

17. If $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are linearly separable then by definition there exists a separating hyperplane

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}+w_0$$

such that

$$\begin{aligned}g(\mathbf{y}_i) &> 0 \text{ if } \mathbf{y}_i \in \omega_1\\ g(\mathbf{y}_i) &< 0 \text{ if } \mathbf{y}_i \in \omega_2.\end{aligned}$$

We augment every $\mathbf{y}_i$ by 1, that is,

$$\tilde{\mathbf{y}}_i = \begin{pmatrix}1\\ \mathbf{y}_i\end{pmatrix}, \quad i=1,\ldots,n$$

and augment the weight vector by appending a bias weight w_0, written as

$$\tilde{\mathbf{w}} = \begin{pmatrix}w_0\\ \mathbf{w}\end{pmatrix}.$$

Then our linear descriminant function can be written as

$$\begin{aligned}\tilde{\mathbf{w}}^t\tilde{\mathbf{y}}_i &> 0 \quad \text{if } \mathbf{y}_i \in \omega_1\\ \tilde{\mathbf{w}}^t\tilde{\mathbf{y}}_i &< 0 \quad \text{if } \mathbf{y}_i \in \omega_2.\end{aligned}$$

We multiply $\tilde{\mathbf{y}}_i$ by -1 if $\mathbf{y}_i \in \omega_2$, and thus there exists a vector $\tilde{\mathbf{w}}$ such that $\tilde{\mathbf{w}}^t\tilde{\mathbf{y}}_i > 0$ for all i.

We use the logic of the Perceptron convergence to find such an $\tilde{\mathbf{w}}$ which requires at most k_o operations for convergence, where

$$k_o = \frac{\|\mathbf{a}(1)-\alpha\tilde{\mathbf{w}}\|^2}{\beta^2},$$

where $\tilde{\mathbf{w}}$ is the first iterate and

$$\begin{aligned}\alpha &= \frac{\beta^2}{\nu}\\ \beta^2 &= \max_i \|\mathbf{y}_i\|^2\\ \nu &= \min_i \tilde{\mathbf{a}}^t\tilde{\mathbf{y}}_i > 0.\end{aligned}$$

The remainder follows from the Perceptron convergence Theorem (Theorem 5.1) in the text.

18. Our criterion function is

$$J_q(\mathbf{a}) = \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{a})} (\mathbf{a}^t\mathbf{y} - b)^2,$$

where $\mathcal{Y}(\mathbf{a})$ is the set of samples for which $\mathbf{a}^t\mathbf{y} \leq b$.

Suppose $\mathbf{y}_1$ is the only sample in $\mathcal{Y}(\mathbf{a}(k))$. Then we have

$$\begin{aligned} J_q(\mathbf{a}(k)) &= (\mathbf{a}^t(k)\mathbf{y}_1 - b)^2 = (\mathbf{a}^t(k)\mathbf{y}_1)^2 + b^2 - 2b\mathbf{a}^t(k)\mathbf{y}_1 \\ &= \left(\sum_{j=1}^{d} a_{kj}y_{1j}\right)^2 + b^2 - 2b\sum_{j=1}^{d} a_{kj}y_{1j}. \end{aligned}$$

The derivative of the criterion function with respect to the components of $\mathbf{a}$ is

$$\begin{aligned} \frac{\partial J_q(\mathbf{a}(k))}{\partial a_{ki}} &= 2\left(\sum_{j=1}^{d} a_{kj}y_{1j}\right) y_{1i} - 2by_{1i} \\ &= 2(\mathbf{a}^t\mathbf{y}_1 - b)y_{1i}, \end{aligned}$$

for $i = 1, \ldots, d$. Thus we have

$$\begin{aligned} \boldsymbol{\nabla} J_q(\mathbf{a}(k)) &= \frac{\partial J_q(\mathbf{a}(k))}{\partial \mathbf{a}(k)} = 2(\mathbf{a}^t\mathbf{y}_1 - b)\mathbf{y}_1 \\ D_{ii'} &= \frac{\partial^2 J_q(\mathbf{a}(k))}{\partial a_{ki'}\partial a_{ki}} \\ &= \frac{\partial}{\partial a_{ki'}}\left[2\left(\sum_{j=1}^{d} a_{kj}y_{1j}\right) y_{1i} - 2by_{1i}\right] \\ &= 2y_{1i'}y_{1i} \end{aligned}$$

for $i, i' = 1, \ldots, d$. This implies

$$\mathbf{D} = 2\mathbf{y}_1\mathbf{y}_i^t.$$

From Eq. 34 in the text, we have that the basic gradient descent algorithm is given by

$$\begin{aligned} \mathbf{a}(1) &= \text{arbitrary} \\ \mathbf{a}(k+1) &= \mathbf{a}(k) + \eta_k \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{a}(k))} \frac{b - \mathbf{a}^t(k)\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}_1, \end{aligned}$$

and thus

$$\mathcal{Y}(\mathbf{a}(k)) = \{\mathbf{y}_1\},$$

which implies

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \eta_k \frac{b - \mathbf{a}^t(k)\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}_1.$$

We also have

$$\mathbf{a}^t(k+1)\mathbf{y}_1 - b = (1-\eta_k)(\mathbf{a}^t(k)\mathbf{y}_1 - b).$$

If we choose $\eta_k = 1$, then $\mathbf{a}(k+1)$ is moved exactly to the hyperplane $\mathbf{a}^t\mathbf{y}_1 = b$. Thus, if we optimally choose $\eta_k = 1$, we have Eq. 35 in the text:

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \frac{b - \mathbf{a}^t(k)\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}_1.$$

19. We begin by following the central equation and proof of Perceptron Convergence given in the text:

$$\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\|^2 = \|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\|^2 + 2(\mathbf{a}(k) - \alpha\hat{\mathbf{a}})^t\eta\mathbf{y}^k + \eta^2\|\mathbf{y}^k\|^2,$$

where $\mathbf{y}^k$ is the kth training presentation. (We suppress the k dependence of η for clarity.) An update occurs when $\mathbf{y}^k$ is misclassified, that is, when $\mathbf{a}^t(k)\mathbf{y}^k \leq b$. We define $\beta^2 = \max_i \|\mathbf{y}^i\|^2$ and $\gamma = \min_i[\hat{\mathbf{a}}^t\mathbf{y}^i] > b > 0$. Then we can write

$$\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\|^2 \leq \|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\|^2 + 2(b - \alpha\gamma)\eta + \eta^2\beta^2,$$

where α is a positive free parameter. We can define

$$\alpha(k) = \frac{1}{\gamma}\left(\sum_{k=1}^{m-1}\eta(k) + b\right)$$

at the mth update. Then we have

$$\|\mathbf{a}(k+1) - \alpha\hat{\mathbf{a}}\|^2 \leq \|\mathbf{a}(k) - \alpha\hat{\mathbf{a}}\|^2 - 2\eta(k)\sum_{l=1}^{m-1}\eta(l) + \eta^2\beta^2.$$

Summing all such equations for the mth update gives

$$\|\mathbf{a}(k+1) - \alpha(k)\hat{\mathbf{a}}\|^2 \leq \|\mathbf{a}(0) - \alpha(0)\hat{\mathbf{a}}\|^2 - 2\sum_{k=1}^{m}\eta(k)\sum_{l=1}^{k}\eta(l) + \sum_{k=1}^{m}\eta^2(k)\beta^2,$$

or equivalently

$$\|\mathbf{a}(k+1) - \alpha(k)\hat{\mathbf{a}}\|^2 \leq \|\mathbf{a}(0) - \alpha(0)\hat{\mathbf{a}}\|^2 - 2\sum_{k,l\neq k}^{m}\eta(k)\eta(l) + \beta^2\sum_{k=1}^{m}\eta^2(k). \qquad (*)$$

On the other hand, we are given that

$$\lim_{m\to\infty}\frac{\sum_{k=1}^{m}\eta^2}{\left(\sum_{k=1}^{m}\eta(k)\right)^2} = 0$$

and we note

$$\left(\sum_{k=1}^{m}\eta(k)\right)^2 = \sum_{k=1}^{m}\eta^2(k) + 2\sum_{(kl)}^{m}\eta(k)\eta(l).$$

We also have

$$\lim_{m\to\infty} \frac{1}{\left(\sum_{k=1}^{m} \eta(k)\right)^2} \left[\sum_{k=1}^{m} \eta^2(k) + 2\sum_{(kl)}^{m} \eta(k)\eta(l) - 2\sum_{(kl)}^{m} \eta(k)\eta(l)\right] = 0$$

which implies

$$\lim_{m\to\infty} \left[1 - \frac{2\sum_{(kl)}^{m} \eta(k)\eta(l)}{\left(\sum_{k=1}^{m} \eta(k)\right)^2}\right] = 0$$

or

$$\lim_{m\to\infty} \frac{2\sum_{(kl)}^{m} \eta(k)\eta(l)}{\left(\sum_{k=1}^{m} \eta(k)\right)^2} = 1. \quad (**)$$

Now we can reconsider $(*)$:

$$\|\mathbf{a}(k+1) - \alpha(k)\hat{\mathbf{a}}\|^2$$
$$\leq \|\mathbf{a}(0) - \alpha(0)\hat{\mathbf{a}}\|^2 - \left(\sum_{k=1}^{m} \eta(k)\right)^2 \left[\frac{2\sum_{(kl)}^{m} \eta(k)\eta(l)}{\left(\sum_{k=1}^{m} \eta(k)\right)^2} - \beta^2 \frac{\sum_{k=1}^{n} \eta^2(k)}{\left(\sum_{k=1}^{n} \eta(k)\right)^2}\right].$$

But we know from $(**)$ that the first term in the brackets goes to 1 and the coefficient of β^2 goes to 0, and all the $\eta(k)$ terms (and their sum) are positive. Moreover, the corrections will never cease, since $\sum_{k=1}^{M} \eta(k) \to \infty$, as long as there are incorrectly classified samples. But the distance term on the left cannot be negative, and thus we conclude that the corrections must cease. This implies that all the samples will be classified correctly.

Section 5.6

20. As shown in the figure, in this case the initial weight vector is marked **0**, and after the successive updates $\mathbf{1}, \mathbf{2}, \ldots \mathbf{12}$, where the updates cease. The sequence of presentations is $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$, $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$, $\mathbf{y}_1$, $\mathbf{y}_3$, $\mathbf{y}_1$, $\mathbf{y}_3$, $\mathbf{y}_2$, and $\mathbf{y}_1$.

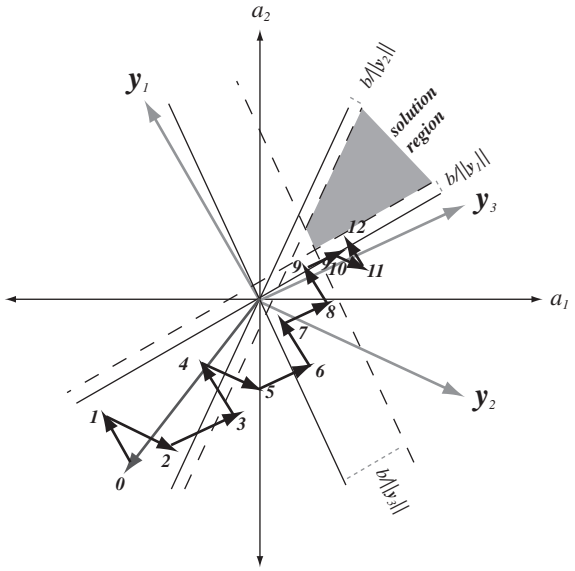
Section 5.7

21. From Eq. 54 in the text we have the weight vector

$$\mathbf{w} = \alpha n \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2),$$

where $\mathbf{w}$ satisfies

$$\left[\frac{1}{n}\mathbf{S}_w + \frac{n_1 n_2}{n^2}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\right]\mathbf{w} = \mathbf{m}_1 - \mathbf{m}_2.$$



We substitute $\mathbf{w}$ into the above and find

$$\left[\frac{1}{n}\mathbf{S}_w + \frac{n_1 n_2}{n^2}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\right]\alpha n\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2) = \mathbf{m}_1 - \mathbf{m}_2,$$

which in turn implies

$$\alpha\left[\mathbf{m}_1 - \mathbf{m}_2 + \frac{n_1 n_2}{n^2}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right] = \mathbf{m}_1 - \mathbf{m}_2.$$

Thus we have

$$\alpha\theta(\mathbf{m}_1 - \mathbf{m}_2) = \mathbf{m}_1 - \mathbf{m}_2$$

where

$$\theta = 1 + \frac{n_1 n_2}{n^2}(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2).$$

This equation is valid for all vectors $\mathbf{m}_1$ and $\mathbf{m}_2$, and therefore we have $\alpha\theta = 1$, or

$$\alpha = \left[1 + \frac{n_1 n_2}{n^2}(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right]^{-1}.$$

22. We define the discriminant function

$$g_0(\mathbf{x}) = (\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) - (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x}).$$

Our goal is to show that the vector $\mathbf{a}$ that minimizes

$$J_s'(\mathbf{a}) = \sum_{\mathbf{y}\in\mathcal{Y}_1}(\mathbf{a}^t\mathbf{y} - (\lambda_{21} - \lambda_{11}))^2 + \sum_{\mathbf{y}\in\mathcal{Y}_2}(\mathbf{a}^t\mathbf{y} + (\lambda_{12} - \lambda_{22}))^2$$

is asymptotically equivalent to the vector $\mathbf{a}$ that minimizes

$$\varepsilon^2 = \int[\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x})]^2 p(\mathbf{x})d\mathbf{x},$$

as given in Eq. 57 in the text. We proceed by calculating the criterion function

$$\begin{aligned} J_s'(\mathbf{a}) &= \sum_{\mathbf{y}\in\mathcal{Y}_1}\left(\mathbf{a}^t\mathbf{y}-(\lambda_{21}-\lambda_{11})\right)^2+\sum_{\mathbf{y}\in\mathcal{Y}_2}\left(\mathbf{a}^t\mathbf{y}-(\lambda_{12}-\lambda_{22})\right)^2 \\ &= n\left[\frac{n_1}{n}\frac{1}{n}\sum_{\mathbf{y}\in\mathcal{Y}_1}\left(\mathbf{a}^t\mathbf{y}-(\lambda_{21}-\lambda_{11})\right)^2+\frac{n_2}{n}\frac{1}{n}\sum_{\mathbf{y}\in\mathcal{Y}_2}\left(\mathbf{a}^t\mathbf{y}+(\lambda_{12}-\lambda_{22})\right)^2\right]. \end{aligned}$$

By the law of large numbers, as $n\to\infty$ we have with probability 1

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{n}J_s'(\mathbf{a}) &= J'(\mathbf{a}) \\ &= P(\omega_1)\mathcal{E}_1\left[(\mathbf{a}^t\mathbf{y}-(\lambda_{21}-\lambda_{11}))^2\right]+P(\omega_2)\mathcal{E}_2\left[(\mathbf{a}^t\mathbf{y}+(\lambda_{12}-\lambda_{22}))^2\right]. \end{aligned}$$

We expand the terms involving expectations as

$$\begin{aligned} \mathcal{E}_1\left[(\mathbf{a}^t\mathbf{y}-(\lambda_{21}-\lambda_{11}))^2\right] &= \int\left(\mathbf{a}^t\mathbf{y}-(\lambda_{21}-\lambda_{11})\right)^2 p(\mathbf{x}|\omega_1)d\mathbf{x}, \\ \mathcal{E}_2\left[(\mathbf{a}^t\mathbf{y}+(\lambda_{21}-\lambda_{11}))^2\right] &= \int(\mathbf{a}^t\mathbf{y}+(\lambda_{12}-\lambda_{22}))^2 p(\mathbf{x}|\omega_2)d\mathbf{x}. \end{aligned}$$

We substitute these into the above equations and find that our criterion function is

$$\begin{aligned} J'(\mathbf{a}) &= \int\left(\mathbf{a}^t\mathbf{y}-(\lambda_{21}-\lambda_{11})\right)^2 p(\mathbf{x}|\omega_1)d\mathbf{x} \\ &\quad+\int\left(\mathbf{a}^t\mathbf{y}+(\lambda_{12}-\lambda_{22})\right)^2 p(\mathbf{x}|\omega_2)d\mathbf{x} \\ &= \int\left(\mathbf{a}^t\mathbf{y}-(\lambda_{21}-\lambda_{11})\right)^2 p(\mathbf{x},\omega_1)d\mathbf{x} \\ &\quad+\int\left(\mathbf{a}^t\mathbf{y}+(\lambda_{12}-\lambda_{22})\right)^2 p(\mathbf{x},\omega_2)d\mathbf{x} \\ &= \int\left(\mathbf{a}^t\mathbf{y}\right)^2[p(\mathbf{x},\omega_1)+p(\mathbf{x},\omega_2)]d\mathbf{x} \\ &\quad-2\int\mathbf{a}^t\mathbf{y}\left[(\lambda_{21}-\lambda_{11})p(\mathbf{x},\omega_1)+(\lambda_{12}-\lambda_{22})p(\mathbf{x},\omega_2)\right]d\mathbf{x} \\ &\quad+(\lambda_{21}-\lambda_{11})^2\int p(\mathbf{x},\omega_1)d\mathbf{x}+(\lambda_{12}-\lambda_{22})^2\int p(\mathbf{x},\omega_2)d\mathbf{x} \\ &= \int(\mathbf{a}^t\mathbf{y})^2p(\mathbf{x})d\mathbf{x}+2\int\mathbf{a}^t\mathbf{y}g_0(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &\quad+(\lambda_{21}-\lambda_{11})^2P(\omega_1)+(\lambda_{12}-\lambda_{22})^2P(\omega_2), \end{aligned}$$

where we have used the fact that

$$p(\mathbf{x})=p(\mathbf{x},\omega_1)+p(\mathbf{x},\omega_2)$$

and

$$\begin{aligned} (\lambda_{21}-\lambda_{11})p(\mathbf{x},\omega_1)+(\lambda_{12}-\lambda_{22})p(\mathbf{x},\omega_2) &= (\lambda_{21}-\lambda_{11})p(\mathbf{x}|\omega_1)+(\lambda_{12}-\lambda_{22})p(\mathbf{x}|\omega_2) \\ &= g_0(\mathbf{x})p(\mathbf{x}). \end{aligned}$$

Furthermore, the probability of finding a particular category ω_i is simply

$$P(\omega_i)=\int p(\mathbf{x},\omega_i)d\mathbf{x},$$

and thus our criterion function can be written

$$\begin{aligned} J'(\mathbf{a}) &= \int \left[\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x})\right]^2 p(\mathbf{x})d\mathbf{x} \\ &\quad + \underbrace{\int [(\lambda_{21}-\lambda_{11})^2 P(\omega_1) + (\lambda_{12}-\lambda_{22})^2 P(\omega_2) - g_o^2(\mathbf{x})]p(\mathbf{x})d\mathbf{x}}_{independent\ of\ \mathbf{a}}. \end{aligned}$$

Since the second term in the above expression is independent of $\mathbf{a}$, minimizing $J'(\mathbf{a})$ with respect to $\mathbf{a}$ is equivalent to minimizing

$$\varepsilon^2 = \int [\mathbf{a}^t\mathbf{y} - g_0(\mathbf{x})]^2 p(\mathbf{x})d\mathbf{x}.$$

In summary, the vector $\mathbf{a}$ that minimizes $J_s'(\mathbf{a})$ provides asymptotically a minimum mean-square error approximation to the Bayes discriminant function $g_0(\mathbf{x})$.

23. Given Eqs. 66 and 67 in the text,

$$\begin{aligned} \mathbf{a}(k+1) &= \mathbf{a}(k) + \eta(k)[\theta(k) - \mathbf{a}^t(k)\mathbf{y}_k]\mathbf{y}_k \\ \hat{\mathbf{a}} &= \mathcal{E}[\mathbf{y}\mathbf{y}^t]^{-1}\mathcal{E}[\theta\mathbf{y}] \end{aligned}$$

we need to show

$$\lim_{n\to\infty} \mathcal{E}[\|\mathbf{a}(k) - \hat{\mathbf{a}}\|^2] = 0$$

given the two conditions

$$\begin{aligned} \sum_{k=1}^{\infty} \eta(k) &= \infty \\ \sum_{k=1}^{\infty} \eta^2(k) &< \infty. \end{aligned}$$

We write $\|\mathbf{a} - \hat{\mathbf{a}}\|^2$ using Eq. 67 as

$$\begin{aligned} \|\mathbf{a}(k+1) - \hat{\mathbf{a}}\|^2 &= \|\mathbf{a} - \hat{\mathbf{a}} + \eta(k)(\theta(k) - \mathbf{a}^t(k)\mathbf{y}_k)\mathbf{y}_k\|^2 \\ &= \|\mathbf{a} - \hat{\mathbf{a}}\|^2 + \eta^2(k)\|(\theta(k) - \mathbf{a}^t(k)\mathbf{y}_k)\mathbf{y}_k\|^2 \\ &\quad +2\eta(k)(\mathbf{a}(k) - \hat{\mathbf{a}})^t(\theta_k - \mathbf{a}^t(k)\mathbf{y}_k)\mathbf{y}_k. \end{aligned}$$

We take the expected value of both sides and find

$$\begin{aligned} \mathcal{E}[\|\mathbf{a}(k+1) - \hat{\mathbf{a}}\|^2] &= \underbrace{\mathcal{E}[\|\mathbf{a}(k) - \hat{\mathbf{a}}\|^2]}_{\epsilon_1} + \eta^2(k)\underbrace{\mathcal{E}[(\theta_k - \mathbf{a}^t(k)\mathbf{y}_k)\mathbf{y}_k]}_{\epsilon_2} \\ &\quad +2\eta(k)\underbrace{\mathcal{E}[(\mathbf{a}(k) - \hat{\mathbf{a}})^t(\theta(k) - \mathbf{a}^t(k)\mathbf{y}_k)\mathbf{y}_k]}_{\epsilon_3}. \end{aligned}$$

Note that $\epsilon_1 \geq 0$ and $\epsilon_2 \geq 0$. If we can show that $\epsilon_3 < 0$ for all $k > M$ for some finite M, then we will have proved our needed result. This is because that the update rule must reduce $\mathcal{E}[\|\mathbf{a}(k+1) - \hat{\mathbf{a}}\|^2]$ for $k > R$ for some finite R since $\sum_{k=1}^{\infty} \eta^2(k)$ is finite whereas $\sum_{k=1}^{\infty} \eta(k)$ is infinite. We know that $\mathcal{E}[\|\mathbf{a}(k+1) - \hat{\mathbf{a}}\|^2] \geq 0$, and thus

$$\lim_{k\to\infty} \mathcal{E}[\|\mathbf{a}(k+1) - \hat{\mathbf{a}}\|^2]$$

must vanish.

Now we seek to show that indeed $\epsilon_3 < 0$. We find

$$\begin{aligned}
\epsilon_3 &= \mathcal{E}[(\mathbf{a}(k)-\hat{\mathbf{a}})^t(\theta(k)-\mathbf{a}^t(k)\mathbf{y})\mathbf{y}_k] \\
&= \mathcal{E}[\mathbf{a}^t(k)\theta(k)\mathbf{y}_k] - \mathcal{E}[(\mathbf{a}^t(k)\mathbf{y}_k)^2] - \mathcal{E}[\theta(k)\hat{\mathbf{a}}^t\mathbf{y}_k] + \mathcal{E}[\hat{\mathbf{a}}^t\mathbf{y}_k\mathbf{a}^t\mathbf{y}_k] \\
&= \mathcal{E}[\theta(k)\mathbf{a}^t(k)\mathbf{y}_k] + \mathcal{E}[\hat{\mathbf{a}}^t\mathbf{y}_k\mathbf{a}^t\mathbf{y}_k] - \mathcal{E}[(\mathbf{a}^t(k)\mathbf{y}_k)^2] - \mathcal{E}[(\mathbf{a}^t\mathbf{y}_k)^2] \\
&= -\mathcal{E}[(\mathbf{a}^t\mathbf{y}_k)^2 + (\hat{\mathbf{a}}^t\mathbf{y}_k)^2 - 2\mathbf{a}^t(k)\mathbf{y}_k\hat{\mathbf{a}}^t\mathbf{y}_k] \\
&= -\mathcal{E}[(\mathbf{a}^t(k)\mathbf{y}_k - \hat{\mathbf{a}}^t\mathbf{y}_k)^2] \\
&= -\mathcal{E}[(\mathbf{a}^t(k)\mathbf{y}_k - \hat{\mathbf{a}}^t\mathbf{y}_k)^2] \leq 0.
\end{aligned}$$

In the above, we used the fact that $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k$ is determined, so we can consider $\mathbf{a}^t(k)$ as a non-random variable. This enables us to write $\mathcal{E}[\theta(k)\mathbf{a}^t(k)\mathbf{y}_k]$ as $\mathcal{E}[\mathbf{a}^t(k)\mathbf{y}_k\mathbf{y}_k^t\hat{\mathbf{a}}]$.

Thus we have proved our claim that $\epsilon_3 < 0$, and hence the full proof is complete.

24. Our criterion function here is

$$J_m(\mathbf{a}) = \mathcal{E}[(\mathbf{a}^t\mathbf{y} - z)^2].$$

(a) We expand this criterion function as

$$\begin{aligned}
J_m(\mathbf{a}) &= \mathcal{E}[(\mathbf{a}^t\mathbf{y} - z)^2] \\
&= \mathcal{E}[\{(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x})) - (z - g_o(\mathbf{x}))\}^2] \\
&= \mathcal{E}[(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))^2 - 2(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))(z - g_o(\mathbf{x})) + (z - g_o(\mathbf{x}))^2] \\
&= \mathcal{E}[(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))^2] - 2\mathcal{E}[(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))(z - g_o(\mathbf{x}))] + \mathcal{E}[(z - g_o(\mathbf{x}))^2].
\end{aligned}$$

(b) We use the fact that $\mathcal{E}[z|\mathbf{x}] = g_o(\mathbf{x})$ to note that

$$\mathcal{E}[(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))(z - g_o(\mathbf{x}))] = \mathcal{E}[(\mathbf{a}^t\mathbf{y}(\mathbf{x}) - g_o(\mathbf{x}))(z - g_o(\mathbf{x}))].$$

This leads to

$$\mathcal{E}\{\mathcal{E}[(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))(z - g_o(\mathbf{x}))|\mathbf{x}]\} = \mathcal{E}\{(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))\mathcal{E}[(z - g_o(\mathbf{x}))|\mathbf{x}]\},$$

since conditioned on $\mathbf{x}, \mathbf{a}^t y(\mathbf{x}) - g_o(\mathbf{x})$ is a constant. Thus we have

$$\mathcal{E}\{(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))[\mathcal{E}(z|\mathbf{x}) - g_o(\mathbf{x})]\} = 0$$

as $\mathcal{E}[z|\mathbf{x}] = g_o(\mathbf{x})$. We put these results together to get

$$J_m(\mathbf{a}) = \mathcal{E}[(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))^2] + \underbrace{\mathcal{E}[(z - g_o(\mathbf{x}))^2]}_{independent\ of\ \mathbf{a}},$$

where the second term in the expression does not depend on $\mathbf{a}$. Thus the vector $\mathbf{a}$ that minimizes J_m also minimizes $\mathcal{E}[(\mathbf{a}^t\mathbf{y} - g_o(\mathbf{x}))^2]$.

25. We are given that

$$\eta_{k+1}^{-1} = \eta_k^{-1} + y_k^2.$$

(a) We solve for η_k^{-1} as

$$\begin{aligned}\eta_k^{-1} &= \eta_{k-1}^{-1} + y_{k-1}^2 \\ &= \eta_{k-2}^{-1} + y_{k-2}^2 + y_{k-1}^2 \\ &= \eta_1^{-1} + y_1^2 + y_2^2 + \ldots + y_{k-1}^2 \\ &= \frac{1 + \eta_1 \sum_{i=1}^{k-1} y_i^2}{\eta_1},\end{aligned}$$

and thus

$$\eta_k = \frac{\eta_1}{1 + \eta_1 \sum_{i=1}^{k-1} y_i^2}$$

for $\eta_1 > 0$ and $0 < a \leq y_i^2 \leq b$.

(b) With these ranges we have

$$a(k-1) \leq \sum_{i=1}^{k-1} y_i^2 \leq b(k-1),$$

which implies

$$\begin{aligned}\eta_k &= \frac{\eta_1}{1 + \eta_1 \sum_{i=1}^{k-1} y_i^2} \\ &\leq \frac{\eta_1}{1 + a(k-1)} \\ &\geq \frac{\eta_1}{1 + b(k-1)}.\end{aligned}$$

This is true for each k, so we can sum to find

$$\sum_k \eta_k \geq \sum_k \frac{\eta_1}{1 + b(k-1)} \rightarrow \infty$$

as

$$\sum_{k=1}^{\infty} \frac{1}{k} = \infty \text{ and } b > 0.$$

Moreover, we have

$$\sum_k \eta_k^2 \leq \sum_k \frac{\eta_1^2}{[1 + a(k-1)]^2} \rightarrow L_1 < \infty$$

as

$$\sum_{k=1}^{\infty} \frac{1}{k^2} < \infty \text{ and } a > 0.$$

Consequently, the learning rates obey

$$\sum_k \eta_k \rightarrow \infty \text{ and } \sum_k \eta_k^2 \rightarrow L < \infty.$$

26. We rewrite the LMS rule using the following notation:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^t \\ \mathbf{y}_2^t \\ \vdots \\ \mathbf{y}_n^t \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

Then the LMS rule before Eq. 61 in the text is

$$\begin{aligned} \mathbf{a}(1) &\quad \text{arbitrary} \\ \mathbf{a}(k+1) &= \mathbf{a}(k) + \eta(k) \sum_{i=1}^{n} y_i(b_i - \mathbf{a}_k^t \mathbf{y}_i). \end{aligned}$$

Note that the condition we are looking for the limiting vector $\hat{\mathbf{a}}$ now reads

$$\sum_{i=1}^{n} \mathbf{y}_i(\mathbf{y}_i^t \hat{\mathbf{a}} - b_i) = \mathbf{0},$$

or equivalently

$$\sum_{i=1}^{n} \mathbf{y}_i(\hat{\mathbf{a}}^t \mathbf{y}_i) = \sum_{i=1}^{n} \mathbf{y}_i b_i.$$

Now we consider how the distance between $\mathbf{a}(k)$ and $\hat{\mathbf{a}}$ changes during an update:

$$\begin{aligned} \|\mathbf{a}(k+1) - \hat{\mathbf{a}}\|^2 &= \|\mathbf{a}(k) - \hat{\mathbf{a}}\|^2 + \frac{\eta^2(1)}{k^2} \underbrace{\left\| \sum_{i=1}^{n} \mathbf{y}_i(b_i - \mathbf{a}^t(k)\mathbf{y}_i) \right\|^2}_{C_k} \\ &\quad + \frac{2\eta(1)}{k} \underbrace{(\mathbf{a}(k) - \hat{\mathbf{a}})^t \sum_{i=1}^{n} \mathbf{y}_i(b_i - \mathbf{a}^t(k)\mathbf{y}_i)}_{D_k} \\ &= \|\mathbf{a}(k) - \hat{\mathbf{a}}\|^2 + \frac{\eta^2(1)}{k^2} C_k + \frac{2\eta(1)}{k} D_k, \end{aligned}$$

where for convenience we have defined C_k and D_k. Clearly $C_k \geq 0$ as it is the sum of non-negative values. Consider now D_k, which we expand as

$$D_k = \sum_{i=1}^{n} \left[-\mathbf{a}^t(k)\mathbf{y}_i\mathbf{a}^t(k)\mathbf{y}_i + \mathbf{a}^t(k)\mathbf{y}_i b_i + \hat{\mathbf{a}}^t\mathbf{y}_i\mathbf{a}^t(k)\mathbf{y}_i - \hat{\mathbf{a}}^t\mathbf{y}_i b_i \right]$$

We can substitute $\sum_{i=1}^{n} \mathbf{y}_i b_i$ with $\sum_{i=1}^{n} \mathbf{y}_i\hat{\mathbf{a}}^t\mathbf{y}_i$, from our definition of $\hat{\mathbf{a}}$. Then D_k can be written

$$\begin{aligned} D_k &= -\sum_{i=1}^{n} (\mathbf{a}^t(k)\mathbf{y}_i)^2 - \sum_{i=1}^{n} (\hat{\mathbf{a}}^t(k)\mathbf{y}_i)^2 + \mathbf{a}^t(k) \sum_{i=1}^{n} \mathbf{y}_i b_i + \hat{\mathbf{a}}^t \sum_{i=1}^{n} \mathbf{y}_i\mathbf{a}^t(k)\mathbf{y}_i \\ &= \sum_{i=1}^{n} \left[-(\mathbf{a}^t(k)\mathbf{y}_i)^2 - (\hat{\mathbf{a}}^t\mathbf{y}_i)^2 + \mathbf{a}^t(k)\mathbf{y}_i\hat{\mathbf{a}}^t\mathbf{y}_i + \hat{\mathbf{a}}^t\mathbf{y}_i\mathbf{a}^t(k)\mathbf{y}_i \right] \end{aligned}$$

and thus

$$D_k = -\sum_{i=1}^{n}(\mathbf{a}^t(k)\mathbf{y}_i - \hat{\mathbf{a}}^t\mathbf{y}_i)^2 \leq 0.$$

Adding the m update equations for the LMS rule we find

$$\|\mathbf{a}(m+1) - \hat{\mathbf{a}}\|^2 = \|\mathbf{a}(1) - \hat{\mathbf{a}}\|^2 + \eta^2(1)\sum_{k=1}^{m}\frac{C_k}{k^2} + 2\eta(1)\sum_{k=1}^{m}\frac{D_k}{k}.$$

Now we take the limit $m \to \infty$ for both sides:

$$\begin{aligned}\lim_{m\to\infty}\|\mathbf{a}(m+1) - \hat{\mathbf{a}}\|^2 &= \|\mathbf{a}(1) - \hat{\mathbf{a}}\|^2 + \eta^2(1)\sum_{k=1}^{\infty}\frac{C_k}{k^2} + 2\eta(1)\sum_{k=1}^{\infty}\frac{D_k}{k} \\ &= \|\mathbf{a}(1) - \hat{\mathbf{a}}\|^2 + \eta^2(1)C_L + 2\eta(1)\sum_{k=1}^{\infty}\frac{D_k}{k},\end{aligned}$$

where if C_k is bounded we have

$$C_L = \sum_{k=1}^{\infty}\frac{C_k}{k^2} < \infty.$$

We also know that if $D_k < 0$

$$\sum_{k=1}^{\infty}\frac{D_k}{k} \to -\infty$$

for all k. But $D_k < 0$ cannot be true for all k (except for finite occassions), otherwise the right-hand side of the equation will be negative while the left-hand side will be non-negative. Thus D_k must be arbitrarily close to zero for some choice N, for all $k > N$. This tells us that

$$\lim_{k\to\infty}\mathbf{a}^t(k)\mathbf{y}_i^k = \hat{\mathbf{a}}^t\mathbf{y}_i^k$$

for $i = 1, 2, \ldots, N$.

To see this, we start with arbitrarily smally positive λ. Then, we know $|D_k| < \lambda$ for all $k > N$ for some integer N. This implies

$$|D_k| = \sum_{i=1}^{n}(\mathbf{a}^t(k)\mathbf{y}_i - \hat{\mathbf{a}}^t\mathbf{y}_i)^2 \leq \lambda,$$

in particular $|\mathbf{a}^t(k)\mathbf{y}_i - \hat{\mathbf{a}}^t\mathbf{y}_i| \leq \sqrt{\lambda}$ will also be true. Now we let

$$\begin{aligned}\mathbf{v}(k) &= \sum_{i=1}^{n}\mathbf{y}_i(b_i - \mathbf{y}_i^t\mathbf{a}(k)) \\ &= \sum_{i=1}^{n}\mathbf{y}_i(b_i - [\hat{\mathbf{a}}^t\mathbf{y}_i \mp \sqrt{\lambda_i}]) \\ &= \sum_{i=1}^{n}\mathbf{y}_i(b_i - \hat{\mathbf{a}}^t\mathbf{y}_i) \mp \mathbf{y}_i\sqrt{\lambda_i} \\ &= \sum_{i=1}^{n}\mp\mathbf{y}_i\sqrt{\lambda_i}\end{aligned}$$

where $\sqrt{\lambda_i} = |\mathbf{a}^t(k)\mathbf{y}_i - \hat{\mathbf{a}}^t\mathbf{y}_i|$.

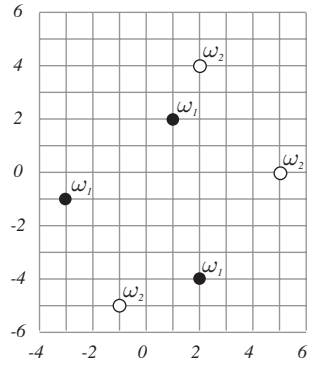
Since λ was arbitrarily close to zero and $|\lambda_i| \leq |\lambda|$, we have

$$\lim_{k\to\infty} \mathbf{v}(k) = \mathbf{0},$$

proving that $\mathbf{a}(k)$ satisfies the limit condition.

Section 5.9

27. The six points are shown in the figure, labeled by their category membership.



(a) By inspecting the plot of the points, we see that ω_1 and ω_2 are not linearly separable, that is, there is no line that can separate all the ω_1 points from all the ω_2 points.

(b) Equation 85 in the text shows how much the $\|\mathbf{e}\|^2$ value is reduced from step k to step $k+1$. For faster convergence we want the maximum possible reduction. We can find an expression for the learning rate η by solving the maxima problem for the right-hand side of Eq. 85,

$$\frac{1}{4}\left(\|\mathbf{e}(k)\|^2 - \|\mathbf{e}(k+1)\|^2\right) = \eta(1-\eta)\|\mathbf{e}^+(k)\|^2 + \eta^2\mathbf{e}^{+t}(k)\mathbf{Y}\mathbf{Y}^\dagger\mathbf{e}^+(k). \qquad (*)$$

We take the derivative with respect to η of the right-hand side, set it to zero and solve:

$$\|\mathbf{e}^+(k)\|^2 - 2\eta\|\mathbf{e}^+(k)\|^2 + 2\eta\mathbf{e}^+(k)\mathbf{Y}\mathbf{Y}^t\mathbf{e}^+(k) = 0,$$

which gives the optimal learning rate

$$\eta_{opt}(k) = \frac{\|\mathbf{e}\|^2}{2\left[\|\mathbf{e}^+(k)\|^2 - \mathbf{e}^+(k)\mathbf{Y}\mathbf{Y}^t\mathbf{e}^+(k)\right]}.$$

The maximum eigenvalue of $\mathbf{Y}^t\mathbf{Y}$ is $\lambda_{max} = 66.4787$. Since all the η values between 0 and $2/\lambda_{max}$ ensure convergence to take the largest step in each iteration we should choose η as large as possible. That is, since $0 < \eta < 2/66.4787$, we can choose $\eta_{opt} = 0.0301 - \epsilon$, where ϵ is an arbitrarily small positive constant.

However, $(*)$ is not suitable for fixed learning rate Ho-Kashyap algorithms. To get a *fixed* optimal η, be proceed as discussed in pages 254–255 in the text: If

we use the algorithm expressed in Eq. 90,

$$\begin{aligned} \mathbf{b}(1) &> 0 \text{ but otherwise arbitrary} \\ \mathbf{a}(1) & \quad \text{arbitrary} \\ \mathbf{b}(k+1) &= \mathbf{b}(k) + (\mathbf{e}(k) + |\mathbf{e}(k)|) \\ \mathbf{a}(k+1) &= \mathbf{a}(k) + \eta \mathbf{R}\mathbf{Y}^t|\mathbf{e}(k)|, \end{aligned}$$

where $\mathbf{R}$ is arbitrary positive definite. As discussed on page 255 of the text, if $\mathbf{R} = \mathbf{I}$, then $\mathbf{A} = 2\eta\mathbf{I} - \eta^2\mathbf{Y}^t\mathbf{Y}$ will be positive definite, thereby ensuring convergence if $0 < \eta < 2/\lambda_{max}$, where λ_{max} is the largest eigenvalue of $\mathbf{Y}^t\mathbf{Y}$. For our case, we have

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & -4 \\ 1 & -3 & -1 \\ -1 & -2 & -4 \\ -1 & 1 & 5 \\ -1 & -5 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{Y}^t\mathbf{Y} = \begin{bmatrix} 6 & 6 & -4 \\ 6 & 44 & 10 \\ -4 & 10 & 62 \end{bmatrix}.$$

The eigenvalues of $\mathbf{Y}^t\mathbf{Y}$ are $\{4.5331, 40.9882, 66.4787\}$.

Section 5.10

28. The linear programming problem on pages 256–257 in the text involved finding

$$\min \{t : t \geq 0, \mathbf{a}^t\mathbf{y}_i + t > b_i \text{ for all } i\}.$$

Our goal here is to show that the resulting weight vector $\mathbf{a}$ minimizes the criterion function

$$J_t(\mathbf{a}) = \max_{i:\mathbf{a}^t\mathbf{y}_i \leq b_i} (b_i - \mathbf{a}^t\mathbf{y}_i).$$

There are two cases to consider: linearly separable and non-linearly separable.
Case 1: Suppose the samples are linearly separable. Then there exists a vector, say $\mathbf{a}_o$, such that

$$\mathbf{a}_o^t\mathbf{y}_i = b_i.$$

Then clearly $\mathbf{a}_o^t\mathbf{y}_i + t > b_i$, for any $t > 0$ and for all i. Thus we have for all i

$$\begin{aligned} 0 &\leq \min\{t : t \geq 0, \mathbf{a}^t\mathbf{y}_i + t > b_i\} \\ &\leq \min\{t : t \geq 0, \mathbf{a}^t\mathbf{y}_i + t > b_i\} = 0. \end{aligned}$$

Therefore, we have

$$\min\{t : t \geq 0, \mathbf{a}^t\mathbf{y}_i + t > b_i\} = 0,$$

and the resulting weight vector is $\mathbf{a}_o$. The fact that $J_t(\mathbf{a}) \geq 0$ for all $\mathbf{a}$ and $J_t(\mathbf{a}_o) = 0$ implies

$$\arg\min_{\mathbf{a}} J_t(\mathbf{a}) = \mathbf{a}_o.$$

This proves that the $\mathbf{a}$ minimizes $J_t(\mathbf{a})$ is the same as the one for solving the modified variable problem.

Case 2: If the samples are *not* linearly separable, then there is no vector $\mathbf{a}_o$ such that

$$\mathbf{a}_o^t \mathbf{y}_i = b_i.$$

This means that for all i

$$\begin{aligned}
\min_{t,\mathbf{a}}\{t : t \geq 0, \mathbf{a}^t\mathbf{y}_i + t > b_i\} &= \min_{t,\mathbf{a}}\{t : t \geq 0, t > b_i - \mathbf{a}^t\mathbf{y}_i\} \\
&= \min_{t,\mathbf{a}}\{t : t \geq 0, t > \max_i(b_i - \mathbf{a}^t\mathbf{y}_i)\} \\
&= \min_{t,\mathbf{a}}\{t : t > \max_{i:\mathbf{a}^t\mathbf{y}_i \leq b_i}(b_i - \mathbf{a}^t\mathbf{y}_i)\} \\
&= \min_{\mathbf{a}}\{\max_{i:\mathbf{a}^t\mathbf{y}_i \leq b_i}(b_i - \mathbf{a}^t\mathbf{y}_i)\} \\
&= \min_{\mathbf{a}} J_t(\mathbf{a}).
\end{aligned}$$

Section 5.11

29. Given n patterns $\mathbf{x}_k$, in d-dimensional space, we associate z_k where

$$\begin{aligned}
\text{If } \mathbf{x}_k \in \omega_1, \text{then } z_k = 1 \\
\text{If } \mathbf{x}_k \in \omega_2, \text{then } z_k = -1.
\end{aligned}$$

We also define a mapping $\boldsymbol{\phi} : \mathbf{R}^d \rightarrow \mathbf{R}^{d+1}$ as

$$\boldsymbol{\phi}(\mathbf{x}) = (\mathbf{x}_k, z_k)$$

where

$$\mathbf{x}_k = \arg\min_{\mathbf{x}_k}(\|\mathbf{x}_k - \mathbf{x}\|).$$

In other words, $\boldsymbol{\phi}$ returns the nearest-neighbor prototype $\mathbf{x}_k$. Then clearly $\boldsymbol{\phi}(\mathbf{x}) = (\mathbf{x}, 0)$ represents a separating hyperplane with normal

$$\mathbf{a} = (\underbrace{0, 0, \ldots, 0}_{d \text{ zeros}}, 1)$$

We verify this by considering any pattern in ω_1. For such a pattern

$$\mathbf{a}^t\boldsymbol{\phi}(\mathbf{x}_k) = (0, 0, \ldots, 0, 1)^t(\mathbf{x}_k, z_k) = z_k = +1.$$

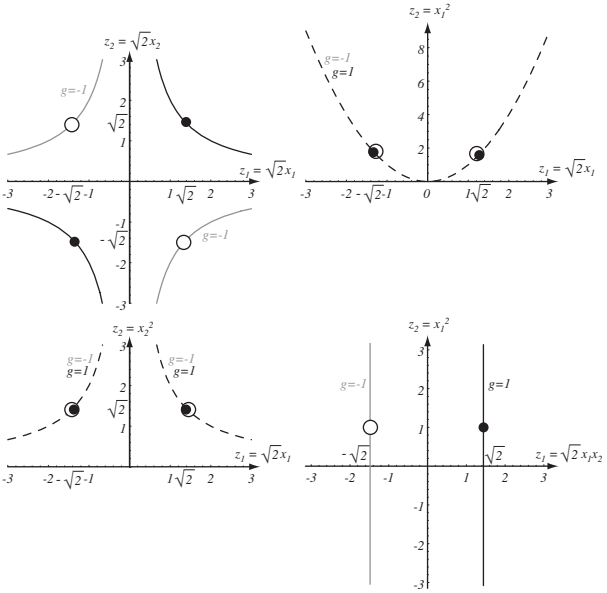
Conversely, for any pattern in ω_2, we have

$$\mathbf{a}^t\boldsymbol{\phi}(\mathbf{x}_k) = (0, 0, \ldots, 0, 1)^t(\mathbf{x}_k, z_k) = z_k = -1.$$

Intuitively speaking, this construction tells us that if we can label the samples unambiguously, then we can always find a mapping to transform class ω_1 and class ω_2 to points into the half spaces.

30. The points are mapped to

$$\begin{aligned}
\omega_1 \quad &: \quad (1, \sqrt{2}, \sqrt{2}, \sqrt{2}, 1, 1)^t, (1, -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 1, 1)^t \\
\omega_2 \quad &: \quad (1, \sqrt{2}, -\sqrt{2}, -\sqrt{2}, 1, 1)^t, (1, -\sqrt{2}, \sqrt{2}, -\sqrt{2}, 1, 1)^t
\end{aligned}$$



as shown in the figure.
The margins are not the same, simply because the real margin is the distance of the support vectors to the optimal hyperplane in $\mathbf{R}^6$ space, and their projection to lower dimensional subspaces does not necessarily preserve the margin.
31. The Support Vector Machine algorithm can be written:

Algorithm 0 (SVM)

```
1  begin initialize a; worst1 ← ∞; worst2 ← ∞; b ← ∞
2    i ← 0
3    do  i ← i + 1
4        if z_i = −1 and a^t y_i z_i < worst1, then worst1 ← a^t y_i z_i; kworst1 ← k
5        if z_i = 1 and a^t y_i z_i < worst2, then worst2 ← a^t y_i z_i; kworst2 ← k
6    until i = n
7    a ← a + y_kworst2 − y_kworst1
8    a_0 ← a^t(y_kworst2 + y_kworst1)/2
9    oldb ← b; b ← a^t y_kworst1 / ‖ a ‖
10 until | b − oldb | < ε
11 return a_0, a
12 end
```

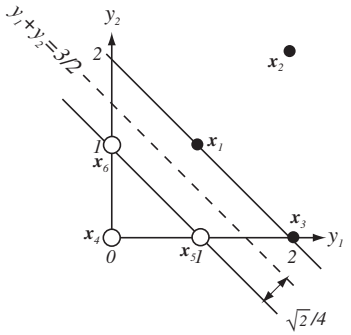
Note that the algorithm picks the worst classified patterns from each class and adjusts $\mathbf{a}$ such that the hyperplane moves toward the center of the worst patterns and rotates so that the angle between the hyperplane and the vector connecting the worst points increases. Once the hyperplane separates the classes, all the updates will involve support vectors, since the **if** statements can only pick the vectors with the smallest $|\mathbf{a}^t\mathbf{y}_i|$.
32. Consider Support Vector Machines for classification.

(a) We are given the following six points in two categories:

$$\omega_1 \quad : \quad \mathbf{x}_1 = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}2\\2\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}2\\0\end{pmatrix}$$

$$\omega_2 \quad : \quad \mathbf{x}_4 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{x}_6 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

with $z_1 = z_2 = z_3 = -1$ and $z_4 = z_5 = z_6 = +1$.



The optimal hyperplane is $y_1 + y_2 = 3/2$, or $(3/2 \ -1 \ -1)^t(1 \ y_1 \ y_2) = 0$. To ensure $z_k \mathbf{a}^t \mathbf{y} \geq 1$, we have to scale $(3/2 \ -1 \ -1)^t$ by 2, and thus the weight vector is $(3 \ -2 \ -2)^t$. The optimal margin is the shortest distance from the patterns to the optimal hyperplane, which is $\sqrt{2}/4$, as can be seen in the figure.

(b) Support vectors are the samples on the margin, that is, the ones with the shortest distance to the separating hyperplane. In this case, the support vectors are $\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_6\} = \{(1,1)^t, (2,0)^t, (1,0)^t, (0,1)^t\}$.

(c) We seek to maximize the criterion given in Eq. 109 in the text,

$$L(\boldsymbol{\alpha}) = \sum_{k=1}^{n} \alpha_k - \frac{1}{2} \sum_{k,j}^{n} \alpha_k \alpha_j z_k z_j \mathbf{y}_j^t \mathbf{y}_k$$

subject to the constraints

$$\sum_{k=1}^{n} z_k \alpha_k = 0$$

for $\alpha_k \geq 0$. Using the constraint, we can substitute $\alpha_6 = \alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 - \alpha_5$ in the expression for $L(\boldsymbol{\alpha})$. Then we can get a system of linear equations by setting the partial derivatives, $\partial L / \partial \alpha_i$ to zero. This yields:

$$\begin{bmatrix} -1 & -2 & -2 & 0 & 1 \\ -2 & -5 & 2 & -1 & 1 \\ -2 & 2 & -5 & 1 & 1 \\ 0 & -1 & 1 & -1 & -1 \\ 1 & 1 & 3 & -1 & -2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \\ -2 \\ 0 \\ 0 \end{bmatrix}.$$

Unfortunately, this is an inconsistent set of equations. Therefore, the maxima must be achieved on the boundary (where some α_i vanish). We try each $\alpha_i = 0$ and solve $\partial L / \partial \alpha_i = 0$:

$$\frac{\partial L(0, \alpha_2, \alpha_3, \alpha_4, \alpha_5)}{\partial \alpha_i} = 0$$

implies $\boldsymbol{\alpha} = 1/5(0, -2\ -2\ 8\ -8\ -4)^t$, which violates the constraint $\alpha_i \geq 0$. Next, both of the following vanishing derivatives,

$$\frac{\partial L(\alpha_1, 0, \alpha_3, \alpha_4, \alpha_5)}{\partial \alpha_i} = \frac{\partial L(\alpha_1, \alpha_2, 0, \alpha_4, \alpha_5)}{\partial \alpha_i} = 0$$

lead to inconsistent equations. Then the derivative

$$\frac{\partial L(\alpha_1, \alpha_2, \alpha_3, 0, \alpha_5)}{\partial \alpha_i} = 0$$

implies $\boldsymbol{\alpha} = 1/5(16\ 0\ 4\ 0\ 14\ 6)^t$, which does not violate the constraint $\alpha_i \geq 0$. In this case the criterion function is $L(\boldsymbol{\alpha}) = 4$. Finally, we have

$$\frac{\partial L(\alpha_1, \alpha_2, \alpha_3, \alpha_4, 0)}{\partial \alpha_i} = 0$$

which implies $\boldsymbol{\alpha} = 1/5(2\ 2\ 2\ 0\ 0\ 6)^t$, and the constraint $\alpha_i \geq 0$ is obeyed. In this case the criterion function is $L(\boldsymbol{\alpha}) = 1.2$.

Thus $\boldsymbol{\alpha} = 1/5(16\ 0\ 4\ 0\ 14\ 6)^t$ is where the criterion function L reaches its maximum within the constraints. Now we seek the weight vector $\mathbf{a}$. We seek to minimize $L(\mathbf{a}, \boldsymbol{\alpha})$ of Eq. 108 in the text,

$$L(\mathbf{a}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{a}\|^2 - \sum_{k=1}^{n} \alpha_k [z_k \mathbf{a}^t \mathbf{y}_k - 1],$$

with respect to $\mathbf{a}$. We take the derivative of the criterion function,

$$\frac{\partial L}{\partial \mathbf{a}} = \mathbf{a} - \sum_{k=1}^{n} \alpha_k z_k \mathbf{y}_k = \mathbf{0},$$

which for the α_k found above has solution

$$\begin{aligned} \mathbf{a} &= -(16/5)\mathbf{y}_1 - 0\mathbf{y}_2 - 4/5\mathbf{y}_3 + 0\mathbf{y}_4 + 14/5\mathbf{y}_5 + 6/5\mathbf{y}_6 \\ &= \begin{pmatrix} 0 \\ -2 \\ -2 \end{pmatrix}. \end{aligned}$$

Note that $\partial L/\partial \mathbf{a} = 0$ here is not sufficient to allow us to find the bias a_0 directly since the $\|\mathbf{a}\|^2$ term does not include the augmented vector $\mathbf{a}$ and $\sum_k \alpha_k z_k = 0$. We determine a_0, then, by using one of the support vectors, for instance $\mathbf{y}_1 = (1\ 1\ 1)^t$. Since $\mathbf{y}_1$ is a support vector, $\mathbf{a}^t \mathbf{y}_1 z_1 = 1$ holds, and thus

$$-\begin{pmatrix} 0 \\ -2 \\ -2 \end{pmatrix} (1\ 1\ 1) = -a_0 + 4 = 1.$$

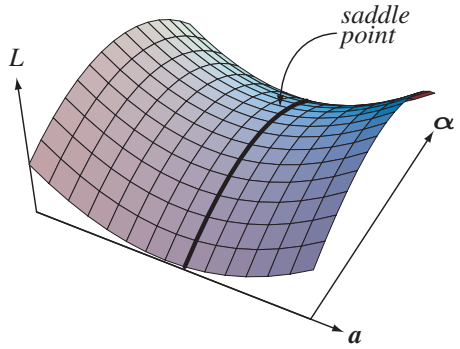
This, then, means $a_0 = 3$, and the full weight vector is $\mathbf{a} = (3\ -2\ -2)^t$.

33. Consider the Kuhn-Tucker theorem and the conversion of a constrained optimization problem for support vector machines to an unconstrained one.

(a) Here the relevant functional is given by Eq. 108 in the text, that is,

$$L(\mathbf{a}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{a}\|^2 - \sum_{k=1}^{n} \alpha_k [z_k \mathbf{a}^t \mathbf{y}_k - 1].$$

We seek to maximize L with respect to $\boldsymbol{\alpha}$ to guarantee that all the patterns are correctly classified, that is $z_k \mathbf{a}^t \mathbf{y}_k \geq 1$, and we want to minimize L with respect to the (un-augmented) $\mathbf{a}$. This will give us the optimal hyperplane. This solution corresponds to a saddle point in $\boldsymbol{\alpha}$–$\mathbf{a}$ space, as shown in the figure.



(b) We write the augmented vector $\mathbf{a}$ as $(a_0 \ \mathbf{a}_r)^t$, where a_0 is the augmented bias. Then we have

$$L(\mathbf{a}_r, \boldsymbol{\alpha}, a_0) = \frac{1}{2}\|\mathbf{a}_r\|^2 - \sum_{k=1}^{n} \alpha_k [z_k \mathbf{a}_r^t \mathbf{y}_k + z_k a_0 - 1].$$

At the saddle point, $\partial L/\partial a_0 = 0$ and $\partial L/\partial \mathbf{a}_r = \mathbf{0}$. The first of these derivative vanishing implies

$$\sum_{k=1}^{n} \alpha_k^* z_k = 0.$$

(c) The second derivative vanishing implies

$$\frac{\partial L}{\partial \mathbf{a}_r} = \mathbf{a}_r - \sum_{k=1}^{n} \alpha_k^* z_k \mathbf{y}_k$$

and thus

$$\mathbf{a}_r = \sum_{k=1}^{n} \alpha_k^* z_k \mathbf{y}_k.$$

Since $\sum_{k=1}^{n} \alpha_k z_k = 0$, we can thus write the solution in augmented form as

$$\mathbf{a} = \sum_{k=1}^{n} \alpha_k^* z_k \mathbf{y}_k.$$

(d) If $\alpha_k^*(z_k\mathbf{a}^{*t}\mathbf{y}_k - 1) = 0$ and if $z_k\mathbf{a}^{*t}\mathbf{y}_k \neq 0$, then $\boldsymbol{\alpha}_k^*$ must be zero. Respectively, call the predicates as h, NOTp and q, then the above states h $AND\ NOT$ p $\rightarrow$ q, which is equivalent to h $AND\ NOT$ q $\rightarrow$ p. In other words, given the expression

$$\alpha_k^*(z_k\mathbf{a}^{*t}\mathbf{y} - 1) = 0,$$

then $\boldsymbol{\alpha}^*$ is non-zero if and only if $z_k\mathbf{a}^{*t}\mathbf{y}_k = 1$.

(e) Here we have

$$\begin{aligned}\bar{L} &= \frac{1}{2}\left\|\sum_{k=1}^{n}\alpha_k z_k\mathbf{y}_k\right\|^2 - \sum_{k=1}^{n}\alpha_k\left[z_k\left(\sum_{l=1}^{n}\alpha_l z_l\mathbf{y}_l\right)\mathbf{y}_k - 1\right] \\ &= \frac{1}{2}\left(\sum_{k=1}^{n}\alpha_k z_k\mathbf{y}_k\right)^t\left(\sum_{k=1}^{n}\alpha_k z_k\mathbf{y}_k\right) - \sum_{kl}^{n}\alpha_k\alpha_l z_k z_l\mathbf{y}_k^t\mathbf{y}_l + \sum_{k=1}^{n}\alpha_k.\end{aligned}$$

Thus we have

$$\bar{L} = \sum_{k=1}^{n}\alpha_k - \frac{1}{2}\sum_{kl}\alpha_k\alpha_l z_k z_l\mathbf{y}_k^t\mathbf{y}_l.$$

(f) See part (e).

34. We repeat Example 2 in the text but with the following four points:

$$\mathbf{y}_1 = (1\ \sqrt{2}\ 5\sqrt{2}\ 5\sqrt{2}\ 1\ 25)^t, \quad \mathbf{y}_2 = (1\ -2\sqrt{2}\ -4\sqrt{2}\ 8\sqrt{2}\ 4\ 16)^t, z_1 = z_2 = -1$$
$$\mathbf{y}_3 = (1\ \sqrt{2}\ 3\sqrt{2}\ 6\sqrt{2}\ 4\ 9)^t, \quad \mathbf{y}_4 = (1\ -2\sqrt{2}\ 5\sqrt{2}\ -5\sqrt{2}\ 1\ 25)^t, z_3 = z_4 = +1$$

We seek the optimal hyperplane, and thus want to maximize the functional given by Eq. 109 in the text:

$$L(\boldsymbol{\alpha}) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\sum_{kl}^{4}\alpha_l\alpha_k z_k z_l\mathbf{y}_k^t\mathbf{y}_l,$$

with constraints $\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4$ and $\alpha_1 \geq 0$. We substitute $\alpha_4 = \alpha_1 + \alpha_2 - \alpha_3$ into $L(\boldsymbol{\alpha})$ and take the partial derivatives with respect to α_1, α_2 and α_3 and set the derivatives to zero:

$$\begin{aligned}\frac{\partial L}{\partial \alpha_1} &= 2 - 208\alpha_1 - 256\alpha_2 + 232\alpha_3 = 0 \\ \frac{\partial L}{\partial \alpha_2} &= 2 - 256\alpha_1 - 592\alpha_2 + 496\alpha_3 = 0 \\ \frac{\partial L}{\partial \alpha_3} &= 232\alpha_1 + 496\alpha_2 - 533\alpha_3 = 0.\end{aligned}$$

The solution to these equations — $\alpha_1 = 0.0154$, $\alpha_2 = 0.0067$, $\alpha_3 = 0.0126$ — indeed satisfy the constraint $\alpha_i \geq 0$, as required.

Now we compute $\mathbf{a}$ using Eq. 108 in the text:

$$\frac{\partial L}{\partial \mathbf{a}} = \mathbf{a} - \sum_{k=1}^{4}\alpha_k z_k\mathbf{y}_k = 0,$$

which has solution

$$\begin{aligned}\mathbf{a} &= 0.0154(-\mathbf{y}_1) + 0.0067(-\mathbf{y}_2) + 0.01261\mathbf{y}_3 + 0.095\mathbf{y}_4 \\ &= (0 \;\; 0.0194 \;\; 0.0496 \;\; -0.145 \;\; 0.0177 \;\; -0.1413)^t.\end{aligned}$$

Note that this process cannot determine the bias term, a_0 directly; we can use a support vector for this in the following way: We note that $\mathbf{a}^t\mathbf{y}_k z_k = 1$ must hold for each support vector. We pick $\mathbf{y}_1$ and then

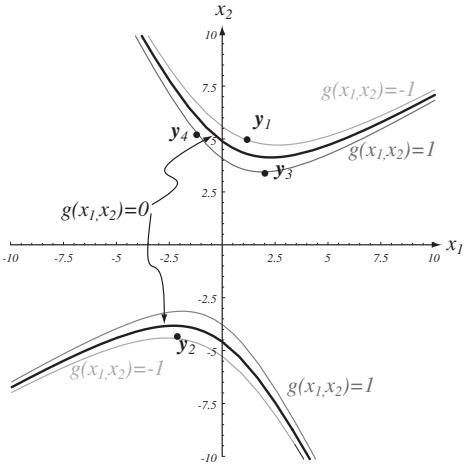
$$-(a_0 \;\; 0.0194 \;\; 0.0496 \;\; -0.145 \;\; 0.0177 \;\; -0.1413) \cdot \mathbf{y}_1 = 1,$$

which gives $a_0 = 3.1614$, and thus the full weight vector is $\mathbf{a} = (3.1614 \;\; 0.0194 \;\; 0.0496 \;\; -0.145 \;\; 0.0177 \;\; -0.1413)^t$.

Now we plot the discriminant function in x_1–x_2 space:

$$\begin{aligned}g(x_1, x_2) &= \mathbf{a}^t(1 \;\; \sqrt{2}x_1 \;\; \sqrt{2}x_2 \;\; \sqrt{2}x_1x_2 \;\; x_1^2 \;\; x_2^2) \\ &= 0.0272x_1 + 0.0699x_2 - 0.2054x_1x_2 + 0.1776x_1^2 - 0.1415x_2^2 + 3.17.\end{aligned}$$

The figure shows the hyperbola corresponding to $g(x_1, x_2) = 0$ as well as the margins, $g(x_1, x_2) = \pm 1$, along with the three support vectors $\mathbf{y}_2$, $\mathbf{y}_3$ and $\mathbf{y}_4$.



Section 5.12

35. Consider the LMS algorithm and Eq. 61 in the text.

(a) The LMS algorithm minimizes $J_s(\mathbf{a}) = \|\mathbf{Ya} - \mathbf{b}\|^2$ and we are given the problem

$$J_s(\mathbf{a}) = \left\| \begin{bmatrix} \mathbf{1}_n & \mathbf{Y}_1 \\ \mathbf{1}_n & \mathbf{Y}_1 \end{bmatrix} \begin{bmatrix} a_0 \\ \mathbf{a}_r \end{bmatrix} - \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \right\|^2$$

We assume we have $2n$ data points. Moreover, we let $\mathbf{b}_i = (b_i^1 \;\; b_i^2 \;\; \ldots b_i^n)^t$ for $i = 1, 2$ be arbitrary margin vectors of size n each, and $\mathbf{Y}_1$ is an n-by-2 matrix containing the data points as the rows, and $\mathbf{a} = (a_0, a_1, a_2)^t = (a_0, \mathbf{a}_r)^t$. Then we have

$$J_s(\mathbf{a}) = \left\| \begin{array}{c} a_0\mathbf{1}_n + \mathbf{Y}_1\mathbf{a}_r - \mathbf{b}_1 \\ -a_0\mathbf{1}_n + \mathbf{Y}_1\mathbf{a}_r - \mathbf{b}_1 \end{array} \right\|^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}(a_0 + \mathbf{y}_i^t\mathbf{a}_r - b_1^i)^2 + \sum_{i=1}^{n}(-a_0 + \mathbf{y}_i^t\mathbf{a}_r - b_2^i)^2 \\
&= \sum_{i=1}^{n}(a_0 - b_1^i)^2 + \sum_{i=1}^{n}(-a_0 - b_2^i)^2 + 2\sum_{i=1}^{n}\mathbf{y}_i^t\mathbf{a}_r(a_0 - b_1^i - a_0 - b_2^i) \\
&= \sum_{i=1}^{n}(a_0 - b_1^i)^2 + \sum_{i=1}^{n}(a_0 + b_2^i)^2 - 2\sum_{i=1}^{n}\mathbf{y}_i^t\mathbf{a}_r(b_1^i + b_2^i).
\end{aligned}
$$

Thus $\partial J_s/\partial a_0 = 0$ implies $a_0 = 0$ and the minimum of J must be at $(0, \mathbf{a}_r)^t$ for some $\mathbf{a}_r$. Hence we showed that $a_0 = 0$ which tells us that the separating plane must go through the origin.

(b) We know from part (a) that the LMS hyperplane must pass through the origin. Thus ω_1 samples must lie on the same side of a separating hyperplane in order to ensure ω_2 samples lie in the other half space. This is guaranteed in the shaded (union) region. We are asked to define the region where there is no shading. We can state this as: if $y \in \{(x_1, x_2)^t | 2x_1 < |x_2|\}$, then the LMS solution to separate ω_1 and ω_2 will not give a separating hyperplane, where

$$
\begin{aligned}
\omega_1 &: \begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}2\\-4\end{pmatrix}, y \\
\omega_2 &: \begin{pmatrix}-1\\-2\end{pmatrix}, \begin{pmatrix}-2\\4\end{pmatrix}, -y
\end{aligned}
$$

as shown in the figure.

(c) Part (b) can be generalized by noting that the feasible region for y (so that LMS will give a separating hyperplane) is the union of the half spaces $\mathcal{H}_i$ determined by $\mathbf{y}_i$ as:

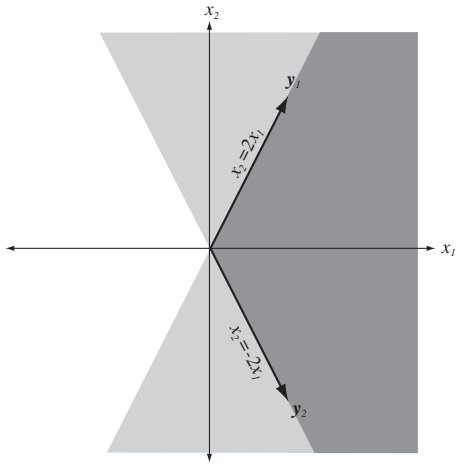
$$
\begin{aligned}
\mathcal{H}_1 &= \{\text{half space induced by the separating hyperplane } \mathbf{y}_1 \text{ and containing } \mathbf{y}_2\} \\
\mathcal{H}_2 &= \{\text{half space induced by the separating hyperplane } \mathbf{y}_2 \text{ and containing } \mathbf{y}_1\}.
\end{aligned}
$$

The LMS will give separating hyperplanes if and only if $\mathbf{y}_3 \in \mathcal{H}_1 \cup \mathcal{H}_2$. Thus, to ensure that the LMS solution not to separate $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$ from $\{-\mathbf{y}_1, -\mathbf{y}_2, -\mathbf{y}_3\}$, we must have $\mathbf{y}_3 \in \bar{\mathcal{H}}_1 \cap \bar{\mathcal{H}}_2$.

36. The algorithm is as follows:

Algorithm 0 (Fixed increment multicategory Perceptron)

```
1  begin initialize a_i  k ← 0, n ← number of samples, c ← number of classes
2        do  k ← (k + 1) mod n
3            i ← class of y^k
4            do  j ← j + 1;  (j ≠ i)
5            until a_i^t y^k < a_j^t y^k  ∨  j > c
6            if  j ≤ c  then  a_i ← a_i + y^k
7                             a_j ← a_j − y^k
8              until no more misclassifications
9              return a_1, a_2, ..., a_c
10       end
```



Some of the advantages are that the algorithm is simple and admits a simple parallel implementation. Some of the disadvantages are that there may be numerical instability and that the search is not particularly efficient.

37. Equation 122 in the text gives the Bayes discriminant function as

$$g_{0i} = -\sum_{j=1}^{c} \lambda_{ij} P(\omega_i|\mathbf{x}).$$

The definition of λ_{ij} from Eq. 121 in the text ensures that for a given i only one λ_{ij} will be non-zero. Thus we have $g_{0i} = P(\omega_i|\mathbf{x})$. We apply Bayes rule to find

$$p(\mathbf{x})g_{0i}(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i). \qquad (*)$$

On the other hand, analogous to Eq. 58 in the text, the criterion function $J_{s_1 i}$ can be written as

$$\begin{aligned} J_{s_1 i}(\mathbf{a}) &= \sum_{\mathbf{y}\in\mathcal{Y}_i} (\mathbf{a}_i^t\mathbf{y} - 1)^2 + \sum_{\mathbf{y}\notin\mathcal{Y}_i} (\mathbf{a}_i^t\mathbf{y})^2 \\ &= n\left[\frac{n_1}{n}\frac{1}{n_1}\sum_{\mathbf{y}\in\mathcal{Y}_i} (\mathbf{a}_i^t\mathbf{y} - 1)^2 + \frac{n_2}{n}\frac{1}{n_2}\sum_{\mathbf{y}\notin\mathcal{Y}_i} (\mathbf{a}_i^t\mathbf{y})^2\right]. \end{aligned}$$

By the law of large numbers, as $n \to \infty$, we have that $1/nJ_{s_1 i}(\mathbf{a}_i)$ approaches $\bar{J}_i(\mathbf{a}_i)$ with probability 1, where $\bar{J}_i(\mathbf{a}_i)$ is given by

$$\bar{J}_i(\mathbf{a}_i) = P(\omega_i)\mathcal{E}_i[(\mathbf{a}_i^t\mathbf{y} - 1)^2] + P(NOT\ \omega_i)\mathcal{E}_2[(\mathbf{a}_i^t\mathbf{y})^2],$$

where

$$\begin{aligned} \mathcal{E}_1[(\mathbf{a}_i^t\mathbf{y} - 1)^2] &= \int (\mathbf{a}_i^t\mathbf{y} - 1)^2 p(\mathbf{x}|\omega_i)d\mathbf{x} \\ \mathcal{E}_2[(\mathbf{a}_i^t\mathbf{y})^2] &= \int (\mathbf{a}_i^t\mathbf{y})^2 p(\mathbf{x}|NOT\ \omega_i)d\mathbf{x}. \end{aligned}$$

Thus we have

$$\bar{J}_i(\mathbf{a}_i) = \int (\mathbf{a}_i^t\mathbf{y} - 1)^2 p(\mathbf{x}, \omega_i)d\mathbf{x} + \int (\mathbf{a}_i^t\mathbf{y})^2 p(\mathbf{x}, NOT\ \omega_i)d\mathbf{x}.$$

We expand the square to find:

$$\bar{J}(\mathbf{a}_i) = \int (\mathbf{a}_i^t\mathbf{y})^2 p(\mathbf{x},\omega_i)d\mathbf{x} - 2\int (\mathbf{a}_i^t\mathbf{y}) p(\mathbf{x},\omega_i)d\mathbf{x} + \int p(\mathbf{x},\omega_i)d\mathbf{x} + \int (\mathbf{a}_i^t\mathbf{y})^2 p(\mathbf{x}, NOT\ \omega_i)d\mathbf{x}.$$

We collect the terms containing $(\mathbf{a}_i^t\mathbf{y})^2$ and find

$$\bar{J}(\mathbf{a}_i) = \int (\mathbf{a}_i^t\mathbf{y})^2 p(\mathbf{x})d\mathbf{x} - 2\int (\mathbf{a}_i^t\mathbf{y}) p(\mathbf{x},\omega_i)d\mathbf{x} + \int p(\mathbf{x},\omega_i)d\mathbf{x}.$$

We use the fact that $p(\mathbf{x})g_{0i}(\mathbf{x}) = p(\mathbf{x},\omega_i)$ from $(*)$, above, we find

$$\bar{J}(\mathbf{a}_i) = \int (\mathbf{a}_i^t\mathbf{y})^2 p(\mathbf{x})d\mathbf{x} - 2\int (\mathbf{a}^t\mathbf{y}) p(\mathbf{x}) g_{0i}(\mathbf{x})d\mathbf{x} + \int p(\mathbf{x},\omega_i)d\mathbf{x},$$

which can be written as

$$\bar{J}(\mathbf{a}_i) = \underbrace{\int [\mathbf{a}_i^t\mathbf{y} - g_{0i}(\mathbf{x})]^2 p(\mathbf{x}d\mathbf{x})}_{\epsilon_i^2, \text{the mean-squared approx. error}} - \underbrace{\int [g_{0i}^2(\mathbf{x})p(\mathbf{x}) - g_{0i}(\mathbf{x})p(\mathbf{x})]d\mathbf{x}}_{\text{independent of } \mathbf{a}_i}.$$

The second term in the sum is independent of weight vector $\mathbf{a}_i$. Hence the $\mathbf{a}_i$ that minimizes $J_{s_1 i}$ also minimizes ϵ_i^2. Thus the MSE solution $\mathbf{A} = \mathbf{Y}^t\mathbf{B}$ (where $\mathbf{A} = \begin{matrix}\mathbf{a}_i & \mathbf{a}_2 & \mathbf{a}_c\end{matrix}$ and $\mathbf{Y} = [\mathbf{Y}_1 \quad \mathbf{Y}_2 \cdots \mathbf{Y}_c]^t$, and $\mathbf{B}$ is defined by Eq. 119 in the text) for the multiclass case yields discriminant functions $\mathbf{a}_i^t\mathbf{y}$ that provide a minimum-mean-square error approximation to the Bayes discriminant functions g_{0i}.

38. We are given the multi-category classification problem with sets $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_c$ to be classified as $\omega_1, \omega_2, \ldots, \omega_c$, respectively. The aim is to find $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_c$ such that if $\mathbf{y} \in \mathcal{Y}_i$, then $\mathbf{a}_i^t\mathbf{y} \geq \mathbf{a}_j^t\mathbf{y}$ for all $j \neq i$. We transform this problem into a binary classification problem using Kesler's construction. We define $G = G_1 \cup G_2 \cup \cdots G_c$ where

$$G_i = \{\eta_{ij} | \mathbf{y} \in \mathcal{Y}_i, j \neq i\}$$

and η_{ij} is as given in Eq. 115 in the text. Moreover we define

$$\boldsymbol{\alpha} = (\begin{matrix}\mathbf{a}_1^t & \mathbf{a}_2 & \cdots & \mathbf{a}_c^t\end{matrix})^t.$$

Perceptron case We wish to show what the Fixed-increment single sample Perceptron algorithm given in Eq. 20 in the text does to our transformed problem. We rewrite Eq. 20 as

$$\begin{aligned} \boldsymbol{\alpha}(1) &= \text{arbitrary} \\ \boldsymbol{\alpha}(k+1) &= \boldsymbol{\alpha}(k) + \mathbf{g}^k \end{aligned}$$

where $\mathbf{g}^k$ is misclassified. The condition $\mathbf{g}^k$ being misclassified at step k implies $\boldsymbol{\alpha}^t(k)\mathbf{g}^k \leq 0$. Since the G_is are disjoint, $\mathbf{g}^k$ must belong to one and only one of the G_i. We shall use a subscript to denote which G_i that $\mathbf{g}^k$ belongs to; for instance, $\mathbf{g}_i^k$ is in G_i. Given this notation, the inequality $\boldsymbol{\alpha}^t(k)\mathbf{g}_i^k \leq 0$ implies $\mathbf{a}_i^t\mathbf{y} - \mathbf{a}_j^t\mathbf{y} \leq 0$ for some $j \neq i$. Thus there is an equivalence between $\boldsymbol{\alpha}^t(k)\mathbf{g}_i^k \leq 0$ and $\mathbf{a}_i^t(k)\mathbf{y} \leq \mathbf{a}_j^t(k)\mathbf{y}$.

Consider the update rule $\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \mathbf{g}_i^k$. At once we see the equivalence:

Multi-category		Two-category
$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \mathbf{g}_i^k$	$\Leftrightarrow$	$\mathbf{a}_i(k+1) = \mathbf{a}_i(k) + \mathbf{y}_i^k$ $\mathbf{a}_j(k+1) = \mathbf{a}_j(k) - \mathbf{y}_j^k$,

where, as defined above, $\boldsymbol{\alpha}$ is the concatenation of $\mathbf{a}$s.

Relaxation rule case The single sample relaxation rule becomes

$$\begin{aligned} \boldsymbol{\alpha}(1) &= \text{arbitrary} \\ \boldsymbol{\alpha}(k+1) &= \boldsymbol{\alpha}(k) + \boldsymbol{\eta}\frac{b - \boldsymbol{\alpha}^t\mathbf{g}^k}{\|\mathbf{g}^k\|^2}\mathbf{g}^k, \end{aligned}$$

where b is the margin. An update takes place when the sample $\mathbf{g}^k$ is incorrectly classified, that is, when $\boldsymbol{\alpha}^t(k)\mathbf{g}^k < b$. We use the same definition of $\mathbf{g}$ as in the Perceptron case above, and can then write

$$\boldsymbol{\alpha}^t(k)\boldsymbol{\eta}_{ij} < b \Leftrightarrow (\mathbf{a}_i^t(k)\mathbf{y} - \mathbf{a}_j^t(k)\mathbf{y}) < b.$$

In the update rule, $\boldsymbol{\alpha}$ can be decomposed into its sub-component $\mathbf{a}$s, yielding the following equivalence:

Multi-category		Two-category
$\boldsymbol{\alpha}(k+1) = \boldsymbol{\alpha}(k) + \eta\frac{b-\boldsymbol{\alpha}^t(k)\mathbf{g}^k}{\|\mathbf{g}^k\|^2}\mathbf{g}^k$	$\Leftrightarrow$	$\mathbf{a}_i(k+1) = \mathbf{a}_i(k) + \frac{b-(\mathbf{a}_i^t(k)-\mathbf{a}_j^t(k))\mathbf{y}^k}{2\|\mathbf{y}^k\|^2}\mathbf{y}^k$ $\mathbf{a}_j(k+1) = \mathbf{a}_j(k) + \frac{b-(\mathbf{a}_i^t(k)-\mathbf{a}_j^t(k))\mathbf{y}^k}{2\|\mathbf{y}^k\|^2}\mathbf{y}^k$.

Computer Exercises

Section 5.4

1. Computer exercise not yet solved

Section 5.5

2. Computer exercise not yet solved
3. Computer exercise not yet solved
4. Computer exercise not yet solved
5. Computer exercise not yet solved
6. Computer exercise not yet solved

Section 5.6

7. Computer exercise not yet solved

Section 5.8

8. Computer exercise not yet solved

Section 5.9

9. Computer exercise not yet solved

Section 5.10

10. Computer exercise not yet solved

Section 5.11

11. Computer exercise not yet solved

Section 5.12

12. Computer exercise not yet solved

Chapter 6

Multilayer neural networks

Problem Solutions

Section 6.2

1. Consider a three-layer network with linear units throughout, having input vector $\mathbf{x}$, vector at the hidden units $\mathbf{y}$, and output vector $\mathbf{z}$. For such a linear system we have $\mathbf{y} = \mathbf{W}_1\mathbf{x}$ and $\mathbf{z} = \mathbf{W}_2\mathbf{y}$ for two matrices $\mathbf{W}_1$ and $\mathbf{W}_2$. Thus we can write the output as

$$\begin{aligned} \mathbf{z} &= \mathbf{W}_2\mathbf{y} = \mathbf{W}_2\mathbf{W}_1\mathbf{x} \\ &= \mathbf{W}_3\mathbf{x} \end{aligned}$$

for some matrix $\mathbf{W}_3 = \mathbf{W}_2\mathbf{W}_1$. But this equation is the same as that of a two-layer network having connection matrix $\mathbf{W}_3$. Thus a three-layer network with linear units throughout can be implemented by a two-layer network with appropriately chosen connections.

Clearly, a non-linearly separable problem cannot be solved by a three-layer neural network with linear hidden units. To see this, suppose a non-linearly separable problem can be solved by a three-layer neural network with hidden units. Then, equivalently, it can be solved by a two-layer neural network. Then clearly the problem is linearly separable. But, by assumption the problem is only non-linearly separable. Hence there is a contradiction and the above conclusion holds true.

2. Fourier's theorem shows that a three-layer neural network with sigmoidal hidden units can act as a universal approximator. Consider a two-dimensional input and a single output $z(x_1, x_2) = z(\mathbf{x})$. Fourier's theorem states

$$z(\mathbf{x}) \simeq \sum_{f_1} \sum_{f_2} A_{f_1 f_2} \cos(f_1 x_1) \cos(f_2 x_2).$$

(a) Fourier's theorem, as stated above, can be rewritten with the trigonometric identity:

$$\cos(\alpha)\cos(\beta) = \frac{1}{2}\cos(\alpha + \beta) + \frac{1}{2}\cos(\alpha - \beta),$$

to give

$$z(x_1, x_2) \simeq \sum_{f_1} \sum_{f_2} \frac{A_{f_1 f_2}}{z2} \left[\cos(f_1 x_1 + f_2 x_2) + \cos(f_1 x_1 - f_2 x_2)\right].$$

(b) We want to show that $\cos(x)$, or indeed any continuous function, can be approximated by the following linear combination

$$f(x) \simeq f(x_0) + \sum_{i=0}^{n} [f(x_{i+1} - f(x_i)] \left[\frac{\text{Sgn}[x - x_i]}{2}\right].$$

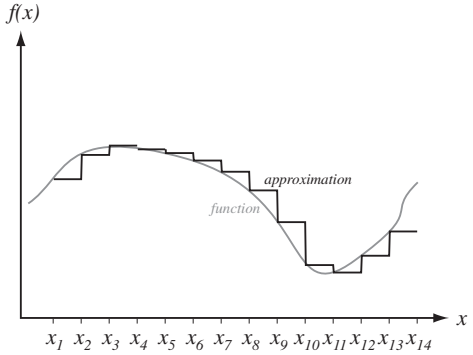
The Fourier series of a function $f(x)$ at a point x_0 converges to $f(x_0)$ if $f(x)$ is of bounded variation in some interval $(x_0 - h, x_0 + h)$ centered on x_0. A function of bounded variation is a follows, given a partition on the interval $a = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = b$ form the sum

$$\sum_{k=1}^{n} |f(x_k) - f(x_{k-1})|.$$

The least upper bound of these sums is called the total variation. For a point $f(x)$ in the neighborhood of $f(x_0)$, we can rewrite the variation as

$$\sum_{i=1}^{n} [f(x_{i+1} - f(x_i)] \left[\frac{\text{Sgn}[x - x_i]}{2}\right],$$

which sets the interval to be $(x, x + 2h)$. Note the function has to be either continuous at $f(x_0)$ or have a discontinuity of the first kind.



(c) As the effective width of the sigmoid vanishes, i.e., as $\sigma \to 0$, the sigmoids become step functions. Then the functions $\cos(f_1 x_1 + f_2 x_2)$ and $\cos(f_1 x_1 - f_2 x_2)$ can be approximated as

$$\begin{aligned}\cos(f_1 x_1 + f_2 x_2) \quad &\simeq \quad \cos(f_1 x_{1_0} + f_2 x_{2_0}) \\ &\quad + \left(\sum_{i=0}^{n} [\cos(x_{1_{i+1}} f_1 + x_{2_{i+1}} f_2) - \cos(x_{1_i} f_1 + x_{2_i} f_2)] \right. \\ &\quad \left. \times \left[\frac{\text{Sgn}[x_1 - x_{1i}]\text{Sgn}[x_2 - x_{2i}]}{2}\right] \right),\end{aligned}$$

and similarly for $\cos[f_1 x_1 - f_2 x_2]$.

(d) The construction does not necessarily guarantee that the derivative is approximated, since there might be discontinuities of the first order, that is, non-continuous first derivative. Nevertheless, the one-sided limits of $f(x_0 + 0)$ and $f(x_0 - 0)$ will exist.

Section 6.3

3. Consider a $d - n_H - c$ network trained with n patterns for m_e epochs.

(a) Consider the space complexity of this problem. The total number of adjustable weights is $dn_H + n_Hc$. The amount of storage for the n patterns is nd.

(b) In stochastic mode we choose pattern randomly and compute

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \Delta\mathbf{w}(t).$$

until stopping criterion is met. Each iteration involves computing $\Delta\mathbf{w}(t)$ and then adding it to $\mathbf{w}(t)$. From Eq. 17 in the text, for hidden-to-output unit weights we have

$$\Delta w_{jk} = \eta(t_k - z_k)f'(net_k)y_j,$$

where $net_k = \sum_{j=1}^{n_H} w_{jk}y_j$ is computed by n_H multiplications and n_H additions. So, Δw_{jk} is computed with $c(2n_H + 1)$ operations.

From Eq. 17 in the text, for input-to-hidden unit weights we have

$$\Delta w_{ji} = \eta x_i f'(net_k) \sum_{k=1}^{c} w_{kj}\delta_k$$

where, net_j is computed in $2d$ operations. Moreover, $\sum_k w_{jk}\delta_k$ is computed in $2c$ operations. Thus we have w_{ij}'s are computed in $[2d + 2c]n_H$ time.

Thus, the time for one iteration is the time to compute $\Delta\mathbf{w}$ plus the time to add $\mathbf{w}$ to $\Delta\mathbf{w}$, that is,

$$\begin{aligned} T &= c(2n_H + 10 + 2(d + c + 1)n_H + (dn_H + n_Hc) \\ &= 3dn_H + 5n_Hc + c + 2n_H. \end{aligned}$$

In summary, the time complexity is $(3dn_H + 5n_Hc + c + 2n_H)m_e$.

(c) Here, the number of iterations $= nm_e$ and thus the time complexity is $(3dn_H + 5n_Hc + c + 2n_H)n\ m_e$.

4. Equation 20 in the text gives the sensitivity at a hidden unit

$$\delta_j = f'(net_j) \sum_{k=1}^{c} w_{kj}\delta_k.$$

For a four-layer or higher-layer neural network, the sensitivity of a hidden unit is likewise given by

$$\delta_j \equiv -\frac{\partial E}{\partial net_j} = -\left[\frac{\partial E}{\partial o_j}\frac{\partial O_j}{\partial net_j}\right].$$

By the chain rule we have

$$\delta_j = -f'(net_j)\sum_k \frac{\partial E}{\partial net_k}\frac{\partial net_k}{\partial o_j}.$$

Now, for a unit k in the next higher layer

$$net_k = \sum_{j'} w_{j'k}o_{j'} = \sum_{j'} w_{j'k}o_{j'} + w_{jk}o_j,$$

where

$$\frac{\delta net_k}{\delta o_j} = w_{jk}.$$

Thus we have

$$\delta_j = f'(net_j)\sum_{k=1}^{c} w_{jk}\underbrace{\Big[-\frac{\partial E}{\partial net_k}\Big]}_{\delta_k} = f'(net_j)\sum_{k=1}^{c} w_{jk}\delta_k.$$

5. From Eq. 21 in the text, the backpropagation rule for training input-to-hidden weights is given by

$$\Delta w_{ji} = \eta x_i f'(net_j)\sum_{k=1} w_{kj}\delta_k = \eta x_i \delta_j.$$

For a fixed hidden unit j, Δw_{ji}'s are determined by x_i. The larger the magnitude of x_i, the larger the weight change. If $x_i = 0$, then there is no input from the ith unit and hence no change in w_{ji}. On the other hand, for a fixed input unit i, $\Delta w_{ji} \propto \delta_j$, the sensitivity of unit j. Change in weight w_{ji} determined how the overall error changes with the activation of unit j. For a large magnitude of δ_j (i.e., large sensitivity), Δw_{ji} is large. Also note that the sensitivities of the subsequent layers propagate to the layer j through weighted sum of sensitivities.

6. There is no reason to expect that the backpropagation rules should be inversely related to $f'(net)$. Note, the backpropagation rules are defined to provide an algorithm that minimizes

$$J = \frac{1}{2}\sum_{k=1}^{c}(t_k - z_k)^2,$$

where $z_k = f(net_k)$. Now, the weights are chosen to minimize J. Clearly, a large change in weight should occur only to significantly decrease J, to ensure convergence in the algorithm, small changes in weight should correspond to small changes in J. But J can be significantly decreased only by a large change in the output o_k. So, large changes in weight should occur where the output varies the most and least where the output varies the least. Thus a change in weight should not be inversely related to $f'(net)$.

7. In a three-layer neural network with bias, the output can be written

$$z_k = f\left(\sum_j w_{jk} f\left(\sum_i w_{ji}x_i\right)\right)$$

can be equivalently expressed as

$$z_k = f\left(\sum_j w_{kj} f\left(\sum_i w_{ji} x_i\right)\right)$$

by increasing the number of input units by 1 and the number of hidden inputs by 1,

$$x_o = 1, \omega_{oj} = \alpha_j, j \neq o, \omega_{io} = 0,$$

and f_h is adjusted such that $o_{bias} = f_h(o) = 1$ and $w_{ok} = \alpha_k$.

8. We denote the number of input units as d, the number of hidden units as n_H and the number of category (output) units as c. There is a single bias unit.

(a) The number of weights is $dn_H + (n_H + 1)c$, where the first term is the number of input-to-hidden weights and the second term the number of hidden-to-output weights, including the bias unit.

(b) The output at output node k is given by Eq. 7 in the text:

$$z_k = \sum_{j=1}^{n_H} f\left(w_{kj} f\left(\sum_{i=1}^{d} w_{ji} x_i + w_{j0}\right) + w_{k0}\right).$$

We assume that the activation function $f(\cdot)$ is odd or antisymmetric, giving

$$f\left(\sum_{j=1}^{d} -w_{ji} x_i\right) \leftrightarrow -f\left(\sum_{j=1}^{d} w_{ji} x_i\right),$$

but since the weighting on this summation also flips, we have

$$(-w_{kj})\left[-f\left(\sum_{j=1}^{d} w_{ji} x_i\right)\right] = (w_{kj})\left[f\left(\sum_{j=1}^{d} w_{ji} x_i\right)\right],$$

and the original output is unchanged.

(c) The hidden units can be exchanged along with corresponding weights and this can leave the network unaffected. The number of subsets that can be constructed over the set n_H is of course 2^{n_H}. Now, because the corresponding weights for each subset n_H different weight orderings can be constructed, so the total hidden unit symmetry is $n_H! 2^{n_H}$. For the case $n_H = 10$, this is a factor of 3,715,891,200.

9. The on-line version of the backpropagation algorithm is as follows:

Algorithm 0 (On-line backpropagation)

```
1 begin initialize n_H, w, η
2         do
3             x ← next input pattern
4             w_ji ← w_ji + ηδ_j x_i;  w_kj ← w_kj + ηδ_k y_i
5         until no more patterns available
6         return w
7   end
```

10. We express the derivative of a sigmoid in terms of the sigmoid itself for positive constants a and b for the following cases.

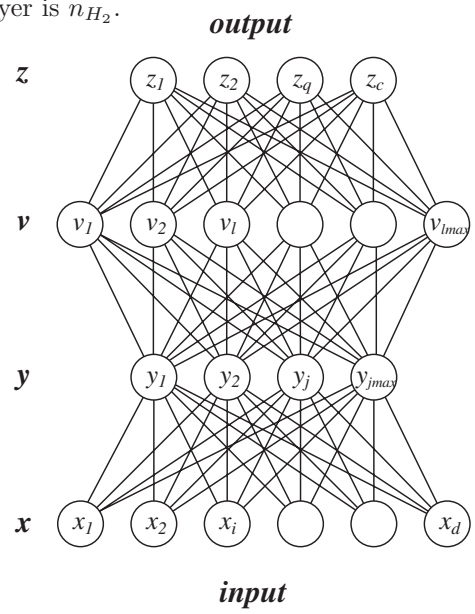
(a) For $f(net) = 1/(1 + e^{a\ net})$, the derivative is

$$\begin{aligned}\frac{df(net)}{d\ net} &= -a\left(\frac{1}{1+e^{a\ net}}\right)^2 e^{a\ net} \\ &= -af(net)(1-f(net)).\end{aligned}$$

(b) For $f(net) = a\tanh(b\ net)$, the derivative is

$$\frac{df(net)}{d\ net} = -2b^2 a\tanh(b\ net)(1-\tanh^2(b\ net)).$$

11. We use the following notation: The activations at the first (input), second, third, and fourth (output) layers are x_i, y_j, v_l, and z_k, respectively, and the indexes are clear from usage. The number of units in the first hidden layer is n_{H_1} and the number in the second hidden layer is n_{H_2}.



Algorithm 0 (Four-layer backpropagation)

```
1 begin initialize xxx
2         xxx
3         xxx
4         return xxx
5 end
```

Section 6.4

12. Suppose the input to hidden weights are set equal to the same value, say w_o, then $w_{ij} = w_o$. Then we have

$$net_j = f(net_j) = \sum_{i=1}^{d} w_{ji}x_i = w_o \sum_i x_i = w_o\mathbf{x}.$$

This means that $o_j = f(net_j)$ is constant, say y_o. Clearly, whatever the topology of the original network, setting the w_{ji} to be a constant is equivalent to changing the topology so that there is only a single input unit, whose input to the next layer is x_o. As, a result of this loss of one-layer and number of input units in the next layer, the network will not train well.

13. If the labels on the two hidden units are exchanged, the shape of the error surface is unaffected. Consider a $d - n_H - c$ three-layer network. From Problem 8 part (c), we know that there are $n_H!2^{n_H}$ equivalent relabelings of the hidden units that do not affect the network output. One can also flip weights for each of these configurations, Thus there should be $n_H!2^{n_H+1}$ possible relabelings and weight changes that leave the error surface unaffected.

14. Consider a simple $2-1$ network with bias. Suppose the training data come from two Gaussians, $p(x|\omega_1) \sim N(-0.5, 1)$ and $p(x|\omega_2) \sim N(0.5, 1)$. The teaching values are ± 1.

(a) The error is a sum over the n patterns as a function of the transfer function $f(\cdot)$ as

$$J(\mathbf{w}) = \sum_{k=1}^{n} (t_k - f(\mathbf{w}^t \mathbf{x}_k + w_0))^2,$$

where t_k is the teaching signal and w_0 the bias weight.

(b) We differentiate twice with respect to the weights to compute the Hessian matrix, which has components

$$H_{ij} = \frac{\partial^2 J(\mathbf{w})}{\partial w_i \partial w_j}.$$

We use the outer product approximation of the Hessian and find

$$H_{ij} = \sum_{k=1}^{n} f'(\mathbf{w}^t \mathbf{x}_k + w_0) x_{ki} f'(\mathbf{w}^t \mathbf{x}_k + w_0) x_{kj}$$

where x_{kj} is the jth component of sample $\mathbf{x}_k$.

(c) Consider two data sets drawn from $p(x|\omega_1) \sim N(\mathbf{u}_i, \mathbf{I})$. The Hessian matrix then has components

$$H_{ij} = n^2(\mu_{1i} + \mu_{2i})(\mu_{1j} + \mu_{2j}),$$

where μ_{1i} is the ith element of the mean fector of class 1, and analogously for class 2.

(d) Problem not yet solved

(e) Problem not yet solved

(f) Problem not yet solved

15. We assume that the error function can be well described by a Hessian matrix $\mathbf{H}$ having d eigenvalues and where λ_{max} and λ_{min} are the maximum and minimum of these eigenvalues.

(a) The optimimum rate is $\eta < 2/\lambda_{max}$.

(b) The convergence criterion is $|1 - \eta\lambda_i| < 1$ for all eigenvalues $i = 1, \ldots, d$.

(c) The time for the system to meet the convergence criterion θ is

$$\theta = (1 - \eta\lambda_i)^T$$

where T is the total number of steps. This factor is dominated by the smallest eigenvalue, so we seek the value T such that

$$\theta = (1 - \eta\lambda_{min})^T.$$

This, in turn, implies that

$$T = \frac{\ln\theta}{\ln(1 - \eta\lambda)}.$$

16. Assume the criterion function $J(\mathbf{w})$ is well described to second order by a Hessian matrix $\mathbf{H}$.

(a) Stepping along the gradient gives at step T

$$\alpha_i^T = (1 - \eta\lambda_i)^T \alpha_i^0.$$

To get a local minimum, we need $|1 - \eta_i| < 1$, and this implies $\eta < 2/\lambda_{max}$.

(b) Consider $(1 - (2/\lambda_{max})\lambda_i)$. More steps will be required along the direction corresponding to the smallest eigenvalue λ_i, and thus the learning rate is governed by

$$1 - \frac{2\lambda_{min}}{\lambda_{max}}.$$

Standardization helps reduce the learning time by making the ratio $\lambda_{max}/\lambda_{min} = 1$.

(c) Standardization is, in essence, a whitening transform.

Section 6.6

17. From Eq. 25 in the text, we have

$$J(\mathbf{w}) = n \left[\frac{n_k}{n} \frac{1}{n_k} \sum_{\mathbf{x} \in \omega_k} [g_k(\mathbf{x}, \mathbf{w}) - 1]^2 + \frac{n - n_k}{n} \frac{1}{n - n_k} \sum_{\mathbf{x} \notin \omega_k} g_k(\mathbf{x}, \mathbf{w})^2 \right].$$

As $n \to \infty$, the proportion of all samples that are in ω_k approaches $P(\omega_k)$. By the law of large numbers, then,

$$\frac{1}{n_k} \sum_{\mathbf{x} \in \omega_k} [g_k(\mathbf{x}, \mathbf{w}) - 1]^2$$

approaches

$$\mathcal{E}([g_k(\mathbf{x}, \mathbf{w}) - 1]^2 | \mathbf{x} \in \omega_k) = \int [g_k(\mathbf{x}, \mathbf{w}) - 1]^2 p(\mathbf{x}|\omega_k) d\mathbf{x}.$$

Likewise, we have

$$\frac{1}{n-n_k}\sum_{x\notin\omega_k}[F_k(\mathbf{x},\omega)]^2,$$

which approaches

$$\mathcal{E}([g_k(\mathbf{x},\mathbf{w})]^2|\mathbf{x}\in\omega_{i\neq k})=\int[g_k(\mathbf{x},\mathbf{w})]^2p(\mathbf{x}|\omega_{i\neq k})d\mathbf{x}.$$

Thus we see, in the limit of infinite data

$$\begin{aligned}
J(\mathbf{w}) &= P(\omega_k)\int[g_k(\mathbf{x},\mathbf{w})-1]^2p(\mathbf{x}|\omega_k)d\mathbf{x}+P(\omega_{i\neq k})\int[g_k(\mathbf{x},\mathbf{w})]^2p(\mathbf{x}|\omega_{i\neq k})d\mathbf{x}\\
&= \int[g_k(\mathbf{x},\mathbf{w})-1]^2p(\mathbf{x},\omega_k)d\mathbf{x}+\int[g_k(\mathbf{x},\mathbf{w})]^2p(\mathbf{x}|\omega_{i\neq k})d\mathbf{x}\\
&= \int[g_k^2(\mathbf{x},\mathbf{w})+1-2g_k(\mathbf{x},\mathbf{w})]p(\mathbf{x},\mathbf{w})]p(\mathbf{x},\omega_k)d\mathbf{x}+\int[g_k(\mathbf{x},\mathbf{w})]^2p(\mathbf{x}|\omega_{i\neq k})d\mathbf{x}\\
&= \int[g_k(\mathbf{x},\mathbf{w})-P(\omega_k|\mathbf{x})]^2p(\mathbf{x})d\mathbf{x}+\int P(\omega_k|\mathbf{x})[1-P(\omega_k|\mathbf{x})]p(\mathbf{x})d\mathbf{x}\\
&= \int[g_k(\mathbf{x},\mathbf{w})-P(\omega_k|\mathbf{x})]^2p(\mathbf{x})dx+\int P(\omega_k|\mathbf{x})P(\omega_{i\neq k}|\mathbf{x})p(\mathbf{x})d\mathbf{x}.
\end{aligned}$$

18. Consider how one of the solutions to the minimum-squared-error condition indeed yields outputs that are estimates of posterior probabilities.

(a) From Eq. 25 in the text, we have

$$\begin{aligned}
J(\mathbf{w}) &= \int g_k^2(\mathbf{x},\mathbf{w})p(\mathbf{x})d\mathbf{x}-2\int g_k(\mathbf{x},\mathbf{w})p(\mathbf{x},\omega_k)d\mathbf{x}+\int p(\mathbf{x},\omega_k)\\
\frac{\partial J(\mathbf{w})}{\partial\mathbf{w}} &= 2\int g_k(\mathbf{x},\mathbf{w})\frac{\partial g_k(\mathbf{x},\mathbf{w})}{\partial\mathbf{w}}p(\mathbf{x})d\mathbf{x}-2\int\frac{\partial g_k(\mathbf{x},\mathbf{w})}{\partial\mathbf{w}}p(\mathbf{x},\omega_k)d\mathbf{x}.
\end{aligned}$$

We set $\frac{\partial J(\mathbf{w})}{\partial\mathbf{w}}=0$ and find

$$\int F_k(\mathbf{x},\omega)\frac{\partial F_k(\mathbf{x},\boldsymbol{\omega})}{\partial\boldsymbol{\omega}}p(\mathbf{x})d\mathbf{x}=\int\frac{\partial F_k(\mathbf{x},\boldsymbol{\omega})}{\partial\boldsymbol{\omega}}p(\mathbf{x},\omega_k)d\mathbf{x}$$

Clearly, $\mathbf{w}^*\equiv\mathbf{w}^*(p(\mathbf{x}))$, the solution of the above equation, depends on the choice of $p(\mathbf{x})$. But, for all $p(\mathbf{x})$, we have

$$\int g_k(\mathbf{x},\mathbf{w}^*(p(\mathbf{x}))\frac{\partial g_k(\mathbf{x},\mathbf{w})}{\partial\mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}^*(p(\mathbf{x}))}p(\mathbf{x})d\mathbf{x}=\int\frac{\partial g_k(\mathbf{x},\mathbf{w})}{\partial\mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}^*(p(\mathbf{x}))}p(\mathbf{x},\omega_k)d\mathbf{x}.$$

Thus we have

$$g_k(\mathbf{x},\mathbf{w}^*(p(\mathbf{x})))p(\mathbf{x})=p(\mathbf{x},\omega_k)$$

with probability 1 on the set of $\mathbf{x}:p(\mathbf{x},\omega_k)>0$. This implies

$$g_k(\mathbf{x},\mathbf{w}^*)=\frac{p(\mathbf{x},\omega_k)}{p(\mathbf{x})}=P(\omega_k|\mathbf{x}).$$

(b) Already shown above.

19. The assumption that the network can represent the underlying true distribution is not used before Eq. 28 in the text. For Eq. 29, however, we invoke $g_k(\mathbf{x};\ \mathbf{w}) \simeq p(\omega_k|\mathbf{x})$, which is used for

$$\sum_{k=1}^{c} \int [g_k(\mathbf{x};\ \mathbf{w}) - P(\omega_k|\mathbf{x})]\, p(\mathbf{x}) d\mathbf{x} = 0.$$

This is true only when the above assumption is met. If the assumption is not met, the gradient descent procedure yields the closest projection to the posterior probability in the class spanned by the network.

20. Recall the equation

$$p(\mathbf{y}|\omega_k) = e^{A(\tilde{\mathbf{w}}_k)+B(\mathbf{y},\phi)+\tilde{\mathbf{w}}^t\mathbf{y}}.$$

(a) Given $p(\mathbf{y}|\omega_k)$, we use Bayes' Theorem to write the posterior as

$$p(\omega_k|\mathbf{y}) = \frac{p(\mathbf{y}|\omega_k)P(\omega_k)}{p(\mathbf{y})}.$$

(b) We interpret $A(\cdot)$, $\tilde{\mathbf{w}}_k$ and ϕ as follows:

$$\begin{aligned} P(\omega_k) &= e^{-A(\tilde{\mathbf{w}}_k)} \\ p(\omega_k|\mathbf{y}) &= \frac{\exp\left[A(\tilde{\mathbf{w}}_k) + B(\mathbf{y},\phi) + \tilde{\mathbf{w}}_k^t\mathbf{y}\right] P(\omega_k)}{\sum_{m=1}^{c} \exp\left[A(\tilde{\mathbf{w}}_m) + B(\mathbf{y},\phi) + \tilde{\mathbf{w}}_m^t\mathbf{y}\right] P(\omega_m)} \\ &= \frac{net_k}{\sum_{m=1}^{c} e^{net_m}}, \end{aligned}$$

where $net_k = b(\mathbf{y},\phi) + \tilde{\mathbf{w}}^t\mathbf{y}$. Thus, $B(\mathbf{y},\phi)$ is the bias, $\tilde{\mathbf{w}}$ is the weight vector describing the separating plane, and $e^{-A(\tilde{\mathbf{w}}_k)}$ is $P(\omega_k)$.

21. Backpropagation with softmax is done in the usual manner; all that must be evaluated differently are the sensitivities at the output and hidden layer units, that is,

$$z_h = \frac{e^{net_h}}{\sum_h e^{net_h}} \qquad \text{and} \qquad \frac{\partial z_h}{\partial net_h} = z_h(1 - z_h).$$

(a) We are given the following terms:

$$\begin{aligned} net_j &= \sum_{i=1}^{d} w_{ji}x_i \\ net_k &= \sum_{j=1}^{n_H} w_{kj}y_j \\ y_j &= f(net_j) \\ z_k &= \frac{e^{net_k}}{\sum_{m=1}^{c} e^{net_m}}, \end{aligned}$$

and the error function

$$J = \frac{1}{2}\sum_{k=1}^{c}(t_k - z_k)^2.$$

To derive the learning rule we have to compute $\partial J/\partial w_{kj}$ and $\partial J/\partial w_{ji}$. We start with the former:

$$\frac{\partial J}{\partial w_{kj}} = \frac{\partial J}{net_k}\frac{\partial net_k}{\partial w_{kj}}.$$

We next compute

$$\frac{\partial J}{\partial net_k} = \sum_{s=1}^{c}\frac{\partial J}{\partial z_s}\frac{\partial z_s}{\partial net_k}.$$

We also have

$$\frac{\partial z_s}{\partial net_k} = \begin{cases} \frac{e^{net_s}(-1)e^{net_k}}{\left(\sum_{m=1}^{c} e^{net_m}\right)^2} = -z_s z_k & \text{if } s \neq k \\ \frac{e^{net_s}}{\left(\sum_{m=1}^{c} e^{net_m}\right)} + (-1)\frac{e^{net_s}e^{net_s}}{\left(\sum_{m=1}^{c} e^{net_m}\right)^2} = z_k - z_k^2 & \text{if } s = k \end{cases}$$

and finally

$$\frac{\partial J}{\partial z_s} = (-1)(t_s - z_s).$$

Putting these together we get

$$\frac{\partial J}{\partial net_k} = \sum_{s\neq k}^{c}(-1)(t_s - z_s)(-z_s z_k) + (-1)(t_k - z_k)(z_k - z_k^2).$$

We use $\partial net_k/\partial w_{kj} = y_j$ and obtain

$$\frac{\partial J}{\partial w_{kj}} = y_j\sum_{s\neq k}^{c}(t_s - z_s)(z_s z_k) - y_j(t_k - z_k)(z_k - z_k^2).$$

Now we have to find input-to-hidden weight contribution to J. By the chain rule we have

$$\frac{\partial J}{\partial w_{ji}} = \frac{\partial J}{\partial y_j}\frac{\partial y_j}{\partial net_j}\frac{\partial net_j}{\partial w_{ji}}.$$

We can at once find out the last two partial derivatives.

$$\frac{\partial y_j}{\partial net_j} = f'(net_j) \text{ and } \frac{\partial net_j}{\partial w_{ji}} = x_i.$$

Now we also have

$$\frac{\partial J}{\partial y_j} = \sum_{s=1}^{c}\frac{\partial J}{\partial z_s}\frac{\partial z_s}{\partial y_j}$$

$$
\begin{aligned}
&= -\sum_{s=1}^{c}(t_s - z_s)\frac{\partial z_s}{\partial y_j} \\
&= -\sum_{s=1}^{c}(t_s - z_s)\left[\sum_{r=1}^{c}\frac{\partial z_s}{\partial net_r}\frac{\partial net_r}{\partial y_j}\right] \\
&= -\sum_{s=1}^{c}(t_s - z_s)\left[\underbrace{\frac{\partial z_s}{\partial net_s}}_{z_s - z_s^2}\underbrace{\frac{\partial net_s}{\partial y_j}}_{w_{sj}} + \sum_{r\neq s}^{c}\underbrace{\frac{\partial z_s}{\partial net_r}}_{-z_s z_r}\underbrace{\frac{\partial net_r}{\partial y_j}}_{w_{rj}}\right] \\
&= -\sum_{s=1}^{c}(t_s - z_s)(z_s - z_s^2)w_{sj} + \sum_{s=1}^{c}\sum_{r\neq s}^{c} z_s z_r w_{rj}(t_s - z_s).
\end{aligned}
$$

We put all this together and find

$$
\begin{aligned}
\frac{\partial J}{\partial w_{ji}} &= x_i f'(net_j)\sum_{s=1}^{c}(t_s - z_s)\sum_{r\neq s}^{c} w_{rj} z_s z_r \\
&\quad -x_i f'(net_j)\sum_{s=1}^{c}(t_s - z_s)w_{sj}(z_s - z_s^2).
\end{aligned}
$$

Of course, the learning rule is then

$$
\begin{aligned}
\Delta w_{ji} &= -\eta\frac{\partial J}{\partial w_{ji}} \\
\Delta w_{kj} &= -\eta\frac{\partial J}{\partial w_{kj}},
\end{aligned}
$$

where the derivatives are as given above.

(b) We are given the cross-entropy criterion function

$$
J_{CE} = \sum_{k=1}^{c} t_k \ln\frac{t_k}{z_k}.
$$

The learning rule derivation is the same as in part (a) except that we need to replace $\partial J/\partial z$ with $\partial J_{ce}/\partial z$. Note that for the cross-entropy criterion we have $\partial J_{ce}/\partial z_k = -t_k/z_k$. Then, following the steps in part (a), we have

$$
\begin{aligned}
\frac{\partial J_{ce}}{\partial w_{kj}} &= y_j\sum_{s\neq k}^{c}\frac{t_k}{z_k}z_s z_k - y_j\frac{t_k}{z_k}(z_k - z_k^2) \\
&= y_j\sum_{s\neq k}^{c} t_k z_k - y_j t_k(1 - z_k).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
\frac{\partial J_{ce}}{\partial w_{ji}} &= x_i f'(net_j)\sum_{s=1}^{c}\frac{t_s}{z_s}\sum_{r\neq x}^{c} w_{rj} z_s z_r \\
&\quad -x_i f'(net_j)\sum_{s=1}^{c}\frac{t_s}{z_s}w_{sj}(z_s - z_s^2).
\end{aligned}
$$

Thus we find

$$\begin{aligned}\frac{\partial J_{ce}}{\partial w_{ji}} &= x_i f'(net_j)\sum_{s=1}^{c} t_s \sum_{r\neq s}^{c} w_{rj} z_r \\ &\quad -x_i f'(net_j)\sum_{s=1}^{c} t_s w_{sj}(1-z_s).\end{aligned}$$

Of course, the learning rule is then

$$\begin{aligned}\Delta w_{ji} &= -\eta\frac{\partial J_{ce}}{\partial w_{ji}} \\ \Delta w_{kj} &= -\eta\frac{\partial J_{ce}}{\partial w_{kj}},\end{aligned}$$

where the derivatives are as given above.

22. In the two-category case, if $g_1 \simeq P(\omega_1|x)$, then $1-g_1 \simeq P(\omega_2|x)$, since we can assume the categories are mutually exclusive and exhaustive. But $1-g_1$ can be computed by a network with input-to-hidden weights identical to those used to compute g_1. From Eq. 27 in the text, we know

$$\sum_{k_1}\int \left[g_{k_1}(\mathbf{x},\mathbf{w}) - P(\omega_{k_1}|\mathbf{x})\right]^2 d\mathbf{x} + \sum_{k_2}\int \left[g_{k_2}(\mathbf{x},\mathbf{w}) - P(\omega_{k_2}|\mathbf{x})\right] d\mathbf{x}$$

is a minimum. this implies that every term in the above equation is minimized.

Section 6.7

23. Consider the weight update rules given by Eqs. 12 and 23 in the text.

(a) The weight updates are have the factor $\eta f'(net)$. If we take the sigmoid $f_b(h) = \tanh(bh)$, we have

$$f_b'(h) = 2b\frac{e^{-bh}}{(1+e^{-bh})^2} - 1 = 2b\underbrace{\frac{1}{e^{bh}+e^{-bh}+2}}_{D} - 1 = 2bD = 1.$$

Clearly, $0 < D < 0.25$ for all b and h. If we assume D is constant, then clearly the product $\eta/\gamma f'_{\gamma b}(h)$ will be equal to $\eta f'_b(h)$, which tells us the increment in the weight values will be the same, preserving the convergence time. The assumption will be approximately true as long as $|bh|$ is very small or kept constant in spite of the change in b.

(b) If the input data is scaled by $1/\alpha$, then the increment of weights at each step will be exactly the same. That is,

$$\frac{\eta}{\gamma} f'_{\gamma b}(h/\gamma) = \eta f'(h).$$

Therefore the convergence time will be kept the same. (However, if the network is a multi-layer one, the input scaling should be applied to the hidden units' outputs as well.)

24. A general additive model is given by Eq. 32 in the text:

$$z_k = f\left(\sum_{j=1}^{d} f_j^k(x_j) + w_o^k\right).$$

The functions f_k act on single components, namely the x_i. This actually restricts the additive models. To have the full power of three-layer neural networks we must assume that f_i is multivariate function of the inputs. In this case the model will become:

$$z_k = f\left(\sum_{j=1}^{d} f_j^k(x_1, x_2, \ldots, x_d) + w_0\right).$$

With this modification it is trivial to implement any three-layer neural network. We let

$$f_j^k = w_{kj} g\left(\sum_{j=1}^{d} w_{ji} x_i + w_{j0}\right)$$

and $w_0^k = w_{k0}$, where g, w_{kj}, w_{ji}, w_{j0}, w_{k0} are defined as in Eq. 32 in the text. We substitute f_j^k into the above equation and at once arrive at the three-layer neural network function shown in Eq. 7 in the text.

25. Let $p_x(x)$ and $p_y(y)$ be probability density functions of x and y, respectively.

(a) From the definition of entropy given by Eq. 37 in Chapter 1 of the text (or Eq. 118 in Section A.7.1 in the Appendix), and the one-dimensional Gaussian,

$$p_x(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right].$$

we have the entropy

$$H(p_x(x)) = -\int_{-\infty}^{\infty} p_x(x) \log p_x(x) dx = \frac{1}{2} + \log\sqrt{2\pi}\sigma \simeq 1.447782 + \log\sigma \text{ bits.}$$

(b) From Sect. 6.8.2 in the text, if $a = 2/(3\sigma)$, then we have $f'(x) \simeq 1$ for $-\sigma < x < \sigma$.

(c) Suppose $y = f(x)$. If there exists a unique inverse mapping $x = f^{-1}(y)$, then $p_y(y)dy = f'(x)dx$. Thus we have

$$\begin{aligned}
H(p_y(y)) &= -\int_{-\infty}^{\infty} p_y(y) \log p_y(y) dy \\
&= -\int_{-\infty}^{\infty} p_x(x) \log p_x(x) dx + \int_{-\infty}^{\infty} p_x(x) \log|f'(x)| dx \\
&= H(p_x(x)) + \log 2ab - 2\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right] \log[1 + \exp(-bx)] dx \\
&= 1.418939 + \log\sigma \text{ bits.}
\end{aligned}$$

(d) We define the function

$$y = f(x) = \begin{cases} a & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

then we have $p_y(y) = \delta(y)$ because

$$\Pr[Y = y] = \begin{cases} 1 & \text{if } y = 0 \\ 0 & \text{if } y = a \\ 0 & \text{otherwise.} \end{cases}$$

his gives our entropy to be $H(y) = -\infty$.

(e) Information possessed by the original data is transmitted with litle loss through sigmoid functions, whereas it is completely lost through through a Diract delta function.

26. The transfer function is given by Eq. 34 in the text:

$$f(net) = \frac{2a}{1 + e^{-b\ net}} - a.$$

(a) We calculate the derivative as

$$\begin{aligned} f'(net) &= \frac{2a}{(1 + e^{-b\ net})^2} b e^{-b\ net} \\ &= \left[\frac{2a}{1 + e^{-b\ net}}\right] \frac{b e^{-b\ net}}{1 + e^{-b\ net}} \\ &= b \frac{2a}{1 + e^{-b\ net}} \left[1 - \frac{1}{1 + e^{-b\ net}}\right] \\ &= \frac{b}{2a}[a^2 - (f(net))^2]. \end{aligned}$$

(b) Note that $f''(net) = \frac{-b}{a} f(net) f'(net)$, where $b > 0$. At $net = \infty$, we have

$$\begin{aligned} f(\infty) &= \frac{2a}{1 + e^{-b\ net}} - a = 2a - a = a \\ f'(\infty) &= \frac{b}{2a}[a^2 - (f(net))^2] = \frac{b}{2a}(a^2 - a^2) = 0 \\ f''(\infty) &= -\frac{b}{a} f(net) f'(net) = 0. \end{aligned}$$

At $net = 0$, we have

$$\begin{aligned} f(0) &= \frac{2a}{1 + e^{-b\ net}} - a = \frac{2a}{a} - a = 0 \\ f'(0) &= \frac{b}{2a}(a^2 - (f(net))^2) = \frac{b}{2a} a^2 = \frac{ab}{2} \\ f''(0) &= 0. \end{aligned}$$

At $net = -\infty$, we have

$$\begin{aligned} f(-\infty) &= \frac{2a}{1 + e^{-b\ net}} - a = 0 - a = -a \\ f'(-\infty) &= \frac{b}{2a}(a^2 - (f(net))^2) = \frac{b}{2a}(a^2 - a^2) = 0 \\ f''(-\infty) &= 0. \end{aligned}$$

27. We consider the computational burden for standardizing data, as described in the text.

(a) The data should be shifted and scaled to ensure zero mean and unit variance (standardization). To compute the mean, the variance we need $2nd$ steps, and thus the complexity is thus $O(nd)$.

(b) A full forward pass of the network activations requires $n_H d + n_H c$ steps. A backpropagation of the error requires $n_H d + n_H dc$ steps. The first term is for the output-to-hidden and the second term is for hidden-to-input weights. The extra c factor comes from the backpropagation formula given in Eq. 21 in the text. Thus nd epochs of training require

$$(nd)n[n_h d + n_H c + n_H c + n_H dc]$$

steps. Assuming a single output netowrk we can set $c = 1$ and use the approximation given in Sec. 6.8.7 to get $n/10 = n_H d + n_H$. We use this result in the above and find

$$n^2 d[n/10 + n/10] = \frac{n^3 d}{5},$$

and thus is $O(n^3 d)$ complexity.

(c) We use the results of parts (a) and (b) above and see that the ratio of steps can be approximated as

$$\frac{nd}{n^3 d/5} = \frac{5}{n^2}.$$

This tells us that the burden of standardizing data gets negligible as n geets larger.

28. The Minkowski error per pattern is defined by

$$J = \sum_{k=1}^{c} |z_k - t_k|^R = \sum_{k=1}^{c} |t_k - z_k|^R.$$

We proceed as in page 290 of the text for the usual sum-square-error function:

$$\frac{\partial J}{\partial w_{kj}} = \frac{\partial J}{\partial net_k} \frac{\partial net_k}{\partial w_{kj}}$$

where $z_k = f\left(\sum_{j=1}^{n_H} w_{kj} y_j\right)$. We also have

$$\frac{\partial J}{\partial net_k} = -R|t_k - z_k|^{R-1} f'(net_k)\mathrm{Sgn}(t_k - z_k),$$

where the signum function can be written $\mathrm{Sgn}(x) = x/|x|$. We also have

$$\frac{\partial net_k}{\partial w_{kj}} = y_j.$$

Thus we have

$$\frac{\partial J}{\partial w_{kj}} = -R|t_k - z_k|^{R-1} f'(net_k) y_j \text{Sgn}(t_k - z_k),$$

and so the update rule for w_{kj} is

$$\Delta w_{kj} = \eta |t_k - z_k|^{R-1} f'(net_k) y_j \text{Sgn}(t_k - z_k).$$

Now we compute $\partial J/\partial w_{ji}$ by the chain rule:

$$\frac{\partial J}{\partial w_{ji}} = \frac{\partial J}{\partial y_j} \frac{\partial y_j}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}}.$$

The first term on the right-hand side is

$$\begin{aligned} \frac{\partial J}{\partial y_j} &= \sum_{k=1}^{c} \frac{\partial J}{\partial z_k} \frac{\partial z_k}{\partial y_j} \\ &= \sum_{k=1}^{c} -R|t_k - z_k|^{R-1} f'(net_k) w_{kj} \text{Sgn}(t_k - z_k). \end{aligned}$$

The second term is simply $\partial y_j/\partial net_j = f'(net_j)$. The third term is $\partial net_j/\partial w_{ji} = x_i$. We put all this together and find

$$\frac{\partial J}{\partial w_{ji}} = -\left[\sum_{k=1}^{c} R|t_k - z_k|^{R-1} \text{Sgn}(t_k - z_k) f'(net_k) w_{kj}\right] f'(net_j) x_i.$$

and thus the weight update rule is

$$\Delta w_{ji} = \eta \left[\sum_{k=1}^{c} w_{kj} f'(net_k) \text{Sgn}(t_k - z_k) R|t_k - z_k|^{R-1}\right] f'(net_j) x_i.$$

For the Euclidean distance measure, we let $R = 2$ and obtain Eqs. 17 and 21 in the text, if we note the identity

$$\text{Sgn}(t_k - z_k) \cdot |t_k - z_k| = t_k - z_k.$$

29. Consider a $d - n_H - c$ three-layer neural network whose input units are liner and output units are sigmoidal but each hidden unit implements a particular polynomial function, trained on a sum-square error criterion. Here the output of hidden unit j is given by

$$o_j = w_{ji} x_i + w_{jm} x_m + q_j x_i x_m$$

for two prespecified inputs, i and $m \neq i$.

(a) We write the network function as

$$z_k = f\left(\sum_{j=1}^{n_H} w_{kj} o_j(\mathbf{x})\right)$$

where $o_j(\mathbf{x})$ is

$$o_j(\mathbf{x}) = w_{ji_j}x_{i_j} + w_j m_j x_{m_j} + q_j x_{i_j} w_{m_j}.$$

Since each hidden unit has prespecified inputs i and m, we use a subscript on i and m to denote their relation to the hidden unit. The criterion function is

$$J = \sum_{k=1}^{c}(t_k - z_k)^2.$$

Now we calculate the derivatives:

$$\frac{\partial J}{\partial w_{kj}} = \frac{\partial J}{\partial net_k}\frac{\partial net_k}{\partial w_{kj}} = -(t_k - z_k)f'(net_k)o_j(\mathbf{x}).$$

Likewise we have

$$\begin{aligned}\frac{\partial J}{\partial w_{ji}} &= \frac{\partial J}{\partial o_j}\frac{\partial o_j}{\partial w_{ji}} = \underbrace{\left[\sum_{k=1}^{c}\frac{\partial J}{\partial z_k}\frac{\partial z_k}{\partial o_j}\right]}_{\partial J/\partial o_j}\frac{\partial o_j}{\partial w_{ji}} \\ &= -\left[\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}\right]\frac{\partial o_j}{\partial w_{ji}}.\end{aligned}$$

Now we turn to the second term:

$$\frac{\partial o_j}{\partial w_{ji}} = \begin{cases} x_i + q_j x_{m_j} & \text{if } i = i_j \\ x_i + q_j x_{i_j} & \text{if } i = m_j \\ 0 & \text{otherwise.}\end{cases}$$

Now we can write $\partial J/\partial w_{ji}$ for the weights connecting inputs to the hidden layer. Remember that each hidden unit j has three weights or parameters for the two specified input units i_j and m_j, namely w_{ji_j}, w_{jm_j}, and q_j. Thus we have

$$\begin{aligned}\frac{\partial J}{\partial w_{ji_j}} &= -\left[\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}\right](x_{i_j} + q_j x_{m_j}) \\ \frac{\partial J}{\partial w_{jm_j}} &= -\left[\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}\right](x_{i_m} + q_j x_{i_m}) \\ \frac{\partial J}{\partial wjr} &= 0 \text{ for } r \neq i_j \text{ and } r \neq m_j.\end{aligned}$$

We must also compute

$$\frac{\partial J}{\partial q_j} = \frac{\partial J}{\partial o_j}\frac{\partial o_j}{\partial q_j},$$

where $\partial J/\partial q_j = x_{i_j}x_{m_j}$ and

$$\frac{\partial J}{\partial o_j} = -\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}.$$

Thus we have

$$\frac{\partial J}{\partial q_j} = -\left[\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}\right] x_{i_j}x_{m_j}.$$

Thus the gradient descent rule for the input-to-hidden weights is

$$\begin{aligned}
\Delta w_{ji_j} &= \eta\left[\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}\right](x_{i_j} + q_j x_{m_j}) \\
\Delta w_{jm_j} &= \eta\left[\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}\right](x_{i_m} + q_j x_{i_j}) \\
\Delta q_j &= \eta\left[\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}\right] x_{i_j}x_{m_j}.
\end{aligned}$$

(b) We observe that $y_i = o_i$, from part (a) we see that the hidden-to-output weight update rule is the same as the standard backpropagation rule.

(c) The most obvious weakness is that the "receptive fields" of the network are a mere two units. Another weakness is that the hidden layer outputs are not bounded anymore and hence create problems in convergence and numerical stability. A key fundamental weakness is that the network is no more a universal approximator/classifier, because the function space F that the hidden units span is merely a subset of all polynomials of degree two and less. Consider a classification task. In essence the network does a linear classification followed by a sigmoidal non-linearity at the output. In order for the network to be able to perform the task, the hidden unit outputs must be linearly separable in the F space. However, this cannot be guaranteed; in general we need a much larger function space to ensure the linear separability.

The primary advantage of this scheme is due to the task domain. If the task domain is known to be solvable with such a network, the convergence will be faster and the computational load will be much less than a standard backpropagation learning network, since the number of input-to-hidden unit weights is $3n_H$ compared to dn_H of standard backpropagation.

30. The solution to Problem 28 gives the backpropagation learning rule for the Minkowski error. In this problem we will derive the learning rule for the Manhattan metric directly. The error per pattern is given by

$$J = \sum_{k=1}^{c}|t_k - z_k|.$$

We proceed as in page 290 of the text for the sum-squared error criterion function:

$$\begin{aligned}
\frac{\partial J}{\partial w_{kj}} &= \frac{\partial J}{\partial net_k}\frac{\partial net_k}{\partial w_{kj}} \\
\frac{\partial J}{\partial net_k} &= -f'(net_k)\mathrm{Sgn}(t_k - z_k),
\end{aligned}$$

where $\text{Sgn}(x) = x/|x|$. Finally, we have $\partial net_k/\partial w_{kj} = y_j$. Thus we have

$$\frac{\partial J}{\partial w_{kj}} = -f'(net_k)y_j\text{Sgn}(t_k - z_k),$$

and so the update rule for w_{kj} is:

$$\Delta w_{kj} = \eta f'(net_k)y_j\text{Sgn}(t_k - z_k).$$

Now we turn to $\partial J/\partial w_{ji}$. By the chain rule we have

$$\frac{\partial J}{\partial w_{ji}} = \frac{\partial J}{\partial y_j}\frac{\partial y_j}{\partial net_j}\frac{\partial net_j}{\partial w_{ji}}.$$

We compute each of the three factors on the right-hand-side of the equation:

$$\begin{aligned}
\frac{\partial J}{\partial y_j} &= \sum_{k=1}^{c}\frac{\partial J}{\partial z_k}\frac{\partial z_k}{\partial y_j} \\
&= \sum_{k=1}^{c} -f'(net_k)w_{kj}\text{Sgn}(t_k - z_k) \\
\frac{\partial y_j}{\partial net_j} &= f'(net_j) \\
\frac{\partial net_j}{\partial w_{ji}} &= x_i.
\end{aligned}$$

Thus we put this together and find

$$\frac{\partial J}{\partial w_{ji}} = -\left[\sum_{k=1}^{c}\text{Sgn}(t_k - z_k)f'(net_k)w_{kj}\right]f'(net_j)x_i.$$

Thus the weight update rule for w_{ji} is:

$$\Delta w_{ji} = \eta\left[\sum_{k=1}^{c} w_{kj}f'(net_k)\text{Sgn}(t_k - z_k)\right]f'(net_j)x_i.$$

The above learning rules define the backpropagation learning rule with the Manhattan metric. Note that the Manhattan metric uses the direction (signum) of $t_k - z_k$ instead of the actual $t_k - z_k$ as in the sum-squared-error criterion.

Section 6.9

31. We are given the criterion function

$$J = \frac{1}{2}\sum_{m=1}^{n}(t_m - z_m)^2.$$

The Hessian matrix is then

$$\frac{\partial^2 J}{\partial w_{ji}\partial w_{lk}} = \frac{1}{2}\left[\sum_{m=1}^{n}\frac{\partial z_m}{\partial w_{ji}}\frac{\partial z_m}{\partial w_{lk}} + \underbrace{\sum_{m=1}^{n}(z_m - t_m)\frac{\partial^2 z_m}{\partial w_{ji}\partial w_{lk}}}_{\simeq 0}\right].$$

We drop the last term in the outer-product approximation, as given in Eq. 45 in the text. Let W_j be any hidden-to-output weight in a single output network. Then we have $\partial z/\partial W_j = f'(net)y_j$. Now we let w_{ji} be any input-to-hidden weight. Then since

$$z = f\left(\sum_j W_j f(\sum_i w_{ji}x_i)\right),$$

the derivative is

$$\frac{\partial z}{\partial w_{ji}} = f'(net)W_j f'(net_j)x_i.$$

Thus the derivatives with respect to the hidden-to-output weights are:

$$\mathbf{X}_v^t = (f'(net)y_1, \ldots, f'(net)y_{n_H}),$$

as in Eq. 47 in the text. Further, the derivatives with respect to the intput-to-hidden weights are

$$\begin{aligned}\mathbf{X}_u^t &= (f'(net)f'(net_j)W_1x_1, \ldots, f'(net)f'(net_{n_H})W_{n_H}x_1,\\ &\quad f'(net)f'(net_1)W_1x_d, \ldots, f'(net)f'(net_{n_H})W_{n_H}x_d).\end{aligned}$$

32. We are given that $J_{CE} = \sum_{k=1}^{c} t_k \ln(t_k/z_k)$. For the calculation of the Hessian matrix, we need to find the expressions for the second derivatives:

$$\begin{aligned}\frac{\partial^2 J_{CE}}{\partial w_{ji}\partial w_{lk}} &= \frac{\partial\left[\frac{\partial J_{CE}}{\partial w_{ji}}\right]}{\partial w_{lk}} = \frac{\partial\left[-\frac{t_k}{z_k}\frac{\partial z}{\partial w_{ji}}\right]}{\partial w_{lk}}\\ &= \frac{t_k}{z_k^2}\left[\frac{\partial z_k}{\partial w_{ji}}\frac{\partial z_k}{\partial w_{lk}} - z_k\frac{\partial^2 z_k}{\partial w_{ji}\partial w_{lk}}\right].\end{aligned}$$

We arrive at the outer product approximation by dropping the second-order terms. We compare the above formula to Eq. 44 in the text and see that the Hessian matrix, $\mathbf{H}_{CE}$ differs from the Hessian matrix in the traditional sum-squared error case in Eq. 45 by a scale factor of t_k/z_k^2. Thus, for a $d - n_H - 1$ network, the Hessian matrix can be approximated by

$$\mathbf{H}_{CE} = \frac{1}{n}\sum_{m=1}^{n}\frac{t_1}{z_1^2}\mathbf{X}^{[m]t}\mathbf{X}^{[m]},$$

where $\mathbf{X}^{[m]}$ is defined by Eqs. 46, 47 and 48 in the text, as well as in Problem 31.

33. We assume that the current weight value is $\mathbf{w}(n-1)$ and we are doing a line search.

(a) The next $\mathbf{w}$ according to the line search will be found via

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \lambda\nabla J(\mathbf{w}(n-1))$$

by minimizing $J(\mathbf{w})$ using only λ. Since we are given that the Hessian matrix is proportional to the identity, $\mathbf{H} \propto \mathbf{I}$, the Taylor expansion of $J(\mathbf{w})$ up to second

degree will give an exact representation of J. We expand around $\mathbf{w}(n-1)$ and find

$$\begin{aligned} J(\mathbf{w}) &= J(\mathbf{w}(n-1)) + (\mathbf{w}-\mathbf{w}(n-1))^t \nabla J(\mathbf{w}(n-1)) \\ &\quad +1/2(\mathbf{w}-\mathbf{w}(n-1))^t \mathbf{H}(\mathbf{w}-\mathbf{w}(n-1)). \end{aligned}$$

We know that $\mathbf{H} = k\mathbf{I}$ for some scalar k. Further, we can write $\mathbf{w} - \mathbf{w}(n-1) = \lambda \nabla J(\mathbf{w}(n-1))$, which can be plugged into the expression for $J(\mathbf{w})$ to give

$$\begin{aligned} J(\mathbf{w}) &= J(\mathbf{w}(n-1)) + \lambda \nabla J^t(\mathbf{w}(n-1)) \nabla J(\mathbf{w}(n-1)) \\ &\quad +(k/2)\lambda^2 \nabla J^t(\mathbf{w}(n-1)) \nabla J(\mathbf{w}(n-1)) \\ &= J(\mathbf{w}(n-1)) + \left(\lambda + \frac{k}{2}\lambda^2)\right) \|\nabla J(\mathbf{w}(n-1))\|^2. \end{aligned}$$

Now we see the result of the line search in J by solving $\partial J/\partial\lambda = 0$ for λ, that is,

$$\frac{\partial J}{\partial \lambda} = \|\nabla J(\mathbf{w}(n-1))\|^2 (1 + k\lambda) = 0,$$

which has solution $\lambda = -1/k$. Thus the next $\mathbf{w}$ after the line search will be

$$\mathbf{w}(n) = \mathbf{w}(n-1) - 1/k \nabla J(\mathbf{w}(n-1)).$$

We can also rearrange terms and find

$$\mathbf{w}(n) - \mathbf{w}(n-1) = \Delta \mathbf{w}(n-1) = -\frac{1}{k} \nabla J(\mathbf{w}(n-1)).$$

Using the Taylor expansion for $\nabla J(\mathbf{w}(n)$ around $\mathbf{w}(n-1)$ we can write

$$\nabla J(\mathbf{w}(n)) = \nabla J(\mathbf{w}(n-1)) + k(\mathbf{w}(n) - \mathbf{w}(n-1)),$$

then substitute $\mathbf{w}(n) - \mathbf{w}(n-1)$ from above into the right-hand-side and find

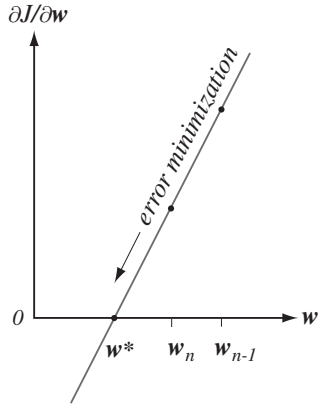
$$\nabla J(\mathbf{w}(n)) = \nabla J(\mathbf{w}(n-1)) + k(-1/k \nabla J(\mathbf{w}(n-1))) = 0.$$

Indeed, Eqs. 56 and 57 in the text give $\beta_n = 0$, given that $\nabla J(\mathbf{w}_n) = 0$, as shown. Equation 56 in the text is trivially satisfied since the numerator is $\|\nabla J(\mathbf{w}(n))\|^2$, and thus the Fletcher-Reeves equation gives $\beta_n = 0$. Equation 57 also vanishes because the numerator is the inner product of two vectors, one of which is $\nabla J(\mathbf{w}(n)) = \mathbf{0}$, and thus the Polak-Ribiere equation gives $\beta_n = 0$.

(b) The above proves that application of a line search reult with $\mathbf{w}(n-1)$ takes us to the minimum of J, and hence no further weight update is necessary.

34. Problem not yet solved

35. The Quickprop algorithm assumes the weights are independent and the error surface is quadratic. Whith these assumptions, $\partial J/\partial \mathbf{w}$ is linear. The graph shows $\partial J/\partial \mathbf{w}$ versus $\mathbf{w}$ and a possible error minimization step. For finding a minimum in the criterion function, we search for $\partial J/\partial \mathbf{w} = 0$. That is the value $\mathbf{w}$ where the line crosses the $\mathbf{w}$ axis.



We can easilty compute the intercept in this one-dimensional illustration. The equation of the line passing through $(x_0, y_0$ and $x_1, y_1)$ is $(x-x_1)/(x_1-x_0) = (y-y_1)/(y_1-y_0)$. The x axis crossing is found by setting $y = 0$ and solving for x, that is

$$x - x_1 = (x_1 - x_0)\frac{-y_1}{y_1 - y_0} = (x_1 - x_0)\frac{y_1}{y_0 - y_1}.$$

Now we can plug in our point shown on the graph, with $x_0 = \mathbf{w}_{n-1}$, $y_0 = \partial J/\partial \mathbf{w}|_{n-1}$, $x_1 = \mathbf{w}_n$, $y_1 = \partial J/\partial \mathbf{w}|_n$, and by noting $\Delta\mathbf{w}(m) = (x_1-x_0)$ and $\Delta\mathbf{w}(n+1) = (x-x_1)$. The final result is

$$\Delta\mathbf{w}(n+1) = \frac{\partial J/\partial \mathbf{w}|_n}{\partial J/\partial \mathbf{w}|_{n-1} - \partial J/\partial \mathbf{w}|_n}\Delta\mathbf{w}(n).$$

Section 6.10

36. Consider the matched filter described by Eq. 66 in the text.

(a) The trial weight function $w(t) = w^*(t) + h(t)$ constrained of Eq. 63 must be obeyed:

$$\int_{-\infty}^{\infty} w^2(t)dt = \int_{-\infty}^{\infty} w^{*2}(t)dt,$$

and we expand to find

$$\int_{-\infty}^{\infty} w^{*2}dt + \int_{-\infty}^{\infty} h^2(t)dt + 2\int_{-\infty}^{\infty} w^*(t)h(t)dt = \int_{-\infty}^{\infty} w^{*2}(t)dt$$

and this implies

$$-2\int_{-\infty}^{\infty} h(t)w^*(t)dt = \int_{-\infty}^{\infty} h^2(t)dt \geq 0.$$

(b) We have

$$\int_{-\infty}^{\infty} x(t)w(t)dt = \underbrace{\int_{-\infty}^{\infty} x(t)w^*(t)dt}_{output\ z^*} + \int_{-\infty}^{\infty} x(t)h(t)dt.$$

From Eq. 66 in the text we know that $w^*(t) = -\lambda/2x(t)$, and thus $x(t) = 2w^*(t)/(-\lambda)$. We plug this form of $x(t)$ into the right-hand side of the above equation and find

$$\int_{-\infty}^{\infty} x(t)w(t)dt = \int_{-\infty}^{\infty} x(t)[w^*(t) + h(t)]dt = z^* + \frac{2}{-\lambda}\int_{-\infty}^{\infty} w^*(t)h(t)dt.$$

From part (a) we can substitute

$$2\int_{-\infty}^{\infty} w^*(t)h(t)dt = -\int_{-\infty}^{\infty} h^2(t)dt,$$

and thus the output z of the matched filter is given by:

$$z = \int_{-\infty}^{\infty} x(t)w(t)dt = z^* - \frac{1}{-\lambda}\int_{-\infty}^{\infty} h^2(t)dt.$$

(c) Plugging the above into Eq. 66 in the text,

$$z^* = \int_{-\infty}^{\infty} w^*(t)x(t)dt,$$

we get

$$z^* = \int_{-\infty}^{\infty} -\frac{\lambda}{2}x(t)x(t)dt.$$

Now we check on the value of λz^*:

$$\lambda z^* = -\frac{\lambda^2}{2}\int_{-\infty}^{\infty} x^2(t)dt < 0.$$

Thus z^* and λ have opposite signs.

(d) Part (b) ensures $z = z^*$ if and only if

$$\frac{1}{\lambda}\int_{-\infty}^{\infty} h^2(t)dt = 0,$$

which occurs if and only if $h(t) = 0$ for all t. That is, $w^*(t)$ ensures the maximum output. Assume $w_1(t)$ also assures maximum output z^* as well, that is

$$\begin{aligned}\int_{-\infty}^{\infty} x(t)w^*(t)dt &= z^* \\ \int_{-\infty}^{\infty} x(t)w_1(t)dt &= z^*.\end{aligned}$$

We subtract both sides and find

$$\int_{-\infty}^{\infty} x(t)[w_1(t) - w^*(t)]dt = 0.$$

Now we can let $h(t) = w_1(t) - w^*(t)$. We have just seen that given the trial weight function $w_1(t) = w^*(t) + h(t)$, that $h(t) = 0$ for all t. This tells us that $w_1(t) - w^*(t) = 0$ for all t, or equivalently $w_1(t) = w^*(t)$ for all t.

37. Consider OBS and OBD algorithms.

(a) We have from Eq. 49 in the text

$$\begin{aligned}\delta J &= \underbrace{\left(\frac{\partial J}{\partial \mathbf{w}}\right)^t \cdot \delta \mathbf{w}}_{\simeq 0} + \frac{1}{2}\delta \mathbf{w}^t \cdot \frac{\partial^2 J}{\partial \mathbf{w}^2} \cdot \delta \mathbf{w} + \underbrace{O(\|\delta \mathbf{w}^3\|)}_{\simeq 0} \\ &= \frac{1}{2}\delta \mathbf{w}^t \mathbf{H} \delta \mathbf{w}.\end{aligned}$$

So, it is required to minimize $\delta \mathbf{w}^t \mathbf{H} \delta \mathbf{w}$, subject to the constraint of deleting one weight, say weight q. Let $\mathbf{u}_q$ be the unit vector along the qth direction in weight space. Then, pre- and post-operating "picks out" the qq entry of a matrix, in particular,

$$\mathbf{u}_q^t \mathbf{H}^{-1} \mathbf{u}_q = [\mathbf{H}^{-1}]_{qq}.$$

Now we turn to the problem of minimizing $\frac{1}{2}\delta \mathbf{w}^t \mathbf{H} \delta \mathbf{w}$ subject to $\delta \mathbf{w}^t \mathbf{u}_q = -w_q$. By the method of Lagrangian multipliers, this is equivalent to minimizing

$$f(\delta \mathbf{w}) = \frac{1}{2}\delta \mathbf{w}^t \mathbf{H} \delta \mathbf{w} - \lambda \underbrace{(\delta \mathbf{w}^t \mathbf{u}_q + w_q)}_{0}.$$

This in turn implies

$$\frac{\partial f}{\partial \delta \mathbf{w}} = \mathbf{H}\delta \mathbf{w} - \lambda(\mathbf{u}_q) = 0.$$

Now since the derivative of matrix products obey

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^t \mathbf{A} \mathbf{x} = 2A\mathbf{x}$$

and

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^t \mathbf{b} = \mathbf{b},$$

we have

$$\delta \mathbf{w} = +\lambda \mathbf{H}^{-1} \mathbf{u}_q.$$

To solve for λ we compute

$$\frac{\delta f}{\delta \lambda} = 0,$$

and this implies

$$\delta \mathbf{w}^t \mathbf{u}_q + w_q = 0.$$

This implies

$$\lambda \mathbf{u}_q^t \mathbf{H}^{-1} \mathbf{u}_q = -\omega_q,$$

which in turn gives the value of the Lagrange undetermined multiplier,

$$\lambda = \frac{-w_q}{\mathbf{u}_q^t \mathbf{H}^{-1} \mathbf{u}_q} = \frac{-w_q}{[\mathbf{H}^{-1}]_{qq}}.$$

So, we have

$$\delta \mathbf{w} = +\lambda \mathbf{H}^{-1} \mathbf{u}_q = -\frac{w_q}{[\mathbf{H}^{-1}]_{qq}} \mathbf{H}^{-1} \mathbf{u}_q.$$

Moreover, we have the saliency of weight q is

$$\begin{aligned} L_q &= \frac{1}{2}[-\frac{w_q}{[\mathbf{H}^{-1}]_{qq}} \mathbf{H}^{-1} \mathbf{u}_q]^t \mathbf{H}[-\frac{w_q}{[\mathbf{H}^{-1}]_{qq}} \mathbf{H}^{-1} \mathbf{u}_q] \\ &= \frac{1}{2} \frac{w_q^2}{([\mathbf{H}^{-1}]_{qq})^2} \mathbf{u}_q^t \mathbf{H}^{-1} \mathbf{H} \mathbf{H}^{-1} \mathbf{u}_q \\ &= \frac{1}{2} \frac{w_q^2}{(\mathbf{H}^{-1})_{qq}^2} \mathbf{u}_q^t \mathbf{H}^{-1} \mathbf{u}_q = \frac{1}{2} \frac{w_{qq}}{([\mathbf{H}^{-1}]_{qq})^2} [\mathbf{H}^{-1}]_{qq} \\ &= \frac{1}{2} \frac{w_q^2}{[\mathbf{H}^{-1}]_{qq}}. \end{aligned}$$

(b) If $\mathbf{H}$ is diagonal, then

$$[\mathbf{H}^{-1}]_{qq} = \frac{1}{\mathbf{H}_{qq}}.$$

So, from part (a), the saliency of weight q for OBD is

$$L_q = \frac{1}{2} \frac{w_q^2}{[\mathbf{H}^{-1}]_{qq}} = \frac{1}{2} w_q^2 \mathbf{H}_{qq}.$$

38. The three-layer radial basis function network is characterized by Eq. 61 in the text:

$$z_k = f\left(\sum_{j=0}^{n_H} w_{kj} \varphi_j(\mathbf{x}) \right),$$

where we can let $y_j = \varphi(\mathbf{x})$. The quadratic error function we use is

$$J = \frac{1}{2} \sum_{k=1}^{c} (t_k - z_k)^2.$$

We begin by finding the contribution of the hidden-to-output weights to the error:

$$\frac{\partial J}{\partial w_{kj}} = \frac{\partial J}{\partial net_k} \frac{\partial net_k}{\partial w_{kj}},$$

where

$$\frac{\partial J}{\partial net_k} = \frac{\partial J}{\partial z_k}\frac{\partial z_k}{\partial net_k} = -(t_k - z_k)f'(net_k).$$

We put these together and find

$$\frac{\partial J}{\partial w_{kj}} = -(t_k - z_k)f'(net_k)\varphi_j(\mathbf{x}).$$

Now we turn to the update rule for $\boldsymbol{\mu}_j$ and b_j. Here we take

$$\varphi_j(\mathbf{x}) = e^{-b_j\|\mathbf{x}-\boldsymbol{\mu}_j\|^2}.$$

Note that $\boldsymbol{\mu}_j$ is a vector which can be written $\mu_{1j}, \mu_{2j}, \ldots, \mu_{dj}$. We now compute $\partial J/\partial\mu_{ij}$:

$$\frac{\partial J}{\partial \mu_{ij}} = \frac{\partial J}{\partial y_j}\frac{\partial y_j}{\partial \mu_{ij}}.$$

We look at the first term and find

$$\begin{aligned}\frac{\partial J}{\partial y_j} &= \sum_{k=1}^{c} -(t_k - z_k)\frac{\partial z_k}{\partial net_k}\frac{\partial net_k}{\partial y_j} \\ &= -\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}.\end{aligned}$$

The second term is

$$\begin{aligned}\frac{\partial y_j}{\partial \mu_{ji}} &= e^{-b\|\mathbf{x}-\boldsymbol{\mu}_j\|^2}(-1)(-1)2(x_i - \mu_{ij}) \\ &= 2\varphi(\mathbf{x}).\end{aligned}$$

We put all this together to find

$$\frac{\partial J}{\partial \mu_{ji}} = -2\varphi(\mathbf{x})(x_i - \mu_{ij})\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}.$$

Finally, we find $\partial J/\partial b_j$ such that the size of the spherical Gaussians be adaptive:

$$\frac{\partial J}{\partial b_j} = \frac{\partial J}{\partial y_j}\frac{\partial y_j}{\partial b_j}.$$

While the first term is given above, the second can be calculated directly:

$$\begin{aligned}\frac{\partial y_j}{\partial b_j} &= \frac{\partial \varphi(\mathbf{x})}{\partial b_j} = e^{-b\|\mathbf{x}-\boldsymbol{\mu}_j\|^2}\|\mathbf{x} - \boldsymbol{\mu}_j\|^2(-1) \\ &= -\varphi(\mathbf{x})\|\mathbf{x} - \boldsymbol{\mu}_j\|^2.\end{aligned}$$

Thus the final derivative is

$$\frac{\partial J}{\partial b_j} = -\varphi(\mathbf{x})\|\mathbf{x} - \boldsymbol{\mu}_j\|^2\sum_{k=1}^{c}(t_k - z_k)f'(net_k)w_{kj}.$$

Now it is a simple matter to write the learning rules:

$$\begin{aligned}\Delta w_{kj} &= -\eta\frac{\partial J}{\partial w_{kj}}\\ \Delta \mu_{ij} &= -\eta\frac{\partial J}{\partial \mu_{ij}}\\ \Delta b_j &= -\eta\frac{\partial J}{\partial b_j},\end{aligned}$$

with the derivatives as given above.

Section 6.11

39. Consider a general d-by-d matrix $\mathbf{K}$ and variable d-dimensional vector $\mathbf{x}$.

(a) We write in component form:

$$\begin{aligned}\mathbf{x}^t\mathbf{K}\mathbf{x} &= \sum_{j=1}^{d}\sum_{i=1}^{d} x_i K_{ij} x_j\\ &= \sum_{i=1}^{d} x_i^2 K_{ii} + \sum_{i\neq j}^{d} x_i x_j K_{ij}.\end{aligned}$$

The derivatives are, then,

$$\begin{aligned}\frac{d}{dx_r}\mathbf{x}^t\mathbf{K}\mathbf{x} &= \frac{d}{dx_r}\left[\sum_{i=1}^{d} x_i^2 K_{ii} + \sum_{i\neq j}^{d} x_i x_j K_{ij}\right]\\ &= \frac{d}{dx_r}\Big[\sum_{i\neq r} x_i^2 K_{ii} + x_r^2 K_{rr}\\ &\quad + \sum_{i}^{d}\sum_{j}^{d} x_i x_j K_{ij} + x_r\sum_{j\neq r} x_j K_{rj} + x_r\sum_{i\neq r} x_i K_{ir}\Big]\\ &= 0 + 2x_r K_{rr} + 0 + \sum_{j\neq r}^{d} x_j K_{rj} + \sum_{i\neq r}^{d} x_i K_{ir}\\ &= \sum_{j=1}^{d} x_j K_{rj} + \sum_{i=1}^{d} x_i K_{ir}\\ &= (\mathbf{K}\mathbf{x})_{r\text{th element}} + (\mathbf{K}^t\mathbf{x})_{t\text{th element}}\end{aligned}$$

This implies

$$\frac{d}{d\mathbf{x}}\mathbf{x}^t\mathbf{K}\mathbf{x} = \mathbf{K}\mathbf{x} + \mathbf{K}^t\mathbf{x} = (\mathbf{K}+\mathbf{K}^t)\mathbf{x}.$$

(b) So, when $\mathbf{K}$ is symmetric, we have

$$\frac{d}{d\mathbf{x}}\mathbf{x}^t\mathbf{K}\mathbf{x} = (\mathbf{K}+\mathbf{K}^t)\mathbf{x} = 2\mathbf{K}\mathbf{x}.$$

40. Consider weight decay and regularization.

(a) We note that the error can be written

$$E = E_{pat} + \gamma \sum_{ij} w_{ij}^2.$$

Then, gradient descent in the error function is

$$w_{ij}^{new} = w_{ij}^{old} - \eta \frac{\delta E}{\delta w_{ij}}$$

where η is the learning rate. We also have

$$\begin{aligned} \frac{\partial E}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}} \left[E_{pat} + \gamma \sum_{ij} w_{ij}^2 \right] \\ &= \gamma 2 w_{ij}. \end{aligned}$$

This implies

$$\begin{aligned} w_{ij}^{new} &= w_{ij}^{old} - 2\gamma\eta w_{ij}^{old} \\ &= w_{ij}^{old}(1 - 2\gamma\eta). \end{aligned}$$

(b) In this case we have $w_{ij}^{new} = w_{ij}^{old}(1 - \xi)$ where ξ is a weight decay constant. Thus

$$w_{ij}^{new} = w_{ij}^{old}(1 - 2\gamma\eta)$$

from part (a). This implies $2\gamma\eta = \xi$ or

$$\gamma = \frac{\xi}{2\eta}.$$

(c) Here we have the error function

$$\begin{aligned} E &= E_{pat} + \frac{\gamma w_{ij}^2}{1 + w_{ij}^2} \\ &= E_{pat} + \gamma \left[1 - \frac{1}{1 + w_{ij}^2} \right] \\ \frac{\partial E}{\partial w_{ij}} &= \frac{\partial E_{pat}}{\partial w_{ij}} + \gamma \frac{\partial}{\partial w_{ij}} \left[1 - \frac{1}{1 + w_{ij}^2} \right] = 0 + \frac{\gamma 2 w_{ij}}{(1 + w_{ij}^2)^2} \\ &= \frac{2\gamma w_{ij}}{(1 + w_{ij}^2)^2}. \end{aligned}$$

The gradient descent procedure gives

$$\begin{aligned} w_{ij}^{new} &= w_{ij}^{old} - \eta \frac{\partial E}{\partial w_{ij}} = w_{ij}^{old} - \eta \frac{2\gamma w_{ij}^{old}}{(1 + w_{ij}^2)^2} \\ &= w_{ij}^{old} \left[1 - \frac{2\gamma\eta}{(1 + w_{ij}^2)^2} \right] \\ &= w_{ij}^{old}(1 - \xi_{ij}) \end{aligned}$$

where

$$\xi_{ij} = \frac{2\gamma\eta}{(1+w_{ij}^2)^2}.$$

This implies

$$\gamma = \frac{\xi_{ij}(1+w_{ij}^2)^2}{2\eta}.$$

(d) Now consider a network with a wide range of magnitudes for weights. Then, the constant weight decay method $w_{ij}^{new} = w_{ij}^{old}(1-\xi)$ is equivalent to regularization via a penalty on $\sum_{ij}\omega_{ij}^2$. Thus, large magnitude weights are penalized and will be pruned to a greater degree than weights of small magnitude. On the other hand, the variable weight decay method $w_{ij}^{new} = w_{ij}^{old}(1-\xi_{ij})$ penalized on $\sum_{ij}\left[1-\frac{1}{1+w_{ij}^2}\right]$, which is less susceptible to large weights. Therefore the pruning will not be as severe on large magnitude weights as in the first case.

41. Suppose that the error E_{pat} is the negative log-likelihood up to a constant, that is,

$$E_{pat} = -\ln p(\mathbf{x}|\mathbf{w}) + \text{constant}.$$

Then we have

$$p(\mathbf{x}|\mathbf{w}) \propto e^{-E_{pat}}.$$

Consider the following prior distribution over weights

$$p(\omega) \propto e^{-\lambda\sum_{i,j} w_{ij}^2}$$

where $\lambda > 0$ is some parameter. Clearly this prior favors small weights, that is, $p(\mathbf{w})$ is large if $\sum_{ij} w_{ij}^2$ is small. The joint density of $\mathbf{x}, \mathbf{w}$ is

$$p(\mathbf{x},\mathbf{w}) = p(\mathbf{x}|\mathbf{w})p(\mathbf{w}) \propto \exp\left[-E_{pat} - \lambda\sum_{ij} w_{ij}^2\right].$$

The posterior density is

$$p(\mathbf{w}|\mathbf{x}) = \frac{p(\mathbf{x},\mathbf{w})}{p(\mathbf{x})} \propto \exp\left[-E_{pat} - \lambda\sum_{ij} w_{ij}^2\right].$$

So, maximizing the posterior density is equivalnet to minimizing

$$E = E_{pat} + \lambda\sum_{ij} w_{ij}^2,$$

which is the E of Eq. 41 in the text.

42. Given the definition of the criterion function

$$J_{ef} = J(\mathbf{w}) + \frac{\epsilon}{2\eta}\mathbf{w}^t\mathbf{w},$$

we can compute the derivative $\partial J_{ef}/\partial w_i$, and thus determine the learning rule. This derivative is

$$\frac{\partial J_{ef}}{\partial w_i} = \frac{\partial J}{\partial w_i} + \frac{\epsilon}{\eta}w_i$$

for any w_i. Thus the learning rule is

$$\begin{aligned} w_i(k+1) &= w_i(k) - \eta\frac{\partial J}{\partial w_i}\bigg|_{w_i(k)} - \epsilon w_i(k) \\ &= -\eta\frac{\partial J}{\partial w_i}\bigg|_{w_i(k)} + (1-\epsilon)w_i(k). \end{aligned}$$

This can be written in two steps as

$$\begin{aligned} w_i^{new}(k) &= w_i^{old}(k)(1-\epsilon) \\ w_i^{old}(k+1) &= -\eta\frac{\partial J}{\partial w_i}\bigg|_{w_i^{old}(k)} + w_i^{new}(k). \end{aligned}$$

Continuing with the next update we have

$$\begin{aligned} w_i^{new}(k+1) &= w_i^{old}(k+1)(1-\epsilon) \\ w_i^{old}(k+2) &= -\eta\frac{\partial J}{\partial w_i}\bigg|_{w_i^{old}(k+1)} + w_i^{new}(k+1). \end{aligned}$$

The above corresponds to the description of weight decay, that is, the weights are updated by the standard learning rule and then shrunk by a factor, as according to Eq. 38 in the text.

43. Equation 70 in the text computes the inverse of $(\mathbf{H} + \alpha\mathbf{I})$ when $\mathbf{H}$ is initialized to $\alpha\mathbf{I}$. When the value of α is chosen small, this serves as a good approximation

$$[\mathbf{H} + \alpha\mathbf{I}]^{-1} \simeq \mathbf{H}^{-1}.$$

We write the Taylor expansion of the error function around $\mathbf{w}^*$ up to second order. The approximation will be exact if the error function is quadratic:

$$J(\mathbf{w}) \simeq J(\mathbf{w}^*) + \left(\frac{\partial J}{\partial \mathbf{w}}\bigg|_{\mathbf{w}^*}\right)^t (\mathbf{w}-\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^t\mathbf{H}(\mathbf{w}-\mathbf{w}^*).$$

We plug in $\mathbf{H} + \alpha\mathbf{I}$ for $\mathbf{H}$ and get

$$J'(\mathbf{w}) \simeq \underbrace{J(\mathbf{w}^*) + \left(\frac{\partial J}{\partial \mathbf{w}}\bigg|_{\mathbf{w}^*}\right)^t (\mathbf{w}-\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^t\mathbf{H}(\mathbf{w}-\mathbf{w}^*)}_{J(\mathbf{w})} + \cdots + (\mathbf{w}-\mathbf{w}^*)^t\alpha\mathbf{I}(\mathbf{w}-\mathbf{w}^*).$$

Thus we have

$$J'(\mathbf{w}) = J(\mathbf{w}) + \alpha \mathbf{w}^t \mathbf{w} - \underbrace{\alpha \mathbf{w}^{*t} \mathbf{w}^*}_{constant}.$$

Thus our calculation of $\mathbf{H}$ modifies the error function to be

$$J'(\mathbf{w}) = J(\mathbf{w}) + \alpha \|\mathbf{w}\|^2 - constant.$$

This constant is irrelevant for minimization, and thus our error function is equivalent to

$$J''(\mathbf{w}) + \alpha \|\mathbf{w}\|^2,$$

which is equivalent to a criterion leading to weight decay.

44. The predicted functional increase in the error for a weight change $\delta\mathbf{w}$ is given by Eq. 68 in the text:

$$\delta J \simeq \frac{1}{2} \delta \mathbf{w}^t \mathbf{H} \delta \mathbf{w}.$$

We want to find which component of $\mathbf{w}$ to set to zero that will lead to the smallest increase in the training error. If we choose the qth component, then $\delta\mathbf{w}$ must satisfy $\mathbf{u}_q^t \delta \mathbf{w} + w_q = 0$, where $\mathbf{u}_q$ is the unit vector parallel to the qth axis. Since we want to minimize δJ with the constraint above, we write the Lagrangian

$$L(\mathbf{w}, \lambda) = \frac{1}{2} \delta \mathbf{w}^t \mathbf{H} \delta \mathbf{w} + \lambda \underbrace{(\mathbf{u}_q^t \delta \mathbf{w} + w_q)}_{0},$$

and then solve the system of equations

$$\frac{\partial L}{\delta \lambda} = 0, \qquad \frac{\partial L}{\partial \mathbf{w}} = 0.$$

These become

$$\begin{aligned}
\frac{\partial L}{\partial \lambda} &= \mathbf{u}_q^t \delta \mathbf{w} + w_q = 0 \\
\frac{\partial L}{\partial w_j} &= \sum_k H_{kj} \delta \mathbf{w}_k \qquad \text{for } j \neq i \\
\frac{\partial L}{\partial w_j} &= \sum_k H_{ij} \delta \mathbf{w}_k + \lambda \qquad \text{for } j = i.
\end{aligned}$$

These latter equations imply $\mathbf{H}\delta\mathbf{w} = -\lambda \mathbf{u}_q$. Thus $\delta \mathbf{w} = \mathbf{H}^{-1}(-\lambda)\mathbf{u}_q$. We substitute this result into the above and solve for the undetermined multiplier λ:

$$\mathbf{u}_q^t \mathbf{H}^{-1} \mathbf{u}_q + w_q = 0$$

which yields

$$\lambda = \frac{-w_q}{[\mathbf{H}^{-1}]_{qq}}.$$

Putting these results together gives us Eq. 69 in the text:

$$\delta\mathbf{w} = -\frac{w_q}{[\mathbf{H}^{-1}]_q q}\mathbf{H}^{-1}\mathbf{u}_q.$$

For the second part of Eq. 69, we just need to substitute the expression for $\delta\mathbf{w}$ we just found for the expression for δJ given in Eq. 68 in the text:

$$\begin{aligned} L_q &= \frac{1}{2}\left[-\frac{w_q}{[\mathbf{H}^{-1}]_{qq}}\mathbf{H}^{-1}\mathbf{u}_q\right]^t \mathbf{H}\left[-\frac{w_q}{[\mathbf{H}^{-1}]_{qq}}\mathbf{H}^{-1}\mathbf{u}_q\right] \\ &= \frac{1}{2}\left(\frac{w_q}{[\mathbf{H}^{-1}]_{qq}}\mathbf{u}_q^t \underbrace{\mathbf{H}\mathbf{H}^{-1}}_{\mathbf{I}} \frac{w_q}{[\mathbf{H}^{-1}]_{qq}}\mathbf{H}^{-1}\mathbf{u}_q\right) \\ &= \frac{1}{2}\frac{w_q^2}{[\mathbf{H}^{-1}]_{qq}}. \end{aligned}$$

45. Consider a simple 2-1 network with bias. First let $\mathbf{w} = (w_1, w_2)^t$ be the weights and $\mathbf{x} = (x_1, x_2)^t$ be the inputs, and w_0 the bias.

(a) Here the error is

$$E = \sum_{p=1}^{n} E_p = \sum_{p=1}^{n}[t_p(\mathbf{x}) - z_p(\mathbf{x})]^2$$

where p indexes the patterns, and t_p is the teaching or desired output and z_p the actual output. We can rewrite this error as

$$E = \sum_{\mathbf{x}\in\mathcal{D}_1}[f(w_1x_1 + w_2x_2 + w_0) - 1]^2 + \sum_{\mathbf{x}\in\mathcal{D}_2}[f(w_1x_1 + w_2x_2 + w_0) + 1]^2.$$

(b) To compute the Hessian matrix $\mathbf{H} = \frac{\partial^2 E}{\partial \mathbf{w}^2}$, we proceed as follows:

$$\begin{aligned} \frac{\partial E}{\partial w_i} &= \sum_{\mathbf{x}\in\omega_1} 2\Big[f(w_1x_1 + w_2x_2 + w_0) - 1]f'(w_1x_1 + w_2x_2 + w_0)x_i \\ &\quad + \sum_{\mathbf{x}\in\omega_2} 2[f(w_1x_1 + w_2x_2 + w_0) + 1]f'(w_1x_1 + w_2x_2 + w_0)x_i, i = 1, 2 \\ \frac{\partial E}{\partial w_j \partial w_i} &= 2\Bigg[\sum_{\mathbf{x}\in\omega_1}\{f'(w_1x_1 + w_2x_2 + w_0)\}^2 x_j x_i \\ &\quad + \sum_{\mathbf{x}\in\omega_1}[f(w_1x_1 + w_2x_2 + w_0) - 1]f''(w_1x_1 + w_2x_2 + w_0)x_j x_i \\ &\quad + \sum_{\mathbf{x}\in\omega_2}\{f'(w_1x_1 + w_2x_2 + w_0)\}^2 x_j x_i \\ &\quad + \sum_{\mathbf{x}\in\omega_2}[f(w_1x_1 + w_2x_2 + w_0) + 1]f''(w_1x_1 + w_2x_2 + w_0)x_j x_i\Bigg]. \end{aligned}$$

We let the net activation at the output unit be denoted

$$net = w_1x_1 + w_2x_2 + w_0 = (w_0\ w_1\ w_2)\begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix} = \tilde{\mathbf{w}}^t\tilde{\mathbf{x}} = \mathbf{w}^t\mathbf{x} + w_0.$$

Then we have

$$\begin{aligned}\frac{\partial^2 E}{\partial w_j \partial w_i} &= 2\sum_{\mathbf{x}\in\omega_1}[(f'(net))^2 + (f(net)1)f''(net)]x_j x_i \\ &\quad +2\sum_{\mathbf{x}\in\omega_2}[(f'(net))^2 + (f(net)+1)f''(net)]x_j x_i \\ &= \sum_{bfx} u(\mathbf{x})x_i x_j\end{aligned}$$

where

$$u(\mathbf{x}) = \begin{cases} (f'(net))^2 + (f(net)-1)f''(net) & \text{if } \mathbf{x}\in\omega_1 \\ (f'(net))^2 + (f(net)+1)f''(net) & \text{if } \mathbf{x}\in\omega_2 \end{cases}$$

So, we have

$$\mathbf{H} = \begin{pmatrix} \sum\limits_{\mathbf{x}} u(\mathbf{x})x_1^2 & \sum\limits_{\mathbf{x}} u(\mathbf{x})x_1x_2 \\ \sum\limits_{\mathbf{x}} u(\mathbf{x})x_1x_2 & \sum\limits_{\mathbf{x}} u(\mathbf{x})x_2^2 \end{pmatrix}.$$

(c) Consider two data sets $N(\mathbf{x}|\omega_i) \sim N(\mu_i, 1), i = 1, 2$. Then, $\mathbf{H}$ in terms of μ_1, μ_2 is $\mathbf{H} = ((H_{ij}))_{2\times 2}$ where,

$$\begin{aligned}H_{ij} &= 2[(f'(net_1)^2) + (f(net_1)-1)f''(net_1)]\mu_{1i}\mu_{1j} \\ &\quad +2[(f'(net_2)^2) + (f(net_2)+1)f''(net_2)]\mu_{2i}\mu_{2j}\end{aligned}$$

for $i, j = 1, 2$, where $net_1 = \mathbf{w}^t\mu_1 + w_0$ and $net_2 = \mathbf{w}^t\mu_2 + w_0$. Clearly, the Hessian matrix is symmetric, that is, $H_{12} = H_{21}$.

(d) We have the Hessian matrix

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{pmatrix}.$$

Then, the two eigenvalues of $\mathbf{H}$ are the roots of $|\mathbf{H} - \lambda\mathbf{I}| = 0$, that is,

$$\begin{vmatrix} h_{11}-\lambda & h_{12} \\ h_{12} & h_{22}-\lambda \end{vmatrix} = 0.$$

Now we have

$$(H_{11}-\lambda)(H_{22}-\lambda) - H_{12}^2 = 0$$

and

$$\lambda^2 - \lambda(H_{11}+H_{22}) + H_{11}H_{22} - H_{12}^2 = 0,$$

and this implies the eigenvalue is

$$\begin{aligned}\lambda &= \frac{H_{11}+H_{22} \pm \sqrt{(H_{11}+H_{22})^2 - 4(H_{11}H_{22}-H_{12}^2)}}{2} \\ &= \frac{H_{11}+H_{22} \pm \sqrt{(H_{11}-H_{22})^2 + 4H_{12}^2}}{2}.\end{aligned}$$

Thus the minimum and maximum eigenvalues are

$$\begin{aligned}\lambda_{min} &= \frac{H_{11}+H_{22}-\sqrt{(H_{11}-H_{22})^2+4H_{12}^2}}{2}\\ \lambda_{max} &= \frac{H_{11}+H_{22}+\sqrt{(H_{11}-H_{22})^2+4H_{12}^2}}{2}.\end{aligned}$$

(e) Suppose $\mu_1 = (1,0)$ and $\mu_2 = (0,1)$. Then from part (c) we have

$$\begin{aligned}H_{12} &= 0\\ H_{11} &= 2[(f'(w_1+w_0))^2 + (f(w_1+w_0)-1)f''(w_1+w_0)]\\ H_{22} &= 2[(f'(w_2+w_0))^2 + (f(w_2+w_0)+1)f''(w_2+w_0)].\end{aligned}$$

From part (d) we have

$$\begin{aligned}\lambda_{min} &= \frac{H_{11}+H_{22}-|H_{11}-H_{22}|}{2}\\ \lambda_{max} &= \frac{H_{11}+H_{22}+|H_{11}-H_{22}|}{2}\end{aligned}$$

and therefore in this case the minimum and maximum eigenvalues are

$$\begin{aligned}\lambda_{min} &= \min(H_{11},H_{22})\\ \lambda_{max} &= \max(H_{11},H_{22}),\end{aligned}$$

and thus their ratio is

$$\begin{aligned}\frac{\lambda_{max}}{\lambda_{min}} &= \frac{\max(H_{11},H_{22})}{\min(H_{11},H_{22})}\\ &= \frac{H_{11}}{H_{22}} \quad \text{or} \quad \frac{H_{22}}{H_{11}}.\end{aligned}$$

Computer Exercises

Section 6.2

1. Computer exercise not yet solved

Section 6.3

2. Computer exercise not yet solved
3. Computer exercise not yet solved
4. Computer exercise not yet solved
5. Computer exercise not yet solved
6. Computer exercise not yet solved
7. Computer exercise not yet solved

Section 6.4

8. Computer exercise not yet solved

Section 6.5

9. Computer exercise not yet solved

Section 6.6

10. Computer exercise not yet solved

Section 6.7

11. Computer exercise not yet solved

Section 6.8

12. Computer exercise not yet solved

Chapter 7

Stochastic methods

Problem Solutions

Section 7.1

1. First, the total number of typed characters in the play is approximately

$$m = 50\ pages \times 80\ \frac{lines}{page} \times 40\ \frac{characters}{line} = 160000\ characters.$$

We assume that each character has an equal chance of being typed. Thus, the probability of typing a specific charactor is $r = 1/30$.

We assume that the typing of each character is an independent event, and thus the probability of typing any particular string of length m is r^m. Therefore, a rough estimate of the length of the total string that must be typed before one copy of **Hamlet** appears is $\frac{1}{r^m} = 30^{160000} \simeq 2.52 \times 10^{236339}$. One year is

$$365.25\ days \times 24\frac{hours}{day} \times 60\frac{minutes}{hour} \times 60\frac{seconds}{minute} = 31557600\ seconds.$$

We are given that the monkey types two characters per second. Hence the expected time needed to type **Hamlet** under these circumstances is

$$2.52 \times 10^{236339} characters\ \times \frac{1}{2}\frac{second}{character} \times \frac{1}{31557600}\frac{year}{seconds} \simeq 4 \times 10^{236331} years,$$

which is *much* larger than 10^{10} years, the age of the universe.

Section 7.2

2. Assume we have an optimization problem with a non-symmetric connection matrix, $\mathbf{W}$, where $w_{ij} \neq w_{ji}$.

$$E \quad = \quad \frac{1}{2}(2E) = \frac{1}{2}\left(-\frac{1}{2}\sum_{i,j=1}^{N} w_{ij}s_is_j - \frac{1}{2}\sum_{i,j=1}^{N} w_{ji}s_is_j\right)$$

$$= -\frac{1}{2}\sum_{i,j=1}^{N}\frac{1}{2}(w_{ij}+w_{ji})s_i s_j$$

We define a new weight matrix

$$\hat{w}_{ij} = \frac{w_{ij}+w_{ji}}{2},$$

which has the property $\hat{w}_{ij} = \hat{w}_{ji}$. So the original optimization problem is equivalent to an optimization problem with a symmetric connection matrix, $\hat{\mathbf{W}}$.

3. Consider Fig. 7.2 in the text.

(a) In the discrete optimization problem, the variables only take the values ± 1, yielding isolated points in the space. Therefore, the energy is only defined on these points, and there is no continuous landscape as shown in Fig. 7.2 in the text.

(b) In the discrete space, all variables take ± 1 values. In other words, all of the feasible solutions are at the corners of a hypercube, and are of equal distance to the "middle" (the center) of the cube. Their distance from the center is $\sqrt{N}$.

(c) Along any "cut" in the feature space parallel to one of the axes the energy will be monotonic; in fact, it will be linear. This is because all other features are held constant, and the energy depends monotonically on a *single* feature. However, for cuts not parallel to a single axis (that is, when two or more features are varied), the energy need not be monotonic, as shown in Fig. 7.6.

4. First, we recall each variable can take two values, ± 1. The total number of configurations is therefore 2^N. An exhaustive search must calculate energy for all configurations. For $N = 100$, the time for exhaustive search is

$$2^{100}\ configurations \times \frac{10^{-8} second}{configuration} \simeq 1.27 \times 10^{22} seconds.$$

We can express this time as

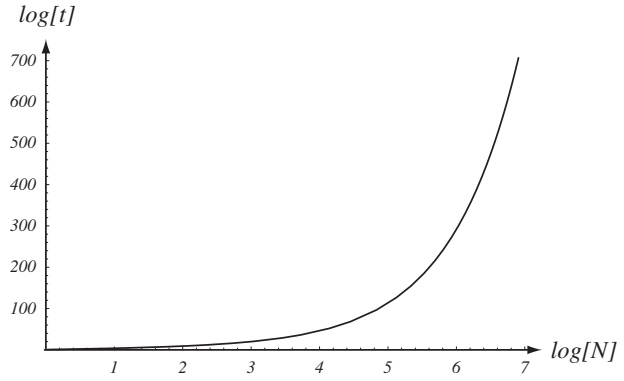
$$1.27 \times 10^{22} seconds \times \frac{1\ minute}{60\ seconds} \times \frac{1\ hour}{60\ minutes} \times \frac{1\ day}{24\ hours} \times \frac{1\ year}{365\ days} \simeq 4.0 \times 10^{14} years,$$

which is ten thousand times larger than 10^{10} *years*, the age of the universe. For $N = 1000$, a simliar argument gives the time as 3.4×10^{285} *years*.

5. We are to suppose that it takes 10^{-10} *seconds* to perform a single multiply-accumulate.

(a) Suppose the network is fully connected. Since the connection matrix is symmetric and $w_{ii} = 0$, we can compute E as $-\sum_{i>j} w_{ij}s_i s_j$ and thus the number of multiply-accumulate operations needed for calculating E for a single configuration is $N(N-1)/2$. We also know that the total number of configurations is 2^N for a network of N binary units. Therefore, the total time $t(N)$ for an exhaustive search on a network of size N (in seconds) can be expressed as

$$\begin{aligned} t(N) &= 10^{-10}\frac{second}{operation} \times \frac{N(N-1)}{2} \times \frac{operations}{configuration} \times 2^N configurations \\ &= 10^{-10}N(N-1)2^{N-1} seconds. \end{aligned}$$



(b) See figure.

(c) We need to find the largest possible N such that the total time is less than or equal to the specified time. Unfortunately, we cannot solve the inverse of $t(N)$ in a closed form. However, since $t(N)$ is monotonically increasing with respect to N, a simple search can be used to solve the problem; we merely increase N until the corresponding time exceeds the specified time. A straightforward calculation for a reasonable N would overflow on most computers if no special software is used. So instead we use the logarithm of the expression, specifically

$$t(N) = 10^{\log 10^{-10} N(N-1)2^{(N-1)}} = 10^{(N-1)\log 2 + \log N + \log(N-1) - 10}.$$

A day is 86400 *seconds*, within which an exhaustive search can solve a network of size 20 (because $5.69 \times 10^4 \simeq t(20) < 86400 < t(21) \simeq 6.22 \times 10^5$). A year is $31557600 \simeq 3.16 \times 10^7$ *seconds*, within which an exhaustive search can solve a network of size 22 (because $7.00 \times 10^6 \simeq t(22) < 3.16 \times 10^7 < t(23) \simeq 8.10 \times 10^7$). A century is $3155760000 \simeq 3.16 \times 10^9$ *seconds*, within which an exhaustive search can solve a network of size 24 (because $9.65 \times 10^8 \simeq t(24) < 3.16 \times 10^9 < t(25) \simeq 1.18 \times 10^{10}$).

6. Notice that for any configuration γ, E_γ does not change with T. It is sufficient to show that for any pair of configurations γ and γ', the ratio of the probabilities $P(\gamma)$ and $P(\gamma')$ goes to 1 as T goes to infinity:

$$\lim_{T\to\infty} \frac{P(\gamma)}{P(\gamma')} = \lim_{T\to\infty} \frac{e^{-E_\gamma/T}}{e^{-E_{\gamma'}/T}} = \lim_{T\to\infty} e^{(E_{\gamma'} - E_\gamma)/T}$$

Since the energy of a configuration does not change with T, E_γ and $E_{\gamma'}$ are constants, so we have:

$$\begin{aligned} \lim_{T\to\infty} \frac{(E_{\gamma'} - E_\gamma)}{T} &= 0 \\ \lim_{T\to\infty} e^{(E_{\gamma'} - E_\gamma)/T} &= 1, \end{aligned}$$

which means the probabilities of being in different configurations are the same if T is sufficiently high.

7. Let k_u^N denote the number pointing up in the subsystem of N magnets and k_d^N denote the number down.

(a) The number of configurations is given by the number of ways of selecting k_u^N magnets out of N and letting them point up, that is,

$$K(N, E_N) = \binom{N}{k_u^N}$$

We immediately have the following set of equations for k_u^N and k_d^N:

$$\begin{aligned} k_u^N + k_d^N &= N \\ k_u^N - k_d^N &= E_N \end{aligned}$$

The solution is given by $k_u^N = \frac{1}{2}(N + E_N)$ and $k_d^N = \frac{1}{2}(N - E_N)$. Therefore we have

$$K(N, E_N) = \binom{N}{k_u^N} = \binom{N}{\frac{1}{2}(N + E_N)}.$$

(b) We follow the approach in part (a), and find

$$K(N, E_M) = \binom{M}{k_u^M} = \binom{M}{\frac{1}{2}(M + E_M)}.$$

(c) Problem not yet solved

(d) Problem not yet solved

(e) Problem not yet solved

8. Consider a single magnet in an applied field.

(a) The two states $\gamma' = 0$ and $\gamma' = 1$ correspond to the two states of the binary magnet, $s = +1$ and $s = -1$, respectively. For $\gamma' = 0$ or $s = +1$, $E_{\gamma'} = E_0$; for $\gamma' = 1$ or $s = -1$, $E_{\gamma'} = -E_0$. Therefore, according to Eq. 3 in the text, the partition function is

$$Z(T) = e^{-E_0/T} + e^{-(-E_0)/T} = e^{-E_0/T} + e^{E_0/T}.$$

(b) According to Eq. 2 in the text, the probability of the magnet pointing up and pointing down are

$$P(s = +1) = \frac{e^{-E_0/T}}{Z(T)}, \qquad P(s = -1) = \frac{e^{E_0/T}}{Z(T)}$$

respectively. We use the partition function obtained in part (a) to calculate the expected value of the state of the magnet, that is,

$$\mathcal{E}[s] = P(s = +1) - P(s = -1) = \frac{e^{-E_0/T} - e^{E_0/T}}{e^{-E_0/T} + e^{E_0/T}} = \tanh(-E_0/T),$$

which is indeed in the form of Eq. 5 in the text.

(c) We want to calculate the expected value of the state, $\mathcal{E}[s]$. We assume the other $N-1$ magnets produce an average magnetic field. In another word, we consider the magnet in question, i, as if it is in an external magnetic field of the same strength as the average field. Therefore, we can calculate the probabilities in a similar way to that in part (b). First, we need to calculate the energy associated with the states of i. Consider the states of the other $N-1$ magnets are fixed. Let γ_i^+ and γ_i^- denote the configuration in which $s_i = +1$ and $s_i = -1$, respectively, while the other $N-1$ magnets take the fixed state values. Therefore, the energy for configuration γ_i^+ is

$$E_{\gamma_i^+} = -\frac{1}{2}\sum_{j,k=1}^{N} w_{jk}s_j s_k = -\frac{1}{2}\left[\sum_{k,j\neq i} w_{kj}s_j s_k + \sum_{k=1}^{N} w_{ki}s_k + \sum_{j=1}^{N} w_{ij}s_j\right]$$

since $s_i = +1$. Recall that the weight matrix is symmetric, that is, $w_{ki} = w_{ik}$, and thus we can write

$$\sum_{k=1}^{N} w_{ki}s_k = \sum_{k=1}^{N} w_{ik}s_k = \sum_{j=1}^{N} w_{ij}s_j,$$

where the last equality is by renaming dummy summation indices. Therefore the energy is

$$E_{\gamma_i^+} = -\sum_{j=1}^{N} w_{ij}s_j - \frac{1}{2}\sum_{j,k\neq i} w_{jk}s_j s_k = -l_i + C,$$

where, as given by Eq. 5 in the text, $l_i = \sum_j w_{ij}s_j$, and $C = -\frac{1}{2}\sum_{j,k\neq i} w_{jk}s_j s_k$ is a constant independent of i. Similarly, we can obtain

$$E_{\gamma_i^-} = \sum_{j=1}^{N} w_{ij}s_j - \frac{1}{2}\sum_{j,k\neq i} w_{jk}s_j s_k = l_i + C$$

Next, we can apply the results from parts (a) and (b). The partition function is thus

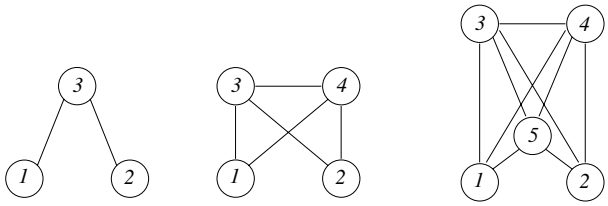
$$Z(T) = e^{-E_{\gamma_i^+}/T} + e^{-E_{\gamma_i^-}/T} = e^{(l_i - C)/T} + e^{(-l_i - C)/T} = (e^{l_i/T} + e^{-l_i/T})/E^{C/T}.$$

Therefore, according to Eq. 2 in the text, the expected value of the state of magnet i in the average field is

$$\begin{aligned}\mathcal{E}[s_i] &= Pr(s_i = +1)(+1) + Pr(s_i = -1)(-1) = \frac{e^{-E_{\gamma_i^+}/T} - e^{-E_{\gamma_i^-}/T}}{Z(T)} \\ &= \frac{e^{(l_i-C)/T} - e^{(-l_i-C)/T}}{(e^{l_i/T} + e^{-l_i/T})/E^{C/T}} = \frac{e^{l_i/T} - e^{-l_i/T}}{e^{l_i/T} + e^{-l_i/T}} = \tanh(l_i/T).\end{aligned}$$

9. The two input nodes are indexed as 1 and 2, as shown in the figure. Since the input units are supposed to have fixed state values, we can ignore the connection between the two input nodes. For this assignment, we use a version of Boltzmann networks with biases, in another word, there is a node with constant value +1 in the network.

Suppose the constant node is numbered as 0. The representation power of Boltzmann networks with biases is the same as those without biases. This can be shown as follows. First, each network without the constant unit can be converted to a network with the constant unit by including the constant node 0 and assigning the connection weights w_{0j} a zero value. Second, each network with the constant unit can also be converted to a network without the constant unit by replacing the constant unit with a pair of units, numbered -1 and -2. Node -1 assumes the same connection weights to other units as the constant unit 0, while node -2 is only connected to node -1. The connection weights between nodes -1 and -2 are a vary large positive number M. In this way, the nodes -1 and -2 are forced to have the same value in the configurations with the lowest energy, one of which requires the value be $+1$.



Next, we are going to determine the set of connection weights between pairs of units. In the follows, we use a vector of state values of all units to indicate the corresponding configuration.

(a) For the exclusive-OR problem, the input-output relationship is as follows:

Input 1	Input 2	Output (3)
$+1$	$+1$	-1
$+1$	-1	$+1$
-1	$+1$	$+1$
-1	-1	-1

Since the Boltzmann network always converges to a minimum energy configuration, it solves the exclusive-OR problem if and only if the following inequalities hold:

$$\begin{aligned} E_{\{+1,+1,-1\}} &< E_{\{+1,+1,+1\}} \\ E_{\{+1,-1,+1\}} &< E_{\{+1,-1,-1\}} \\ E_{\{-1,+1,+1\}} &< E_{\{-1,+1,-1\}} \\ E_{\{-1,-1,-1\}} &< E_{\{-1,-1,+1\}}. \end{aligned}$$

Note that the energy of the three-unit system can be written as

$$E_{\{s_1,s_2,s_3\}} = -w_{13}s_1s_3 - w_{23}s_2s_3 - w_{03}s_3$$

The above set of inequalities reduce to

$$\begin{aligned} +w_{13} + w_{23} + w_{03} &< -w_{13} - w_{23} - w_{03} \\ -w_{13} + w_{23} - w_{03} &< +w_{13} - w_{23} + w_{03} \\ +w_{13} - w_{23} - w_{03} &< -w_{13} + w_{23} + w_{03} \\ -w_{13} - w_{23} + w_{03} &< +w_{13} + w_{23} - w_{03} \end{aligned}$$

or equivalently,

$$\begin{aligned} w_{13} + w_{23} + w_{03} &< 0 && (1)\\ w_{23} &< w_{13} + w_{03} && (2)\\ w_{13} &< w_{23} + w_{03} && (3)\\ w_{03} &< w_{13} + w_{23} && (4) \end{aligned}$$

From (1) + (4), we have $w_{03} < 0$, and from (2) + (3), we have $0 < w_{03}$, which is a contradiction. Therefore, a network of the specified form cannot solve the exclusive-OR problem.

(b) For the two-output network, the input-output relationship is as follows:

Input 1	Input 2	Output 3	Output 4
+1	+1	−1	+1
+1	−1	+1	−1
−1	+1	+1	−1
−1	−1	−1	+1

This network solves the exclusive-OR problem if and only if

$$\begin{aligned} E_{\{+1,+1,-1,+1\}} &< E_{\{+1,+1,-1,-1\}}\\ E_{\{+1,+1,-1,+1\}} &< E_{\{+1,+1,+1,-1\}}\\ E_{\{+1,+1,-1,+1\}} &< E_{\{+1,+1,+1,+1\}}\\ E_{\{+1,-1,+1,-1\}} &< E_{\{+1,-1,-1,-1\}}\\ E_{\{+1,-1,+1,-1\}} &< E_{\{+1,-1,-1,+1\}}\\ E_{\{+1,-1,+1,-1\}} &< E_{\{+1,-1,+1,+1\}}\\ E_{\{-1,+1,+1,-1\}} &< E_{\{-1,+1,-1,-1\}}\\ E_{\{-1,+1,+1,-1\}} &< E_{\{-1,+1,-1,+1\}}\\ E_{\{-1,+1,+1,-1\}} &< E_{\{-1,+1,+1,+1\}}\\ E_{\{-1,-1,-1,+1\}} &< E_{\{-1,-1,-1,-1\}}\\ E_{\{-1,-1,-1,+1\}} &< E_{\{-1,-1,+1,-1\}}\\ E_{\{-1,-1,-1,+1\}} &< E_{\{-1,-1,+1,+1\}}. \end{aligned}$$

Recall the energy for the four-unit network:

$$E_{\{s_1,s_2,s_3,s_4\}} = -w_{13}s_1s_3 - w_{14}s_1s_4 - w_{23}s_2s_3 - w_{24}s_2s_4 - w_{34}s_3s_4 - w_{03}s_3 - w_{04}s_4.$$

The above set of inequalities reduce to:

$$\begin{aligned} +w_{13} - w_{14} + w_{23} - w_{24} + w_{34} + w_{03} - w_{04} &< +w_{13} + w_{14} + w_{23} + w_{24} - w_{34} + w_{03} + w_{04}\\ +w_{13} - w_{14} + w_{23} - w_{24} + w_{34} + w_{03} - w_{04} &< -w_{13} + w_{14} - w_{23} + w_{24} + w_{34} - w_{03} + w_{04}\\ +w_{13} - w_{14} + w_{23} - w_{24} + w_{34} + w_{03} - w_{04} &< -w_{13} - w_{14} - w_{23} - w_{24} - w_{34} - w_{03} - w_{04}\\ -w_{13} + w_{14} + w_{23} - w_{24} + w_{34} - w_{03} + w_{04} &< +w_{13} + w_{14} - w_{23} - w_{24} - w_{34} + w_{03} + w_{04}\\ -w_{13} + w_{14} + w_{23} - w_{24} + w_{34} - w_{03} + w_{04} &< +w_{13} - w_{14} - w_{23} + w_{24} + w_{34} + w_{03} - w_{04}\\ -w_{13} + w_{14} + w_{23} - w_{24} + w_{34} - w_{03} + w_{04} &< -w_{13} - w_{14} + w_{23} + w_{24} - w_{34} - w_{03} - w_{04}\\ +w_{13} - w_{14} - w_{23} + w_{24} + w_{34} - w_{03} + w_{04} &< -w_{13} - w_{14} + w_{23} + w_{24} - w_{34} + w_{03} + w_{04}\\ +w_{13} - w_{14} - w_{23} + w_{24} + w_{34} - w_{03} + w_{04} &< -w_{13} + w_{14} + w_{23} - w_{24} + w_{34} + w_{03} - w_{04} \end{aligned}$$

$$
\begin{aligned}
+w_{13} - w_{14} - w_{23} + w_{24} + w_{34} - w_{03} + w_{04} &< +w_{13} + w_{14} - w_{23} - w_{24} - w_{34} - w_{03} - w_{04} \\
-w_{13} + w_{14} - w_{23} + w_{24} + w_{34} + w_{03} - w_{04} &< -w_{13} - w_{14} - w_{23} - w_{24} - w_{34} + w_{03} + w_{04} \\
-w_{13} + w_{14} - w_{23} + w_{24} + w_{34} + w_{03} - w_{04} &< +w_{13} - w_{14} + w_{23} - w_{24} + w_{34} - w_{03} + w_{04} \\
-w_{13} + w_{14} - w_{23} + w_{24} + w_{34} + w_{03} - w_{04} &< +w_{13} + w_{14} + w_{23} + w_{24} - w_{34} - w_{03} - w_{04}
\end{aligned}
$$

or equivalently

$$
\begin{aligned}
w_{34} &< w_{14} + w_{24} + w_{04} && (1)\\
w_{13} + w_{23} + w_{03} &< w_{14} + w_{24} + w_{04} && (2)\\
w_{13} + w_{23} + w_{34} + w_{03} &< 0 && (3)\\
w_{23} + w_{34} &< w_{13} + w_{03} && (4)\\
w_{14} + w_{23} + w_{04} &< w_{13} + w_{24} + w_{03} && (5)\\
w_{14} + w_{34} + w_{04} &< w_{24} && (6)\\
w_{13} + w_{34} &< w_{23} + w_{03} && (7)\\
w_{13} + w_{24} + w_{04} &< w_{14} + w_{23} + w_{03} && (8)\\
w_{24} + w_{34} + w_{04} &< w_{14} && (9)\\
w_{14} + w_{24} + w_{34} &< w_{04} && (10)\\
w_{14} + w_{24} + w_{03} &< w_{13} + w_{23} + w_{04} && (11)\\
w_{34} + w_{03} &< w_{13} + w_{23}. && (12)
\end{aligned}
$$

From (2) + (11), we have the constraint $w_{03} < w_{04}$, and from (5) + (8), we have $w_{04} < w_{03}$. These two inequalities form a contradiction and therefore the network cannot solve the exclusive-OR problem.

(c) We consider the assignments as follows:

Input 1	Input 2	Output 3	Output 4	Hidden 5
$+1$	$+1$	-1	$+1$	$+1$
$+1$	-1	$+1$	-1	-1
-1	$+1$	$+1$	-1	-1
-1	-1	-1	$+1$	-1

For the network to solve the exclusive-OR problem, it must be the case that

$$
\begin{aligned}
E_{\{+1,+1,-1,+1,+1\}} &\leq E_{\{+1,+1,-1,+1,-1\}} \\
E_{\{+1,+1,-1,+1,+1\}} &< E_{\{+1,+1,-1,-1,\pm 1\}} \\
E_{\{+1,+1,-1,+1,+1\}} &< E_{\{+1,+1,+1,-1,\pm 1\}} \\
E_{\{+1,+1,-1,+1,+1\}} &< E_{\{+1,+1,+1,+1,\pm 1\}} \\
E_{\{+1,-1,+1,-1,-1\}} &\leq E_{\{+1,-1,+1,-1,+1\}} \\
E_{\{+1,-1,+1,-1,-1\}} &< E_{\{+1,-1,-1,-1,\pm 1\}} \\
E_{\{+1,-1,+1,-1,-1\}} &< E_{\{+1,-1,-1,+1,\pm 1\}} \\
E_{\{+1,-1,+1,-1,-1\}} &< E_{\{+1,-1,+1,+1,\pm 1\}} \\
E_{\{-1,+1,+1,-1,-1\}} &\leq E_{\{-1,+1,+1,-1,+1\}} \\
E_{\{-1,+1,+1,-1,-1\}} &< E_{\{-1,+1,-1,-1,\pm 1\}} \\
E_{\{-1,+1,+1,-1,-1\}} &< E_{\{-1,+1,-1,+1,\pm 1\}} \\
E_{\{-1,+1,+1,-1,-1\}} &< E_{\{-1,+1,+1,+1,\pm 1\}} \\
E_{\{-1,-1,-1,+1,-1\}} &\leq E_{\{-1,-1,-1,+1,+1\}} \\
E_{\{-1,-1,-1,+1,-1\}} &< E_{\{-1,-1,-1,-1,\pm 1\}}
\end{aligned}
$$

$$\begin{aligned} E_{\{-1,-1,-1,+1,-1\}} &< E_{\{-1,-1,+1,-1,\pm 1\}} \\ E_{\{-1,-1,-1,+1,-1\}} &< E_{\{-1,-1,+1,+1,\pm 1\}} \end{aligned}$$

The set of inequalities can be simplified as follows

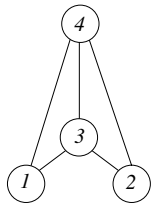
$$\begin{aligned}
w_{35} &\leq w_{05}+w_{15}+w_{25}+w_{45} && (1)\\
w_{34} &< w_{04}+w_{14}+w_{24}+w_{45} && (2)\\
w_{34}+w_{35} &< w_{04}+w_{05}+w_{14}+w_{15}+w_{24}+w_{25} && (3)\\
w_{03}+w_{13}+w_{23}+w_{35} &< w_{04}+w_{14}+w_{24}+w_{45} && (4)\\
w_{03}+w_{13}+w_{23} &< w_{04}+w_{05}+w_{14}+w_{15}+w_{24}+w_{25} && (5)\\
w_{03}+w_{13}+w_{23}+w_{34}+w_{35} &< 0 && (6)\\
w_{03}+w_{13}+w_{23}+w_{34} &< w_{05}+w_{15}+w_{25}+w_{45} && (7)\\
w_{05}+w_{15}+w_{35} &\leq w_{25}+w_{45} && (8)\\
w_{05}+w_{15}+w_{23}+w_{34} &< w_{03}+w_{13}+w_{25}+w_{45} && (9)\\
w_{23}+w_{34}+w_{35} &< w_{03}+w_{13} && (10)\\
w_{04}+w_{05}+w_{14}+w_{15}+w_{23} &< w_{03}+w_{13}+w_{24}+w_{25} && (11)\\
w_{04}+w_{14}+w_{23}+w_{35} &< w_{03}+w_{13}+w_{24}+w_{45} && (12)\\
w_{04}+w_{05}+w_{14}+w_{15}+w_{34}+w_{35} &< w_{24}+w_{25} && (13)\\
w_{04}+w_{14}+w_{34} &< w_{24}+w_{45} && (14)\\
w_{05}+w_{25}+w_{35} &\leq w_{15}+w_{45} && (15)\\
w_{05}+w_{13}+w_{25}+w_{34} &< w_{03}+w_{15}+w_{23}+w_{45} && (16)\\
w_{13}+w_{34}+w_{35} &< w_{03}+w_{23} && (17)\\
w_{04}+w_{05}+w_{13}+w_{24}+w_{25} &< w_{03}+w_{14}+w_{15}+w_{23} && (18)\\
w_{04}+w_{13}+w_{24}+w_{35} &< w_{03}+w_{14}+w_{23}+w_{45} && (19)\\
w_{04}+w_{05}+w_{24}+w_{25}+w_{34}+w_{35} &< w_{14}+w_{15} && (20)\\
w_{04}+w_{24}+w_{34} &< w_{14}+w_{45} && (21)\\
w_{05}+w_{45} &\leq w_{15}+w_{25}+w_{35} && (22)\\
w_{05}+w_{14}+w_{24}+w_{34} &< w_{04}+w_{15}+w_{25}+w_{35} && (23)\\
w_{14}+w_{24}+w_{34}+w_{45} &< w_{04} && (24)\\
w_{03}+w_{05}+w_{14}+w_{24} &< w_{04}+w_{13}+w_{15}+w_{23}+w_{25} && (25)\\
w_{03}+w_{14}+w_{24}+w_{45} &< w_{04}+w_{13}+w_{23}+w_{35} && (26)\\
w_{03}+w_{05}+w_{34}+w_{45} &< w_{13}+w_{15}+w_{23}+w_{25} && (27)\\
w_{03}+w_{34} &< w_{13}+w_{23}+w_{35} && (28)
\end{aligned}$$

Consider the set of weights $w_{05}=w_{35}=-1, w_{15}=w_{25}=1, w_{03}=-1/2, w_{13}=w_{23}=w_{45}=1/2, w_{14}=w_{24}=-1/4, w_{04}=1/4, w_{34}=0$. The above set of inequalities reduce to

-1	$\leq$	$3/2$	(1)	0	$<$	$1/4$	(2)	-1	$<$	$3/4$	(3)
$-1/2$	$<$	$1/4$	(4)	$1/2$	$<$	$3/4$	(5)	$-1/2$	$<$	0	(6)
$1/2$	$<$	$3/2$	(7)	-1	$\leq$	$3/2$	(8)	$1/2$	$<$	$3/2$	(9)
$-1/2$	$<$	0	(10)	$1/2$	$<$	$3/4$	(11)	$-1/2$	$<$	$1/4$	(12)
-1	$<$	$3/4$	(13)	0	$<$	$1/4$	(14)	-1	$\leq$	$3/2$	(15)
$1/2$	$<$	$3/2$	(16)	$-1/2$	$<$	0	(17)	$1/2$	$<$	$3/4$	(18)
$-1/2$	$<$	$1/4$	(19)	-2	$<$	$3/4$	(20)	0	$<$	$1/4$	(21)
$-1/2$	$\leq$	1	(22)	$-3/2$	$<$	$5/4$	(23)	0	$<$	$1/4$	(24)
-2	$<$	$13/4$	(25)	$-1/2$	$<$	$1/4$	(26)	-1	$<$	3	(27)
$-1/2$	$<$	0	(28)								

Since all these inequalities hold, the network can indeed implement the exclusive-OR problem.

10. As with the solution to Problem 9, we consider networks with biases (that is, a constant unit). The two input nodes are indexed as 1 and 2, the output is 4, and the hidden unit is 3, as shown in the figure.



We consider the assignments of all nodes as follows:

Input 1	Input 2	Hidden (3)	Output (4)
$+1$	$+1$	$+1$	-1
$+1$	-1	-1	$+1$
-1	$+1$	-1	$+1$
-1	-1	-1	-1

For the network to solve the exclusive-OR problem, it must be the case that

$$\begin{aligned}
E_{\{+1,+1,+1,-1\}} &\leq E_{\{+1,+1,-1,-1\}} \\
E_{\{+1,+1,+1,-1\}} &< E_{\{+1,+1,+1,+1\}} \\
E_{\{+1,+1,+1,-1\}} &< E_{\{+1,+1,-1,+1\}} \\
E_{\{+1,-1,-1,+1\}} &\leq E_{\{+1,-1,+1,+1\}} \\
E_{\{+1,-1,-1,+1\}} &< E_{\{+1,-1,+1,-1\}} \\
E_{\{+1,-1,-1,+1\}} &< E_{\{+1,-1,-1,-1\}} \\
E_{\{-1,+1,-1,+1\}} &\leq E_{\{-1,+1,+1,+1\}} \\
E_{\{-1,+1,-1,+1\}} &< E_{\{-1,+1,+1,-1\}} \\
E_{\{-1,+1,-1,+1\}} &< E_{\{-1,+1,-1,-1\}} \\
E_{\{-1,-1,-1,-1\}} &\leq E_{\{-1,-1,+1,-1\}} \\
E_{\{-1,-1,-1,-1\}} &< E_{\{-1,-1,+1,+1\}} \\
E_{\{-1,-1,-1,-1\}} &< E_{\{-1,-1,-1,+1\}}
\end{aligned}$$

Consider the set of weights $w_{03} = w_{34} = -1, w_{13} = w_{23} = 1, w_{04} = -1/2$, and $w_{14} = w_{24} = 1/2$. The energies involved in the above set of inequalities are

$$\begin{aligned}
E_{\{+1,+1,+1,-1\}} &= -\frac{3}{2} \\
E_{\{+1,+1,-1,-1\}} &= +\frac{5}{2} \\
E_{\{+1,+1,+1,+1\}} &= -\frac{1}{2} \\
E_{\{+1,+1,-1,+1\}} &= -\frac{1}{2} \\
E_{\{+1,-1,-1,+1\}} &= -\frac{3}{2}
\end{aligned}$$

$$
\begin{aligned}
E_{\{+1,-1,+1,+1\}} &= +\frac{5}{2} \\
E_{\{+1,-1,+1,-1\}} &= -\frac{1}{2} \\
E_{\{+1,-1,-1,-1\}} &= -\frac{1}{2} \\
E_{\{-1,+1,-1,+1\}} &= -\frac{3}{2} \\
E_{\{-1,+1,+1,+1\}} &= +\frac{5}{2} \\
E_{\{-1,+1,+1,-1\}} &= -\frac{1}{2} \\
E_{\{-1,+1,-1,-1\}} &= -\frac{1}{2} \\
E_{\{-1,-1,-1,-1\}} &= -\frac{7}{2} \\
E_{\{-1,-1,+1,-1\}} &= +\frac{1}{2} \\
E_{\{-1,-1,+1,+1\}} &= +\frac{11}{2} \\
E_{\{-1,-1,-1,+1\}} &= -\frac{5}{2}
\end{aligned}
$$

So, the set of inequalities hold, and the network can indeed implement the solution to the exclusive-OR problem.

Section 7.3

11. We use the definition of Kullback-Leibler divergence from Eq. 12 in the text as follows:

$$\overline{D}_{KL}(Q(\alpha^o|\alpha^i), P(\alpha^o|\alpha^i)) = \sum_{\alpha^i} Q(\alpha^i) \sum_{\alpha^o} Q(\alpha^o|\alpha^i) \log \frac{Q(\alpha^o|\alpha^i)}{P(\alpha^o|\alpha^i)}.$$

Recall that from Bayes' rule, we have

$$Q(\alpha^o|\alpha^i) = \frac{Q(\alpha^o, \alpha^i)}{Q(\alpha^i)} \qquad \text{and} \qquad P(\alpha^o|\alpha^i) = \frac{P(\alpha^o, \alpha^i)}{P(\alpha^i)}.$$

Thus the Kullback-Leibler divergence in this case can be written

$$
\begin{aligned}
& \overline{D}_{KL}(Q(\alpha^o|\alpha^i), P(\alpha^o|\alpha^i)) \\
= & \sum_{\alpha^i} Q(\alpha^i) \sum_{\alpha^o} Q(\alpha^o|\alpha^i) \log \frac{Q(\alpha^o|\alpha^i)}{P(\alpha^o|\alpha^i)} \\
= & \sum_{\alpha^i} Q(\alpha^i) \sum_{\alpha^o} Q(\alpha^o|\alpha^i) \log \frac{Q(\alpha^o, \alpha^i) P(\alpha^i)}{P(\alpha^o, \alpha^i) Q(\alpha^i)} \\
= & \sum_{\alpha^i} Q(\alpha^i) \sum_{\alpha^o} Q(\alpha^o|\alpha^i) \left(\log \frac{Q(\alpha^o, \alpha^i)}{P(\alpha^o, \alpha^i)} - \log \frac{Q(\alpha^i)}{P(\alpha^i)} \right) \\
= & \sum_{\alpha^i} \sum_{\alpha^o} Q(\alpha^i) Q(\alpha^o|\alpha^i) \log \frac{Q(\alpha^o, \alpha^i)}{P(\alpha^o, \alpha^i)} - \sum_{\alpha^i} Q(\alpha^i) \log \frac{Q(\alpha^i)}{P(\alpha^i)} \sum_{\alpha^o} Q(\alpha^o|\alpha^i)
\end{aligned}
$$

$$= \sum_{\alpha^i,\alpha^o} Q(\alpha^o,\alpha^i)\log\frac{Q(\alpha^o,\alpha^i)}{P(\alpha^o,\alpha^i)} - \sum_{\alpha^i} Q(\alpha^i)\log\frac{Q(\alpha^i)}{P(\alpha^i)}.$$

Using the definition of D_{KL} in Eq. 7 in the text, we have

$$\overline{D}_{KL}(Q(\alpha^o|\alpha^i),P(\alpha^o|\alpha^i)) = D_{KL}(Q(\alpha^o,\alpha^i),P(\alpha^o,\alpha^i)) - D_{KL}(Q(\alpha^i),P(\alpha^i))$$

According to Eq. 13 in the text, the learning rule is given by

$$\begin{aligned}\Delta w_{ij} &= -\eta\frac{\partial\overline{D}_{KL}(Q(\alpha^o|\alpha^i),P(\alpha^o|\alpha^i))}{\partial w_{ij}} \\ &= -\eta\left[\frac{\partial D_{KL}(Q(\alpha^o,\alpha^i),P(\alpha^o,\alpha^i))}{\partial w_{ij}} - \frac{\partial D_{KL}(Q(\alpha^i),P(\alpha^i))}{\partial w_{ij}}\right]\end{aligned}$$

Then we can use Eqs. 8, 9 and 10 in the text to get the learning rule as follows:

$$\begin{aligned}\Delta w_{ij} &= \frac{\eta}{T}\left[(\mathcal{E}_Q[s_is_j]_{\alpha^i\alpha^o\ clamped} - \mathcal{E}[s_is_j]_{free}) - (\mathcal{E}_Q[s_is_j]_{\alpha^i\ clamped} - \mathcal{E}[s_is_j]_{free})\right] \\ &= \frac{\eta}{T}\left[\mathcal{E}_Q[s_is_j]_{\alpha^i\alpha^o\ clamped} - \mathcal{E}_Q[s_is_j]_{\alpha^i\ clamped}\right]\end{aligned}$$

which is indeed in the form of Eq. 14 in the text.

12. Let the constant factor in Eq. 15 in the text be 1. We have that, if $i \neq j$,

$$w_{ij} = \frac{1}{K}\sum_{k=1}^{K} s_i(\mathbf{x}^k)s_j(\mathbf{x}^k).$$

Here, we have $K = 3$. The weight matrix can be calculated as

$$\mathbf{W} = \frac{1}{3}\begin{pmatrix} 0 & -1 & 3 & -1 & -1 & -3 \\ -1 & 0 & -1 & -1 & -1 & 1 \\ 3 & -1 & 0 & -1 & -1 & -3 \\ -1 & -1 & -1 & 0 & -1 & 1 \\ -1 & -1 & -1 & -1 & 0 & 1 \\ -3 & 1 & -3 & 1 & 1 & 0 \end{pmatrix}.$$

(a) First, we determine the energy change due to the perturbation. Let $\mathbf{x}_l^k$ denote the pattern obtained by perturbing unit l in the pattern $\mathbf{x}^k$. According to Eq. 1, the energy change is

$$\begin{aligned}\Delta E(\mathbf{x}^k,l) &= E_{\mathbf{x}_l^k} - E_{\mathbf{x}^k} \\ &= -\frac{1}{2}\sum_{ij} w_{ij}s_i(\mathbf{x}_l^k)s_j(\mathbf{x}_l^k) + \frac{1}{2}\sum_{ij} w_{ij}s_i(\mathbf{x}^k)s_j(\mathbf{x}^k) \\ &= -\frac{1}{2}\sum_{ij} w_{ij}(s_i(\mathbf{x}_l^k)s_j(\mathbf{x}_l^k) - s_i(\mathbf{x}^k)s_j(\mathbf{x}^k))\end{aligned}$$

Since each unit only takes 2 possible values ± 1, we have $s_l(\mathbf{x}_l^k) = -s_l(\mathbf{x}^k)$. We also have $s_i(\mathbf{x}_l^k) = s_i(\mathbf{x}^k)$ if $i \neq l$ since we only perturb one unit at a time. So, we can cancel most of the terms in the above summation.

Also noting $w_{ij} = w_{ji}$, we have

$$\begin{aligned}\Delta E(\mathbf{x}^k, l) &= -\sum_{i\neq l} w_{il}(s_i(\mathbf{x}_l^k)s_l(\mathbf{x}_l^k) - s_i(\mathbf{x}^k)s_l(\mathbf{x}^k)) \\ &= -\sum_{i\neq l} w_{il}s_i(\mathbf{x}^k)(-s_l(\mathbf{x}^k) - s_l(\mathbf{x}^k)) \\ &= 2s_l(\mathbf{x}^k)\sum_{i\neq l} w_{il}s_i(\mathbf{x}^k)\end{aligned}$$

Therefore, the change in energy obeys

$\Delta E(\mathbf{x}^k, l)$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$
$\mathbf{x}^1$	$\frac{14}{3}$	$-\frac{2}{3}$	$\frac{14}{3}$	2	2	$\frac{14}{3}$
$\mathbf{x}^2$	$\frac{14}{3}$	2	$\frac{14}{3}$	2	$-\frac{2}{3}$	$\frac{14}{3}$
$\mathbf{x}^3$	$\frac{14}{3}$	2	$\frac{14}{3}$	$-\frac{2}{3}$	2	$\frac{14}{3}$

Therefore, none of the three patterns give local minima in energy.

(b) Let $\hat{\mathbf{x}}$ denote the pattern in which $s_i(\hat{\mathbf{x}}) = -s_i(\mathbf{x})$. We have

$$E_{\hat{\mathbf{x}}} = -\frac{1}{2}\sum_{ij} w_{ij}s_i(\hat{\mathbf{x}})s_j(\hat{\mathbf{x}}) = -\frac{1}{2}\sum_{ij} w_{ij}(-s_i(\mathbf{x}))(-s_j(\mathbf{x})) = -\frac{1}{2}\sum_{ij} w_{ij}s_i(\mathbf{x})s_j(\mathbf{x}) = E_{\mathbf{x}}$$

Therefore, if $\mathbf{x}$ gives a local minimum in energy, $\hat{\mathbf{x}}$ also gives a local minimum in energy.

13. As in Problem 12, the symmetric connection matrix is

$$\mathbf{W} = \frac{1}{3}\begin{pmatrix} 0 & -1 & 3 & 1 & -1 & -3 & 1 & -1 \\ -1 & 0 & -1 & -3 & -1 & 1 & -3 & 3 \\ 3 & -1 & 0 & 1 & -1 & -3 & 1 & -1 \\ 1 & -3 & 1 & 0 & 1 & -1 & 3 & -3 \\ -1 & -1 & -1 & 1 & 0 & 1 & 1 & -1 \\ -3 & 1 & -3 & -1 & 1 & 0 & -1 & 1 \\ 1 & -3 & 1 & 3 & 1 & -1 & 0 & -3 \\ -1 & 3 & -1 & -3 & -1 & 1 & -3 & 0 \end{pmatrix}.$$

We thus need merely to check whether the patterns give local minima in energy. We have

$\Delta E(\mathbf{x}^k, l)$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$	$l=7$	$l=8$
$\mathbf{x}^1$	2	$\frac{14}{3}$	2	$\frac{14}{3}$	$\frac{14}{3}$	2	$\frac{14}{3}$	$\frac{14}{3}$
$\mathbf{x}^2$	6	$\frac{26}{3}$	6	$\frac{26}{3}$	$\frac{2}{3}$	6	$\frac{26}{3}$	$\frac{26}{3}$
$\mathbf{x}^3$	$\frac{26}{3}$	$\frac{22}{3}$	$\frac{26}{3}$	$\frac{22}{3}$	$-\frac{2}{3}$	$\frac{26}{3}$	$\frac{22}{3}$	$\frac{22}{3}$

Therefore, $\mathbf{x}^1$ and $\mathbf{x}^2$ give local minima in energy, but $\mathbf{x}^3$ does not.

14. For simplicity, we consider there is only one pattern with a single missing feature. Suppose the deficient input pattern is α^d, the missing feature is represented by s_k, and a pattern of the hidden units is β. If after annealing with clamped input pattern α^d, the values of s_k and β are s_k^* and β^*, respectively. We need to show that $P(s_k = s_k^*|\alpha^d) > P(s_k = -s_k^*|\alpha^d)$. We consider the ratio of the two probabilities

$$\frac{P(s_k = s_k^*|\alpha^d)}{P(s_k = -s_k^*|\alpha^d)} = \frac{\sum\limits_\beta P(s_k = s_k^*, \beta|\alpha^d)}{\sum\limits_\beta P(s_k = -s_k^*, \beta|\alpha^d)} = \frac{\sum\limits_\beta P(\alpha^d, s_k = s_k^*, \beta|\alpha^d)}{\sum\limits_\beta P(\alpha^d, s_k = -s_k^*, \beta|\alpha^d)}.$$

Since the pattern $\{\alpha^d, s_k = s_k^*, \beta^*\}$ gives the minimal energy due to annealing, supposing there is only one global minimal corresponding to clamped α^d, we have for any β, that

$$E_{\{\alpha^d, s_k = s_k^*, \beta^*\}} < E_{\{\alpha^d, s_k = s_k^*, \beta \neq \beta^*\}} \quad \text{and} \quad E_{\{\alpha^d, s_k = s_k^*, \beta^*\}} < E_{\{\alpha^d, s_k = -s_k^*, \beta\}}.$$

Since for any configuration γ including α^d as a subpattern, it holds that

$$P(\gamma|\alpha^d) = \frac{e^{-E_\gamma/T}}{\sum_{s_k,\beta} e^{-E_{\{\alpha^d, s_k, \beta\}}/T}}.$$

Therefore, the following three limits hold:

$$\begin{aligned} \lim_{T\to 0} P(\alpha^d, s_k = s_k^*, \beta^*|\alpha^d) &= 1 \\ \lim_{T\to 0} P(\alpha^d, s_k = s_k^*, \beta \neq \beta^*|\alpha^d) &= 0 \\ \lim_{T\to 0} P(\alpha^d, s_k = -s_k^*, \beta|\alpha^d) &= 0. \end{aligned}$$

So, for any $0 < \epsilon < 1/2$, there exists a temperature T_0, such that if $T < T_0$,

$$P(\alpha^d, s_k = s_k^*, \beta^*|\alpha^d) > 1 - \epsilon.$$

Therefore, under the same conditions we have the two inequalities:

$$\begin{aligned} \sum_{\beta \neq \beta^*} P(\alpha^d, s_k = s_k^*, \beta|\alpha^d) &< \epsilon \\ \sum_\beta P(\alpha^d, s_k = -s_k^*, \beta|\alpha^d) &< \epsilon. \end{aligned}$$

Therefore the ratio of the probability of $s_k = s_k^*$ to that for $s_k = -s_k^*$, all given the deficient information α^d, is

$$\begin{aligned} \frac{P(s_k = s_k^*|\alpha^d)}{P(s_k = -s_k^*|\alpha^d)} &= \frac{\sum\limits_\beta P(\alpha^d, s_k = s_k^*, \beta|\alpha^d)}{\sum\limits_\beta P(\alpha^d, s_k = -s_k^*, \beta|\alpha^d)} \\ &= \frac{P(\alpha^d, s_k = s_k^*, \beta^*|\alpha^d) + \sum\limits_{\beta \neq \beta^*} P(\alpha^d, s_k = s_k^*, \beta|\alpha^d)}{\sum\limits_\beta P(\alpha^d, s_k = -s_k^*, \beta|\alpha^d)} \\ &> \frac{1-\epsilon}{\epsilon} > 1. \end{aligned}$$

So s_k^* is the most probable value given α^d.

15. Assume the correct category is represented by s_k, the input pattern is α^i, and the known incorrect catagory is represented by s_j, $s_j = -1, j \neq k$. We consider the ratio between probabilities $P(s_k = +1|\alpha^i, s_j = -1)$ and $P(s_k = +1|\alpha^i)$:

$$\frac{P(s_k = +1|\alpha^i, s_j = -1)}{P(s_k = +1|\alpha^i)} = \frac{P(s_k = +1, \alpha^i, s_j = -1)}{P(s_k = +1, \alpha^i)} \frac{P(\alpha^i)}{P(\alpha^i, s_j = -1)}$$

For a successfully trained network for classification, $s_k = +1$ implies $s_j = -1$. Therefore, $P(s_k = +1, \alpha^i, s_j = -1) = P(s_k = +1, \alpha^i)$, and

$$\frac{P(s_k = +1|\alpha^i, s_j = -1)}{P(s_k = +1|\alpha^i)} = \frac{P(\alpha^i)}{P(\alpha^i, s_j = -1)} = \frac{1}{P(s_j = -1|\alpha^i)} \geq 1,$$

and hence the claim is verified.

16. Problem not yet solved

17. Here we have the variance in response to the force is

$$\text{Var}[l_i] = \text{Var}\left[\sum_{j=1}^{N} w_{ij} s_j\right] = \sum_{j=1}^{N} \text{Var}[w_{ij} s_j] = N\text{Var}[w_{ij} s_j].$$

Since $\text{Var}[x] = \mathcal{E}[x^2] - (\mathcal{E}[x])^2$, we can write

$$\text{Var}[w_{ij} s_j] = \mathcal{E}[w_{ij}^2 s_j^2] - (\mathcal{E}[w_{ij} s_j])^2 = \mathcal{E}[w_{ij}^2]\mathcal{E}[s_j^2] - (\mathcal{E}[w_{ij}]\mathcal{E}[s_j])^2.$$

Since $P(s_j = +1) = P(s_j = -1) = 0.5$, the expected value of s_j is $\mathcal{E}[s_j] = 0$, and $s_j^2 = 1$. Therefore, we have

$$\text{Var}[w_{ij} s_j] = \mathcal{E}[w_{ij}^2],$$

and thus

$$\text{Var}[l_i] = N\mathcal{E}[w_{ij}^2].$$

We seek to have the variance of the response to be 1, that is $\text{Var}[l_i] = 1$. which implies $\mathcal{E}[w_{ij}^2] = 1/N$. If we randomly initialize w_{ij} according to a uniform distribution from $[-h, h]$, we have

$$\mathcal{E}[w_{ij}^2] = \int_{-h}^{h} \frac{1}{2h} t^2 dt = \frac{1}{2h} \left.\frac{t^3}{3}\right|_{-h}^{h} = \frac{h^2}{3}.$$

Under these conditions, then,

$$\frac{1}{N} = \frac{h^2}{3}, \qquad h = \sqrt{3/N}$$

So, the weights are initialized uniformly in the range $-\sqrt{3/N} < w_{ij} < +\sqrt{3/N}$.

18. We consider the problem of setting the learning rate for a Boltzmann network having N units.

(a) In the solution to Problem 11, we showed that the Kullback-Leibler divergence obeyed

$$\overline{D}_{KL}(Q(\alpha^o|\alpha^i), P(\alpha^o|\alpha^i)) = D_{KL}(Q(\alpha^i, \alpha^o), P(\alpha^i, \alpha^o)) - D_{KL}(Q(\alpha^i), P(\alpha^i)).$$

As such, we need merely derive the result for D_{KL}. From Eq. 7 in the text, we obtain

$$\begin{aligned} D_{KL}(Q(\alpha), P(\alpha)) &= \sum_{\alpha} Q(\alpha) \log \frac{Q(\alpha)}{P(\alpha)} \\ &= \sum_{\alpha} Q(\alpha) \log Q(\alpha) - \sum_{\alpha} Q(\alpha) \log P(\alpha). \end{aligned}$$

Therefore, we have

$$\frac{\partial D_{KL}(Q(\alpha), P(\alpha))}{\partial w_{uv}} = -\sum_{\alpha} \frac{Q(\alpha)}{P(\alpha)} \frac{\partial P(\alpha)}{\partial w_{uv}},$$

which yields the Hessian matrix as

$$\begin{aligned} \mathbf{H} &= \frac{\partial^2 D_{KL}(Q(\alpha), P(\alpha))}{\partial w_{ij} \partial w_{uv}} \\ &= \sum_{\alpha} \frac{Q(\alpha)}{P(\alpha)^2} \frac{\partial P(\alpha)}{\partial w_{ij}} \frac{\partial P(\alpha)}{\partial w_{uv}} - \sum_{\alpha} \frac{Q(\alpha)}{P(\alpha)} \frac{\partial^2 P(\alpha)}{\partial w_{ij} \partial w_{uv}}. \end{aligned}$$

If the network has sufficient parameters, as the number of training examples goes to infinity, the weights converge to the optimal one, $\mathbf{w}^*$, where $P(\alpha) = Q(\alpha)$. At this limit, the second term in the above expression can be omitted, because

$$\begin{aligned} \sum_{\alpha} \frac{Q(\alpha)}{P(\alpha)} \frac{\partial^2 P(\alpha)}{\partial w_{ij} \partial w_{uv}} &= \sum_{\alpha} \frac{\partial^2 P(\alpha)}{\partial w_{ij} \partial w_{uv}} \\ &= \frac{\partial^2}{\partial w_{ij} \partial w_{uv}} \left(\sum_{\alpha} P(\alpha) \right) \\ &= \frac{\partial^2 1}{\partial w_{ij} \partial w_{uv}} = 0. \end{aligned}$$

Therefore, the Hessian is simply

$$\mathbf{H} \simeq \sum_{\alpha} \frac{Q(\alpha)}{P(\alpha)^2} \frac{\partial P(\alpha)}{\partial w_{ij}} \frac{\partial P(\alpha)}{\partial w_{uv}}$$

From Eq. 9, we have

$$\frac{1}{P(\alpha)} \frac{\partial P(\alpha)}{\partial w_{ij}} = \frac{1}{T} \left[\sum_{\beta} s_i(\alpha\beta) s_j(\alpha\beta) P(\beta|\alpha) - \mathcal{E}[s_i s_j] \right],$$

and thus we can write the Hessian as

$$\begin{aligned} \mathbf{H} &\simeq \sum_{\alpha} \frac{Q(\alpha)}{P(\alpha)^2} \frac{\partial P(\alpha)}{\partial w_{ij}} \\ &= \frac{1}{T^2} \sum_{\alpha} Q(\alpha) \left[\sum_{\beta} s_i(\alpha\beta) s_j(\alpha\beta) P(\beta|\alpha) - \mathcal{E}[s_i s_j] \right] \left[\sum_{\beta'} s_u(\alpha\beta') s_v(\alpha\beta') P(\beta'|\alpha) - \mathcal{E}[s_u s_v] \right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{T^2}\Bigg[\sum_{\alpha} Q(\alpha) \sum_{\beta} s_i(\alpha\beta) s_j(\alpha\beta) P(\beta|\alpha) \sum_{\beta'} s_u(\alpha\beta') s_v(\alpha\beta') P(\beta'|\alpha) \\
&\quad -\mathcal{E}_Q[s_i s_j]_\alpha \mathcal{E}[s_u s_v] - \mathcal{E}[s_i s_j]\mathcal{E}_Q[s_u s_v]_\alpha + \mathcal{E}[s_i s_j]\mathcal{E}[s_u s_v]\Bigg] \\
&= \frac{1}{T^2}\Bigg[\sum_{\alpha} Q(\alpha) \sum_{\beta,\beta'} s_i(\alpha\beta) s_j(\alpha\beta) s_u(\alpha\beta') s_v(\alpha\beta') P(\beta,\beta'|\alpha) \\
&\quad -\mathcal{E}_Q[s_i s_j]_\alpha \mathcal{E}[s_u s_v] - \mathcal{E}[s_i s_j]\mathcal{E}_Q[s_u s_v]_\alpha + \mathcal{E}[s_i s_j]\mathcal{E}[s_u s_v]\Bigg] \\
&= \frac{1}{T^2}\big[\mathcal{E}_Q[s_i s_j s_u s_v]_\alpha - \mathcal{E}_Q[s_i s_j]_\alpha \mathcal{E}[s_u s_v] - \mathcal{E}[s_i s_j]\mathcal{E}_Q[s_u s_v]_\alpha + \mathcal{E}[s_i s_j]\mathcal{E}[s_u s_v]\big],
\end{aligned}
$$

where $\mathcal{E}_Q[\cdot]_\alpha$ indicates the expected value with clamped α averaged according to Q. We can further simplify the above expression by noticing that when the weights converge to $\mathbf{w}^*$, $P(\alpha) = Q(\alpha)$. Thus, the Hessian is

$$
\begin{aligned}
\mathbf{H} &\simeq \frac{1}{T^2}\big[\mathcal{E}_Q[s_i s_j s_u s_v]_\alpha - \mathcal{E}_Q[s_i s_j]_\alpha \mathcal{E}[s_u s_v] - \mathcal{E}[s_i s_j]\mathcal{E}_Q[s_u s_v]_\alpha + \mathcal{E}[s_i s_j]\mathcal{E}[s_u s_v]\big] \\
&\simeq \frac{1}{T^2}\big[\mathcal{E}[s_i s_j s_u s_v] - \mathcal{E}[s_i s_j]\mathcal{E}[s_u s_v] - \mathcal{E}[s_i s_j]\mathcal{E}[s_u s_v] + \mathcal{E}[s_i s_j]\mathcal{E}[s_u s_v]\big] \\
&\simeq \frac{1}{T^2}\big[\mathcal{E}[s_i s_j s_u s_v] - \mathcal{E}[s_i s_j]\mathcal{E}[s_u s_v]\big].
\end{aligned}
$$

Recall that under general conditions

$$\mathcal{E}[xy] - \mathcal{E}[x]\mathcal{E}[y] = \mathcal{E}[(x - \mathcal{E}[x])(y - \mathcal{E}[y])].$$

In this case, then, the Hessian matrix is

$$\mathbf{H} = \frac{1}{T^2}\boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the random vector $\mathbf{s}^\star = \{s_i s_j\}$.

(b) We calculate the curvature as

$$
\begin{aligned}
\mathbf{w}^t\mathbf{H}\mathbf{w} &= \frac{1}{T^2}\mathbf{w}^t\boldsymbol{\Sigma}\mathbf{w} = \frac{1}{T^2}\mathbf{w}^t\mathcal{E}[(\mathbf{s}^\star - \mathcal{E}[\mathbf{s}^\star])(\mathbf{s}^\star - \mathcal{E}[\mathbf{s}^\star])^t]\mathbf{w} \\
&= \frac{1}{T^2}\mathcal{E}[\mathbf{w}^t(\mathbf{s}^\star - \mathcal{E}[\mathbf{s}^\star])(\mathbf{s}^\star - \mathcal{E}[\mathbf{s}^\star])^t\mathbf{w}] \\
&= \frac{1}{T^2}\mathcal{E}[(\mathbf{w}\cdot(\mathbf{s}^\star - \mathcal{E}[\mathbf{s}^\star]))^2] = \frac{1}{T^2}\mathcal{E}[(\mathbf{w}\cdot\mathbf{s}^\star - \mathbf{w}\cdot\mathcal{E}[\mathbf{s}^\star])^2] \\
&= \frac{1}{T^2}\mathcal{E}[(\mathbf{w}\cdot\mathbf{s}^\star)^2 - 2(\mathbf{w}\cdot\mathbf{s}^\star)(\mathbf{w}\cdot\mathcal{E}[\mathbf{s}^\star]) + (\mathbf{w}\cdot\mathcal{E}[\mathbf{s}^\star])^2] \\
&= \frac{1}{T^2}\left[\mathcal{E}[(\mathbf{w}\cdot\mathbf{s}^\star)^2] - 2\mathcal{E}[(\mathbf{w}\cdot\mathbf{s}^\star)(\mathbf{w}\cdot\mathcal{E}[\mathbf{s}^\star])] + \mathcal{E}[(\mathbf{w}\cdot\mathcal{E}[\mathbf{s}^\star])^2]\right] \\
&= \frac{1}{T^2}\left[\mathcal{E}[(\mathbf{w}\cdot\mathbf{s}^\star)^2] - 2(\mathbf{w}\cdot\mathcal{E}[\mathbf{s}^\star])^2 + (\mathbf{w}\cdot\mathcal{E}[\mathbf{s}^\star])^2\right] \\
&= \frac{1}{T^2}\left[\mathcal{E}[(\mathbf{w}\cdot\mathbf{s}^\star)^2] - (\mathbf{w}\cdot\mathcal{E}[\mathbf{s}^\star])^2\right]
\end{aligned}
$$

$$\begin{aligned} &\leq \frac{1}{T^2}\mathcal{E}[(\mathbf{w}\cdot\mathbf{s}^\star)^2] = \frac{1}{T^2}\mathcal{E}\left[\left(\sum_{ij} w_{ij}s_i s_j\right)^2\right] \\ &\leq \frac{1}{T^2}\mathcal{E}\left[\left(\sum_{ij}|w_{ij}|\right)^2\right] = \frac{1}{T^2}\left(\sum_{ij}|w_{ij}|\right)^2 \end{aligned}$$

(c) According to the inequality

$$\left(\frac{a_1+a_2+\cdots+a_n}{n}\right)^2 \leq \frac{a_1^2+a_2^2+\cdots+a_n^2}{n},$$

we can write

$$\begin{aligned} \left(\frac{\sum_{ij}|w_{ij}|}{N(N-1)}\right)^2 &\leq \frac{\sum_{ij}|w_{ij}|^2}{N(N-1)} = \frac{1}{N(N-1)} \\ \left(\sum_{ij}|w_{ij}|\right)^2 &\leq N(N-1). \end{aligned}$$

Therefore, the curvature is bounded by

$$\mathbf{w}^t\mathbf{H}\mathbf{w} \leq \frac{N(N-1)}{T^2}.$$

(d) The optimal learning rate, η, is inverse to the curvature, and hence

$$\eta = \frac{1}{\mathbf{w}^t\mathbf{H}\mathbf{w}} \geq \frac{T^2}{N(N-1)}.$$

Section 7.4

19. It is sufficient to show that A_{ij} and B_{ij} do not have a property which can be translated into the constraint that the probabilities sum to 1. For the transition probalities a_{ij} and b_{jk} in HMM, we have

$$\sum_j a_{ij} = 1, \qquad \sum_k b_{jk} = 1$$

According to Eq. 23, we have

$$a_{ij} = e^{A_{ij}/T}, \qquad b_{jk} = e^{B_{jk}/T}$$

Therefore, if there exists a HMM equivalent for any Boltzmann chain, A_{ij} and B_{jk} must satisfy the constraints

$$\sum_j e^{A_{ij}/T} = 1, \qquad \sum_k e^{B_{jk}/T} = 1,$$

which are not required for a general Boltzmann chain.

20. Since at each time step, there are exactly one hidden unit and one visible unit which are on, the number of configurations for each time step is ck, and total number of legal paths is $(ck)^{T_f}$.

21. With a known initial ($t = 1$) hidden unit, the Boltzmann chain is annealed with all visible units and the known initial hidden units clamped to $+1$, and other initial hidden units clamped to 0. If the initial hidden unit is unknown, we can hypothetically add an "always-on" hidden unit at $t = 0$, and connect it to all hidden units at $t = 1$, and the connection weights are C_i. Since in a Boltzmann chain exactly one hidden unit and one visible unit can be on, it follows that exactly one initial ($t = 1$) hidden unit will be on, which is consistent to the semantics of Hidden Markov Models. The generalized energy then is

$$E_{\boldsymbol{\omega}\mathbf{V}} = E[\boldsymbol{\omega}^{T_f}, \mathbf{V}^{T_f}] = -C_i - \sum_{t=1}^{T_f - 1} A_{ij} - \sum_{t=1}^{T_f} B_{jk},$$

where C_i indicates the connection weight between the constant unit and the initial hidden unit which is on.

22. In a Boltzmann zipper, the hidden units are clustered into "cells." For example, in the one shown in Fig. 7.12 in the text, the "fast" hidden units at the first two time steps for the "fast" chain and the "slow" hidden units at the first time step for the "slow" chain are interconnected through **E**, forming a "cell" consisting of three groups of hidden units. Since we require that exactly one hidden unit is on in each chain. In a "cell," three units are on after annealing, one in each group, and they are all interconnected. Therefore, the connection matrix **E** cannot be simply related to transition probabilities for two reasons. First, since each chain represents behavior at different time scales, there is no time ordering between the on-unit in the "slow" chain and the two on-units in the "fast" chain in the same "cell" (otherwise the "slow" chain can be "squeezed" into the "fast" chain and forming a single chain). However, it is necessary to have explicit time ordering for a state transition interpretation to make sense. Second, even though we can exert a time ordering for the on-units, the first on-unit in the ordering is connected to two other on-units, which cannot have a state transition interpretation where only a single successor state is allowed.

As an extreme case, an all-zero **E** means the two chains are not correlated. This is completely meaningful, and **E** is not normalized. The weights in **E** can also be negative to indicate a negative or suppressive correlation.

Section 7.5

23. We address the problem of a population of size L of N-bit chromosomes.

(a) The total number of different N-bit chromosomes is 2^N. So the number of different populations of size L is the number of different ways of selecting L chromosomes out of 2^N possible choices, allowing repetition. Since the number of combinations of selection n items out of m types ($n < m$) with repetition is $\binom{m+n-1}{n}$, we have the number of different populations of size L is

$$\binom{2^N + L - 1}{L} = \binom{L + 2^N - 1}{2^N - 1}.$$

Next, we show that the number of combinations of selection n items out of m types ($n < m$) with repetition is $\binom{m+n-1}{n}$. Denote the number as $A(m, n)$.

Suppose there are k distinctive types in the result, $1 \leq k \leq n$. The number of ways of choosing the types is $\binom{m}{k}$, and the number of ways of choosing n items of exactly k types is $\binom{m}{k}B(n,k)$, where $B(n,k)$ is the number of ways of parition n into k positive integers. Thus the total number of ways of choosing n items out of m types ($n < m$) is the sum of the above number over k

$$\sum_{k=1}^{n} \binom{m}{k} B(n,k).$$

Here $B(n,k)$ can be calculated recursively as follows:

$$\begin{aligned} B(n,1) &= 1 \\ B(n,n) &= 1 \\ B(n,k) &= \sum_{i=1}^{n-k+1} B(n-i,k-1). \end{aligned}$$

The first two equations are obvious. The third one is due to the fact that the last element can take values ranging from 1 to $n-(k-1)$ and thus

$$\begin{aligned} B(n,k) &= \sum_{i=1}^{n-k+1} B(n-i,k-1) \\ &= B(n-1,k-1) + \sum_{i=2}^{n-k+1} B(n-i,k-1) \\ &= B(n-1,k-1) + \sum_{j=1}^{n-k} B(n-1-j,k-1) \\ &= B(n-1,k-1) + B(n-1,k). \end{aligned}$$

Recall the following relations:

$$\begin{aligned} \binom{n}{k} &= \binom{n-1}{k-1} + \binom{n-1}{k} \\ \binom{n}{0} &= \binom{n}{n} = 1. \end{aligned}$$

Using these relations we can now prove $B(n,k) = \binom{n-1}{k-1}$ by induction. If $k = 1$ or $k = n$, the result holds trivially. Suppose $B(n-1,k-1) = \binom{n-2}{k-2}$ and $B(n-1,k) = \binom{n-2}{k-1}$. Then we have

$$B(n,k) = B(n-1,k-1) + B(n-1,k) = \binom{n-2}{k-2} + \binom{n-2}{k-1} = \binom{n-1}{k-1}.$$

Therefore we have

$$\begin{aligned} \sum_{k=1}^{n} \binom{m}{k} B(n,k) &= \sum_{k=1}^{n} \binom{m}{k}\binom{n-1}{k-1} \\ &= \sum_{k=1}^{n} \binom{m}{k}\binom{n-1}{n-k} = \binom{m+n-1}{n}, \end{aligned}$$

as desired, and the proof is complete.

(b) Assume each chromosome has a distinctive fitness score. Since chromosomes with the highest L_s fitness scores are selected as parents, the number of all possible sets of parents is equivalent to the number of ways of selection L_s chromosomes from all possible chromosomes excluding those with the lowest $L - L_s$ fitness scores. Therefore, the number is

$$\binom{L_s + (2^N - (L - L_s)) - 1}{(2^N - (L - L_s)) - 1} = \binom{2^N - L + 2L_s - 1}{2^N - L + L_s - 1}.$$

(c) If $L_s = L$, the above expression is simply

$$\binom{2^N + L - 1}{2^N - 1},$$

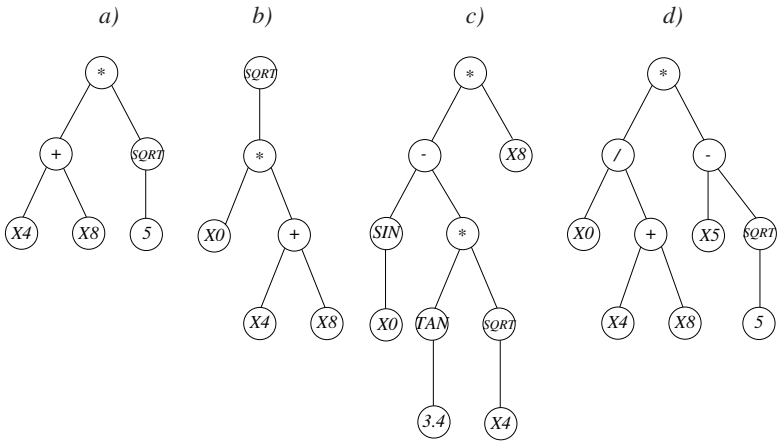
the same as the result in part (a).

(d) If $L_s = 1$, the expression is

$$\binom{2^N - L + 1}{2^N - L} = \binom{2^N - L + 1}{1} = 2^N - L + 1.$$

Section 7.6

24. See figure.



(a) `(* (+ X4 X8) (SQRT 5))`

(b) `(SQRT (* X0 (+ X4 X8)))`

(c) `(* (- (SIN X0) (* (TAN 3.4) (SQRT X4))) X8)`.

(d) `(* (/ X0 (+ X4 X8)) (- X5 (SQRT 5)))`.

(e) Constants and variables {`X3, X0, 5.5, X5, -4.5, 2.7`}, binary operators {`+, NOR, *, /, -, OR`}, and unary operators {`+, SQRT, SIN, -, NOT, TAN`}. Note `+` and `-` can act as both binary and unary operators.

Computer Exercises

Section 7.2

1. Computer exercise not yet solved
2. Computer exercise not yet solved
3. Computer exercise not yet solved

Section 7.3

4. Computer exercise not yet solved
5. Computer exercise not yet solved
6. Computer exercise not yet solved

Section 7.4

7. Computer exercise not yet solved

Section 7.5

8. Computer exercise not yet solved

Section 7.6

9. Computer exercise not yet solved

Chapter 8

Nonmetric methods

Problem Solutions

Section 8.2

1. Assume a particular query occurs more than once along some path through a given tree. The branch taken through the uppermost (closest to the root) occurence of this query must also be taken in all subsequent occurences along this path. As a result, branch(es) not taken in the uppermost instance of the query may be pruned from lower occurences of that query along the path without altering the classification of any pattern. After such pruning, the lower occurences of the query are left with only a single child node. At this point, these queries serve no purpose and may be deleted by directly connecting their parent node to their child.

By repeatedly applying the above algorithm to every possible path through a tree and to every query which is duplicated along a path, we can convert any given tree to an equivalent tree in which each path consists of distinct queries.

Section 8.3

2. Consider a non-binary tree, where the number of branches at nodes can vary throughout the tree.

(a) Suppose our tree has root node R, with (local) branching ratio B. Each of the B children is itself a root of a subtree having arbitrary depth. If we can prove that we can replace the root node with binary nodes, then by induction we can do it for its children — the roots of the subtrees. In this way, an arbitrary tree can be expressed as a binary tree.

We represent a tree with a root node R and B children nodes as $\{R, (a_1, a_2, \ldots a_B)\}$.

$B = 2$ This is a binary tree.

$B = 3$ The tree $\{R, (a_1, a_2, a_3)\}$ can be represented as a binary tree $\{R, (a_1, R_2)\}$ where the subtree is $\{R_2, (a_2, a_3)\}$.

$B = k \geq 3$ We can keep replacing each root of the subtree as follows: the first $\{R, (a_1, a_2, \ldots, a_k)\}$ becomes $\{R, (a_1, R_2)\}$, with subtree $\{R_2, (a_2, R_3)\}$, ..., $\{R_{k-1}, (a_{k-1}, a_k)\}$. By induction, then, any tree can be replaced by an equivalent binary tree.

While the above shows that any node with $B \geq 2$ can be replaced by a node with binary decisions, we can apply this to all nodes in the expanded tree, and thereby make the entire tree binary.

(b) The number of levels depends on the number of classes. If the number of classes is 2, then the functionally equivalent tree is of course 2 levels in depth. If the number of categories is c, we can create a tree with c levels though this is not needed. Instead we can split the root node sending $c/2$ categories to the left, and $c/2$ categories to the right. Likewise, each of these can send $c/4$ to the left and $c/4$ to the right. Thus a *tight* upper bound is $\lceil \log c \rceil$.

(c) The lower bound is 3 and the upper bound is $2B - 1$.

3. PROBLEM NOT YET SOLVED

4. PROBLEM NOT YET SOLVED

5. We use the entropy impurity given in Eq. 1,

$$i(N) = -\sum_i P(\omega_i) \log_2 \left(P(\omega_i)\right) = H(\omega_i).$$

(a) After splitting on a binary feature $F \in \{R, L\}$, the weighted impurity at the two child nodes is

$$\begin{aligned} P(L)i(L) + P(R)i(R) &= -P(L)\sum_i P(\omega_i|L) \log_2 \left(P(\omega_i|L)\right) \\ &\quad -P(R)\sum_i P(\omega_i|R) \log_2 \left(P(\omega_i|R)\right) \\ &= -\sum_i P(\omega_i, L) \log_2 \left(\frac{P(\omega_i, L)}{P(L)}\right) \\ &\quad -\sum_i P(\omega_i, R) \log_2 \left(\frac{P(\omega_i, R)}{P(R)}\right) \\ &= -\sum_{i,F} P(\omega_i, F) \log_2 \left(P(\omega_i, F)\right) \\ &\quad +P(L) \log_2 \left(P(L)\right) + P(R) \log_2 \left(P(R)\right) \\ &= H(\omega, F) - H(F). \end{aligned}$$

Therefore, the drop in impurity is

$$\Delta i(N) = H(\omega) - H(\omega, F) + H(F).$$

But $H(\omega) \leq H(\omega, F) \leq H(\omega) + H(F)$, and therefore we have

$$0 \leq \Delta i(N) \leq H(F) \leq 1 \text{ bit}.$$

(b) At each node, the weighted impurity at the child nodes will be less than that at the parent, even though individual descendandant may have a greater impurity

than their parent. For example at the ($x_2 < 0.61$) node in the upper tree of example 1, the impurity is 0.65, while that at the left child is 0.0, and that at the right is 1.0. The left branch is taken $\frac{2}{3}$ of the time, however, so the weighted impurity at the children is $\frac{2}{3} \times 0 + \frac{1}{3} \times 1 = 0.33$. Similarly, at the ($x_1 < 0.6$) node of the lower tree, the right child has higher impurity (0.92) than the parent (0.76), but the weighted average at the children is $\frac{2}{3} \times 0 + \frac{1}{3} \times 0.92 = 0.304$. In each case, the reduction in impurity is between 0 and 1 bit, as required.

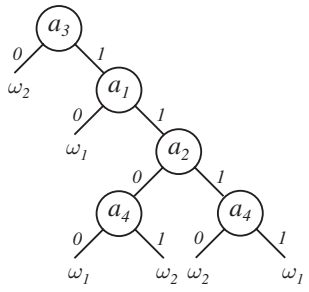
(c) For B-way branches, we have $0 \leq \Delta i(N) \leq \log_2(B)$ bits.

6. PROBLEM NOT YET SOLVED
7. PROBLEM NOT YET SOLVED
8. There are four attributes, $\{a_1, a_2, a_3, a_4\}$ to be used in our decision tree.

(a) To select the query at the root node, we investigate queries on each of the four attributes. The following shows the number of patterns sent to the "left" and "right" for each value, and the entropy at the resulting children nodes:

query	sent left	left entropy	sent right	right entropy
a_1	$2\omega_1, 2\omega_2$	1	$2\omega_1, 2\omega_2$	1
a_2	$2\omega_1, 2\omega_2$	1	$2\omega_1, 2\omega_2$	1
a_3	$0\omega_1, 2\omega_2$	0	$4\omega_1, 2\omega_2$	0.9183
a_4	$2\omega_1, 3\omega_2$	0.9710	$2\omega_1, 1\omega_2$	0.9183

Because query a_3 leads to the greatest weighted reduction in impurity, a_3 should be the query at the root node. We continue and grow the tree shown in the figure.



(b) We can expand the tree into rules as

$$\begin{aligned}
\omega_1 &= (a_3\ AND\ NOTa_1)\ OR\ (a_3\ AND\ a_1\ AND\ NOT\ a_2\ AND\ NOT\ a_4) \\
&\quad OR\ (a_3\ AND\ a_1\ AND\ a_2\ AND\ a_4) \\
&= a_3\ AND\ (NOT\ a_1\ OR\ a_1\ AND\ (NOT\ a_2\ AND\ NOT\ a_4)\ OR\ (a_2\ AND\ a_4)). \\
\omega_2 &= NOT\ a_3\ OR\ (a_3AND\ a_1\ AND\ NOT\ a_2\ AND\ a_4) \\
&\quad OR\ (a_3\ AND\ a_1\ AND\ a_2\ AND\ NOT\ a_4) \\
&= NOT\ a_3OR(a_3\ AND\ a_1)\ AND\ ((NOT\ a_2\ AND\ a_4)\ OR\ (a_2\ AND\ NOT\ a_4)).
\end{aligned}$$

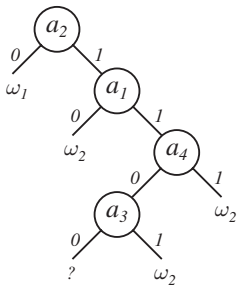
9. PROBLEM NOT YET SOLVED
10. PROBLEM NOT YET SOLVED
11. PROBLEM NOT YET SOLVED

12. PROBLEM NOT YET SOLVED
13. PROBLEM NOT YET SOLVED
14. PROBLEM NOT YET SOLVED
15. PROBLEM NOT YET SOLVED
16. The four attributes are denoted $\{a_1, a_2, a_3, a_4\}$.

(a) To select the query at the root node, we investigate queries on each of the four attributes. The following shows the number of patterns sent to the "left" and "right" for each value, and the entropy at the resulting children nodes:

query	sent left	left entropy	sent right	right entropy
a_1	$2\omega_1, 1\omega_2$	0.9183	$2\omega_1, 3\omega_2$	0.9710
a_2	$3\omega_1, 0\omega_2$	0	$1\omega_1, 4\omega_2$	0.7219
a_3	$2\omega_1, 1\omega_2$	0.9183	$2\omega_1, 3\omega_2$	0.9710
a_4	$3\omega_1, 2\omega_2$	0.9710	$1\omega_1, 2\omega_2$	0.9183

Because query a_2 leads to the greatest weighted reduction in impurity, a_2 should be the query at the root node. We continue and grow the tree shown in the figure, where the "?" denotes the leaf having equal number of ω_1 and ω_2 patterns.



(b) To select the query at the root node, we investigate queries on each of the four attributes. The following shows the number of patterns sent to the "left" and "right" for each value, and the entropy at the resulting children nodes:

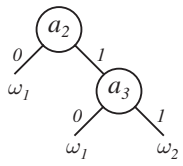
query	sent left	left entropy	sent right	right entropy
a_1	$4\omega_1, 1\omega_2$	0.7219	$4\omega_1, 3\omega_2$	0.9852
a_2	$6\omega_1, 0\omega_2$	0	$2\omega_1, 4\omega_2$	0.9183
a_3	$4\omega_1, 1\omega_2$	0.7219	$4\omega_1, 3\omega_2$	0.9852
a_4	$6\omega_1, 2\omega_2$	0.8113	$2\omega_1, 2\omega_2$	1

Because query a_2 leads to the greatest weighted reduction in impurity, a_2 should be the query at the root node. We continue and grow the tree shown in the figure.

Section 8.4

17. PROBLEM NOT YET SOLVED

Section 8.5



18. PROBLEM NOT YET SOLVED

19. Here our strings are composed of letters in the alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$.

(a) Consider the following string (and shift positions)

“	a	c	a	c	c	a	c	b	a	c”
	1	2	3	4	5	6	7	8	9	10

The last-occurence function gives $\mathcal{F}(\mathtt{a}) = 9$, $\mathcal{F}(\mathtt{b}) = 8$, $\mathcal{F}(\mathtt{c}) = 10$, and 0 otherwise. Likewise, the good-suffix function gives $\mathcal{G}(\mathtt{c}) = 7$, $\mathcal{G}(\mathtt{ac}) = 6$, $\mathcal{G}(\mathtt{bac}) = 0$, and 0 otherwise.

(b) Consider the following string (and shift positions)

“	a	b	a	b	a	b	c	b	c	b	a	a	a	b	c	b	a	a”
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

The last-occurence function gives $\mathcal{F}(\mathtt{a}) = 18$, $\mathcal{F}(\mathtt{b}) = 16$, $\mathcal{F}(\mathtt{c}) = 15$ and 0 otherwise. Likewise, the good-suffix function gives $\mathcal{G}(\mathtt{a}) = 17$, $\mathcal{G}(\mathtt{aa}) = 12$, $\mathcal{G}(\mathtt{baa}) = 10$, $\mathcal{G}(\mathtt{cbaa}) = 9$, $\mathcal{G}(\mathtt{bcbaa}) = 8$, $\mathcal{G}(\mathtt{abcbaa}) = 0$, and 0 otherwise.

(c) Consider the following string (and shift positions)

“	c	c	c	a	a	a	b	a	b	a	c	c	c”
	1	2	3	4	5	6	7	8	9	10	11	12	13

The last-occurence function gives $\mathcal{F}(\mathtt{a}) = 10$, $\mathcal{F}(\mathtt{b}) = 9$, $\mathcal{F}(\mathtt{c}) = 13$, and 0 otherwise. Likewise, $\mathcal{G}(\mathtt{c}) = 12$, $\mathcal{G}(\mathtt{cc}) = 11$, $\mathcal{G}(\mathtt{ccc}) = 1$, $\mathcal{G}(\mathtt{accc}) = 0$, and 0 otherwise.

(d) Consider the following string (and shift positions)

“	a	b	b	a	b	b	a	b	b	c	b	b	a	b	b	c	b	b	a”
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

The last-occurence function gives $\mathcal{F}(\mathtt{a}) = 19$, $\mathcal{F}(\mathtt{b}) = 18$, $\mathcal{F}(\mathtt{c}) = 16$, and 0 otherwise. Likewise, $\mathcal{G}(\mathtt{a}) = 13$, $\mathcal{G}(\mathtt{ba}) = 12$, $\mathcal{G}(\mathtt{bba}) = 11$, $\mathcal{G}(\mathtt{cbba}) = 10$, $\mathcal{G}(\mathtt{bcbba}) = 9$, $\mathcal{G}(\mathtt{bbcbba}) = 8$, $\mathcal{G}(\mathtt{abbcbba}) = 7$, $\mathcal{G}(\mathtt{babbcbba}) = 6$, $\mathcal{G}(\mathtt{bbabbcbba}) = 5$, and 0 otherwise.

20. We use the information from Fig. 8.8 in the text.

(a) The string and the number of comparisons at each shift are:

“	p	r	o	b	a	b	i	l	i	t	i	e	s	_	f	o	r	_	e	s	t	i	m	a	t	e	s”
	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	9

The sum is the total number of character comparisons: 28.

(b) The test string and index positions are

```
"e  s  t  i  m  a  t  e  s"
 1  2  3  4  5  6  7  8  9
```

Here the last-occurence function gives $\mathcal{F}(\texttt{s}) = 9$, $\mathcal{F}(\texttt{e}) = 8$, $\mathcal{F}(\texttt{t}) = 7$, $\mathcal{F}(\texttt{a}) = 6$, $\mathcal{F}(\texttt{m}) = 5$, $\mathcal{F}(\texttt{i}) = 4$, and 0 otherwise. Likewise, the good-suffix function gives $\mathcal{G}(\texttt{s}) = 2$, $\mathcal{G}(\texttt{es}) = 1$, and 0 otherwise.

(c) The four relevant shifts are shown below, the number of letter comparison is shown in bold at the right.

```
"p r o b a b i l i t i e s _ f o r _ e s t i m a t e s"
 e s t i m a t e s                                      1
           e s t i m a t e s                            1
                             e s t i m a t e s           1
                                   e s t i m a t e s     9
                                                        12
```

The first shift is $9 - \mathcal{F}(\texttt{i}) = 5$. The second shift is $9 - \mathcal{F}(\texttt{\_}) = 9$, and the last shift is $9 - \mathcal{F}(\texttt{m}) = 4$. The total number of character comparisons is 12, as shown at the lower right.

21. Here the test string is "abcca".

(a) Naive string-matching progresses left-to-right. Each letter comparison is indicated by an underline, and the total number of such comparisons at each shift is shown at the right. The naive string-matching algorithm requires 20 letter comparisons in this case.

```
"a b c c c d a b a c a b b c a"
 a b c c a                     5
   a b c c a                   1
     a b c c a                 1
       a b c c a               1
         a b c c a             1
           a b c c a           1
             a b c c a         3
               a b c c a       1
                 a b c c a     2
                   a b c c a   1
                     a b c c a 3
                               20
```

The Boyer-Moore string-matching algorithm requires 9 letter comparisons, in this case.

```
"a b c c c d a b a c a b b c a"
 a b c c a                     1
     a b c c a                 2
           a b c c a           3
               a b c c a       3
                               9
```

(b) By an analysis similar to that in part (a), we find that the naive string-matching algorithm requires 16 letter comparisons in this case.

The Boyer-Moore string-matching algorithm requires 5 letter comparisons in this case.

```
"d a d a d a d a d a d a d a d"
 a b c c a                                  1
           a b c c a                        2
                     a b c c a              2
                                            5
```

(c) By an analysis similar to that in part (a), we find that the naive string-matching algorithm requires 19 letter comparisons in this case.

The Boyer-Moore string-matching algorithm requires 7 letter comparisons in this case.

```
"a b c b c a b c a b c a b c"
 a b c c a                                  1
   a b c c a                                3
     a b c c a                              1
           a b c c a                        1
                 a b c c a                  1
                                            7
```

(d) By an analysis similar to that in part (a), we find that the naive string-matching algorithm requires 18 letter comparisons in this case.

The Boyer-Moore string-matching algorithm requires 4 letter comparisons in this case.

```
"a c c a b c a b a b a c c a"
 a b c c a                                  1
       a b c c a                            1
             a b c c a                      2
                                            4
```

(e) By an analysis similar to that in part (a), we find that the naive string-matching algorithm requires 14 letter comparisons in this case.

The Boyer-Moore string-matching algorithm requires 18 letter comparisons in this case. Note that in this unusual case, the naive string-matching algorithm is more efficient than Boyer-Moore algorithm.

```
"b b c c a c b c c a b b c c a"
 a b c c a                                  5
   a b c c a                                1
     a b c c a                              1
           a b c c a                        5
             a b c c a                      1
                   a b c c a                1
                     a b c c a              5
                                           18
```

22. Here is pseudocode for the last-occurrence function.

Algorithm 0 (Last-occurrence)

```
1 begin initialize F(x)
2 i ← m + 1
3 do i ← i − 1
4     if [F[x(i)] = 0] then F[x(i)] ← i
5 until i = 1
6 end
```

(a) The time complexity in this serial implementation is based on the sum of the d computations for initializing $\mathcal{F}$, and the m calculations of $\mathcal{F}$, that is, $O(d+m)$.

(b) The space complexity is $O(m)$ since we need to store the final m values.

(c) For "bonbon" the number of comparisons is 26 + 6 = 32. For "marmalade" the number of comparisons is 26 + 9 = 35. For "abcdabdabcaabcda" the number of comparisons is 26 + 16 = 42.

23. Here the class is determined by the minimum edit distance to any pattern in a category.

(a) By straightforward edit distance calculation, we find that the edit distance d from the test pattern $\mathbf{x}$ = "abacc" to each of the patterns in the categories are as shown in the table.

ω_1	d
aabbc	3
ababcc	**1**
babbcc	2

ω_2	d
bccba	4
bbbca	3
cbbaaaa	5

ω_3	d
caaaa	4
cbcaab	4
baaca	3

The minimum distance, 1, occurs for the second pattern in ω_1, and thus we the test pattern should be assigned to ω_1.

(b) As in part (a), the table shows the edit distance d from the test $\mathbf{x}$ = "abca" to each of the nine patterns.

ω_1	d
aabbc	3
ababcc	3
babbcc	3

ω_2	d
bccba	3
bbbca	**2**
cbbaaaa	5

ω_3	d
caaaa	3
cbcaab	3
baaca	**2**

The minimum distance, 2, occurs for patterns in two categories. Hence, here there is a tie between cateogories ω_2 and ω_3.

(c) As in part (a), the table shows the edit distance d from the test $\mathbf{x}$ = "ccbba" to each of the nine patterns.

ω_1	d
aabbc	3
ababcc	5
babbcc	4

ω_2	d
bccba	**2**
bbbca	3
cbbaaaa	4

ω_3	d
caaaa	3
cbcaab	4
baaca	4

Here the minimum distance, 2, occurs for the first pattern in ω_2, and thus the test pattern should be assigned to cateogory ω_2.

(d) As in part (a), the table shows the edit distance d from the test $\mathbf{x}$ = "bbaaac" to each of the nine patterns.

ω_1	d
aabbc	4
ababcc	3
babbcc	4

ω_2	d
bccba	4
bbbca	3
cbbaaaa	**2**

ω_3	d
caaaa	3
cbcaab	3
baaca	3

Here the minimum distance, 2, occurs for the third pattern in ω_2, and thus the test pattern should be assigned to category ω_2.

24. Here the class is determined by the minimum edit distance to any pattern in a category.

(a) By straightforward edit distance calculation, we find that the edit distance d from the test pattern $\mathbf{x}$ = "ccab" to each of the patterns in the categories are as shown in the table.

ω_1	d
aabbc	4
ababcc	4
babbcc	5

ω_2	d
bccba	3
bbbca	4
cbbaaaa	5

ω_3	d
caaaa	3
cbcaab	**2**
baaca	4

Here the minimum distance, 2, occurs for the second pattern in ω_3, and thus the test pattern should be assigned to ω_3.

(b) As in part (a), the table shows the edit distance d from the test $\mathbf{x}$ = "abdca" to each of the nine patterns.

ω_1	d
aabbc	3
ababcc	3
babbcc	3

ω_2	d
bccba	3
bbbca	**2**
cbbaaaa	5

ω_3	d
caaaa	4
cbcaab	4
baaca	3

Here the minimum distance, 2, occurs for the second pattern in ω_2, and thus we assign the test pattern to ω_2.

(c) As in part (a), the table shows the edit distance d from the test $\mathbf{x}$ = "abc" to each of the nine patterns.

ω_1	d
aabbc	**2**
ababcc	3
babbcc	3

ω_2	d
bccba	4
bbbca	3
cbbaaaa	6

ω_3	d
caaaa	4
cbcaab	4
baaca	3

Here the minimum distance, 2, occurs for the first pattern in ω_1, and thus we assign the test pattern to ω_1.

(d) As in part (a), the table shows the edit distance d from the test $\mathbf{x}$ = "bacaca" to each of the nine patterns.

ω_1	d
aabbc	4
ababcc	4
babbcc	3

ω_2	d
bccba	3
bbbca	3
cbbaaaa	4

ω_3	d
caaaa	3
cbcaab	4
baaca	**1**

Here the minimum distance, 1, occurs for the third pattern in ω_3, and thus we assign the test pattern to ω_3.

25. Here the class is determined by the minimum cost edit distance d_c to any pattern in a category, where we assume interchange is twice as costly as an insertion or a deletion.

(a) By straightforward cost-based edit distance calculation, we find that the edit distance d_c from the test pattern $\mathbf{x}$ = "abacc" to each of the patterns in the categories are as shown in the table.

ω_1	d_c
aabbc	4
ababcc	**1**
babbcc	3

ω_2	d_c
bccba	4
bbbca	6
cbbaaaa	8

ω_3	d_c
caaaa	6
cbcaab	7
baaca	4

The final categorization corresponds to the minimum of these cost-based distances, i.e, the distance = 1 to the second pattern in ω_1. Thus we assign the test pattern to ω_1.

(b) As in part (a), the table shows the cost-based edit distance d_c from the test $\mathbf{x}$ = "abca" to each of the nine patterns.

ω_1	d_c
aabbc	**3**
ababcc	4
babbcc	4

ω_2	d_c
bccba	**3**
bbbca	**3**
cbbaaaa	7

ω_3	d_c
caaaa	5
cbcaab	4
baaca	**3**

The minimum cost distance is 3, which occurs for four patterns. There is a three-way tie between the cateogories ω_1, ω_2 and ω_3. (In the event of a tie, in practice we often consider the next close pattern; in this case this would lead us to classify the test pattern as ω_2.)

(c) As in part (a), the table shows the cost-based edit distance d_c from the test $\mathbf{x}$ = "ccbba" to each of the nine patterns.

ω_1	d_c
aabbc	6
ababcc	7
babbcc	7

ω_2	d_c
bccba	**2**
bbbca	4
cbbaaaa	4

ω_3	d_c
caaaa	6
cbcaab	5
baaca	6

Hence $\mathbf{x}$ should be classified as ω_2.

(d) As in part (a), the table shows the cost-based edit distance d_c from the test $\mathbf{x}$ = "bbaaac" to each of the nine patterns.

ω_1	d_c
aabbc	4
ababcc	3
babbcc	4

ω_2	d_c
bccba	4
bbbca	3
cbbaaaa	**2**

ω_3	d_c
caaaa	3
cbcaab	3
baaca	3

Hence $\mathbf{x}$ should be classified as ω_2.

26. We assume the costs are all positive.

(a) Non-negativity is always guaranteed because all entries for computing edit distance are non-negative. Reflexivity and symmetry are also guaranteed. The triangle inequality is not guaranteed, however.

(b) Example: Suppose the costs for insertion and deletion are 1 and the cost for substitution is 5. Consider the three patterns $\mathbf{x}_1$ = "aba," $\mathbf{x}_2$ = "aab," and $\mathbf{x}_3$ = "aac." In this case, the cost-based distances d_c are as shown in the matrix

$$\begin{array}{c} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{array} \begin{array}{c} \begin{array}{ccc} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{array} \\ \begin{pmatrix} 0 & 2 & 2 \\ 5 & 0 & 5 \\ 2 & 2 & 0 \end{pmatrix} \end{array}.$$

Note that $d_c(\mathbf{x}_2, \mathbf{x}_3) > d_c(\mathbf{x}_1, \mathbf{x}_2) + d_c(\mathbf{x}_1, \mathbf{x}_3)$, violating the triangle inequality.

27. PROBLEM NOT YET SOLVED
28. PROBLEM NOT YET SOLVED
29. PROBLEM NOT YET SOLVED

Section 8.6

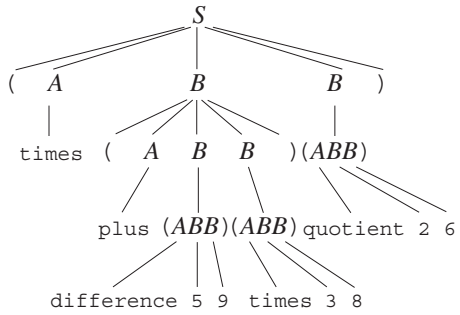
30. Recall that *Lisp* expressions are of the general form `(operation` *operand*$_1$ *operand*$_2$`)`. (More generally, for some operations, such as multiplication, `*`, and addition, `+`, there can be an arbitrary number of operands.)

(a) Here the alphabet is $\mathcal{A} = \{\texttt{0}, \texttt{1}, \texttt{2}, \ldots, \texttt{9}, \texttt{plus}, \texttt{minus}, \texttt{quotient}, (,)\}$, the set of intermediate symbols is $\mathcal{I} = \{A, B\}$, the starting symbol is $\mathcal{S} = S$, and the productions are

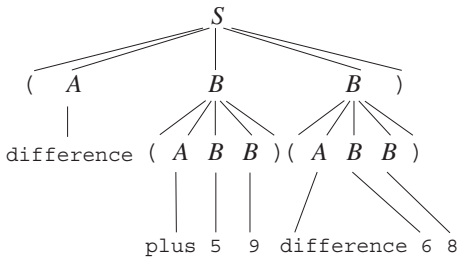
$$\mathcal{P} = \begin{cases} \mathbf{p}_1: & S \rightarrow (ABB) \\ \mathbf{p}_2: & A \rightarrow \texttt{plus}|\texttt{difference}|\texttt{times}|\texttt{quotient} \\ \mathbf{p}_3: & B \rightarrow \texttt{0}|\texttt{1}|\ldots|\texttt{9} \\ \mathbf{p}_4: & B \rightarrow (ABB) \end{cases}$$

(b) An advanced *Lisp* parser could eliminate "excess" parentheses, but for the grammar above, three of the five expressions are ungrammatical:

- `(times (plus (difference 5 9)(times 3 8))(quotient 2 6))` can be expressed in this grammar (see figure).
- `(7 difference 2)` cannot be expressed in this grammar.



- `(quotient (7 plus 2)(plus 6 3))` cannot be expressed in this grammar.
- `((plus)(6 2))` cannot be expressed in this grammar.
- `(difference (plus 5 9)(difference 6 8))` can be expressed in this grammar (see figure).

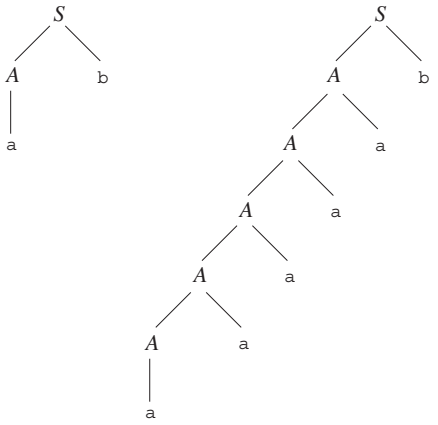


31. Here the language is $\mathcal{L}(G) = \{\mathtt{a}^n b | n \geq 1\}$.

(a) The alphabet is $\mathcal{A} = \{\mathtt{a}, \mathtt{b}\}$, the intermediate symbol is $\mathcal{I} = \{A\}$, the starting symbol is $\mathcal{S} = S$, and the productions are

$$\mathcal{P} = \left\{ \begin{array}{ll} \mathbf{p}_1 & S \rightarrow A\mathtt{b} \\ \mathbf{p}_2 & A \rightarrow A\mathtt{a} \\ \mathbf{p}_3 & A \rightarrow \mathtt{a} \end{array} \right\}$$

(b) See figure.



32. Here the language G has alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, intermediate symbols $\mathcal{I} = \{A, B\}$, starting symbol $\mathcal{S} = S$, and the productions are

$$\mathcal{P} = \left\{ \begin{array}{ll} \mathbf{p}_1 & S \rightarrow \mathtt{c}A\mathtt{b} \\ \mathbf{p}_2 & A \rightarrow \mathtt{a}B\mathtt{a} \\ \mathbf{p}_3 & B \rightarrow \mathtt{a}B\mathtt{a} \\ \mathbf{p}_4 & B \rightarrow \mathtt{cb} \end{array} \right\}$$

(a) We consider the rules in turn:

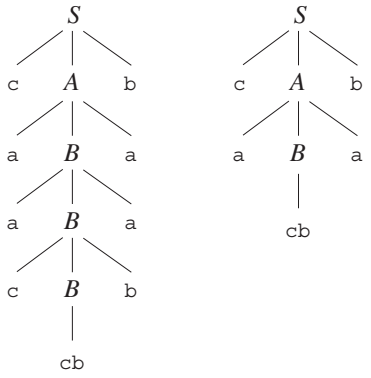
- $S \rightarrow \mathtt{c}A\mathtt{b}$: type 3 (of the form $\alpha \rightarrow z\beta$)
- $A \rightarrow \mathtt{a}B\mathtt{a}$: type 3 (of the form $\alpha \rightarrow z\beta$)
- $B \rightarrow \mathtt{a}B\mathtt{a}$: type 3 (of the form $\alpha \rightarrow z\beta$)
- $B \rightarrow \mathtt{cb}$: type 2 (of the form $I \rightarrow x$) and type 3 (of the form $\alpha \rightarrow z$)

Thus G is type 3 or regular grammar.

(b) We now prove by induction that the language generated by G is $\mathcal{L}(G) = \{\mathtt{c}\mathtt{a}^n\mathtt{c}\mathtt{b}\mathtt{a}^n\mathtt{b} | n \geq 1\}$. Note that for $n = 1$, cacbab can be generated by G: $S \rightarrow \mathtt{c}A\mathtt{b} \rightarrow \mathtt{c}\mathtt{a}B\mathtt{a}\mathtt{b} \rightarrow \mathtt{c}\mathtt{a}\mathtt{c}\mathtt{b}\mathtt{a}\mathtt{b}$.

Now suppose that $\mathtt{c}\mathtt{a}^n\mathtt{c}\mathtt{b}\mathtt{a}^n\mathtt{b}$ is generated by G. Now $\mathtt{c}\mathtt{a}^{n+1}\mathtt{c}\mathtt{b}\mathtt{a}^{n+1}\mathtt{b} = \mathtt{c}\mathtt{a}\mathtt{a}^n\mathtt{c}\mathtt{b}\mathtt{a}^n\mathtt{a}\mathtt{b}$. Given the above, we know that $S \rightarrow \mathtt{c}\tilde{A}\mathtt{b}$, where $\tilde{A} = \mathtt{a}^n\mathtt{c}\mathtt{b}\mathtt{a}^n$ can be generated by G. When we parse $\mathtt{c}\mathtt{a}\mathtt{a}^n\mathtt{c}\mathtt{b}\mathtt{a}^n\mathtt{a}\mathtt{b}$ using G, we find $S \rightarrow \mathtt{c}A\mathtt{b} \rightarrow \mathtt{c}\mathtt{a}B\mathtt{a}\mathtt{b}$, and thus $B = \tilde{A}$. Thus $\mathtt{c}\mathtt{a}^{n+1}\mathtt{c}\mathtt{b}\mathtt{a}^{n+1}\mathtt{b}$ is generated by G.

(c) See figure.

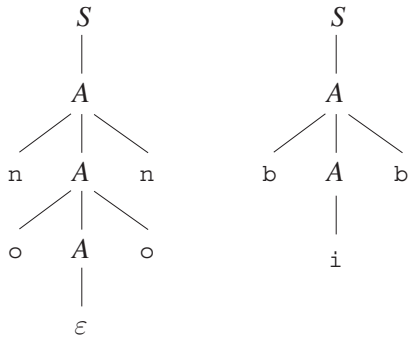


33. A palindrome is a string that reads the same forward as backward.

(a) Here the grammar G has the English alphabet and the null character, $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \ldots, \mathtt{z}, \epsilon\}$, the single internal symbol $\mathcal{I} = A$, and starting symbol $\mathcal{S} = S$. The null symbol is needed for generating palindromes having an even number of letters. The productions are of the form:

$$\mathcal{P} = \left\{ \begin{array}{ll} \mathbf{p}_1 & S \rightarrow A \\ \mathbf{p}_2 & A \rightarrow \mathtt{a}|\mathtt{b}|\ldots|\mathtt{z}|\epsilon \\ \mathbf{p}_3 & A \rightarrow \mathtt{a}A\mathtt{a}|\mathtt{b}A\mathtt{b}|\ldots|\mathtt{z}A\mathtt{z} \end{array} \right\}$$

See figure.

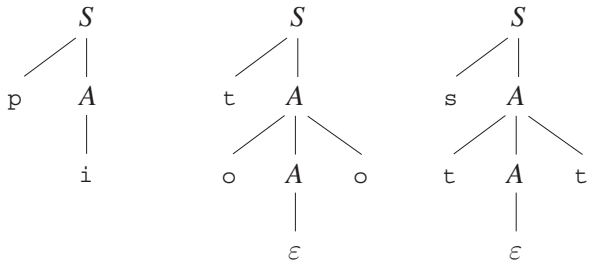


(b) The grammar is of type 3 because every rewrite rule is of the form $\alpha \rightarrow z\beta$ or $\alpha \rightarrow z$.

(c) Here the grammar G has the English alphabet and the null character, $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \ldots, \mathtt{z}, \epsilon\}$, the single internal symbol $\mathcal{I} = A$ and starting symbol $\mathcal{S} = S$. The null symbol is needed for generating palindromes having an even number of letters. The productions are of the form:

$$\mathcal{P} = \left\{ \begin{array}{ll} \mathbf{p}_1 & S \rightarrow \mathtt{a}A|\mathtt{b}A|\ldots|\mathtt{z}A \\ \mathbf{p}_2 & A \rightarrow \mathtt{a}|\mathtt{b}|\ldots|\mathtt{z}|\epsilon \\ \mathbf{p}_3 & A \rightarrow \mathtt{a}A\mathtt{a}|\mathtt{b}A\mathtt{b}|\ldots|\mathtt{z}A\mathtt{z} \end{array} \right\}$$

See figure.



34. Consider the numbers $1, 2, \ldots, 999$.

(a) Here the productions are of the form:

$$\mathcal{P} = \left\{ \begin{array}{ll} \mathbf{p}_1 & \textit{digits}6 \rightarrow \textit{digits}3 \\ \mathbf{p}_2 & \textit{digits}3 \rightarrow \textit{digit}1\ \mathtt{hundred}\ \textit{digits}2 \\ \mathbf{p}_3 & \textit{digits}3 \rightarrow \textit{digit}1\ \mathtt{hundred} \\ \mathbf{p}_4 & \textit{digits}3 \rightarrow \textit{digits}2 \\ \mathbf{p}_5 & \textit{digits}2 \rightarrow \mathtt{teens}|\mathtt{tys}|\mathtt{tys}\ \textit{digit}1|\textit{digit}1 \\ \mathbf{p}_6 & \textit{digit}1 \rightarrow \mathtt{1}|\mathtt{2}|\ldots|\mathtt{9} \end{array} \right\}$$

Thus there are 9 possible derivations for *digit1*, 10 possible derivations for teens, 8 possible derivations for tys. Thus there are $(10 + 8 + 8 \times 9 + 9) = 99$ possible derivations for *digits2*. Thus there are $(9 \times 99 + 9 + 99)$ possible derivations for *digits3*. So there are 999 possible derivations for *digits6* under these restrictive conditions.

(b) For the numbers $1, 2, \ldots, 999,999$ we have

$$\mathcal{P} = \left\{ \begin{array}{ll} \mathbf{p}_1 & digits6 \rightarrow digits3 \texttt{ thousand } digits3 \\ \mathbf{p}_2 & digits6 \rightarrow digits3 \texttt{ thousand} \\ \mathbf{p}_3 & digits6 \rightarrow digits3 \end{array} \right\}$$

We know from part (a) that $|digits3| = 999$. Thus $|digits6| = 999 \times 999 + 999 + 999 = 999,999$ — indeed, one for each number.

(c) The grammar does not allow any pronunciation in more than one way because there are only 999,999 numbers and there are only 999,999 possible pronunciations and these are in a one-to-one correspondence. In English, however, there is usually more than one way to pronounce a single number. For instance, 2000 can be pronounced `two thousand` or `twenty hundred`, however the grammar above does not allow this second pronunciation.

35. PROBLEM NOT YET SOLVED

36. Consider grammars and the Chomsky normal form.

(a) Here, most of the rewrite rules in G are not in Chomsky normal form:

$$\mathcal{P} = \left\{ \begin{array}{rl} S \rightarrow \texttt{b}A & \text{not CNF} \\ S \rightarrow \texttt{a}B & \text{not CNF} \\ A \rightarrow \texttt{b}AA & \text{not CNF} \\ A \rightarrow \texttt{a}S & \text{not CNF} \\ A \rightarrow \texttt{a} & \text{CNF} \\ \vdots & \vdots \end{array} \right\}$$

We need only one rule to violate the conditions of being in Chomsky normal form for the entire grammar to be not in CNF. Thus this grammar is not in CNF.

(b) Here the rewrite rules of G' are

$$\mathcal{P} = \left\{ \begin{array}{ll} \text{rewrite rule} & \text{CNF?} \\ \hline S \rightarrow C_bA & \text{yes } A \rightarrow BC \\ S \rightarrow C_aB & \text{yes } A \rightarrow BC \\ A \rightarrow C_aS & \text{yes } A \rightarrow BC \\ A \rightarrow C_bD_1 & \text{yes } A \rightarrow BC \\ A \rightarrow \texttt{a} & \text{yes } A \rightarrow \texttt{z} \\ B \rightarrow C_bS & \text{yes } A \rightarrow BC \\ B \rightarrow C_aD_2 & \text{yes } A \rightarrow BC \\ B \rightarrow \texttt{b} & \text{yes } A \rightarrow \texttt{z} \\ D_1 \rightarrow AA & \text{yes } A \rightarrow BC \\ D_2 \rightarrow BB & \text{yes } A \rightarrow BC \\ C_a \rightarrow \texttt{a} & \text{yes } A \rightarrow \texttt{z} \\ C_b \rightarrow \texttt{b} & \text{yes } A \rightarrow \texttt{z} \end{array} \right\}$$

Thus, indeed all the rewrite rules are of the form $A \rightarrow BC$ or $A \rightarrow \texttt{z}$, and the grammar G is in CNF.

(c) Theorem: Any context-free language without the null symbol ϵ can be generated by a grammar G_1 in which the rewrite rules are of the form $A \rightarrow BC$ or $A \rightarrow \texttt{z}$.

Proof: First, if a production ends in a string containing just a single terminal symbol, then that production is in acceptable CNF form (i.e., $A \rightarrow \mathtt{z}$). Now consider a production of the form $A \rightarrow Z_1 Z_2 \cdots Z_n$, where $n \geq 2$. If Z_i is a terminal symbol $\mathtt{a}$, we introduce a new internal variable C_a and the production $C_a \rightarrow \mathtt{a}$, which is in the permissible form. Now we expand our intermediate variables to include this new C_a, and the rewrite rules accordingly to form G_2. Thus, $\mathcal{L}(G_1) \subseteq \mathcal{L}(G_2)$.

Next we use induction on the number of steps in a derivation that if G_2 give a derivation $A \rightarrow \mathtt{z}$, then G_1 does as well. If so that if G_2 generates a final string, then G_1 does as well. This result is obvious for final strings of length 1. Suppose now that it is true for derivations requiring k steps. Let D^1 be a derivation in G_2 requiring $k+1$ steps. The first must be of the general form $A \rightarrow B_1 B_2 \cdots B_m$ for $m \geq 2$. Then we can write $\mathbf{Z} = Z_1 Z_2 \cdots Z_m$ where in grammar G_2 we have that B_i derives Z_i.

If B_i is the same as C_{a_i} for some terminal a_i, then B_i must be Z_i. By our construction, there exists a production $A \rightarrow Z_1 Z_2 \cdots Z_n$ where $Z_i = B_i$. For those B_i we know that the derivation in G_1 requires no more than k steps, so by induction hypothesis Z_i leads to B_i.

Thus we see that any context-free language can be generated by a grammar where every production is of the form $A \rightarrow \mathtt{z}$ or $A \rightarrow Z_1 Z_2 \cdots Z_n$ for $n \geq 2$.

(d) The string "`aabab`" cannot be derived in either G or G'.

37. [[state that this problem requires the pumping Lemma in Chapter 6 of **Introduction to Automata Theory** by Hopcroft and Ullman.]]

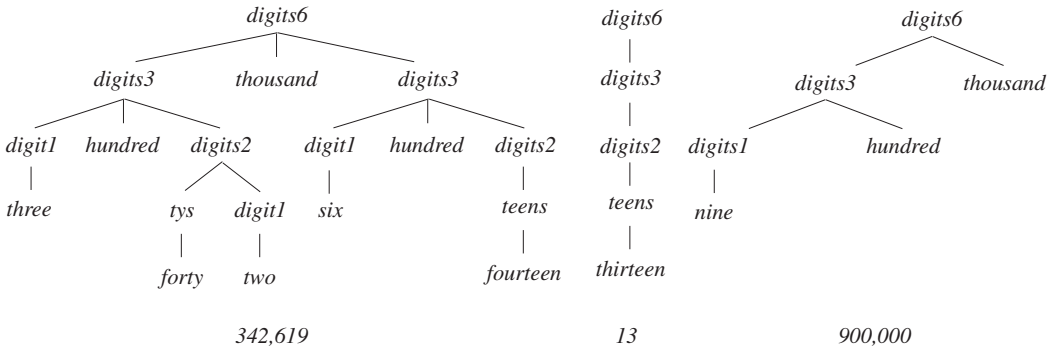
38. PROBLEM NOT YET SOLVED

39. The first three numbers can be derived, but the last two cannot.

(a) See figure.

(b) See figure.

(c) See figure.



(d) `two thousand six hundred thousand five` cannot be derived.

(e) `one hundred sixty eleven` cannot be derived.

Section 8.7

40. Here $\mathcal{D}_1 = \{\texttt{ab}, \texttt{abb}, \texttt{abbb}\}$ and $\mathcal{D}_2 = \{\texttt{ba}, \texttt{aba}, \texttt{babb}\}$ are positive examples from grammars G_1 and G_2, respectively.

(a) Some candidate rewrite rules for G_1 and G_2, are

$$\mathcal{P}_1 = \left\{ \begin{array}{ll} \mathbf{p}_1 & S \to A \\ \mathbf{p}_2 & A \to \texttt{ab} \\ \mathbf{p}_3 & A \to A\texttt{b} \end{array} \right\} \qquad \mathcal{P}_2 = \left\{ \begin{array}{ll} \mathbf{p}_1 & S \to A \\ \mathbf{p}_2 & A \to \texttt{ba} \\ \mathbf{p}_3 & A \to \texttt{a}A \\ \mathbf{p}_4 & A \to A\texttt{b} \end{array} \right\}$$

(b) Here we infer G_1, being sure not to include rules that would yield strings in $\mathcal{D}_2$.

i	$\mathbf{x}_i^+$	$\mathcal{P}$	$\mathcal{P}$ produces $\mathcal{D}_2$?
1	ab	$S \to A$ $S \to \texttt{ab}$	No
2	abb	$S \to A$ $A \to \texttt{ab}$ $A \to A\texttt{b}$	No
3	abbb	$S \to A$ $A \to \texttt{ab}$ $A \to A\texttt{b}$	No

(c) Here we infer G_2, being sure not to include rules that would yield strings in $\mathcal{D}_1$.

i	$\mathbf{x}_i^+$	$\mathcal{P}$	$\mathcal{P}$ produces $\mathcal{D}_1$?
1	ba	$S \to A$ $A \to \texttt{ba}$	No
2	aba	$S \to A$ $A \to \texttt{ba}$ $A \to \texttt{a}A$	No
3	babb	$S \to A$ $A \to \texttt{ba}$ $A \to \texttt{a}A$ $A \to A\texttt{b}$	No

(d) Two of the strings are ambiguous, and one each is in $\mathcal{L}(G_1)$ and $\mathcal{L}(G_2)$:

- bba is ambiguous
- abab $\in \mathcal{L}(G_2)$: $S \to A \to \texttt{a}A \to \texttt{a}A\texttt{b} \to \texttt{abab}$
- bbb is ambiguous
- abbbb $\in \mathcal{L}(G_1)$: $S \to A \to A\texttt{b} \to A\texttt{bb} \to A\texttt{bbb} \to \texttt{abbbb}$.

Section 8.8

41. Problem not yet solved

Computer Exercises

Section 8.3

1. Computer exercise not yet solved
2. Computer exercise not yet solved

Section 8.4

3. Computer exercise not yet solved
4. Computer exercise not yet solved
5. Computer exercise not yet solved

Section 8.5

6. Computer exercise not yet solved
7. Computer exercise not yet solved
8. Computer exercise not yet solved

Section 8.6

9. Computer exercise not yet solved
10. Computer exercise not yet solved

Section 8.7

11. Computer exercise not yet solved

Chapter 9

Algorithm-independent machine learning

Problem Solutions

Section 9.2

1. Problem not yet solved
2. Problem not yet solved
3. Problem not yet solved
4. Problem not yet solved
5. Problem not yet solved
6. Problem not yet solved
7. Problem not yet solved
8. Problem not yet solved
9. We seek to prove the relation

$$2^n = \sum_{r=0}^{n} \binom{n}{r}$$

two different ways.

(a) Recall that the polynomial expression is

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r.$$

We substitute $x = y = 1$ and find the desired result directly, as shown in Eq. 5 in the text:

$$(1+1)^n = 2^n = \sum_{r=0}^{n} \binom{n}{r}.$$

(b) We define

$$K(n) = \sum_{r=0}^{n} \binom{n}{r}$$

and proceed by induction. Clearly, by simple substitution, $K(1) = 1+1 = 2^1 = 2^n$. We seek to show that in general $K(n+1) = 2K(n) = 2 \cdot 2^n = 2^{n+1}$. We write this explicitly

$$\begin{aligned}
K(n+1) &= \sum_{r=0}^{n+1} \binom{n+1}{r} \\
&= \binom{n}{0} + \sum_{r=1}^{n} \binom{n+1}{r} + \binom{n}{n} \\
&= \binom{n}{0} + \sum_{r=1}^{n} \frac{(n+1)!}{r!(n+1-r)!} + \binom{n}{n} \\
&= \binom{n}{0} + \sum_{r=1}^{n} \frac{n!(n-r+1+r)!}{r!(n+1-r)!} + \binom{n}{n} \\
&= \binom{n}{0} + \sum_{r=1}^{n} \left(\frac{n!(n-r+1)}{r!(n+1-r)!} + \frac{n!r}{r!(n+1-r)!} \right) + \binom{n}{n} \\
&= \binom{n}{0} + \sum_{r=1}^{n} \left(\frac{n!}{r!(n-r)!} + \frac{n!}{(r-1)!(n+1-r)!} \right) + \binom{n}{n} \\
&= \binom{n}{0} + \sum_{r=1}^{n} \left(\binom{n}{r} + \binom{n}{r-1} \right) + \binom{n}{n} \\
&= \sum_{r=0}^{n} \left(\binom{n}{r} + \binom{n}{r} \right) \\
&= 2 \sum_{r=0}^{n} \binom{n}{r} \\
&= 2K(n) = 2^{n+1}.
\end{aligned}$$

10. Problem not yet solved

11. Consider the Ugly Duckling Theorem (Theorem 9.2).

(a) If a classification problem has constraints placed on the features, the number of patterns will be less than that of an unconstrained problem. However, the total number o predicates depends only on the number of patterns, not the number of features, and so is still

$$\sum_{r=0}^{d} \binom{d}{r} = 2^d,$$

where d is the number of distinct possible patterns. The total number of predicates shared by two patterns remains

$$\sum_{r=2}^{d} \binom{d-2}{r-2} = 2^{d-2}.$$

In short, there are no changes to the derivation of the Ugly Duckling Theorem.

(b) Since there are only three different types of cars seen, we have three different patterns, $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, with features:

$$\begin{aligned}\mathbf{x}_1 &: f_1 \text{ AND } f_2 \text{ AND } f_4 \text{ AND NOT } f_5 \text{ AND NOT } f_6\\ \mathbf{x}_2 &: f_1 \text{ AND } f_2 \text{ AND } f_3 \text{ AND NOT } f_4 \text{ AND } f_5 \text{ AND NOT } f_6\\ \mathbf{x}_3 &: \text{NOT}[f_1 \text{OR } f_2 \text{ OR } f_3 \text{ OR } f_4 \text{ OR } f_5] \text{ AND } f_6,\end{aligned}$$

where f_1 respresents a car model similar to car A, f_2 represents an engine similar to car A, f_3 represents a four-door car, f_4 represents a red car, f_5 represents a green car, f_6 represents a blue car.

To determine similarity, we count the number of predicates shared by two cars. With car A and car B, the predicates are $\mathbf{x}_1$ OR $\mathbf{x}_2$, and $\mathbf{x}_1$ OR $\mathbf{x}_2$ OR $\mathbf{x}_3$. With car B and car C, the predicates are $\mathbf{x}_2$ OR $\mathbf{x}_3$, and $\mathbf{x}_1$ OR $\mathbf{x}_2$ OR $\mathbf{x}_3$. Since the number of predicates is the same the patterns are "equally similar," according to the definition given in the text.

In effect, we could have described the three patterns with three features, with f_1 representing a four-door car model/engine similar to car C, f_2 representing a red car, f_3 representing a green car. In this framework, the patters would be:

$$\begin{aligned}\mathbf{x}_1 &: \text{NOT } f_1 \text{AND } f_2 \text{ AND NOT } f_3\\ \mathbf{x}_2 &: \text{NOT } f_1 \text{ AND NOT } f_2 \text{ AND } f_3\\ \mathbf{x}_3 &: f_1 \text{ AND NOT } f_2 \text{ AND NOT } f_3.\end{aligned}$$

The result is that the three patterns being different from one another by only one feature, and hence cars A and B are equally similar as cars B and C.

12. PROBLEM NOT YET SOLVED

13. Consider the Kolmogorov complexity of different sequences.

(a) 010110111011110 . . .. This sequence is made up of m sequences of 0 and l 1s, where $l = 1 \ldots m$. Therefore the complexity is $O(\log_2 m)$.

(b) $000 \ldots 100 \ldots 000$, a sequence made up of m 0s, a single 1, and $(n - m - 1)$ 0s. The complexity is just that needed to specify m, i.e., $O(\log_2 m)$.

(c) $e = 10.10110111111000010 \ldots_2$, The complexity of this constant is $O(1)$.

(d) $2e = 101.0110111111000010 \ldots_2$, The complexity of this constant is $O(1)$.

(e) The binary digits of π, but where every 100th digit is changed to the numeral 1. The constant π has a complexity of $O(1)$; it is a simple matter to wrap a program around this to change each 100th digit. This does not change the complexity, and the answer is $O(1)$.

(f) As in part (e), but now we must specify n, with has complexity $O(\log_2 n)$.

14. PROBLEM NOT YET SOLVED

15. For two binary strings x_1 and x_2, the Kolmogorov complexity of the pair can be at worst $K(x_1, x_2) = K(x_1) + K(x_2)$, as we can easily write concatenate two programs, one for computing x_1 and one for x_2. If the two strings share information,

the Kolmogorov complexity will be less than $K(x_1) + K(x_2)$, since some information from one of the strings can be used in the generatation of the other string.

16. PROBLEM NOT YET SOLVED

17. The definition "the least number that cannot be defined in less than twenty words" is already a definition of less than twenty words. The definition of the Kolmogorov complexity is the length of the shortest program to describe a string. From the above paradoxical statement, we can see that it is possible that we are not "clever" enough to determine the shortest program length for a string, and thus we will not be able to determine easily the complexity.

Section 9.3

18. The mean-square error is

$$\begin{aligned}\mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x})^2] &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D}) - 2g(\mathbf{x};\ \mathcal{D})F(\mathbf{x}) + F^2(\mathbf{x})] \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D}] - \mathcal{E}_{\mathcal{D}}[2g(\mathbf{x};\ \mathcal{D}F(\mathbf{x})] + \mathcal{E}_{\mathcal{D}}[F^2(\mathbf{x})] \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - 2F(\mathbf{x})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + F^2(\mathbf{x}).\end{aligned}$$

Note, however, that

$$\begin{aligned}\mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2] &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D}) - 2g(\mathbf{x};\ \mathcal{D})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + [\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]]^2] \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - \mathcal{E}_{\mathcal{D}}[2g(\mathbf{x};\ \mathcal{D})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]] + \mathcal{E}_{\mathcal{D}}[(\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2] \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - 2\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2 \\ &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2.\end{aligned}$$

We put these two results together and find

$$\mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] = \mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2] + (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2.$$

We now apply this result to the function $g(\mathbf{x}) - F(\mathbf{x})$ and obtain

$$\begin{aligned}\mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x}))^2] &= \mathcal{E}_{\mathcal{D}}[g^2(\mathbf{x};\ \mathcal{D})] - 2F(\mathbf{x})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + F^2(\mathbf{x}) \\ &= \mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D} - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]))^2] + (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2 \\ &\quad -2F(\mathbf{x})\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] + F^2(\mathbf{x}) \\ &= \mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2] + (\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x})])^2 \\ &= \underbrace{(\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D}) - F(\mathbf{x})])^2}_{bias^2} + \underbrace{\mathcal{E}_{\mathcal{D}}[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2]}_{variance}.\end{aligned}$$

Since the estimate can be more or less than the function $F(\mathbf{x})$, the bias can be negative. The variance cannot be negative, as it is the expected value of a squared number.

19. For a given data set $\mathcal{D}$, if $g(\mathbf{x};\ \mathcal{D})$ agrees with the Bayes classifier, the expected error rate will be $\text{Min}[F(\mathbf{x}, 1 - F(\mathbf{x})]$; otherwise it will be $\text{Max}[F(\mathbf{x}, 1 - F(\mathbf{x})]$. Thus we have

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y] &= \text{Min}[F(\mathbf{x}), 1 - F(\mathbf{x})]\Pr[g(\mathbf{x};\ \mathcal{D}) = y_B] \\ &\quad +\text{Max}[F(\mathbf{x}), 1 - F(\mathbf{x})]\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B].\end{aligned}$$

However, under these conditions we can write

$$\text{Max}[F(\mathbf{x}), 1 - F(\mathbf{x})] = \text{Min}[F(\mathbf{x}), 1 - F(\mathbf{x})] + \underbrace{\text{Max}[F(\mathbf{x}), 1 - F(\mathbf{x})] - \text{Min}[F(\mathbf{x}), 1 - F(\mathbf{x})]}_{|2F(\mathbf{x})-1}.$$

Thus we conclude

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y &= |2F(\mathbf{x})-1|\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] \\ &\quad +\text{Min}[F(\mathbf{x}), 1-F(\mathbf{x})](\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] + \Pr[g(\mathbf{x};\ \mathcal{D}) = y_B]) \\ &= |2F(\mathbf{x})-1|\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] + \text{Min}[F(\mathbf{x}), 1-F(\mathbf{x})].\end{aligned}$$

20. If we make the convenient assumption that $p(g(\mathbf{x};\ \mathcal{D}))$ is a Gaussian, that is,

$$p(g(\mathbf{x};\ \mathcal{D})) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-(g-\mu)^2/(2\sigma^2)]$$

where $\mu = \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})]$ and $\sigma^2 = \text{Var}[g(\mathbf{x};\ \mathcal{D})]$. From Eq. 19 in the text, then, for $F(\mathbf{x}) < 1/2$ we have

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] &= \int_{1/2}^{\infty} p(g(\mathbf{x};\ \mathcal{D}))dg \\ &= \int_{1/2}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp[-(g-\mu)^2/(2\sigma^2)]dg \\ &= \frac{1}{\sqrt{2\pi}} \int_{(1/2-\mu)/\sigma}^{\infty} \exp[-u^2/2]du,\end{aligned}$$

where $u = (g-\mu)/\sigma$ and $du = dg/\sigma$. For the other case, that is, $F(\mathbf{x}) \geq 1/2$, we have

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D}) \neq y_B] &= \int_{-\infty}^{1/2} p(g(\mathbf{x};\ \mathcal{D}))dg \\ &= \int_{-\infty}^{1/2} \frac{1}{\sqrt{2\pi}\sigma}\exp[-(g-\mu)^2/(2\sigma^2)]dg \\ &= \frac{1}{\sqrt{2\pi}} \int_{-(1/2-\mu)/\sigma}^{\infty} \exp[-u^2/2]du,\end{aligned}$$

where $u = -(g-\mu)/\sigma$ and $du = -dg/\sigma$. Therefore, we have

$$\begin{aligned}\Pr[g(\mathbf{x};\ \mathcal{D})] &= \begin{cases} \frac{1}{\sqrt{2\pi}} \int_{(1/2-\mu)/\sigma}^{\infty} \exp[-u^2/2]du & \text{if } F(\mathbf{x}) < 1/2 \\ \frac{1}{\sqrt{2\pi}} \int_{-(1/2-\mu)/\sigma}^{\infty} \exp[-u^2/2]du & \text{if } F(\mathbf{x}) \geq 1/2 \end{cases} \\ &= \frac{1}{\sqrt{2\pi}} \int_{t}^{\infty} \exp[-u^2/2]du = \frac{1}{2}\left[1 - \text{erf}[t/\sqrt{2}]\right] = \Phi(t),\end{aligned}$$

where

$$t = \begin{cases} \frac{1/2-\mu}{\sigma} & \text{if } F(\mathbf{x}) < 1/2 \\ -\frac{1/2-\mu}{\sigma} & \text{if } F(\mathbf{x}) \geq 1/2. \end{cases}$$

Thus, we can write

$$\begin{aligned}
\Pr[g(\mathbf{x};\ \mathcal{D})] &= \text{sgn}[F(\mathbf{x}) - 1/2]\frac{\mu - 1/2}{\sigma} \\
&= \text{sgn}[F(\mathbf{x}) - 1/2]\frac{\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D}) - 1/2]}{\sqrt{\text{Var}[g(\mathbf{x};\ \mathcal{D})]}} \\
&= \underbrace{\text{sgn}[F(\mathbf{x}) - 1/2][\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})] - 1/2]}_{boundary\ bias}\underbrace{\text{Var}[g(\mathbf{x};\ \mathcal{D})]^{-1/2}}_{variance}.
\end{aligned}$$

21. Problem not yet solved
22. Problem not yet solved

Section 9.4

23. The jackknife estimate of the mean is given by Eq. 25 in the text:

$$\begin{aligned}
\mu_{(\cdot)} &= \frac{1}{n}\sum_{i=1}^{n}\mu_{(i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n-1}\sum_{j\neq i}x_j\right] \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}x_j - x_i\right] \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}[n\hat{\mu} - x_i] \\
&= \frac{n}{n-1}\hat{\mu} - \frac{1}{n(n-1)}\sum_{i=1}^{n}x_i \\
&= \frac{n}{n-1}\hat{\mu} - \frac{1}{n-1}\hat{\mu} \\
&= \hat{\mu}.
\end{aligned}$$

24. Problem not yet solved
25. Problem not yet solved
26. We must verify that Eq. 26 in the text for the jackknife estimate of the variance of the mean is formally equivalent to the variance estimate given by Eq. 23 in the text. From Eq. 26 we have

$$\begin{aligned}
\text{Var}[\hat{\mu}] &= \frac{n-1}{n}\sum_{i=1}^{n}(\mu_{(i)} - \mu_{(\cdot)})^2 \\
&= \frac{n-1}{n}\sum_{i=1}^{n}\left(\left(\frac{n\bar{x} - x_i}{n-1}\right) - \mu_{(\cdot)}\right)^2 \\
&= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{n\bar{x} - x_i}{n-1} - \frac{n-1}{n-1}\bar{x}\right)^2 \\
&= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{n\bar{x} - x_i - (n-1)\bar{x}}{n-1}\right)^2
\end{aligned}$$

$$\begin{aligned} &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{\bar{x}-x_i}{n-1}\right)^2 \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2, \end{aligned}$$

which is Eq. 23 in the text.

27. Consider the computational complexity of different statistics based on resampling.

(a) The jackknife estimate of the mean is

$$\theta_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\theta_{(i)},$$

which requires n summations, and thus has a complexity $O(n)$.

(b) The jackknife estimate of the median has complexity just that required for the sorting operation, which is $O(n\log n)$.

(c) The jackknife estimate of the standard deviation is

$$\begin{aligned} \sqrt{\mathrm{Var}[\hat{\theta}]} &= \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\theta_{(i)}-\theta_{(\cdot)}\right)^2} \\ &= \sqrt{\frac{n-1}{n}\left[\sum_{i=1}^{n}\theta_{(i)}^2-\left(\frac{1}{n}\sum_{i=1}^{n}\theta_{(i)}\right)^2\right]}, \end{aligned}$$

which requires $2n$ summations, and thus has a complexity $O(n)$.

(d) The bootstrap estimate of the mean is

$$\hat{\theta}^{*(\cdot)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*(b)},$$

which requires B summations, and thus has a complexity $O(B)$.

(e) The bootstrap estimate of the medium has complexity the same as that as the sorting operation, which is $O(B\log B)$.

(f) The bootstrap estimate of the standard deviation is

$$\begin{aligned} \sqrt{\mathrm{Var}_{Boot}[\theta]} &= \sqrt{\frac{1}{B}\sum_{i=1}^{n}\left(\hat{\theta}^{*(b)}-\hat{\theta}^{*(\cdot)}\right)^2} \\ &= \sqrt{\frac{1}{B}\left[\sum_{b=1}^{B}(\hat{\theta}^{*(b)})^2-\left(\frac{1}{B}\sum_{b'=1}^{B}\hat{\theta}^{*(b')}\right)^2\right]}, \end{aligned}$$

which requires $2B$ summations, and thus has a complexity $O(B)$.

28. Problem not yet solved

Section 9.5

29. Problem not yet solved

30. In boosting, if none of the patterns in $\mathcal{D}_2$ are correctly classified by C_1, then the classifier to be trained, C_2, will not have good accuracy on the patterns correctly classified by C_1. This means that C_1 and C_2 will disagree on most of the patterns, and this means that C_3 will dominate the full ensemble classifier. Paradoxically, if C_2 classifies *none* of the patterns in $\mathcal{D}_1$ correctly, then C_2 actually shares a great deal of information with C_1. In a c-category problem, $(1 - 1/c)$ of the patterns should be misclassified in a "most informative" set as this will complement C_1 on its slightly bettern than chance, $1/c$ of a correct classification.

31. There are a number of algorithms that are acceptable here. One approach is to perform a simple binary search over the line between $\mathbf{x}_1$ and $\mathbf{x}_2$, as

Algorithm 0 (Basic binary line search)

```
1 begin initialize x1, x2, ε
2         while (true)
3                 x ← (x1 + x2)/2
4                 if ‖g1(x) − g2(x)‖ < ε then return x
5                 if g1(x) < g2(x) then x2 ← x
6         end
```

Another version uses the discriminant functions $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ to heuristically guide the search for the boundary. As the two points under consideration, $\mathbf{x}_1$ and $\mathbf{x}_2$ become closer together, the true discriminant functions can be approximated linearly. In the following, $\mathbf{x}_3$ is the point along the line from $\mathbf{x}1$ to $\mathbf{x}_2$ where g_1 and g_2 cross in the linear approaximation.

Algorithm 0 (Binary search)

```
1 begin initialize x1, x2, ε
2         while (true)
3                 g(x1) ← g2(x1) − g2(x2)
4                 g(x2) ← g1(x2) − g2(x2)
5                 x3 ← x1 + (x2 − x1)/[‖g(x1)‖ − ‖g(x2)‖]
6                 if g(x3) ≥ 0 then x1 ← x3 else x2 ← x3
7                 if ‖g(x3)‖ < ε then return x3
8         end
```

32. Problem not yet solved

Section 9.6

33. Problem not yet solved

34. Our data come from either a uniform distribution or a Gaussian.

(a) Clearly, the maximum-likelihood values for the limits of the uniform distribution are the lower and upper limits within the data set: $x_l = 0.2$ and $x_u = 0.9$.

(b) The maximum-likelihood values of the mean and standard deviation are just those statistics of the data set, which turn out to be $\mu = 0.5143$ and $\sigma = 0.2231$, respectively.

(c) It is necessary to integrate over a region Δx about a given data point to convert from probability density to probability. The probability of the data given a particular model is then the product of the probability of each data point given the model (uniform θ_U or Gaussian θ_G), that is,

$$\begin{aligned} P(\mathcal{D}|\theta_U) &= 12.1427(\Delta x)^7 \\ P(\mathcal{D}|\theta_G) &= 1.7626(\Delta x)^7, \end{aligned}$$

and thus a uniform distribution is a better model for this data.

35. Problem not yet solved
36. Problem not yet solved
37. Problem not yet solved
38. Problem not yet solved
39. Problem not yet solved
40. Problem not yet solved
41. Problem not yet solved
42. Problem not yet solved
43. Problem not yet solved

Section 9.7

44. Problem not yet solved
45. Problem not yet solved

Computer Exercises

Section 9.2

1. Computer exercise not yet solved

Section 9.3

2. Computer exercise not yet solved

Section 9.4

3. Computer exercise not yet solved

Section 9.5

4. Computer exercise not yet solved
5. Computer exercise not yet solved
6. Computer exercise not yet solved

Section 9.6

7. Computer exercise not yet solved
8. Computer exercise not yet solved
9. Computer exercise not yet solved

Section 9.7

10. Computer exercise not yet solved

Chapter 10

Unsupervised learning and clustering

Problem Solutions

Section 10.2

1. We are given that x can assume values $0, 1, \ldots, m$ and that the priors $P(\omega_j)$ are known.

(a) The likelihood is given by

$$P(x|\boldsymbol{\theta}) = \sum_{j=1}^{c} \binom{m}{x} \theta_j^x (1-\theta_j)^{m-x} P(\omega_j),$$

with normalization constraint $\sum_{x=0}^{m} P(x|\boldsymbol{\theta}) = 1$, for all $\boldsymbol{\theta}$. Thus $P(x|\boldsymbol{\theta})$ represents m independent equations in the c unknowns $\theta_1, \ldots, \theta_c$; there are multiple solutions if $c > m$ and we do not have identifiability.

(b) On page 519 in the text, the case of $m = 1, c = 2$ was shown to be completely unidentifiable but that the sum $\theta_1 + \theta_2$ *could* be identified. In the present case, too, an m-dimensional subspace of $(\theta_1, \ldots, \theta_c)$ is identifiable. Whether it is *completely* identifiable or not depends on the actual values for $P(x|\theta)$ that are observed. For example, when $m = 1, c = 2$ if $P(x = 1|\theta) = 1$, then $\theta_1 + \theta_2 = 2$ and thus $\theta_1 = \theta_2 = 1$, and we have complete identifiability. Thus, in general, nothing can be said about the complete identifiability of $(\theta_1, \ldots, \theta_c)$, though an m-dimensional subspace will be completely identifiable.

(c) Suppose now that the priors $P(\omega_j)$ are unknown. Then, there are a total of $c+(c-1) = 2c-1$ unknown parameters — c of them the unknowns $\theta_1, \ldots, \theta_c$ and

$c-1$ them the unknown $P(\omega_j)$ reduced by the single constraint $\sum_{j=1}^{c} P(\omega_j) = 1$. Thus, the problem is not identifiable if $2c - 1 > m$.

2. PROBLEM NOT YET SOLVED

3. We are given the mixture density

$$P(x|\boldsymbol{\theta}) = P(\omega_1)\frac{1}{\sqrt{2\pi}\sigma_1}e^{-x^2/(2\sigma_1^2)} + (1 - P(\omega_1))\frac{1}{\sqrt{2\pi}\sigma_2}e^{-x^2/(2\sigma_2^2)}.$$

(a) When $\sigma_1 = \sigma_2$, then $P(\omega_1)$ can take any value in the range $[0, 1]$, leaving the same mixture density. Thus the density is completely unidentifiable.

(b) If $P(\omega_1)$ is fixed (and known) but not $P(\omega_1) = 0, 0.5$, or 1.0, then the model is identifiable. For those three values of $P(\omega_1)$, we cannot recover parameters for the first distribution. If $P(\omega_1) = 1$, we cannot recover parameters for the second distribution. If $P(\omega_1) = 0.5$, the parameters of the two distributions are interchangeable.

(c) If $\sigma_1 = \sigma_2$, then $P(\omega_1)$ cannot be identified because $P(\omega_1)$ and $P(\omega_2)$ are interchangeable. If $\sigma_1 \neq \sigma_2$, then $P(\omega_1)$ can be determined uniquely.

Section 10.3

4. We are given that $\mathbf{x}$ is a binary vector and that $P(\mathbf{x}|\theta)$ is a mixture of c multivariate Bernoulli distributions:

$$P(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{c} P(\mathbf{x}|\omega_i, \boldsymbol{\theta})P(\omega_i),$$

where

$$P(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = \prod_{j=1}^{d} \theta_{ij}^{x_{ij}}(1 - \theta_{ij})^{1-x_{ij}}.$$

(a) We consider the log-likelihood

$$\ln\, P(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) \;=\; \sum_{j=1}^{d} \left[x_{ij}\ln\, \theta_{ij} + (1 - x_{ij})\ln\, (1 - \theta_{ij})\right],$$

and take the derivative

$$\begin{aligned}
\frac{\partial \ln\, P(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)}{\partial \theta_{ij}} &= \frac{x_{ij}}{\theta_{ij}} - \frac{1 - x_{ij}}{1 - \theta_{ij}} \\
&= \frac{x_{ij}(1 - \theta_{ij}) - \theta_{ij}(1 - x_{ij})}{\theta_{ij}(1 - \theta_{ij})} \\
&= \frac{x_{ij} - x_{ij}\theta_{ij} - \theta_{ij} + \theta_{ij}x_{ij}}{\theta_{ij}(1 - \theta_{ij})} \\
&= \frac{x_{ij} - \theta_{ij}}{\theta_{ij}(1 - \theta_{ij})}.
\end{aligned}$$

We set this to zero, which can be expressed in a more compact form as

$$\sum_{k=1}^{n} \hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i)\frac{x_k - \hat{\boldsymbol{\theta}}_i}{\hat{\boldsymbol{\theta}}_i(1 - \hat{\boldsymbol{\theta}}_i)} = 0.$$

(b) Equation 7 in the text shows that the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_i$ must satisfy

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i} \ln\, P(x_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = 0.$$

We can write the equation from part (a) in component form as

$$\nabla_{\boldsymbol{\theta}_i} \ln\, P(x_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = \frac{x_k \hat{\boldsymbol{\theta}}_i}{\hat{\boldsymbol{\theta}}_i(1-\hat{\boldsymbol{\theta}}_i)},$$

and therefore we have

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\frac{\mathbf{x}_k - \hat{\boldsymbol{\theta}}_i}{\hat{\boldsymbol{\theta}}_i(1-\hat{\boldsymbol{\theta}}_i)} = 0.$$

We assume $\hat{\boldsymbol{\theta}}_i \in (0, 1)$, and thus we have

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)(\mathbf{x}_k - \hat{\boldsymbol{\theta}}_i) = \mathbf{0},$$

which gives the solution

$$\hat{\boldsymbol{\theta}}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)x_k}{\sum_{k=1}^{n} \hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i)}.$$

(c) Thus $\hat{\boldsymbol{\theta}}_i$, the maximum-likelihood estimate of $\boldsymbol{\theta}_i$, is a weighted average of the $\mathbf{x}_k$'s, with the weights being the posteriori probabilities of the mixing weights $\hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)$ for $k = 1, \ldots, n$.

5. We have a c-component mixture of Gaussians with each component of the form

$$p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) \sim N(\boldsymbol{\mu}_i, \sigma_i^2\mathbf{I}),$$

or more explicitly,

$$p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = \frac{1}{(2\pi)^{d/2}\sigma_i^d}\exp\left[-\frac{1}{2\sigma_i^2}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i)\right].$$

We take the logarithm and find

$$\ln\, p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = -\frac{d}{2}\ln\,(2\pi) - \frac{d}{2}\ln\,\sigma_i^2 - \frac{1}{2\sigma_i^2}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i),$$

and the derivative with respect to the variance is

$$\begin{aligned}\frac{\partial \ln\, p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)}{\partial \sigma_i^2} &= -\frac{d}{2\sigma_i^2} + \frac{1}{2\sigma_i^4}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i) \\ &= \frac{1}{2\sigma_i^4}(-d\sigma_i^2 + \|\mathbf{x}-\boldsymbol{\mu}_i\|^2).\end{aligned}$$

The maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_i$ must satisfy Eq. 12 in the text, that is,

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \nabla_{\boldsymbol{\theta}_i} \ln\, p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = \mathbf{0}.$$

We set the derivative with respect to σ_i^2 to zero, that is,

$$\begin{aligned} \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \frac{\partial \ln\, p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i)}{\partial \sigma_i^2} &= \\ \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \frac{1}{2\hat{\sigma}_i^4}(-d\hat{\sigma}_i^2 + \|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2) &= 0, \end{aligned}$$

rearrange, and find

$$d\hat{\sigma}_i^2 \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) = \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2.$$

The solution is

$$\hat{\sigma}_i^2 = \frac{\frac{1}{d} \sum\limits_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2}{\sum\limits_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)},$$

where $\hat{\boldsymbol{\mu}}_i$ and $\hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)$, the maximum-likelihood estimates of $\boldsymbol{\mu}_i$ and $P(\omega_i|\mathbf{x}_k, \boldsymbol{\theta}_i)$, are given by Eqs. 11–13 in the text.

6. Our c-component normal mixture is

$$p(\mathbf{x}|\alpha) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \alpha) P(\omega_j),$$

and the sample log-likelihood function is

$$l = \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\alpha).$$

We take the derivative with respect to α and find

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= \sum_{k=1}^{n} \frac{\partial \ln\, p(\mathbf{x}_k|\alpha)}{\partial \alpha} = \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k, \alpha)} \frac{\partial p(\mathbf{x}_k, \alpha)}{\partial \alpha} \\ &= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k, \alpha)} \frac{\partial}{\partial \alpha} \sum_{l=1}^{c} p(\mathbf{x}_k|\omega_j, \alpha) P(\omega_j) \\ &= \sum_{k=1}^{n} \sum_{j=1}^{c} \frac{p(\mathbf{x}_k|\omega_j, \alpha) P(\omega_j)}{p(\mathbf{x}_k, \alpha)} \frac{\partial}{\partial \alpha} \ln\, p(\mathbf{x}_k|\omega_j, \alpha) \\ &= \sum_{k=1}^{n} \sum_{j=1}^{c} P(\omega_j|\mathbf{x}_k, \alpha) \frac{\partial \ln\, p(\mathbf{x}_k|\omega_j, \alpha)}{\partial \alpha}, \end{aligned}$$

where by Bayes' Theorem we used

$$P(\omega_j|\mathbf{x}_k,\alpha)=\frac{p(\mathbf{x}_k|\omega_j,\alpha)P(\omega_j)}{p(\mathbf{x}_k|\alpha)}.$$

7. Our c-component normal mixture is

$$p(\mathbf{x}|\alpha)=\sum_{j=1}^{c}p(\mathbf{x}|\omega_j,\alpha)P(\omega_j),$$

and the sample log-likelihood function is

$$l=\sum_{k=1}^{n}\ln\, p(\mathbf{x}_k|\alpha).$$

We take the derivative with respect to α and find

$$\begin{aligned}\frac{\partial l}{\partial\alpha} &= \sum_{k=1}^{n}\frac{\partial \ln\, p(\mathbf{x}_k|\alpha)}{\partial\alpha}=\sum_{k=1}^{n}\frac{1}{p(\mathbf{x}_k,\alpha)}\frac{\partial p(\mathbf{x}_k,\alpha)}{\partial\alpha}\\ &= \sum_{k=1}^{n}\frac{1}{p(\mathbf{x}_k,\alpha)}\frac{\partial}{\partial\alpha}\sum_{l=1}^{c}p(\mathbf{x}_k|\omega_j,\alpha)P(\omega_j)\\ &= \sum_{k=1}^{n}\sum_{j=1}^{c}\frac{p(\mathbf{x}_k|\omega_j,\alpha)P(\omega_j)}{p(\mathbf{x}_k,\alpha)}\frac{\partial}{\partial\alpha}\ln\, p(\mathbf{x}_k|\omega_j,\alpha)\\ &= \sum_{k=1}^{n}\sum_{j=1}^{c}P(\omega_j|\mathbf{x}_k,\alpha)\frac{\partial \ln\, p(\mathbf{x}_k|\omega_j,\alpha)}{\partial\alpha},\end{aligned}$$

where by Bayes' Theorem we used

$$P(\omega_j|\mathbf{x}_k,\alpha)=\frac{p(\mathbf{x}_k|\omega_j,\alpha)P(\omega_j)}{p(\mathbf{x}_k|\alpha)}.$$

8. We are given that θ_1 and θ_2 are statistically independent, that is, $p(\theta_1,\theta_2)=p(\theta_1)p(\theta_2)$.

(a) We use this assumption to derive

$$\begin{aligned}p(\theta_1,\theta_2|x_1) &= \frac{p(\theta_1,\theta_2,x_1)}{p(x_1)}=\frac{p(x_1|\theta_1,\theta_2)p(\theta_1,\theta_2)}{p(x_1)}\\ &= \frac{p(x_1|\theta_1,\theta_2)p(\theta_1)p(\theta_2)}{p(x_1)}\\ &= [p(x_1|\omega_1,\theta_1,\theta_2)P(\omega_1)+p(x_1|\omega_2,\theta_1,\theta_2)P(\omega_2)]\frac{p(\theta_1)p(\theta_2)}{p(x_1)}\\ &= [p(x_1|\omega_1,\theta_1)P(\omega_1)+p(x_1|\omega_2,\theta_2)P(\omega_2)]\frac{p(\theta_1)p(\theta_2)}{p(x_1)}.\end{aligned}$$

Therefore, $p(\theta_1,\theta_2|x_1)$ can be factored as $p(\theta_1|x_1)p(\theta_2|x_1)$ if and only if

$$p(x_1|\omega_1,\theta_1)P(\omega_1)+p(x_1|\omega_2,\theta_2)P(\omega_2)$$

can be written as a product of two terms, one involving only θ_1 and the other involving θ_2. Now, for any choice of $P(\omega_1)$, we have $P(\omega_2) = 1 - P(\omega_1)$. This then implies that

$$p(x_1|\omega_1, \theta_1)P(\omega_1) + p(x_1|\omega_2, \theta_2)P(\omega_2)$$

can be factored in terms of θ_1 and θ_2 if and only if $p(x_1|\omega_i, \theta_i)$ is *constant* in θ_i. This, in turn, implies that $p(\theta_1, \theta_2|x_1)$ cannot be factored if

$$\frac{\partial \ln\, p(x_1|\omega_i, \theta_i)}{\partial \theta_i} \neq 0$$

for $i = 1, 2$.

(b) Posteriors of parameters need not be independent even if the priors are assumed independent (unless of course the mixture component densities are parameter free). To see this, we first note $0 < P(\omega_j) < 1$ and $n_j/n \simeq P(\omega_j)$ as $n \to \infty$, and in that case we have

$$\begin{aligned}
&\max_{\mu_1,\ldots,\mu_c} \frac{1}{n} \ln\, p(x_1, \ldots, x_n|\mu_1, \ldots, \mu_c) \\
\simeq\ & \sum_{j=1}^{c} P(\omega_j) \ln\, P(\omega_j) - \frac{1}{2} \ln\, (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{c} P(\omega_j)\sigma^2 \\
=\ & \sum_{j=1}^{c} P(\omega_j) \ln\, P(\omega_j) - \frac{1}{2} \ln\, (2\pi\sigma^2) - \frac{1}{2} \\
=\ & \sum_{j=1}^{c} P(\omega_j) \ln\, P(\omega_j) - \frac{1}{2} \ln\, (2\pi\sigma^2 e).
\end{aligned}$$

9. Problem not yet solved

10. We consider the log-likelihood function for data set $\mathcal{D}$,

$$\begin{aligned}
l &= \ln\, p(\mathcal{D}|\boldsymbol{\theta}) \\
&= \ln \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}) \\
&= \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}),
\end{aligned}$$

where

$$p(\mathbf{x}_k|\boldsymbol{\theta}) = \sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j)P(\omega_j).$$

We seek to maximize l with respect to $\boldsymbol{\theta}$ and $P(\omega_i)$, subject to the constraints that

$$P(\omega_i) \geq 0 \quad \text{and} \quad \sum_{j=1}^{c} P(\omega_j) = 1.$$

We use the method of Lagrange undetermined multipliers, and define the objective function (to be maximized) as

$$f(\boldsymbol{\theta}, P(\omega_1), \ldots, P(\omega_c)) = l + \lambda_o \left(\sum_{i=1}^{c} P(\omega_i) - 1 \right).$$

This form of objective function guarantees that the normalization constraint is obeyed. We thus demand that at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$

$$\frac{\partial f(\boldsymbol{\theta}, P(\omega_1), \ldots, P(\omega_c))}{\partial \boldsymbol{\theta}_i} = 0$$

for $i = 1, \ldots, c$. Furthermore, the derivative obeys

$$\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{\theta}_i} = \frac{\partial l}{\partial \boldsymbol{\theta}_i} = \nabla_{\boldsymbol{\theta}_i} l &= \sum_{k=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_i} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}) \\
&= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}_i} p(\mathbf{x}_k|\boldsymbol{\theta}) \\
&= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}_i} \left[\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j) P(\omega_j) \right] \\
&= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k, \boldsymbol{\theta})} \left[\frac{\partial}{\partial \boldsymbol{\theta}_i} p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) \right] P(\omega_i) \\
&= \sum_{k=1}^{n} \frac{p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) P(\omega_i)}{p(\mathbf{x}_k, \boldsymbol{\theta})} \frac{1}{p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i)} \frac{\partial}{\partial \boldsymbol{\theta}_i} p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) \\
&= \sum_{k=1}^{n} P(\omega_i|\mathbf{x}_k, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \ln\, p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i).
\end{aligned}$$

Thus at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ the derivative obeys $\nabla_{\boldsymbol{\theta}_i} f = \mathbf{0}$, or equivalently

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\theta}_i} \ln\, p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = \mathbf{0},$$

where

$$\begin{aligned}
\hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}) &= \frac{p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) \hat{P}(\omega_i)}{p(\mathbf{x}_k|\hat{\boldsymbol{\theta}})} \\
&= \frac{p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) \hat{P}(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \hat{\boldsymbol{\theta}}_j) \hat{P}(\omega_j)},
\end{aligned}$$

as given by Eq. 13 in the text.

We continue with the derivative with respect to the priors

$$\frac{\partial f}{\partial P(\omega_i)} = \frac{\partial}{\partial P(\omega_i)} \left[\sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}) + \lambda_o \left(\sum_{i=1}^{c} P(\omega_i) - 1 \right) \right]$$

$$\begin{aligned} &= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \frac{\partial}{\partial P(\omega_i)} \sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j) P(\omega_j) + \lambda_o \\ &= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) + \lambda_o. \end{aligned}$$

We evaluate this at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $P(\omega_i) = \hat{P}(\omega_i)$ and find

$$\begin{aligned} \frac{\partial f}{\partial P(\omega_i)} &= \sum_{k=1}^{n} \frac{p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i)}{p(\mathbf{x}_k|\hat{\boldsymbol{\theta}})} + \lambda_o \\ &= \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}})}{\hat{P}(\omega_i)} + \lambda_o = 0. \end{aligned}$$

This implies

$$\hat{P}(\omega_i) = -\frac{1}{\lambda_o} \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}),$$

and since $\sum_{i=1}^{c} \hat{P}(\omega_i) = 1$ we have

$$-\frac{1}{\lambda_o} \sum_{k=1}^{n} \sum_{i=1}^{c} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}) = 1,$$

or $\lambda_o = -n$ where

$$\hat{P}(\omega_i) = \frac{1}{n} \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}).$$

Section 10.4

11. Our Gaussian densities are of the form $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, and we use the following terminology:

$$\begin{aligned} \sigma_{pq} &= pq\text{th element of } \boldsymbol{\Sigma} \\ \sigma^{pq} &= pq\text{th element of } \boldsymbol{\Sigma}^{-1} \\ x_p(k) &= p\text{th element of } \mathbf{x}_k \\ \mu_p(i) &= p\text{th element of } \boldsymbol{\mu}_i. \end{aligned}$$

(a) We write the class-conditional density as a Gaussian,

$$p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_i)\right].$$

The log-likelihood is

$$\ln p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_i),$$

as given by Eq. 21 in the text. We take the derivative and find

$$\frac{\partial \ln\, p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i)}{\partial \sigma^{pq}} = \frac{\partial}{\partial \sigma^{pq}}\left[-\tfrac{d}{2}\ln\,(2\pi) + \tfrac{1}{2}\ln\,|\boldsymbol{\Sigma}^{-1}| \right.$$
$$\left. -\tfrac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_i)\right]. \quad (*)$$

Now we also have that the squared Mahalanobis distance is

$$(\mathbf{x}_k-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_i) = \sum_{p'=1}^{d}\sum_{q'=1}^{d}(x_{p'}(k)-\mu_{p'}(i))\,\sigma^{p'q'}\,(x_{q'}(k)-\mu_{q'}(i))\,,$$

and its derivative is

$$\frac{\partial\left[(\mathbf{x}_k-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_i)\right]}{\partial\sigma^{pq}} = \begin{cases} (x_p(k)-\mu_p(i))(x_q(k)-\mu_q(i)) & \\ +(x_q(k)-\mu_q(i))(x_p(k)-\mu_p(i)) & \text{if} \quad p\neq q \\ (x_p(k)-\mu_p(i))^2 & \text{if} \quad p=q \end{cases}$$
$$= \begin{cases} 2(x_p(k)-\mu_p(i))(x_q(k)-\mu_q(i)) & \text{if} \quad p\neq q \\ (x_p(k)-\mu(i))^2 & \text{if} \quad p=q. \end{cases} \quad (**)$$

In component form, the derivative of the determinant of the inverse covariance is

$$\frac{\partial \ln\,|\boldsymbol{\Sigma}^{-1}|}{\partial\sigma^{pq}} = \frac{1}{|\boldsymbol{\Sigma}^{-1}|}\frac{\partial|\boldsymbol{\Sigma}^{-1}|}{\partial\sigma^{pq}}$$
$$= \frac{1}{|\boldsymbol{\Sigma}^{-1}|}\left[|(\boldsymbol{\Sigma}^{-1})_{pq}| + |(\boldsymbol{\Sigma}^{-1})_{qp}|\right],$$

where we have used the fact that $\sigma^{pq}=\sigma^{qp}$ (i.e., both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ are symmetric), and the notation $(\boldsymbol{\Sigma}^{-1})_{ij}$ = (i,j)th minors of $\boldsymbol{\Sigma}^{-1}$ with ith row and jth column deleted. We can use the above results and

$$\left|\boldsymbol{\Sigma}^{-1}\right| = \sum_{q'=1}^{d}\sigma^{pq'}\left|(\boldsymbol{\Sigma}^{-1})_{pq'}\right|$$

to write

$$\frac{\partial \ln\,\left|\boldsymbol{\Sigma}^{-1}\right|}{\partial\sigma^{pq}} = \begin{cases} \frac{\left|\left(\boldsymbol{\Sigma}^{-1}\right)_{pq}\right|}{\left|\boldsymbol{\Sigma}^{-1}\right|} + \frac{\left|\left(\boldsymbol{\Sigma}^{-1}\right)_{qp}\right|}{\left|\boldsymbol{\Sigma}^{-1}\right|} & \text{if} \quad p\neq q \\ \frac{\left|\left(\boldsymbol{\Sigma}^{-1}\right)_{pq}\right|}{\left|\boldsymbol{\Sigma}^{-1}\right|} & \text{if} \quad p=q \end{cases}$$
$$= \begin{cases} 2\sigma_{pq} & \text{if} \quad p\neq q \\ \sigma_{pq} & \text{if} \quad p=q. \end{cases} \quad (***)$$

We substitute $(**)$ and $(***)$ into $(*)$ and find

$$\frac{\partial \ln\, p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i)}{\partial\sigma^{pq}} = \begin{cases} \sigma_{pq}-(x_p(k)-\mu_p(i))(x_q(k)-\mu_q(i)) & \text{if} \quad p\neq q \\ \frac{1}{2}[\sigma_{pq}-(x_p(k)-\mu_p(i))(x_q(k)-\mu_q(i))] & \text{if} \quad p=q \end{cases}$$

$$\begin{aligned} &= \left(1 - \frac{\delta_{pq}}{2}\right)\left[\sigma_{pq} - (x_p(k) - \mu_p(i))(x_q(k) - \mu_q(i))\right], \\ \text{where } \delta_{pq} &= \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases} \end{aligned}$$

is the Kronecker symbol.

(b) The derivatives of the log-likelihood function are

$$\frac{\partial l}{\partial \alpha} = \sum_{k=1}^{n}\sum_{j=1}^{c} P(\omega_j|\mathbf{x}_k, \alpha)\frac{\partial \ln\, p(\mathbf{x}_k|\omega_j, \alpha)}{\partial \alpha}$$

and

$$\begin{aligned} \frac{\partial l}{\partial \sigma^{pq}} &= \sum_{k=1}^{n}\sum_{j=1}^{c} P(\omega_j|\mathbf{x}_k, \boldsymbol{\mu}_j, \boldsymbol{\Sigma})\frac{\partial \ln\, p(\mathbf{x}_k|\omega_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma})}{\partial \sigma^{pq}} \\ &= \sum_{k=1}^{n}\sum_{j=1}^{c} P(\omega_j|\mathbf{x}_k, \boldsymbol{\theta}_j)\left(1 - \frac{\delta_{pq}}{2}\right)\left[\sigma_{pq} - (x_p(k) - \mu_p(j))(x_q(k) - \mu_q(j))\right]. \end{aligned}$$

At the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_j$ of $\boldsymbol{\theta}_j$, we have $\partial l/\partial \sigma^{pq} = 0$, and therefore

$$\left(1 - \frac{\delta_{pq}}{2}\right)\hat{\sigma}_{pq} = \frac{\left(1 - \frac{\delta_{pq}}{2}\right)\sum_{k=1}^{n}\sum_{j=1}^{c}\hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j)(x_p(k) - \hat{\mu}_p(j))(x_q(k) - \hat{\mu}_q(j))}{\sum_{k=1}^{n}\sum_{j=1}^{c}\hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j)},$$

which yields the maximum-likelihood estimate of the pq-entry to the covariance matrix, that is,

$$\hat{\sigma}_{pq} = \frac{\sum_{k=1}^{n}\sum_{j=1}^{c}\hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j)(x_p(k) - \hat{\mu}_p(j))(x_q(k) - \hat{\mu}_q(j))}{\sum_{k=1}^{n}\sum_{j=1}^{c}\hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j)}.$$

In vector and matrix form, this result can be written just a bit more compactly as

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{k=1}^{n}\sum_{j=1}^{c}\hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j)^t}{\sum_{k=1}^{n}\sum_{j=1}^{c}\hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j)},$$

where we used Eq. 13 in the text,

$$\hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j) = \frac{P(\mathbf{x}_k|\omega_j, \hat{\boldsymbol{\theta}}_j)\hat{P}(\omega_j)}{\sum_{i=1}^{c} p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)},$$

and the normalization constraint

$$\sum_{j=1}^{c}\hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j) = 1.$$

We can also write

$$\begin{aligned}
&\sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_j)(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_j)^t \\
&= \mathbf{x}_k\mathbf{x}_k^t \sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j) - \mathbf{x}_{kj} \sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)\hat{\boldsymbol{\mu}}_j^t \\
&\quad - \left[\sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)\hat{\boldsymbol{\mu}}_j\right]\mathbf{x}_k^t + \sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)\boldsymbol{\mu}_j\boldsymbol{\mu}_j^t \\
&= \mathbf{x}_k\mathbf{x}_k^t - \mathbf{x}_k \sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)\boldsymbol{\mu}_j^t - \left[\sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)\boldsymbol{\mu}_j\right]\mathbf{x}_k^t \\
&\quad + \sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)\boldsymbol{\mu}_j\boldsymbol{\mu}_j^t \\
&= \sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^t - \sum_{j=1}^{c}\left[\sum_{k=1}^{n} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)\mathbf{x}_k\right]\boldsymbol{\mu}_j^t \\
&\quad - \sum_{j=1}^{c}\boldsymbol{\mu}_j\left[\sum_{k=1}^{n} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)\mathbf{x}_k^t\right] + \sum_{j=1}^{c}\left[\sum_{k=1}^{n} \hat{P}(\omega_j|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j)\mathbf{x}_k\right]\boldsymbol{\mu}_j\boldsymbol{\mu}_j^t \\
&= \sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^t - \sum_{j=1}^{c} n\hat{P}(\omega_j)\boldsymbol{\mu}_j\boldsymbol{\mu}_j^t - \sum_{j=1}^{c} n\hat{P}(\omega_j)\boldsymbol{\mu}_j\boldsymbol{\mu}_j^t + \sum_{j=1}^{c} \hat{P}(\omega_j)\boldsymbol{\mu}_j\boldsymbol{\mu}_j^t \\
&= \sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^t - n\sum_{j=1}^{c} \hat{P}(\omega_j)\boldsymbol{\mu}_j\boldsymbol{\mu}_j^t.
\end{aligned}$$

We rearrange this result, substitute it above and find that the maximum-likelihood estimate of the covariance matrix is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^t - \sum_{j=1}^{c} \hat{P}(\omega_j)\boldsymbol{\mu}_j\boldsymbol{\mu}_j^t.$$

12. We are told that the distributions are normal with parameters given by $p(x|\omega_1) \sim N(0,1)$ and $p(x|\omega_2) \sim N(0,1/2)$. Thus the evidence is

$$p(x) = \frac{P(\omega_1)}{\sqrt{2\pi}}e^{-x^2/2} + \frac{1-P(\omega_1)}{\sqrt{\pi}}e^{-x^2}.$$

(a) If one sample x_1 is observed, the likelihood function is then

$$\begin{aligned}
l = p(x_1) &= \frac{P(\omega_1)}{\sqrt{2\pi}}e^{-x_1^2/2} + \frac{1-P(\omega_1)}{\sqrt{\pi}}e^{-x_1^2} \\
&= \left[\frac{e^{-x_1^2/2}}{\sqrt{2\pi}} - \frac{e^{-x_1^2}}{\sqrt{\pi}}\right]P(\omega_1) + \frac{e^{-x_1^2}}{\sqrt{\pi}}.
\end{aligned}$$

Thus l is indeed linear in $P(\omega_1)$, and hence must be extremal at one of the limiting values of $P(\omega_1)$, that is at either $P(\omega_1) = 0$ or $P(\omega_1) = 1$. (We ignore

the case where l is independent of $P(\omega_1)$.) Thus l is maximized at $P(\omega_1) = 0$ if the slope of l is negative, i.e., if

$$\frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} - \frac{1}{\sqrt{\pi}} e^{-x_1^2} < 0.$$

This is equivalent to the inequality

$$\frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} < \frac{1}{\sqrt{\pi}} e^{-x_1^2},$$

or simply $x_1^2 < \ln 2$. Thus if $x_1^2 < \ln 2$ the maximum-likelihood estimate of $P(\omega_1)$ is $\hat{P}(\omega_1) = 0$.

(b) If the slope of l is *positive*, then l is maximized at $P(\omega_1) = 1$. A positive slope implies

$$\frac{e^{-x_1^2/2}}{\sqrt{2\pi}} - \frac{e^{-x_1^2}}{\sqrt{\pi}} > 0,$$

or simply $x_1^2 > \ln 2$. Thus if $x_1^2 > \ln 2$, the maximum-likelihood estimate of $P(\omega_1)$ is $\hat{P}(\omega_1) = 1$.

(c) If there is a single sample, we are forced to infer that the maximum-likelihood estimate for the prior of *one* of the categories is 1.0, and the other is 0.0, depending upon which density is greater at the sampled point.

13. We assume that ther real line $\mathbf{R}^1$ can be divided into c non-overlapping intervals $\Omega_1, \ldots, \Omega_c$ such that $\bigcup_{j=1}^{c} \Omega_j = \mathbf{R}^1$ (the real line) and $\Omega_j \bigcap \Omega_{j'} = \emptyset$ for $j \neq j'$. We are given that at any point x, only *one* of the Gaussian functions differs significantly from zero, and thus

$$p(x|\mu_1, \ldots, \mu_c) \simeq \frac{P(\omega_j)}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x - \mu_j)^2\right].$$

We thus have for some particular j that depends upon x

$$\begin{aligned}
\frac{1}{n} \ln p(x_1, \ldots, x_n|\mu_1, \ldots, \mu_c) &= \frac{1}{n} \sum_{k=1}^{n} \ln p(x_k|\mu_1, \ldots, \mu_c) \\
&= \frac{1}{n} \sum_{j=1}^{c} \sum_{k:x_k \in \Omega_j} \ln p(x_k|\mu_1, \ldots, \mu_c) \\
&\simeq \frac{1}{n} \sum_{j=1}^{c} \sum_{k:x_k \in \Omega_j} \ln \left[\frac{P(\omega_j)}{\sqrt{2\pi}\sigma} e^{-(x_k - \mu_j)^2/(2\sigma^2)}\right].
\end{aligned}$$

We take logarithms on the right-hand side, rearrange, and find

$$\begin{aligned}
\frac{1}{n} \ln p(x_1, \ldots, x_n|\mu_1, \ldots, \mu_c) &\simeq \frac{1}{n} \sum_{j=1}^{c} \sum_{k:x_k \in \Omega_j} \left[\ln P(\omega_j) - \frac{1}{2} \ln (2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_k - \mu_j)^2\right] \\
&= \frac{1}{n} \sum_{j=1}^{c} \ln P(\omega_j) \sum_{k:x_k \in \Omega_j} 1 - \frac{1}{n} \frac{1}{2} \ln (2\pi\sigma^2) \sum_{j=1}^{c} \sum_{k:x_k \in \Omega_j} 1
\end{aligned}$$

$$
\begin{aligned}
& -\frac{1}{n}\sum_{j=1}^{c}\sum_{k:x_k\in\Omega_j}\frac{1}{2\sigma^2}(x_k-\mu_j)^2 \\
= \quad & \frac{1}{n}\sum_{j=1}^{c}P(\omega_j)n_j-\frac{1}{2}\ln\,(2\pi\sigma^2)-\frac{1}{n}\sum_{j=1}^{c}\sum_{k:x_k\in\Omega_j}(x_k-\mu_j)^2,
\end{aligned}
$$

where $n_j=\sum_{k:x_k\in\Omega_j}1$ is the number of points in the interval Ω_j. The result above implies

$$
\begin{aligned}
& \max_{\mu_1,\ldots,\mu_c}\frac{1}{n}\ln\,p(x_1,\ldots,x_n|\mu_1,\ldots,\mu_c) \\
\simeq \quad & \frac{1}{n}\sum_{j=1}^{c}n_j\ln\,P(\omega_j)-\frac{1}{2}\ln\,(2\pi\sigma^2)+\frac{1}{n}\sum_{j=1}^{c}\max_{\mu_j}\sum_{k:x_k\in\Omega_j}[-(x_k-\mu_j)^2].
\end{aligned}
$$

However, we note the fact that

$$
\max_{\mu_j}\sum_{k:x_k\in\Omega_j}[-(x_k-\mu_j)^2]
$$

occurs at

$$
\begin{aligned}
\hat{\mu}_j &= \frac{\sum_{k:x_k\in\Omega_j}x_k}{\sum_{k:x_k\in\Omega_j}1} \\
&= \frac{\sum_{k:x_k\in\Omega_j}x_k}{n_j} \\
&= \bar{x}_j,
\end{aligned}
$$

for some interval, j say, and thus we have

$$
\begin{aligned}
& \max_{\mu_1,\ldots,\mu_c}\frac{1}{n}p(x_1,\ldots,x_n|\mu_1,\ldots,\mu_c) \\
\simeq \quad & \frac{1}{n}\sum_{j=1}^{n}n_j\ln\,P(\omega_j)-\frac{1}{2}\ln\,(2\pi\sigma^2)-\frac{1}{2\sigma^2}\frac{1}{n}\sum_{j=1}^{c}\sum_{k:x_k\in\Omega_j}(x_k-\bar{x}_j)^2 \\
= \quad & \frac{1}{n}\sum_{j=1}^{c}n_j\ln\,P(\omega_j)-\frac{1}{2}\ln\,(2\pi\sigma^2)-\frac{1}{2\sigma^2}\frac{1}{n}\sum_{j=1}^{c}n_j\frac{1}{n_j}\sum_{k':x_k\in\Omega_j}(x_k-\bar{x}_j)^2.
\end{aligned}
$$

Thus if $n\to\infty$ (i.e., the number of independently drawn samples is very large), we have $n_j/n=$ the proportion of total samples which fall in Ω_j, and this implies (by the law of large numbers) that we obtain $P(\omega_j)$.

14. We let the mean value be denoted

$$
\bar{\mathbf{x}}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.
$$

Then we have

$$
\frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{x})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\mathbf{x})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\mathbf{x})
$$

$$\begin{aligned}
&= \frac{1}{n}\Big[\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\bar{\mathbf{x}}) \\
&\quad +2(\bar{\mathbf{x}}-\mathbf{x})^t\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}})+n(\bar{\mathbf{x}}-\mathbf{x})^t\boldsymbol{\Sigma}^{-1})(\bar{\mathbf{x}}-\mathbf{x})\Big] \\
&= \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k\bar{\mathbf{x}})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k\bar{\mathbf{x}})+(\bar{\mathbf{x}}-\mathbf{x})^t\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\mathbf{x}) \\
&\geq \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\bar{\mathbf{x}}),
\end{aligned}$$

where we used

$$\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}})=\sum_{k=1}^{n}\mathbf{x}_k-n\bar{\mathbf{x}}=n\bar{\mathbf{x}}-n\bar{\mathbf{x}}=\mathbf{0}.$$

Since $\boldsymbol{\Sigma}$ is positive definite, we have

$$(\bar{\mathbf{x}}-\mathbf{x})^t\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\mathbf{x})\geq 0,$$

with strict inequality holding if and only if $\mathbf{x}\neq\bar{\mathbf{x}}$. Thus

$$\frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{x})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\mathbf{x})$$

is minimized at $\mathbf{x}=\bar{\mathbf{x}}$, that is, at

$$\mathbf{x}=\bar{\mathbf{x}}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$$

15. PROBLEM NOT YET SOLVED

16. The basic operation of the algorithm is the computation of the distance between a sample and the center of a cluster which takes $O(d)$ time since each dimension needs to be compared seperately. During each iteration of the algorithm, we have to classify each sample with respect to each cluster center, which amounts to a total number of $O(nc)$ distance computations for a total complexity $O(ncd)$. Each cluster center than needs to be updated, which takes $O(cd)$ time for each cluster, therefore the update step takes $O(cd)$ time. Since we have T iterations of the classification and update step, the total time complexity of the algorithm is $O(Tncd)$.

17. We derive the equations as follows.

(a) From Eq. 14 in the text, we have

$$\ln\, p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i)=\ln\,\frac{|\boldsymbol{\Sigma}_i^{-1}|^{1/2}}{(2\pi)^{d/2}}-\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_i).$$

It was shown in Problem 11 that

$$\frac{\partial \ln\, p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i)}{\partial\sigma_{pq}(i)}=\left(1-\frac{\delta_{pq}}{2}\right)[\sigma_{pq}(i)-(x_p(k)-\mu_p(i))(x_q(k)-\mu_q(i))].$$

We write the squared Mahalanobis distance as

$$\begin{aligned}(\mathbf{x}_k-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_i) &= \sum_{p=1}^{d}(x_{p'}(k)-\mu_{p'}(i))^2\sigma^{p'p'}(i)\\ &\quad+\sum_{p'=1}^{d}\sum_{\substack{q'=1\\p\neq q}}^{d}(x_{p'}(k)-\mu_{p'}(i))\sigma^{p'q'}(i)(x_{q'}(k)-\mu_{q'}(i)).\end{aligned}$$

The derivative of the likelihood with respect to the pth coordinate of the ith mean is

$$\begin{aligned}\frac{\partial \ln\, p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i)}{\partial\mu_p(i)} &= \frac{\partial}{\partial\mu_p(i)}\Big[-\frac{1}{2}\sum_{p'=1}^{d}(x_{p'}(k)-\mu_{p'}(i))^2\sigma^{p'p'}(i)\\ &\quad-\frac{1}{2}\sum_{p'=1}^{d}\sum_{\substack{q'=1\\p\neq q}}^{d}(x_{p'}(k)-\mu_{p'}(i))\sigma^{p'q'}(i)(x_{q'}-\mu_{q'}(i))\Big]\\ &= -\frac{1}{2}2(x_p(k)-\mu_p(i))(-1)\sigma^{pp}(i)\\ &\quad-\frac{1}{2}\sum_{\substack{q'=1\\q'\neq p}}^{d}(x_{q'}(k)-\mu_{q'}(i))\sigma^{pq'}(i)(-1)\\ &\quad-\frac{1}{2}\sum_{\substack{p'=1\\p'\neq p}}^{d}(x_{p'}(k)-\mu_{p'}(i))\sigma^{pq'}(i)(-1)\\ &= \sum_{q=1}^{d}(x_q(k)-\mu_q(i))\sigma^{pq}(i).\end{aligned}$$

From Eq. xxx in the text, we know that $\hat{P}(\omega_i)$ and $\hat{\boldsymbol{\theta}}_i$ must satisfy

$$\hat{P}(\omega_i)=\frac{1}{n}\sum_{k=1}^{n}\hat{P}(\omega_i|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_i)$$

and

$$\sum_{k=1}^{n}\hat{P}(\omega_i|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i}\ln\, p(\mathbf{x}_k|\omega_i,\hat{\boldsymbol{\theta}}_i)=\mathbf{0}$$

or

$$\sum_{k=1}^{n}\hat{P}(\omega_i|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_i)\frac{\partial \ln\, p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i)}{\partial\mu_p(i)}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i}=0.$$

We evaluate the derivative, as given above, and find

$$\sum_{k=1}^{n}\hat{P}(\omega_i|\mathbf{x}_k,\hat{\boldsymbol{\theta}}_i)\sum_{q=1}^{d}(x_q(k)-\hat{\mu}_q(i))\hat{\sigma}^{pq}(i)=0,$$

which implies

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_k - \hat{\mu_i}) = 0.$$

We solve for $\hat{\boldsymbol{\mu}}_i$ and find

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\mathbf{x}_k}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)}.$$

Furthermore, we have

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\frac{\partial \ln\, p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i)}{\partial \sigma^{pq}}\Big|_{\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i} = 0,$$

and this gives

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\left(1 - \frac{\delta pq}{2}\right)[\hat{\sigma}_{pq}(i) - (x_p(k) - \hat{\mu}_p(i))(x_q(k) - \hat{\mu}_q(i))] = 0.$$

We solve for $\hat{\sigma}_{pq}(i)$ and find

$$\hat{\sigma}_{pq}(i) = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)[(x_p(k) - \hat{\mu}_p(i))(x_q(k) - \hat{\mu}_q(i))]}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)}$$

and thus

$$\hat{\boldsymbol{\Sigma}}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)^t}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)}.$$

(b) For the conditions of this part, we have $\sigma_i = \sigma_i^2\mathbf{I}$ and thus

$$\ln\, p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) = \ln\, \frac{1}{(2\pi\sigma_i^2)^{\frac{d}{2}}} - \frac{1}{2\sigma_i^2}(\mathbf{x}_k - \boldsymbol{\mu}_i)^t(\mathbf{x}_k - \boldsymbol{\mu}_i),$$

and thus it is easy to verify that

$$\nabla_{\boldsymbol{\mu}_i} \ln\, p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) = \frac{1}{\sigma_i^2}(\mathbf{x}_k - \boldsymbol{\mu}_i).$$

and hence

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\frac{1}{\sigma_i^2}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i) = 0.$$

Thus the mean

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\mathbf{x}_k}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)}$$

does not change. This implies

$$\frac{\partial \ln\, p(\mathbf{x}_k|\omega_i, \theta_i)}{\partial \sigma_i^2} = -\frac{d}{2\sigma_i^2} + \frac{1}{2\sigma_i^4}||\mathbf{x}_k - \boldsymbol{\mu}_i||^2.$$

We have, moreover,

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \left[\frac{\partial \ln\, p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i)}{\partial \sigma_i^2} \Big|_{\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i} \right] = 0$$

and thus

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) \left[-\frac{d}{2\hat{\sigma_i^2}} + \frac{1}{2\hat{\sigma_i^4}||\mathbf{x}_k - \hat{\mu_i}||^2} \right] = 0$$

and this gives

$$\hat{\sigma}_i^2 = \frac{\frac{1}{d}\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)||\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i||^2}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)}.$$

(c) In this case, the covariances are equal: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$. Clearly the maximum-likelihood estimate of $\boldsymbol{\mu}_i$, which does not depend on the choice of $\boldsymbol{\Sigma}_i$, will remain the same. Therefore, we have

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\mathbf{x}_k}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)}.$$

The log-likelihood function is

$$\begin{aligned} l &= \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}) \\ &= \sum_{k=1}^{n} \ln \left[\sum_{j=1}^{c} p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j) P(\omega_j) \right] \\ &= \sum_{k=1}^{n} \ln \left[\sum_{j=1}^{c} P(\omega_j) \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{d/2}} \exp \left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_j) \right] \right] \end{aligned}$$

and hence the derivative is

$$\begin{aligned} \frac{\partial l}{\partial \sigma^{pq}} &= \sum_{k=1}^{n} \frac{\partial \ln\, p(\mathbf{x}_k|\boldsymbol{\theta})}{\partial \sigma^{pq}} \\ &= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \frac{\partial}{\partial \sigma^{pq}} p(\mathbf{x}_k|\boldsymbol{\theta}) \\ &= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \sum_{j=1}^{c} P(\omega_j) \frac{\partial}{\partial \sigma^{pq}} p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j) \\ &= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\boldsymbol{\theta})} \sum_{j=1}^{c} P(\omega_j) p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j) \frac{\partial}{\partial \sigma^{pq}} \ln\, p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j) \end{aligned}$$

However, we have from Problem 17,

$$\frac{\partial \ln\, p(\mathbf{x}_k|\omega_j, \boldsymbol{\theta}_j)}{\partial \sigma^{pq}} = \left(1 - \frac{\delta_{pq}}{2}\right) \left[\sigma_{pq} - (x_p(k) - \mu_p(j))(x_q(k) - \mu_q(j))\right].$$

We take derivatives and set

$$\frac{\partial l}{\partial \sigma^{pq}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$$

and consequently

$$\sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\hat{\boldsymbol{\theta}})} \sum_{j=1}^{c} P(\omega_j) p(\mathbf{x}_k|\omega_j, \hat{\boldsymbol{\theta}}) \left(1 - \frac{\delta_{pq}}{2}\right) [\hat{\sigma_{pq}} - (x_p(k) - \hat{\mu_p}(j))(x_q(k) - \hat{\mu_q}(j))] = 0.$$

This gives the equation

$$\hat{\sigma}_{pq} \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\hat{\boldsymbol{\theta}})} \sum_{j=1}^{c} P(\omega_j) p(\mathbf{x}_k|\omega_j, \hat{\boldsymbol{\theta}}_j)$$

$$= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k|\hat{\boldsymbol{\theta}})} \sum_{j=1}^{c} P(\omega_j) p(\mathbf{x}_k|\omega_j, \hat{\boldsymbol{\theta}}) \left[(x_p(k) - \hat{\mu_p}(i))(x_q(k) - \hat{\mu_q}(i))\right].$$

We have, therefore, $\hat{\sigma}^{pq} = \sum_{k=1}^{n} 1$, since

$$\begin{aligned} \sum_{j=1}^{c} P(\omega_j) p(\mathbf{x}_k|\omega_j, \hat{\boldsymbol{\theta}}) &= p(\mathbf{x}_k, \hat{\boldsymbol{\theta}}) \\ &= \sum_{k=1}^{n} \sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j)[(x_p(k) - \hat{\mu}_q(i))(x_q(k) - \hat{\mu}_q(i))] \end{aligned}$$

and this implies that the estimated covariance matrix is

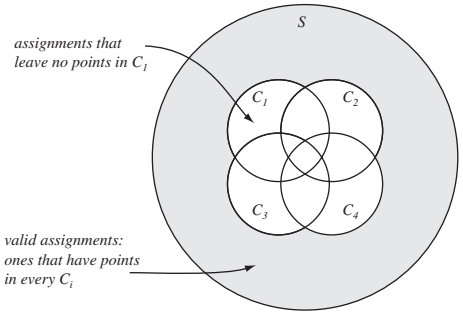
$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=1}^{n} \sum_{j=1}^{c} \hat{P}(\omega_j|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j)^t.$$

Section 10.5

18. We shall make use of the following figure, where the largest circle represents all possible assignments of n points to c clusters, and the smaller circle labeled C_1 represents assignments of the n points that do *not* have points in cluster C_1, and likewise for C_2 up through C_c (the case $c = 4$ is shown). The gray region represents those assigments that are valid, that is, assignments in which none of the C_i are empty. Our goal is then to compute the number of such valid assignments, that is the cardinality of the gray region.

(a) First we find the cardinality of *invalid* assigments, that is the number of assignments in the white region:

$$\begin{aligned} &\sum_{i=1}^{c} |C_i| - \sum_{i \neq j}^{c} |C_i \cap C_j| + \sum_{i \neq j \neq k}^{c} |C_i \cap C_j \cap C_k| - \cdots \pm \sum_{i \neq j \neq k \ldots} \underbrace{|C_i \cap C_j \cap C_k \cdots C_q|}_{c \text{ terms}} \\ &= \binom{c}{1} N_1 - \binom{c}{2} N_2 + \binom{c}{3} N_3 - \cdots \pm \binom{c}{c} N_c \\ &= \sum_{i=1}^{c} \binom{c}{i} (-1)^{c-i+1} N_i, \end{aligned}$$



where we define N_i to be the number of distinct assignments of n points that leave exactly i (labeled) clusters empty. Since each of the n points can be assigned to one of $i-1$ clusters, we have $N_i = (i-1)^n$. We put these intermediate results together and find that the number of invalid assignments (where the clusters are labeled) is

$$\sum_{i=1}^{c} \binom{c}{i} (-1)^{c-i+1} (i-1)^n.$$

But the above assumed that the clusters were labeled, that is, that we could distinguish between C_1 and C_2, for instance. In clustering, no such labeling is given. Thus the above overcounts by the number of ways we can assign c labels to c clusters, that is, by $c!$. Therefore, the cardinality of the white area in the figure (the number of invalid assignments) is the above result divided by $c!$, that is

$$\frac{1}{c!} \sum_{i=1}^{c} \binom{c}{i} (-1)^{c-i+1} (i-1)^n.$$

The number of *valid* assignments (the gray region in the figure) is thus the total number of assignments (indicated by S in the figure) minus the number of invalid assignments (the white region). This total number is $\binom{c}{c} i^n$. To find the total number of valid assignments, we subtract:

$$\binom{c}{c} i^n - \sum_{i=1}^{c} \binom{c}{i} (-1)^{c-i+1} (i-1)^n.$$

We perform a substitution $i \leftarrow i-1$ and regroup to obtain our final answer:

$$\frac{1}{c!} \sum_{i=1}^{c} (-1)^{c-i} i^n.$$

(b) For the case $c = 5$ and $n = 100$ the number of clusterings is

$$\begin{aligned}
&\frac{1}{5!} \sum_{i=1}^{5} \binom{5}{i} (-1)^{5-i} i^{100} \\
&= \frac{1}{5!} \left[5 \cdot 1^{100} + 10 \cdot 2^{100} + 10 \cdot 3^{100} + 5 \cdot 4^{100} + 1 \cdot 5^{100} \right] \\
&= 65738408701461898606895733752711432902699495364788241645840659777500 \\
&\simeq 6.57 \times 10^{67}.
\end{aligned}$$

(c) As given in part (a), the number of distinct clusterings of 1000 points into 10 clusters is

$$\frac{1}{10!}\sum_{i=1}^{10}\binom{10}{i}(-1)^{10-i}i^{1000}.$$

Clearly, the term that dominates this is the case $i = 10$, since the 10^{1000} is larger than any other in the sum. We use Stirling's aproximation, $x! \simeq x^x e^{-x}\sqrt{2\pi x}$, which is valid for large x. Thus we approximate

$$\begin{aligned}\frac{1}{10!}10^{1000} &= \frac{10^{10}}{10!}10^{990}\\ &\simeq \frac{e^{10}}{\sqrt{2\pi 10}}10^{990}\\ &\simeq 2778\times 10^{990}\\ &\simeq 2.778\times 10^{993}.\end{aligned}$$

Incidentally, this result approximates quite well the exact number of clusterings,

275573192239858906525573192239858906525573191758192446064084102276
241562179050768703410910649475589239352323968422047450227142883275
301297357171703073190953229159974914411337998286379384317700201827
699064518600319141893664588014724019101677001433584368575734051087
113663733326712187988280471413114695165527579301182484298326671957
439022946882788124670680940407335115534788347130729348996760498628
758235529529402633330738877542418150010768061269819155260960497017
554445992771571146217248438955456445157253665923558069493011112420
464236085383593253820076193214110122361976285829108747907328896777
952841254396283166368113062488965007972641703840376938569647263827
074806164568508170940144906053094712298457248498509382914917294439
593494891086897486875200023401927187173021322921939534128573207161
833932493327072510378264000407671043730179941023044697448254863059
191921920705755612206303581672239643950713829187681748502794419033
667715959241433673878199302318614319865690114473540589262344989397
5880
$\simeq 2.756\times 10^{993}$.

Section 10.6

19. We show that the ranking of distances between samples is invariant to any monotonic transformation of the dissimilarity values.

(a) We define the *value* v_k at level k to be min $\delta(\mathcal{D}_i, \mathcal{D}_j)$, where $\delta(\mathcal{D}_i, \mathcal{D}_j)$ is the dissimilarity between pairs of clusters $\mathcal{D}_i$ and $\mathcal{D}_j$. Recall too that by definition in hierarchical clustering if two clusters are joined at some level (say k), they

remain joined for all subsequent levels. (Below, we assume that only two clusters are merged at any level.)

Suppose at level k that two clusters, $\mathcal{D}$ and $\mathcal{D}'$ are merged into $\mathcal{D}^*$ at level $k+1$. Then we have

$$\begin{aligned} v_{k+1} &= \min_{\substack{\mathcal{D}_i,\mathcal{D}_j \\ \text{distinct at } k}} [\delta(\mathcal{D}_i,\mathcal{D}_j)] \\ &= \min\left[\min_{\substack{\mathcal{D}_i,\mathcal{D}_j\neq\mathcal{D}^* \\ \text{distinct at } k}} [\delta(\mathcal{D}_i,\mathcal{D}_j)], \min_{\substack{\mathcal{D}_i\neq\mathcal{D}^* \\ \text{is cluster at } k}} [\delta(\mathcal{D}_i,\mathcal{D}_j)]\right] \\ &= \min\left[\min_{\substack{\mathcal{D}_i,\mathcal{D}_j\neq\mathcal{D},\mathcal{D}' \\ \text{are distinct at } k-1}} [\delta(\mathcal{D}_i,\mathcal{D}_j)], \min_{\substack{\mathcal{D}_i\neq\mathcal{D},\mathcal{D}' \\ \text{is cluster at } k-1}} [\delta(\mathcal{D}_i,\mathcal{D}\cup\mathcal{D}')]\right]. \end{aligned}$$

If $\delta = \delta_{min}$, then we have

$$\begin{aligned} \min_{\substack{\mathcal{D}_i\neq\mathcal{D},\mathcal{D}' \\ \text{is cluster at } k-1}} [\delta(\mathcal{D}_i,\mathcal{D}\cup\mathcal{D}')] &= \min_{\substack{\mathcal{D}_i\neq\mathcal{D},\mathcal{D}' \\ \text{is cluster at } k-1}} \min_{\substack{x\in\mathcal{D}_i \\ x'\in\mathcal{D}\cup\mathcal{D}'}} \delta(\mathbf{x},\mathbf{x}') \\ &= \min_{\substack{\mathcal{D}_i\neq\mathcal{D},\mathcal{D}' \\ \text{is cluster at } k-1}} \min\left[\min_{\substack{x\in\mathcal{D}_i \\ x'\in\mathcal{D}}} \delta(\mathbf{x},\mathbf{x}'), \min_{\substack{x\in\mathcal{D}_i \\ x'\in\mathcal{D}'}} \delta(\mathbf{x},\mathbf{x}')\right] \\ &= \min_{\substack{\mathcal{D}_i\neq\mathcal{D},\mathcal{D}' \\ \text{is cluster at } k-1}} \min[\delta(\mathcal{D}_i,\mathcal{D}),\delta(\mathcal{D}_i,\mathcal{D}')] \\ &\geq \min_{\substack{\mathcal{D}_i,\mathcal{D}_j \\ \text{are distance at } k-1}} [\delta(\mathcal{D}_i,\mathcal{D}_j]\,. \end{aligned}$$

Therefore we can conclude that

$$\begin{aligned} v_{k+1} &= \min\left[\min_{\substack{\mathcal{D}_i,\mathcal{D}_j\neq\mathcal{D},\mathcal{D}' \\ \text{are distinct at } k-1}} [\delta(\mathcal{D}_i,\mathcal{D}_j),\delta(\mathcal{D}_i,\mathcal{D}),\delta(\mathcal{D}_i,\mathcal{D}')]\right] \\ &\geq v_k \end{aligned}$$

and thus $v_{k+1} \geq v_k$ for all k. A similar argument goes through for $\delta = \delta_{max}$.

(b) From part (a) we have $v_{k+1} \geq v_k$ for all k, and thus

$$0 = v_1 \leq v_2 \leq v_3 \leq \cdots \leq v_n,$$

so the similarity values retain their ordering.

Section 10.7

20. It is sufficient to derive the equation for one single cluster, which allows us to drop the cluster index. We have

$$\begin{aligned} \sum_{\mathbf{x}\in\mathcal{D}} \|\mathbf{x}-\mathbf{m}\|^2 &= \sum_{\mathbf{x}\in\mathcal{D}} \left\|\mathbf{x} - \frac{1}{n}\sum_{\mathbf{x}'\in\mathcal{D}} \mathbf{x}'\right\|^2 \\ &= \sum_{\mathbf{x}\in\mathcal{D}} \left\|\frac{1}{n}\sum_{\mathbf{x}'\in\mathcal{D}} \mathbf{x}-\mathbf{x}'\right\|^2 \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{\mathbf{x}\in\mathcal{D}}\Big\|\sum_{\mathbf{x}'\in\mathcal{D}}\mathbf{x}-\mathbf{x}'\Big\|^2\\
&= \frac{1}{n^2}\sum_{\mathbf{x}\in\mathcal{D}}\sum_{\mathbf{x}'\in\mathcal{D}}\sum_{\mathbf{x}''\in\mathcal{D}}(\mathbf{x}-\mathbf{x}')^t(\mathbf{x}-\mathbf{x}'')\\
&= \frac{1}{n^2}\frac{1}{2}\Big[\sum_{\mathbf{x}\in\mathcal{D}}\sum_{\mathbf{x}'\in\mathcal{D}}\sum_{\mathbf{x}''\in\mathcal{D}}(\mathbf{x}-\mathbf{x}')^t(\mathbf{x}-\mathbf{x}'')\\
&\qquad\qquad +\sum_{\mathbf{x}\in\mathcal{D}}\sum_{\mathbf{x}'\in\mathcal{D}}\sum_{\mathbf{x}''\in\mathcal{D}}(\mathbf{x}-\mathbf{x}')^t[(\mathbf{x}-\mathbf{x}')+(\mathbf{x}'-\mathbf{x}'')]\Big]\\
&= \frac{1}{n^2}\frac{1}{2}\Big[\sum_{\mathbf{x}\in\mathcal{D}}\sum_{\mathbf{x}'\in\mathcal{D}}\sum_{\mathbf{x}''\in\mathcal{D}}(\mathbf{x}-\mathbf{x}')^t(\mathbf{x}-\mathbf{x}'')\\
&\qquad\qquad +\sum_{\mathbf{x}\in\mathcal{D}}\sum_{\mathbf{x}'\in\mathcal{D}}\sum_{\mathbf{x}''\in\mathcal{D}}\|\mathbf{x}-\mathbf{x}'\|^2+(\mathbf{x}-\mathbf{x}')^t(\mathbf{x}'-\mathbf{x}'')\Big]\\
&= \frac{1}{n^2}\frac{1}{2}\sum_{\mathbf{x}\in\mathcal{D}}\sum_{\mathbf{x}'\in\mathcal{D}}\sum_{\mathbf{x}''\in\mathcal{D}}\|\mathbf{x}-\mathbf{x}'\|^2\\
&\quad +\frac{1}{n^2}\frac{1}{2}\underbrace{\Big[\sum_{\mathbf{x}\in\mathcal{D}}\sum_{\mathbf{x}'\in\mathcal{D}}\sum_{\mathbf{x}''\in\mathcal{D}}(\mathbf{x}-\mathbf{x}')^t(\mathbf{x}-\mathbf{x}'')-\sum_{\mathbf{x}\in\mathcal{D}}\sum_{\mathbf{x}'\in\mathcal{D}}\sum_{\mathbf{x}''\in\mathcal{D}}(\mathbf{x}'-\mathbf{x})^t(\mathbf{x}'-\mathbf{x}'')\Big]}_{=0}\\
&= \frac{1}{2}n\frac{1}{n^2}\sum_{\mathbf{x}\in\mathcal{D}}\sum_{\mathbf{x}'\in\mathcal{D}}\|\mathbf{x}-\mathbf{x}'\|^2\\
&= \frac{1}{2}n\bar{s}.
\end{aligned}
$$

21. We employ proof by contradiction. From Problem 14 with $\mathbf{\Sigma}=\mathbf{I}$, we know that for a non-empty set of samples $\mathcal{D}_i$ that

$$\sum_{\mathbf{x}\in\mathcal{D}_i}\|\mathbf{x}-\boldsymbol{\alpha}_i\|^2$$

is minimized at $\mathbf{x}=\mathbf{m}_i$, the mean of the points in $\mathcal{D}_i$. Now our criterion function is

$$J_e=\sum_{\mathcal{D}_i\neq\phi}\sum_{\mathbf{x}\in\mathcal{D}_i}\|\mathbf{x}-\mathbf{m}_i\|^2.$$

Suppose that the partition minimizing J_e has an empty subset. Then $n\geq c$ and there must exist at least one subset in the partition containing two or more samples; we call that subset $\mathcal{D}_j$. We write this as

$$\mathcal{D}_j=\mathcal{D}_{j1}\bigcup\mathcal{D}_{j2},$$

where $\mathcal{D}_{j1}$ and $\mathcal{D}_{j2}$ are disjoint and non-empty. Then we conclude that

$$J_e=\sum_{\mathcal{D}_i\neq\phi}\sum_{\mathbf{x}\in\mathcal{D}_i}\|\mathbf{x}-\mathbf{m}_i\|^2$$

can be written as

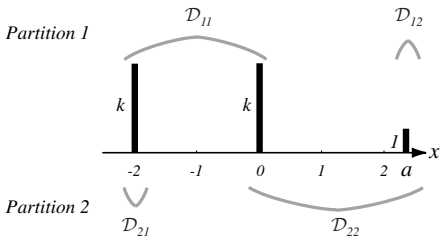
$$J_e=A+\sum_{\mathbf{x}\in\mathcal{D}_j}\|\mathbf{x}-\mathbf{m}_j\|^2,$$

where A is some constant. However, we also have

$$\begin{aligned}\sum_{\mathbf{x}\in\mathcal{D}_j}\|\mathbf{x}-\mathbf{m}_j\|^2 &= \sum_{\mathbf{x}\in\mathcal{D}_{j1}}\|\mathbf{x}-\mathbf{m}_j\|^2+\sum_{\mathbf{x}\in\mathcal{D}_{j2}}\|\mathbf{x}-\mathbf{m}_j\|^2\\ &> \sum_{\mathbf{x}\in\mathcal{D}_{j1}}\|\mathbf{x}-\mathbf{m}_{j1}\|^2+\sum_{\mathbf{x}\in\mathcal{D}_{j2}}\|\mathbf{x}-\mathbf{m}_{j2}\|^2.\end{aligned}$$

Thus, replacing $\mathcal{D}_j$ by $\mathcal{D}_{j1}$ and $\mathcal{D}_{j2}$ and removing the empty subset yields a partition into c disjoint subsets that reduces J_e. But by assumption the partition chosen *minimized* J_e, and hence we have a contradiction. Thus there can be no empty subsets in a partition that minimizes J_e (if $n \geq c$). In short, we should partition into as many subsets are allowed, since fewer subsets means larger dispersion within subsets.

22. The figure shows our data and terminology.



(a) We have $n = 2k+1$ points to be placed in $c = 2$ clusters and of course the clusters should be non-empty. Thus the only two cases to consider are those shown in the figure above.

In **Partition 1**, we have

$$\begin{aligned}m_{11} &= \frac{-2k+0\cdot k}{2k} = -1,\\ m_{12} &= a,\end{aligned}$$

and thus the value of our cluster criterion function is

$$\begin{aligned}J_{e1} &= \sum_{x\in\mathcal{D}_{11}}(x-m_{11})^2+\sum_{x\in\mathcal{D}_{12}}(x-m_{12})^2\\ &= \sum_{i=1}^{k}(-2+1)^2+\sum_{i=1}^{k}(0+1)^2+(a-a)^2\\ &= k+k+0=2k.\end{aligned}$$

In **Partition 2**, we have

$$\begin{aligned}m_{21} &= -2,\\ m_{22} &= \frac{k\cdot 0+a}{k+1}=\frac{a}{k+1},\end{aligned}$$

and thus the value of our cluster criterion function is

$$J_{e2} = \sum_{x\in\mathcal{D}_{21}}(x-m_{21})^2+\sum_{x\in\mathcal{D}_{22}}(x-m_{22})^2$$

$$\begin{aligned}
&= \sum_{x\in\mathcal{D}_{21}}(-2+2)^2+\sum_{i=1}^{k}\left(0-\frac{a}{k+1}\right)^2+\left(a-\frac{a}{k+1}\right)^2\\
&= 0+\frac{a^2}{(k+1)^2}k+\frac{a^2k^2}{(k+1)^2}\\
&= \frac{a^2k(k+1)}{(k+1)^2}=\frac{a^2k}{k+1}.
\end{aligned}$$

Thus if $J_{e2}<J_{e1}$, that is, if $a^2/(k+1)<2k$ or equivalently $a^2<2(k+1)$, then the partition that minimizes J_e is Partition 2, which groups the k samples at $x=0$ with the one sample at $x=a$.

(b) If $J_{e1}<J_{e2}$, i.e., $2k<a^2/(k+1)$ or equivalently $2(k+1)>a^2$, then the partition that minimzes J_e is Partition 1, which groups the k-samples at $x=-2$ with the k samples at $x=0$.

23. Our sum-of-square (scatter) criterion is $\mathrm{tr}[\mathbf{S}_W]$. We thus need to calculate $\mathbf{S}_W$, that is,

$$\begin{aligned}
\mathbf{S}_W &= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t\\
&= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}[\mathbf{x}\mathbf{x}^t-\mathbf{m}_i\mathbf{x}^t-\mathbf{x}\mathbf{m}_i^t+\mathbf{m}_i\mathbf{m}^t]\\
&= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}\mathbf{x}\mathbf{x}^t-\sum_{i=1}^{c}m_1n_i\mathbf{m}_i^t-\sum_{i=1}^{c}n_i\mathbf{m}_i\mathbf{m}_i^t+\sum_{i=1}^{c}n_i\mathbf{m}_i\mathbf{m}_i^t\\
&= \sum_{k=1}^{4}\mathbf{x}_k\mathbf{x}_k^t-\sum_{i=1}^{c}n_i\mathbf{m}_i\mathbf{m}_i^t,
\end{aligned}$$

where n_i is the number of samples in $\mathcal{D}_i$ and

$$\mathbf{m}_i=\frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i}\mathbf{x}.$$

For the data given in the problem we have the following:

$$\begin{aligned}
\sum_{k=1}^{4}\mathbf{x}_k\mathbf{x}_k^t &= \begin{pmatrix}4\\5\end{pmatrix}\begin{pmatrix}4&5\end{pmatrix}+\begin{pmatrix}1\\4\end{pmatrix}\begin{pmatrix}1&4\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}0&1\end{pmatrix}+\begin{pmatrix}5\\0\end{pmatrix}\begin{pmatrix}5&0\end{pmatrix}\\
&= \begin{pmatrix}16&20\\20&25\end{pmatrix}+\begin{pmatrix}1&4\\4&16\end{pmatrix}+\begin{pmatrix}0&0\\0&1\end{pmatrix}+\begin{pmatrix}25&0\\0&0\end{pmatrix}\\
&= \begin{pmatrix}42&24\\24&42\end{pmatrix}.
\end{aligned}$$

Partition 1: Our means are

$$\begin{aligned}
\mathbf{m}_1 &= \frac{1}{2}\left(\begin{pmatrix}4\\5\end{pmatrix}+\begin{pmatrix}1\\4\end{pmatrix}\right)=\begin{pmatrix}5/2\\9/2\end{pmatrix},\\
\mathbf{m}_2 &= \frac{1}{2}\left(\begin{pmatrix}0\\1\end{pmatrix}+\begin{pmatrix}5\\0\end{pmatrix}\right)=\begin{pmatrix}5/2\\1/2\end{pmatrix},
\end{aligned}$$

and thus the matrix products are

$$\mathbf{m}_1\mathbf{m}_1^t = \begin{pmatrix} 25/4 & 45/4 \\ 45/4 & 81/4 \end{pmatrix} \text{ and } \mathbf{m}_2\mathbf{m}_2^t = \begin{pmatrix} 25/4 & 5/4 \\ 5/4 & 1/4 \end{pmatrix}.$$

Our scatter matrix is therefore

$$\mathbf{S}_W = \begin{pmatrix} 42 & 24 \\ 24 & 42 \end{pmatrix} - 2\begin{pmatrix} 25/4 & 45/4 \\ 45/4 & 81/4 \end{pmatrix} - 2\begin{pmatrix} 25/4 & 5/4 \\ 5/4 & 1/4 \end{pmatrix} = \begin{pmatrix} 17 & -1 \\ -1 & 1 \end{pmatrix},$$

and thus our criterion values are the trace

$$\text{tr}[\mathbf{S}_W] = \text{tr}\begin{pmatrix} 17 & -1 \\ -1 & 1 \end{pmatrix} = 17 + 1 = 18,$$

and the determinant

$$|\mathbf{S}_W| = 17 \cdot 1 - (-1) \cdot (-1) = 16.$$

Partition 2: Our means are

$$\begin{aligned} \mathbf{m}_1 &= \frac{1}{2}\left(\begin{pmatrix} 4 \\ 5 \end{pmatrix} + \begin{pmatrix} 5 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 9/2 \\ 5/2 \end{pmatrix}, \\ \mathbf{m}_2 &= \frac{1}{2}\left(\begin{pmatrix} 1 \\ 4 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 1/2 \\ 5/2 \end{pmatrix}, \end{aligned}$$

and thus our matrix products are

$$\mathbf{m}_1\mathbf{m}_1^t = \begin{pmatrix} 81/4 & 45/4 \\ 45/4 & 25/4 \end{pmatrix} \text{ and } \mathbf{m}_2\mathbf{m}_2^t = \begin{pmatrix} 1/4 & 5/4 \\ 5/4 & 25/4 \end{pmatrix}.$$

Our scatter matrix is therefore

$$\mathbf{S}_W = \begin{pmatrix} 42 & 24 \\ 24 & 42 \end{pmatrix} - 2\begin{pmatrix} 81/4 & 45/4 \\ 45/4 & 25/4 \end{pmatrix} - 2\begin{pmatrix} 1/4 & 5/4 \\ 5/4 & 25/4 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 17 \end{pmatrix},$$

and thus our criterion values are the trace

$$\text{tr}[\mathbf{S}_W] = \text{tr}\begin{pmatrix} 17 & -1 \\ -1 & 1 \end{pmatrix} = 1 + 17 = 18,$$

and the determinant

$$|\mathbf{S}_W| = 1 \cdot 17 - (-1) \cdot (-1) = 16.$$

Partition 3: Our means are

$$\begin{aligned} \mathbf{m}_1 &= \frac{1}{3}\left(\begin{pmatrix} 4 \\ 5 \end{pmatrix} + \begin{pmatrix} 1 \\ 4 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 5/3 \\ 3 \end{pmatrix}, \\ \mathbf{m}_2 &= \begin{pmatrix} 5 \\ 0 \end{pmatrix}. \end{aligned}$$

and thus our matrix products are

$$\mathbf{m}_1\mathbf{m}_1^t = \begin{pmatrix} 25/9 & 5 \\ 5 & 9 \end{pmatrix} \text{ and } \mathbf{m}_2\mathbf{m}_2^t = \begin{pmatrix} 25 & 0 \\ 0 & 0 \end{pmatrix}.$$

Our scatter matrix is therefore

$$\mathbf{S}_W = \begin{pmatrix} 42 & 24 \\ 24 & 42 \end{pmatrix} - 3\begin{pmatrix} 25/9 & 5 \\ 5 & 9 \end{pmatrix} - 1\begin{pmatrix} 25 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 26/3 & 22/3 \\ 22/3 & 26/3 \end{pmatrix},$$

and thus our criterion values are

$$\text{tr } \mathbf{S}_W = \text{tr } \begin{pmatrix} 17 & -1 \\ -1 & 1 \end{pmatrix} = 26/3 + 26/3 = 17.33,$$

and

$$|\mathbf{S}_W| = 26/3 \cdot 26/3 - 22/3 \cdot 22/3 = 21.33.$$

We summarize our results as

Partition	tr[$\mathbf{S}_W$]	$\lvert\mathbf{S}_W\rvert$
1	18	16
2	18	16
3	17.33	21.33

Thus for the tr[$\mathbf{S}_W$] criterion Partition 3 is favored; for the $|\mathbf{S}_W|$ criterion Partitions 1 and 2 are equal, and are to be favored over Partition 3.

24. PROBLEM NOT YET SOLVED

25. Consider a non-singular transformation of the feature space: $y = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a d-by-d non-singular matrix.

(a) If we let $\tilde{\mathcal{D}}_i = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{D}_i\}$ denote the data set transformed to the new space, then the scatter matrix in the transformed domain can be written as

$$\begin{aligned} \mathbf{S}_W^y &= \sum_{i=1}^{c} \sum_{\mathbf{y} \in \tilde{\mathcal{D}}_i} (\mathbf{y} - \mathbf{m}_i^y)(\mathbf{y} - \mathbf{m}_i^y)^t \\ &= \sum_{i=1}^{c} \sum_{\mathbf{y} \in \tilde{\mathcal{D}}_i} (\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i)(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i)^t \\ &= \mathbf{A} \sum_{i=1}^{c} \sum_{\mathbf{y} \in \tilde{\mathcal{D}}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \end{aligned}$$

where $\mathbf{A}^t = \mathbf{A}\mathbf{S}_W\mathbf{A}^t$. We also have the between-scatter matrix

$$\begin{aligned} \mathbf{S}_B^y &= \sum_{i=1}^{c} n_i(\mathbf{m}_i^y - \mathbf{m}^y)(\mathbf{m}_i - \mathbf{m}^y)^t \\ &= \sum_{i=1}^{c} n_i(\mathbf{A}\mathbf{m}_i - \mathbf{A}\mathbf{m})(\mathbf{A}\mathbf{m}_i - \mathbf{A}\mathbf{m})^t \\ &= \mathbf{A}\left[\sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t\right]\mathbf{A}^t \\ &= \mathbf{A}\mathbf{S}_B\mathbf{A}^t. \end{aligned}$$

The product of the inverse matrices is

$$\begin{aligned}[\mathbf{S}_W^y]^{-1}[\mathbf{S}_B^y]^{-1} &= (\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_B\mathbf{A}^t)\\ &= (\mathbf{A}^t)^{-1}\mathbf{S}_W^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{S}_B\mathbf{A}^t\\ &= (\mathbf{A}^t)^{-1}\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{A}^t.\end{aligned}$$

We let λ_i for $i = 1, \ldots, d$ denote the eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$. There exist vectors $\mathbf{z}_i, \ldots, \mathbf{z}_d$ such that

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{z}_i = \lambda_i\mathbf{z}_i,$$

for $i = 1, \ldots, d$, and this in turn implies

$$(\mathbf{A}^t)^{-1}\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{A}^t\mathbf{A}^t)^{-1}\mathbf{z}_i = \lambda_i(\mathbf{A}^t)^{-1}\mathbf{z}_i,$$

or

$$\mathbf{S}_W^{y-1}\mathbf{S}_B^y\mathbf{u}_i = \lambda_i\mathbf{u}_i,$$

where $\mathbf{u}_i = (\mathbf{A}^t)^{-1}\mathbf{z}_i$. This implies that $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{S}_W^{y-1}\mathbf{S}_B^y$, and finally that $\lambda_1, \ldots, \lambda_d$ are invariant to non-singular linear transformation of the data.

(b) Our total scatter matrix is $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$, and thus

$$\begin{aligned}\mathbf{S}_T^{-1}\mathbf{S}_W &= (\mathbf{S}_B + \mathbf{S}_W)^{-1}\mathbf{S}_W\\ &= [\mathbf{S}_W^{-1}(\mathbf{S}_B + \mathbf{S}_W)]^{-1}\\ &= [\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]^{-1}.\end{aligned}$$

If $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\mathbf{S}_W^{-1}\mathbf{S}_B$ and the $\mathbf{u}_1, \ldots, \mathbf{u}_d$ are the corresponding eigenvectors, then $\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{u}_i = \lambda_i\mathbf{u}_i$ for $i = 1, \ldots, d$ and hence

$$\mathbf{u}_i + \mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{u}_i = \mathbf{u}_i + \lambda_i\mathbf{u}_i.$$

This equation implies

$$[\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]\mathbf{u}_i = (1 + \lambda_i)\mathbf{u}_i.$$

We multiply both sides of the equation by $(1 + \lambda_i)^{-1}[\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]^{-1}$ and find

$$(1 + \lambda_i)^{-1}\mathbf{u}_i = [\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B]^{-1}\mathbf{u}_i$$

and this implies $\nu_i = 1/(1 + \lambda_i)$ are eigenvalues of $\mathbf{I} + \mathbf{S}_W^{-1}\mathbf{S}_B$.

(c) We use our result from part (a) and find

$$J_d = \frac{|\mathbf{S}_W|}{|\mathbf{S}_T|} = |\mathbf{S}_T^{-1}\mathbf{S}_W| = \prod_{i=1}^{d}\nu_i = \prod_{i=1}^{d}\frac{1}{1 + \lambda_i},$$

which is invariant to non-singular linear transformations described in part (a).

26. Consider a non-singular transformation of the feature space: $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a d-by-d non-singular matrix. We let $\mathcal{D}_i = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{D}^x\}$. We have

$$\begin{aligned}
\mathbf{S}_W^y &= \sum_{i=1}^{c} \sum_{\mathbf{y}\in\mathcal{D}_i} (\mathbf{y} - \mathbf{m}^y)(\mathbf{y} - \mathbf{m}^y)^t \\
&= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i^x)(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i^x)^t \\
&= \mathbf{A}\mathbf{S}_W^x\mathbf{A}^t.
\end{aligned}$$

In a similar way, we have

$$\begin{aligned}
\mathbf{S}_B^y &= \mathbf{A}\mathbf{S}_B^x\mathbf{A}^t \\
\mathbf{S}_t^y &= \mathbf{A}\mathbf{S}_t^x\mathbf{A}^t \\
(\mathbf{S}_t^y)^{-1}\mathbf{S}_W^y &= (\mathbf{A}^t)^{-1}(\mathbf{S}_t^x)^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{S}_W^x\mathbf{A}^t \\
&= (\mathbf{A}^t)^{-1}(\mathbf{S}_t^x)^{-1}\mathbf{S}_W^x\mathbf{A}^t \\
(\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y &= (\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{S}_B^x\mathbf{A}^t \\
&= (\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t.
\end{aligned}$$

(a) From problem 25 (b), we know that

$$\begin{aligned}
\mathrm{tr}[(\mathbf{S}_t^x)^{-1}\mathbf{S}_W^x] &= \sum_{i=1}^{d} \nu_i \\
&= \sum_{i=1}^{d} \frac{1}{1+\lambda_i}
\end{aligned}$$

as well as $\mathrm{tr}[\mathbf{B}^{-1}\mathbf{S}\mathbf{B}]$, because they have the same eigenvalues so long as $\mathbf{B}$ is non-singular. This is because if $\mathbf{S}\mathbf{x} = \nu_i\mathbf{x}$, then $\mathbf{S}\mathbf{B}\mathbf{B}^{-1}\mathbf{x} = \nu_i\mathbf{x}$, then also $\mathbf{B}^{-1}\mathbf{S}\mathbf{B}\mathbf{B}^{-1}\mathbf{x} = \mathbf{B}^{-1}\nu_i\mathbf{x} = \nu_i\mathbf{B}^{-1}\mathbf{x}$. Thus we have

$$\mathbf{B}^{-1}\mathbf{S}\mathbf{B}(\mathbf{B}^{-1}\mathbf{x}) = \nu_i(\mathbf{B}^{-1}\mathbf{x}).$$

We put this together and find

$$\begin{aligned}
\mathrm{tr}[(\mathbf{S}_t^y)^{-1}\mathbf{S}_W^y] &= \mathrm{tr}[(\mathbf{A}^t)^{-1}(\mathbf{S}_t^x)^{-1}\mathbf{S}_W^x\mathbf{A}^t] \\
&= \mathrm{tr}[(\mathbf{S}_t^x)^{-1}\mathbf{S}_W^x] \\
&= \sum_{i=1}^{d} \frac{1}{1+\lambda_i}.
\end{aligned}$$

(b) See Solution to Problem 25 part (c).

(c) Here we have the determinant

$$\begin{aligned}
|(\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y| &= |(\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t| \\
&= \prod \text{eigenvalues of } [(\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t] \\
&= \prod \text{eigenvalues of } [(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x] \\
&= \prod_{i=1}^{d} \lambda_i.
\end{aligned}$$

(d) The typical value of the criterion is zero or close to zero. This is because $\mathbf{S}_B$ is often singular, even if samples are not from a subspace. Even when $\mathbf{S}_B$ is not singular, some λ_i is likely to be very small, and this makes the product small. Hence the criterion is not always useful.

27. Equation 68 in the text defines the criterion $J_d = |\mathbf{S}_W| = \left|\sum_{i=1}^{c} \mathbf{S}_i\right|$, where

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$$

is the scatter matrix for category ω_i, defined in Eq. 61 in the text. We let $\mathbf{T}$ be a non-singular matrix and consider the change of variables $\mathbf{x}' = \mathbf{T}\mathbf{x}$.

(a) From the conditions stated, we have

$$\mathbf{m}_i' = \frac{1}{n_i} \sum_{\mathbf{x}'\in\mathcal{D}_{i'}} \mathbf{x}'$$

where n_i is the number of points in category ω_i. Thus we have the mean of the transformed data is

$$\mathbf{m}_i' = \frac{1}{n_i} \sum_{\mathbf{x}\in\mathcal{D}_i} \mathbf{T}\mathbf{x} = \mathbf{T}\mathbf{m}_i.$$

Furthermore, we have the transformed scatter matrix is

$$\begin{aligned}
\mathbf{S}_i' &= \sum_{\mathbf{x}'\in\mathcal{D}_i'} (\mathbf{x}'-\mathbf{m}_i')(\mathbf{x}'-\mathbf{m}_i')^t \\
&= \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{T}\mathbf{x}-\mathbf{T}\mathbf{m}_i)(\mathbf{T}\mathbf{x}-\mathbf{T}\mathbf{m}_i)^t \\
&= \mathbf{T}\left[\sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t\right]\mathbf{T}^t = \mathbf{T}\mathbf{S}_i\mathbf{T}^t.
\end{aligned}$$

(b) From the conditions stated by the problem, the criterion function of the transformed data must obey

$$\begin{aligned}
J_d' = |\mathbf{S}_W'| &= \left|\sum_{i=1}^{c} \mathbf{S}_i'\right| = \left|\sum_{i=1}^{c} \mathbf{T}\mathbf{S}_i\mathbf{T}^t\right| = \left|\mathbf{T}\left(\sum_{i=1}^{c} \mathbf{S}_i\right)\mathbf{T}^t\right| \\
&= |\mathbf{T}||\mathbf{T}^t|\left|\sum_{i=1}^{c} \mathbf{S}_i\right| \\
&= |\mathbf{T}|^2 J_d.
\end{aligned}$$

Therefore, J_d' differs from J_d only by an overall non-negative scale factor $|\mathbf{T}|^2$.

(c) Since J_d' differs from J_d only by a scale factor of $|\mathbf{T}|^2$ (which does not depend on the partitioning into clusters) J_d' and J_d will rank partitions in the same order. Hence the optimal clustering based on J_d is always the optimal clustering based on J_d'. Optimal clustering is invariant to non-singular linear transformations of the data .

28. Consider a non-singular transformation of the feature space $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a d-by-d non-singular matrix. We let $\mathcal{D}_i = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{D}_i^x\}$ be the transformed data set. We then have the within scatter matrix as

$$\begin{aligned}\mathbf{S}_W^y &= \sum_{i=1}^{c}\sum_{\mathbf{y}\in\mathcal{D}_i}(\mathbf{y}-\mathbf{m}_i^y)(\mathbf{y}-\mathbf{m}_i^y)^t \\ &= \sum_{i=1}^{c}\sum_{\mathbf{y}\in\mathcal{D}_i}(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{m}_i^x)(\mathbf{y}-\mathbf{A}\mathbf{m}_i^x)^t \\ &= \mathbf{A}\mathbf{S}_W^x\mathbf{A}^t.\end{aligned}$$

In a similar way, we have

$$\begin{aligned}\mathbf{S}_B^y &= \mathbf{A}\mathbf{S}_B^x\mathbf{A}^t \\ \mathbf{S}_t^y &= \mathbf{A}\mathbf{S}_t^x\mathbf{A}^t \\ (\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y &= (\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{S}_B^x\mathbf{A}^t \\ &= (\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t.\end{aligned}$$

If λ is an eigenvalue of $(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x$ with corresponding eigenvector $\mathbf{x}$, that is, $(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{x} = \lambda\mathbf{x}$, then we have

$$(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\underbrace{(\mathbf{A}^t(\mathbf{A}^t)^{-1})}_{\mathbf{I}}\mathbf{x} = \lambda\mathbf{x},$$

which is equivalent to

$$(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t((\mathbf{A}^t)^{-1}\mathbf{x}) = \lambda\mathbf{x}.$$

We multiply both sides on the left by $(\mathbf{A}^t)^{-1}$ and find

$$(\mathbf{A}^t)^{-1}(\mathbf{S}_W^x)^{-1}\mathbf{S}_B^x\mathbf{A}^t((\mathbf{A}^t)^{-1}\mathbf{x}) = (\mathbf{A}^t)^{-1}\lambda\mathbf{x},$$

which yields

$$(\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y((\mathbf{A}^t)^{-1}\mathbf{x}) = \lambda((\mathbf{A}^t)^{-1}\mathbf{x}).$$

Thus we see that λ is an eigenvalue of $(\mathbf{S}_W^y)^{-1}\mathbf{S}_B^y$ with corresponding eigenvector $(\mathbf{A}^t)^{-1}\mathbf{x}$.

29. We consider the problems that might arise when using the determinant criterion for clustering.

(a) Here the within-cluster matrix for category i is

$$\mathbf{S}_W^i = \sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t,$$

and is the sum of n_i matrices, each of rank at most 1. Thus $\text{rank}[\mathbf{S}_W^i] \leq n_i$. These matrices are not independent; they satisfy the constraint

$$\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i) = \mathbf{0}.$$

This implies that

$$\text{rank}[\mathbf{S}_W] \leq \sum_{i=1}^{c}(n_i - 1) = n - c.$$

(b) Here we have the between scatter matrix

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$$

is the sum of c matrices of rank at most 1, and thus rank$[\mathbf{S}_B] \leq c$. The matrices are constrained by

$$\sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t = \mathbf{0},$$

and thus rank$[\mathbf{S}_B] \leq c - 1$. The between-cluster scatter matrix, $\mathbf{S}_B$, is always singular if $c \leq d$, the dimension of the space. The determinant criterion for clustering is not useful under such conditions because at least one of the eigenvalues is 0, since as a result the determinant is also 0.

Section 10.8

30. Our generalization of the basic minimum-squared-error criterion function of Eq. 54 in the text is:

$$J_T = \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)^t \mathbf{S}_T^{-1} (\mathbf{x} - \mathbf{m}_i).$$

(a) We consider a non-singular transformation of the feature space of the form $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a d-by-d non-singular matrix. We let $\tilde{\mathcal{D}}_i = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{D}_i\}$ denote the transformed data set. Then, we have the criterion in the transformed space is

$$\begin{aligned} J_T^y &= \sum_{i=1}^{c} \sum_{\mathbf{y}\in\tilde{\mathcal{D}}_i} (\mathbf{y} - \mathbf{m}_i^y)^t \mathbf{S}_T^{y-1} (\mathbf{y} - \mathbf{m}_i^y) \\ &= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i)^t \mathbf{S}_T^{y-1} (\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{m}_i). \end{aligned}$$

As mentioned in the solution to Problem 25, the within- and between-scatter matrices transform according to:

$$\begin{aligned} \mathbf{S}_W^y &= \mathbf{A}\mathbf{S}_W\mathbf{A}^t \quad \text{and} \\ \mathbf{S}_B^y &= \mathbf{A}\mathbf{S}_B\mathbf{A}^t, \end{aligned}$$

respectively. Thus the scatter matrix in the transformed coordinates is

$$\mathbf{S}_T^y = \mathbf{S}_W^y + \mathbf{S}_B^y = \mathbf{A}(\mathbf{S}_W + \mathbf{S}_B)\mathbf{A}^t = \mathbf{A}\mathbf{S}_T\mathbf{A}^t,$$

and this implies

$$[\mathbf{S}_T^y]^{-1} = (\mathbf{A}\mathbf{S}_T\mathbf{A}^t)^{-1} = (\mathbf{A}^t)^{-1}\mathbf{S}_T^{-1}\mathbf{A}^{-1}.$$

Therefore, we have

$$\begin{aligned} J_T^y &= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} \sum (\mathbf{A}(\mathbf{x}-\mathbf{m}_i))^t (\mathbf{A}^t)^{-1} \mathbf{S}_T^{-1} \mathbf{A}^{-1} (\mathbf{A}(\mathbf{x}-\mathbf{m}_i)) \\ &= \sum_{i=1}^{c} (\mathbf{x}-\mathbf{m}_i)^t (\mathbf{A}^t)(\mathbf{A}^t)^{-1} \mathbf{S}_T^{-1} (\mathbf{x}-\mathbf{m}_i) \\ &= \sum_{i=1}^{c} (\mathbf{x}-\mathbf{m}_i) \mathbf{S}_T^{-1} (\mathbf{x}-\mathbf{m}_i) = J_T. \end{aligned}$$

In short, then, J_T is invariant to non-singular linear transformation of the data.

(b) We consider sample $\hat{\mathbf{x}}$ being transferred from $\mathcal{D}_i$ to $\mathcal{D}_j$. Recall that the total scatter matrix $\mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t$, given by Eq. 64 in the text, does not change as a result of changing the partition. Therefore the criterion is

$$J_T^* = \sum_{k=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_k^*} (\mathbf{x}-\mathbf{m}_k^*)^t \mathbf{S}_T^{-1} (\mathbf{x}-\mathbf{m}_k^*),$$

where

$$\mathcal{D}_k^* = \begin{cases} \mathcal{D}_k & \text{if} \quad k \neq i, j \\ \mathcal{D}_i - \{\hat{\mathbf{x}}\} & \text{if} \quad k = i \\ \mathcal{D}_j + \{\hat{\mathbf{x}}\} & \text{if} \quad k = j. \end{cases}$$

We note the following values of the means after transfer of the point:

$$\begin{aligned} \mathbf{m}_k^* &= \mathbf{m}_k \text{ if } k \neq i, j, \\ \mathbf{m}_i^* &= \frac{\sum\limits_{\mathbf{x}\in\mathcal{D}_i^*} \mathbf{x}}{\sum\limits_{\mathbf{x}\in\mathcal{D}_i^*} 1} = \frac{\sum\limits_{\mathbf{x}\in\mathcal{D}_i} \mathbf{x} - \hat{\mathbf{x}}}{n_i - 1} \\ &= \frac{n_i m_i - \hat{\mathbf{x}}}{n_i - 1} = \frac{(n_i-1)\mathbf{m}_i - (\hat{\mathbf{x}} - \mathbf{m}_i)}{n_i - 1} \\ &= \mathbf{m}_i - \frac{\hat{\mathbf{x}} - \mathbf{m}_i}{n_i - 1}, \\ \mathbf{m}_j^* &= \frac{\sum\limits_{\mathbf{x}\in H_i} \mathbf{x} + \hat{\mathbf{x}}}{n_i + 1} \\ &= \frac{n_j m_j + \hat{\mathbf{x}}}{n_j + 1} = \frac{(n_j+1)\mathbf{m}_j + (\hat{\mathbf{x}} - \mathbf{m}_j)}{n_j + 1} \\ &= \mathbf{m}_j + \frac{\hat{\mathbf{x}} - \mathbf{m}_j}{n_j + 1}. \end{aligned}$$

Thus our criterion function is

$$\begin{aligned} J_T^* = &\sum_{k=1, k\neq i,j}^{c} (\mathbf{x}-\mathbf{m}_k)^t \mathbf{S}_T^{-1} (\mathbf{x}-\mathbf{m}_k) + \sum_{\mathbf{x}\in\mathcal{D}_i^*} (\mathbf{x}-\mathbf{m}_i^*)^t \mathbf{S}_T^{-1} (\mathbf{x}-\mathbf{m}_i^*) \\ &+ \sum_{\mathbf{x}\in\mathcal{D}_j^*} (\mathbf{x}-\mathbf{m}_j^*)^t \mathbf{S}_T^{-1} (\mathbf{x}-\mathbf{m}_j). \qquad (*) \end{aligned}$$

We expand the sum:

$$\begin{aligned}
&\sum_{\mathbf{x}\in\mathcal{D}_i^*}(\mathbf{x}-\mathbf{m}_i^*)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i^*)+\sum_{\mathbf{x}\in\mathcal{D}_j^*}(\mathbf{x}-\mathbf{m}_j^*)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_j^*)\\
=&\sum_{\mathbf{x}\in\mathcal{D}_i^*}\mathbf{x}^t\mathbf{S}_T^{-1}\mathbf{x}-n_i^*\mathbf{m}_i^{*t}+\sum_{\mathbf{x}\in\mathcal{D}_j^*}\mathbf{x}^t\mathbf{S}_T^{-1}\mathbf{x}-n_j^*\mathbf{m}_j^{*t}\mathbf{S}_T^{-1}\mathbf{m}_j^*\\
=&\sum_{\mathbf{x}\in\mathcal{D}_i^*}\mathbf{x}^t\mathbf{S}_T^{-1}\mathbf{x}-\hat{\mathbf{x}}\mathbf{S}_T^{-1}\hat{\mathbf{x}}-(n_i-1)\left(\mathbf{m}_i-\frac{\hat{\mathbf{x}}-\mathbf{m}_i}{n_i-1}\right)^t\mathbf{S}_T^{-1}\left(\mathbf{m}_i-\frac{\hat{\mathbf{x}}-\mathbf{m}_i}{n_i-1}\right)\\
&+\sum_{\mathbf{x}\in\mathcal{D}_j^*}\mathbf{x}^t\mathbf{S}_T^{-1}\mathbf{x}+\hat{\mathbf{x}}\mathbf{S}_T^{-1}\hat{\mathbf{x}}-(n_j+1)\left(\mathbf{m}_j+\frac{\hat{\mathbf{x}}-\mathbf{m}_j}{n_j+1}\right)^t\mathbf{S}_T^{-1}\left(\mathbf{m}_j+\frac{\hat{\mathbf{x}}-\mathbf{m}_j}{n_j+1}\right)\\
=&\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i)-\hat{\mathbf{x}}^t\mathbf{S}_T^{-1}\hat{\mathbf{x}}+\mathbf{m}_i\mathbf{S}_T^{-1}\mathbf{m}_i+2\mathbf{m}_i^t\mathbf{S}_T^{-1}\hat{\mathbf{x}}-2\mathbf{m}_i^t\mathbf{S}_T^{-1}\mathbf{m}_i\\
&-\frac{1}{n_i-1}(\hat{\mathbf{x}}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_i)\\
&+\sum_{\mathbf{x}\in\mathcal{D}_j}(\mathbf{x}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_j)+\hat{\mathbf{x}}^t\mathbf{S}_T^{-1}\hat{\mathbf{x}}-\mathbf{m}_j\mathbf{S}_T^{-1}\mathbf{m}_j+2\mathbf{m}_j^t\mathbf{S}_T^{-1}\hat{\mathbf{x}}+2\mathbf{m}_j^t\mathbf{S}_T^{-1}\mathbf{m}_j\\
&-\frac{1}{n_j+1}(\hat{\mathbf{x}}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_j)\\
=&\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_i)-\frac{n_i}{n_i+1}(\hat{\mathbf{x}}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_i)\\
&+\frac{n_j}{n_j+1}(\hat{\mathbf{x}}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_j).
\end{aligned}$$

We substitute this result in $(*)$ and find

$$\begin{aligned}
J_T^* &= \sum_{k=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_k}(\mathbf{x}-\mathbf{m}_k)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_k)\\
&\quad+\left[\frac{n_j}{n_j+1}(\mathbf{x}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\mathbf{x}-\mathbf{m}_j)-\frac{n_i}{n_i-1}(\hat{\mathbf{x}}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_i)\right]\\
&= J_T+\left[\frac{n_j}{n_j+1}(\hat{\mathbf{x}}-\mathbf{m}_j)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_j)-\frac{n_i}{n_i-1}(\hat{\mathbf{x}}-\mathbf{m}_i)^t\mathbf{S}_T^{-1}(\hat{\mathbf{x}}-\mathbf{m}_i)\right].
\end{aligned}$$

(c) If we let $\mathcal{D}$ denote the data set and n the number of points, the algorithm is:

Algorithm 0 (Minimize J_T)

```
1   begin initialize D, c
2       Compute c means m_1, ..., m_c
3       Compute J_T
4       do  Randomly select a sample;  call it x̂
5           Determine closest mean to x̂;  call it m_j
6       if n_i = 1 then go to line 10
7         if j ≠ i then ρ_j ← (n_j/(n_j+1)) (x̂ − m_j)^t S_T^{-1} (x̂ − m_j)
8           if j = 1 then ρ_j ← (n_i/(n_i−1)) (x̂ − m_i)^t S_T^{-1} (x̂ − m_i)
9             if ρ_k ≤ ρ_j for all j  then transfer x̂ to D_k
```

```
10          Update J_T, m_i, and m_k
11      until J_T has not changed in n tries
12  end
```

31. The total scatter matrix, $\mathbf{S}_T = \sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^t$, given by Eq. 64 in the text, does not change. Thus our criterion function is

$$J_e = \text{tr}[\mathbf{S}_W] = \text{tr}[\mathbf{S}_T - \mathbf{S}_B] = \text{tr}[\mathbf{S}_T] - \text{tr}[\mathbf{S}_B].$$

We let J_e^* be the criterion function which results from transferring a sample $\hat{\mathbf{x}}$ from $\mathcal{D}_i$ to $\mathcal{D}_j$. Thus we have

$$\begin{aligned}
J_e^* &= \text{tr}[\mathbf{S}_T] - \text{tr}[\mathbf{S}_B^*] \qquad (*)\\
&= \text{tr}[\mathbf{S}_T] - \sum_k n_k^* \|\mathbf{m}_k^* - \mathbf{m}\|^2\\
&= \text{tr}[\mathbf{S}_T] - \sum_{k\neq i,j} n_k^* \|\mathbf{m}_k^* - \mathbf{m}\|^2 - \sum_{k=i,j} n_k^* \|\mathbf{m}_k^* - \mathbf{m}\|^2\\
&= \text{tr}[\mathbf{S}_T] - \sum_{k\neq i,j} n_k \|\mathbf{m}_k - \mathbf{m}\|^2 - n_i^* \|\mathbf{m}_i^* - \mathbf{m}\|^2 - n_j^* \|\mathbf{m}_j^* - \mathbf{m}\|^2.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
n_i^* \|\mathbf{m}_i^* - \mathbf{m}\|^2 + n_j^* \|\mathbf{m}_j^* - \mathbf{m}\|^2 &= (n_i - 1)\left\|\mathbf{m}_i - \frac{\hat{\mathbf{x}} - \mathbf{m}_i}{n_i - 1} - \mathbf{m}\right\|^2\\
&\quad + (n_j + 1)\left\|\mathbf{m}_j + \frac{\hat{\mathbf{x}} - \mathbf{m}_j}{n_j + 1} - \mathbf{m}\right\|^2,
\end{aligned}$$

as shown in Problem 30. We thus find that the means change by

$$\begin{aligned}
\mathbf{m}_i^* &= \mathbf{m}_i - \frac{\hat{\mathbf{x}} - \mathbf{m}_i}{n_i - 1},\\
\mathbf{m}_j^* &= \mathbf{m}_j + \frac{\hat{\mathbf{x}} - \mathbf{m}_j}{n_j + 1}.
\end{aligned}$$

We substitute these into $(*)$ above and find through a straightforward but tedious calculation:

$$\begin{aligned}
J_e^* &= (n_i - 1)\left[\|\mathbf{m}_i - \mathbf{m}\|^2 + \frac{1}{(n_i - 1)^2}\|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 - \frac{2}{n_i - 1}(\hat{\mathbf{x}} - \mathbf{m}_i)^t(\mathbf{m}_i - \mathbf{m})\right]\\
&\quad + (n_j - 1)\left[\|\mathbf{m}_j - \mathbf{m}\|^2 + \frac{1}{(n_j + 1)^2}\|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 - \frac{2}{n_j + 1}(\hat{\mathbf{x}} - \mathbf{m}_j)^t(\mathbf{m}_j - \mathbf{m})\right]\\
&= n_i\|\mathbf{m}_i - \mathbf{m}\|^2 - \|\mathbf{m}_i - \mathbf{m}\|^2 - 2(\hat{\mathbf{x}} - \mathbf{m}_i)^t(\mathbf{m}_i - \mathbf{m}) + \frac{1}{n_i - 1}\|\hat{\mathbf{x}} - \mathbf{m}_i\|^2\\
&\quad + n_j\|\mathbf{m}_j - \mathbf{m}\|^2 - \|\mathbf{m}_j - \mathbf{m}\|^2 + 2(\hat{\mathbf{x}} - \mathbf{m}_j)^t(\mathbf{m}_j - \mathbf{m}) + \frac{1}{n_j + 1}\hat{\mathbf{x}} - \mathbf{m}_j\|^2\\
&= n_i\|\mathbf{m}_i - \mathbf{m}\|^2 + n_j\|\mathbf{m}_j - \mathbf{m}\|^2 + \|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 - \|\hat{\mathbf{x}} - \mathbf{m}_j\|^2\\
&\quad + \frac{1}{n_i - 1}||\hat{\mathbf{x}} - \mathbf{m}_i||^2 + \frac{1}{n_j + 1}||\hat{\mathbf{x}} - \mathbf{m}_j||^2\\
&= n_i\|\mathbf{m}_i - \mathbf{m}\|^2 + n_j\|\mathbf{m}_j - \mathbf{m}\|^2 + \frac{n_i}{n_i - 1}\|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 - \frac{n_j}{n_j + 1}\|\hat{\mathbf{x}} - \mathbf{m}_j\|^2.
\end{aligned}$$

Therefore our criterion function is

$$\begin{aligned} J_e^* &= \text{tr}[\mathbf{S}_T] - \text{tr}[\mathbf{S}_B^*] \\ &= \text{tr}[\mathbf{S}_T] - \sum_k n_k \|\mathbf{m}_k - \mathbf{m}\|^2 - \frac{n_i}{n_i - 1}\|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 + \frac{n_j}{n_j + 1}\|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 \\ &= J_e + \frac{n_j}{n_j + 1}\|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 - \frac{n_i}{n_i - 1}\|\hat{\mathbf{x}} - \mathbf{m}_i\|^2. \end{aligned}$$

Section 10.9

32. Our similarity measure is given by Eq. 50 in the text:

$$s(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^t \mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}.$$

(a) We have that $\mathbf{x}$ and $\mathbf{x}'$ are d-dimensional vectors with $x_i = 1$ if $\mathbf{x}$ possesses the ith feature and $x_i = -1$ otherwise. The Euclidean length of the vectors obeys

$$\|\mathbf{x}\| = \|\mathbf{x}'\| = \sqrt{\sum_{i=1}^{d} x_i^2} = \sqrt{\sum_{i=1}^{d} 1} = \sqrt{d},$$

and thus we can write

$$\begin{aligned} s(\mathbf{x}, \mathbf{x}') &= \frac{\mathbf{x}^t \mathbf{x}'}{\sqrt{d}\sqrt{d}} = \frac{1}{d}\sum_{i=1}^{d} x_i x_i' \\ &= \frac{1}{d}\,[\text{number of common features} - \text{number of features not common}] \\ &= \frac{1}{d}[\text{number of common features} - (d - \text{number of common features})\,] \\ &= \frac{2}{d}\,(\text{number of common features}) \; - 1. \end{aligned}$$

(b) The length of the difference vector is

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}'\|^2 &= (\mathbf{x} - \mathbf{x}')^t(\mathbf{x} - \mathbf{x}') \\ &= \mathbf{x}^t\mathbf{x} + \mathbf{x}'^t\mathbf{x}' - 2\mathbf{x}^t\mathbf{x}' \\ &= \|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 - 2s(\mathbf{x}, \mathbf{x}')\|\mathbf{x}\|\|\mathbf{x}'\| \\ &= d + d - 2s(\mathbf{x}, \mathbf{x}')\sqrt{d}\sqrt{d} \\ &= 2d[1 - s(\mathbf{x}, \mathbf{x}')], \end{aligned}$$

where, from part (a), we used $\|\mathbf{x}\| = \|\mathbf{x}'\| = \sqrt{d}$.

33. Consider the following candidates for metrics or pseudometrics.

(a) Squared Euclidean distance:

$$s(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2 = \sum_{i=1}^{d} (x_i - x_i')^2.$$

Clearly we have

$$\begin{aligned} s(\mathbf{x},\mathbf{x}') &\geq 0 \qquad\qquad \text{(non-negativity)} \\ s(\mathbf{x},\mathbf{x}') &= 0 \Leftrightarrow \mathbf{x}=\mathbf{x}' \quad \text{(uniqueness)} \\ s(\mathbf{x},\mathbf{x}') &= s(\mathbf{x}',\mathbf{x}) \qquad \text{(symmetry)}. \end{aligned}$$

Now consider the triangle inequality for the particular case $d = 1$, and $x = 0, x' = 1/2$ and $x'' = 1$.

$$\begin{aligned} s(\mathbf{x},\mathbf{x}'') &= (0-1)^2 = 1 \\ s(\mathbf{x},\mathbf{x}') &= (0-1/2)^2 = 1/4 \\ s(\mathbf{x}',\mathbf{x}'') &= (1/2-1)^2 = 1/4. \end{aligned}$$

Thus we have

$$s(\mathbf{x},\mathbf{x}') + s(\mathbf{x}',\mathbf{x}'') = 1/2 < s(\mathbf{x},\mathbf{x}''),$$

and thus $s(\mathbf{x},\mathbf{x}'')$ is not less than or equal to $s(\mathbf{x},\mathbf{x}') + s(\mathbf{x}',\mathbf{x}'')$. In short, the squared Euclidean distance does not obey the triangle inequality and is not a metric. Hence it cannot be an ultrametric either.

(b) Euclidean distance:

$$s(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\| = \sqrt{\sum_{i=1}^{d}(x_i - x_i')^2}.$$

Clearly symmetry, non-negativity and uniqueness hold, as in part (a); now we turn to the triangle inequality. First consider two vectors $\mathbf{x}$ and $\mathbf{x}'$. We will need to show that $\|\mathbf{x}+\mathbf{x}'\| \leq \|\mathbf{x}\| + \|\mathbf{x}'\|$; we do this as follows:

$$\begin{aligned} \|\mathbf{x}+\mathbf{x}'\|^2 &= (\mathbf{x}+\mathbf{x}')^t(\mathbf{x}+\mathbf{x}') \\ &= \mathbf{x}^t\mathbf{x} + \mathbf{x}'^t\mathbf{x}' + 2\mathbf{x}'^t\mathbf{x} \\ &= \|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 + 2\mathbf{x}^t\mathbf{x}' \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 + 2\|\mathbf{x}\|\|\mathbf{x}'\|. \end{aligned}$$

We use the Cauchy-Schwarz Inequality (a special case of the Hölder inequality, see part (c) below), i.e., $|\mathbf{x}^t\mathbf{x}'| \leq \|\mathbf{x}\|\|\mathbf{x}'\|$, and thus find

$$\|\mathbf{x}+\mathbf{x}'\|^2 \leq (\|\mathbf{x}\| + \|\mathbf{x}'\|)^2$$

and thus

$$\|\mathbf{x}+\mathbf{x}'\| \leq \|\mathbf{x}\| + \|\mathbf{x}'\|.$$

We put this together to find

$$\begin{aligned} s(\mathbf{x},\mathbf{x}'') = \|\mathbf{x}-\mathbf{x}''\| &= \|(\mathbf{x}-\mathbf{x}') + (\mathbf{x}'-\mathbf{x}'')\| \\ &\leq \|\mathbf{x}-\mathbf{x}'\| + \|\mathbf{x}'-\mathbf{x}''\| \\ &= s(\mathbf{x},\mathbf{x}') + s(\mathbf{x}',\mathbf{x}''), \end{aligned}$$

and thus the Euclidean metric is indeed a metric. We use the same sample points as in the example of part (a) to test whether the Euclidean metric is an ultrametric:

$$\begin{aligned} s(\mathbf{x}, \mathbf{x}'') &= 1 \\ s(\mathbf{x}, \mathbf{x}') = s(\mathbf{x}, \mathbf{x}'') &= 1/4 \end{aligned}$$

so

$$\begin{aligned} \max[s(\mathbf{x}, \mathbf{x}'), s(\mathbf{x}', \mathbf{x}'')] &= 1/4 \\ &< 1 = s(\mathbf{x}, \mathbf{x}''). \end{aligned}$$

Thus the Euclidean metric is not an ultrametric.

(c) Minkowski metric:

$$s(\mathbf{x}, \mathbf{x}') = \left(\sum_{i=1}^{d} |x_i - x_i'|^q \right)^{1/q}.$$

It is a simple matter to show that the properties of non-negativity, uniqueness and symmetry hold for the Minkowski metric. In order to prove that the triangle inequality also holds, we will first need to prove Hölder's inequality, that is, for p and q positive numbers such that $1/p + 1/q = 1$

$$\sum_{i=1}^{d} |x_i x_i'| \leq \left(\sum_{i=1}^{d} |x_i|^q \right)^{1/q} \cdot \left(\sum_{i=1}^{d} |x_i'|^p \right)^{1/p},$$

with equality holding if and only if

$$\left(\frac{|x_j|}{\left(\sum_{i=1}^{d} x_i \right)^{1/p}} \right)^{1/q} = \left(\frac{|x_j'|}{\left(\sum_{i=1}^{d} x_i' \right)^{1/q}} \right)^{1/p}$$

for all j. The limiting case of $p = 1$ and $q = \infty$ can be easily verified directly. We thus turn to the case $1 < p, q, < \infty$. Consider two real, non-negative numbers a and b, and $0 \leq \lambda \leq 1$; we have

$$a^\lambda b^{(1-\lambda)} \leq \lambda a + (1 - \lambda) b,$$

with equality if and only if $a = b$. To show this, we consider the function

$$f(t) = t^\lambda - \lambda t + \lambda - 1$$

for $t \geq 0$. Thus we have $f'(t) = \lambda(t^{(\lambda-1)} - 1) \geq 0$. Furthermore, $f(t) \leq f(1) = 0$ with equality only for $t = 1$. Thus we have

$$t^\lambda \leq \lambda t + 1 - \lambda.$$

We let $t = a/b$, substitute above and our intermediate inequality is thus proven.

We apply this inequality to the particular case

$$a = \left(\frac{|x_j|}{\left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}} \right)^p, \qquad b = \left(\frac{|x'_j|}{\left(\sum_{i=1}^{d} |x'_i|^q \right)^{1/q}} \right)^q$$

with $\lambda = 1/p$ and $1 - \lambda = 1/q$. Thus for each component j we have

$$\frac{|x_j x'_j|}{\left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{d} |x'_i|^q \right)^{1/q}} \leq \frac{1}{p} \left(\frac{|x_j|}{\left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}} \right)^p + \frac{1}{q} \left(\frac{|x'_j|}{\left(\sum_{i=1}^{d} |x'_i|^q \right)^{1/q}} \right)^q.$$

We sum this inequality over all $j = 1, \ldots, d$ and find

$$\frac{\sum_{j=1}^{d} |x_j x'_j|}{\left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{d} |x'_i|^q \right)^{1/q}} \leq \frac{1}{p} + \frac{1}{q} = 1,$$

and the Hölder inequality is thereby proven.

We now use the Hölder inequality to prove that the triangle inequality holds for the Minkowski metric, a result known as Minkowski's inequality. We follow the logic above and have

$$\sum_{i=1}^{d} |x_i + x'_i|^p \leq \sum_{i=1}^{d} |x_i + x'_i|^{p-1} |x_i| + \sum_{i=1}^{d} |x_i + x'_i|^{p-1} |x'_i|.$$

We apply Hölder's inequality to each summation on the right hand side and find

$$\begin{aligned} \sum_{j=1}^{d} |x_i + x'_i|^p &\leq \left(\sum_{i=1}^{d} |x_i + x'_i|^{(p-1)q} \right)^{1/q} \left[\left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{d} |x'_i|^p \right)^{1/p} \right] \\ &= \left(\sum_{i=1}^{d} |x_i + x'_i|^p \right)^{1/q} \left[\left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{d} |x'_i|^p \right)^{1/p} \right]. \end{aligned}$$

We divide both side by $\left(\sum_{i=1}^{d} \left| x_i + x'_i \right|^p \right)^{1/q}$, recall that $1 - 1/q = 1/p$, and thereby obtain

$$\left(\sum_{i=1}^{d} |x_i + x'_i|^p \right)^{1/p} \leq \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{d} |x'_i|^p \right)^{1/p}.$$

Thus, using the notation of the Minkowski metric above, we have

$$s(\mathbf{x} + \mathbf{x}', \mathbf{0}) \leq s(\mathbf{x}, \mathbf{0}) + s(\mathbf{x}', \mathbf{0}),$$

or with simple substitution and rearrangement

$$s(\mathbf{x}, \mathbf{x}'') \leq s(\mathbf{x}, \mathbf{x}') + s(\mathbf{x}', \mathbf{x}''),$$

for arbitrary $\mathbf{x}$, $\mathbf{x}'$ and $\mathbf{x}''$. Thus our triangle inequality is thereby proven and the Minkowski measure is a true metric.

Ultrametric [[more hereProblem not yet solved]]

(d) Cosine:

$$s(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^t \mathbf{x}'}{\|\mathbf{x}\| \|\mathbf{x}'\|}.$$

We investigate the condition of uniqueness in the particular case $d = 2$ and $\mathbf{x} = \binom{1}{1}, \mathbf{x}' = \binom{1}{-1}$. For these points we have

$$\mathbf{x}^t \mathbf{x}' = 1 - 1 = 0,$$

but note that $\mathbf{x} \neq \mathbf{x}'$ here. Therefore $s(\mathbf{x}, \mathbf{x}')$ does not possess the uniqueness property, and thus is not a metric and cannot be an ultrametric either.

(e) Dot product:

$$s(\mathbf{x}, \mathbf{x}') = \mathbf{x}^t \mathbf{x}'.$$

We use the same counterexample as in part (c) to see that the dot product is neither a metric nor an ultrametric.

(f) One-sided tangent distance:

$$s(\mathbf{x}, \mathbf{x}') = \min_{\boldsymbol{\alpha}} \|\mathbf{x} + \boldsymbol{\alpha}\mathbf{T}(\mathbf{x}) - \mathbf{x}'\|^2.$$

There is no reason why the symmetry property will be obeyed, in general, that is, $s(\mathbf{x}, \mathbf{x}') \neq s(\mathbf{x}', \mathbf{x})$, or

$$\min_{\boldsymbol{\alpha}_1} \|\mathbf{x} + \boldsymbol{\alpha}_1 \mathbf{T}_1(\mathbf{x}) - \mathbf{x}'\|^2 \neq \min_{\boldsymbol{\alpha}_2} \|\mathbf{x}' + \boldsymbol{\alpha}_2 \mathbf{T}_2(\mathbf{x}') - \mathbf{x}\|^2,$$

and thus the one-sided tangent distance is not not a metric and not an ultrametric.

34. Consider merging two clusters $\mathcal{D}_i$ and $\mathcal{D}_j$ and whether various values of the parameters in the function

$$d_{hk} = \alpha d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}|$$

can be used for a range of distance or similarity measures.

(a) We consider d_{min} defined by

$$\begin{aligned}
d_{min}(\mathcal{D}_i, \mathcal{D}_j) &= \min_{\substack{\mathbf{x} \in \mathcal{D}_i \\ \mathbf{x}' \in \mathcal{D}_j}} \|\mathbf{x} - \mathbf{x}'\| \\
d_{hk} &= \min \left[\min_{\substack{\mathbf{x} \in \mathcal{D}_i \\ \mathbf{x}' \in \mathcal{D}_j}} \|\mathbf{x} - \mathbf{x}'\|, \min_{\substack{\mathbf{x} \in \mathcal{D}_j \\ \mathbf{x}' \in \mathcal{D}_h}} \|\mathbf{x} - \mathbf{x}'\| \right] \\
&= \min (d_{hi}, d_{hj}) \\
&= \frac{1}{2} d_{hi} + \frac{1}{2} d_{hj} - \frac{1}{2} |d_{hi} - d_{hj}| \\
&= \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}|
\end{aligned}$$

for $\alpha_i = \alpha_j = 1/2, \beta = 0$ and $\gamma = -1/2$.

(b) We consider d_{max} defined by

$$\begin{aligned} d_{max}(\mathcal{D}_i, \mathcal{D}_j) &= \max_{\substack{\mathbf{x}\in\mathcal{D}_i \\ \mathbf{x}'\in\mathcal{D}_j}} \|\mathbf{x}-\mathbf{x}'\| \\ d_{hk} &= \max_{\substack{\mathbf{x}\in\mathcal{D}_k \\ \mathbf{x}'\in\mathcal{D}_h}} \|\mathbf{x}-\mathbf{x}'\| = \max_{\substack{\mathbf{x}\in\mathcal{D}_i \bigcup \mathcal{D}_j \\ \mathbf{x}'\in\mathcal{D}_h}} \|\mathbf{x}-\mathbf{x}'\| \\ &= \max \left[\max_{\substack{\mathbf{x}\in\mathcal{D}_i \\ \mathbf{x}'\in\mathcal{D}_j}} \|\mathbf{x}-\mathbf{x}'\|, \max_{\substack{\mathbf{x}\in\mathcal{D}_j \\ \mathbf{x}'\in\mathcal{D}_h}} \|\mathbf{x}-\mathbf{x}'\| \right] \\ &= \max\left(d_{hi}, d_{hj}\right) \\ &= \frac{1}{2}d_{hi} + \frac{1}{2}d_{hj} + \frac{1}{2}|d_{hi} - d_{hj}| \\ &= \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}| \end{aligned}$$

for $\alpha_i = d_j = 1/2, \beta = 0$ and $\gamma = 1/2$.

(c) We consider d_{ave} defined by

$$\begin{aligned} d_{ave}(\mathcal{D}_i, \mathcal{D}_j) &= \frac{1}{n_i n_j} \sum_{\substack{\mathbf{x}\in\mathcal{D}_i \\ \mathbf{x}'\in\mathcal{D}_j}} \|\mathbf{x}-\mathbf{x}'\| \\ d_{hk} &= \frac{1}{n_h n_k} \sum_{\substack{\mathbf{x}\in\mathcal{D}_k \\ \mathbf{x}'\in\mathcal{D}_h}} \|\mathbf{x}-\mathbf{x}'\| \\ &= \frac{1}{n_h(n_i+n_j)} \sum_{\substack{\mathbf{x}\in\mathcal{D}_i\cap\mathcal{D}_j \\ \mathbf{x}'\in\mathcal{D}_h}} \|\mathbf{x}-\mathbf{x}'\| \\ &= \frac{1}{n_h(n_i+n_j)} \left[\sum_{\substack{\mathbf{x}\in\mathcal{D}_i \\ \mathbf{x}'\in\mathcal{D}_h}} \|\mathbf{x}-\mathbf{x}'\| + \sum_{\substack{\mathbf{x}\in\mathcal{D}_j \\ \mathbf{x}\in\mathcal{D}_h}} \|\mathbf{x}-\mathbf{x}'\| \right] \\ &= \frac{1}{n_h(n_i+n_j)} [n_h n_i d_{hi} + n_h n_j d_{hj}] \\ &= \frac{n_i}{n_i+n_j} d_{hi} + \frac{n_j}{n_i+n_j} d_{hj} \\ &= \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}| \end{aligned}$$

for $\alpha_i = n_i/(n_i+n_j), \alpha_j = n_j/(n_i+n_j)$ and $\beta = \gamma = 0$.

(d) We consider d^2_{mean} defined by

$$\begin{aligned} d^2_{mean}(\mathcal{D}_i, \mathcal{D}_j) &= \|\mathbf{m}_i - \mathbf{m}_j\|^2 \\ d_{hk} &= \|\mathbf{m}_h - \mathbf{m}_k\|^2 \\ \mathbf{m}_k &= \frac{\sum_{\mathbf{x}\in\mathcal{D}_k} \mathbf{x}}{\sum_{\mathbf{x}\in\mathcal{D}_k} 1} = \frac{\sum_{\mathbf{x}\in\mathcal{D}_i\cap\mathcal{D}_j} \mathbf{x}}{n_i+n_j} \\ &= \frac{\sum_{\mathbf{x}\in\mathcal{D}_i} \mathbf{x} + \sum_{\mathbf{x}\in\mathcal{D}_j} \mathbf{x}}{n_i+n_j} = \frac{n_i\mathbf{m}_i + n_j\mathbf{m}_j}{n_i+n_j}, \end{aligned}$$

$$
\begin{aligned}
d_{hk} &= \Big\|\mathbf{m}_h - \frac{n_i}{n_i+n_j}\mathbf{m}_i - \frac{n_j}{n_i+n_j}\mathbf{m}_j\Big\|^2 \\
&= \Big\|\frac{n_i}{n_i+n_j}\mathbf{m}_h - \frac{n_i}{n_i+n_j}\mathbf{m}_i + \frac{n_j}{n_i+n_j}\mathbf{m}_h - \frac{n_j}{n_i+n_j}\mathbf{m}_j\Big\|^2 \\
&= \Big\|\frac{n_i}{n_i+n_j}(\mathbf{m}_h - \mathbf{m}_i)\Big\|^2 + \Big\|\frac{n_j}{n_i+n_j}(\mathbf{m}_h - \mathbf{m}_j)\Big\|^2 \\
&\quad +2\frac{n_i}{n_i+n_j}(\mathbf{m}_h - \mathbf{m}_i)^t\frac{n_j}{n_i+n_j}(\mathbf{m}_h - \mathbf{m}_j) \\
&= \frac{n_i^2+n_in_j}{(n_i+n_j)^2}\|\mathbf{m}_h - \mathbf{m}_i\|^2 + \frac{n_j^2+n_in_j}{(n_i+n_j)^2}\|\mathbf{m}_h - \mathbf{m}_j\|^2 \\
&\quad +\frac{n_in_j}{(n_i+n_j)^2}[(m_h - m_i)^t(\mathbf{m}_i - \mathbf{m}_j) - (m_h - m_j)^t(\mathbf{m}_i - \mathbf{m}_j)] \\
&= \frac{n_i}{n_i+n_j}\|\mathbf{m}_h - \mathbf{m}_i\|^2 + \frac{n_j}{n_i+n_j}\|\mathbf{m}_h - \mathbf{m}_j\|^2 - \frac{n_in_j}{(n_i+n_j)^2}\|\mathbf{m}_h - \mathbf{m}_j\|^2 \\
&= \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma|d_{hi} - d_{hj}|,
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_i &= \frac{n_i}{n_i+n_j} \\
\alpha_j &= \frac{n_j}{n_i+n_j} \\
\beta &= -\frac{n_in_j}{(n_i+n_j)^2} = -\alpha_i\alpha_j \\
\gamma &= 0.
\end{aligned}
$$

35. The sum-of-squared-error criterion is given by Eq. 72 in the text:

$$
\begin{aligned}
J_e &= \sum_{i'=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_{i'}}\|\mathbf{x} - \mathbf{m}_{i'}\|^2 \\
&= \sum_{i'=1}^{c}\left[\sum_{\mathbf{x}\in\mathcal{D}_{i'}}\mathbf{x}^t\mathbf{x} - n_{i'}\mathbf{m}_{i'}^t\mathbf{m}_{i'}\right] \\
&= \sum_{\mathbf{x}}\mathbf{x}^t\mathbf{x} - \sum_{i'=1}^{c}n_{i'}\mathbf{m}_{i'}^t\mathbf{m}_i.
\end{aligned}
$$

We merge $\mathcal{D}_i$ and $\mathcal{D}_j$ into $\mathcal{D}_k$ and find an increase in the criterion function J_e of

$$
\begin{aligned}
\Delta &\equiv J_e^* - J_e = \sum_{\mathbf{x}}\mathbf{x}^t\mathbf{x} - \sum_{\substack{i'=1\\ i\neq k}}^{c}n_{i'}\mathbf{m}_{i'}^t\mathbf{m}_{i'} - n_k\mathbf{m}_k^t\mathbf{m}_k \\
&\quad - \left[\sum_{\mathbf{x}}\mathbf{x}^t\mathbf{x} - \sum_{\substack{i'=1\\ i'\neq i<j}}^{c}n_{i'}\mathbf{m}_{i'}^t\mathbf{m}_i - n_i\mathbf{m}_i^t\mathbf{m}_i - n_j\mathbf{m}_j^t\mathbf{m}_j\right] \\
&= n_i\mathbf{m}_i^t\mathbf{m}_i + n_j\mathbf{m}_j^t\mathbf{m}_j - n_k\mathbf{m}_k^t\mathbf{m}_k,
\end{aligned}
$$

where

$$
n_k = n_i + n_j
$$

$$\begin{aligned}
\mathbf{m}_k &= \frac{\sum\limits_{\mathbf{x}\in\mathcal{D}_k}\mathbf{x}}{\sum\limits_{\mathbf{x}\in\mathcal{D}_k}1} = \frac{\sum\limits_{\mathbf{x}\in\mathcal{D}_i}\mathbf{x}+\sum\limits_{\mathbf{x}\in\mathcal{D}_j}\mathbf{x}}{n_i+n_j} = \frac{n_i\mathbf{m}_i+n_j\mathbf{m}_j}{n_i+n_j} \\
&= \frac{n_i}{n_i+n_j}\mathbf{m}_i+\frac{n_j}{n_i+n_j}\mathbf{m}_j \\
nn_k\mathbf{m}_k^t\mathbf{m}_k &= (n_i+n_j)\left[\frac{n_i}{n_i+n_j}\mathbf{m}_i+\frac{n_j}{n_i+n_j}\mathbf{m}_j\right]^t\left[\frac{n_i}{n_i+n_j}\mathbf{m}_i+\frac{n_j}{n_i+n_j}\mathbf{m}_j\right] \\
&= \frac{n_i^2}{n_i+n_j}\mathbf{m}_i^t\mathbf{m}_i+\frac{n_j^2}{n_i+n_j}\mathbf{m}_j^t\mathbf{m}_j+\frac{2n_in_j}{n_i+n_j}\mathbf{m}_i^t\mathbf{m}_j.
\end{aligned}$$

thus the difference in criterion function, $\Delta = J_e^* - J_e$, is

$$\begin{aligned}
\Delta &= \left(n_i-\frac{n_i^2}{n_i+n_j}\right)\mathbf{m}_i^t\mathbf{m}_i+\left(n_j-\frac{n_i^2}{n_i+n_j}\right)\mathbf{m}_j^t\mathbf{m}_j-\frac{2n_in_j}{n_i+n_j}\mathbf{m}_i^t\mathbf{m}_j \\
&= \frac{n_in_j}{n_i+n_j}\mathbf{m}_j^t\mathbf{m}_i+\frac{n_in_j}{n_i+n_j}\mathbf{m}_j^t\mathbf{m}_j-\frac{2n_in_j}{n_i+n_j}\mathbf{m}_j^t\mathbf{m}_j \\
&= \frac{n_in_j}{n_i+n_j}(\mathbf{m}_i^t\mathbf{m}_i+\mathbf{m}_j^t\mathbf{m}_j-2\mathbf{m}_i^t\mathbf{m}_j) \\
&= \frac{n_in_j}{n_i+n_j}\|\mathbf{m}_i-\mathbf{m}_j\|^2.
\end{aligned}$$

The smallest increase in J_e corresponds to the smallest value of Δ, and this arises from the smallest value of

$$\frac{n_in_j}{n_i+n_j}\|\mathbf{m}_i-\mathbf{m}_j\|^2.$$

36. PROBLEM NOT YET SOLVED
37. PROBLEM NOT YET SOLVED
38. PROBLEM NOT YET SOLVED
39. Given a set of points, we can define a fully connected graph by connecting each pair of points with an edge that has the distance betyween the end points as its weight. Assume (without loss of generality) that all the weights are mutually different, that is $w(e) \neq w(e')$ if $e \neq e'$. The nearest-neighbor cluster algorithm merges at each iteration the clusters of minimal distance and this corresponds to adding the edge of least weight that joins two different clusters to the spanning tree. In other words, the algorithm examines all the edges in the graph in order of increasing weight and adds them to the spanning tree if the edge joins two disjoint clusters.

To see the optimality of this algorithm with respect to the sum of the edge lengths, we use *reductio ad absurdum*: we will assume that there exists a minimum spanning tree T' which has not been constructed by this algorithm and then arrive at a contradiction, thus showing that the algorithm indeed constructs a minimum spanning tree. We call the tree constructed by the algorithm T and assume that there exists a tree $T' \neq T$ with a sum of weights less than that of T. Say the choice of $e_1, e_2, e_3, \ldots$ with $w(e_1) < w(e_2) < w(e_3) < \ldots$ leads to the tree T. Sort the edges of T' such that $w(e'_1) < w(e'_2) < w(e'_3) < \ldots$. Now, there must exist a first index j where these sequences of edges differ, that is $e_j \neq e'_j$ and $e_i = e'_i$ for all $1 \leq i < j$. Add e_j to T'. This leads to a circle in the graph T'. In this circle, there exists an edge e with $w(e_j) < w(e)$ because j was the first index with a difference in the sequences and e_j is in T (and T has been constructed by the nearest-neighbor clustering algorithm).

Delete e from T'. The result is a tree with lower sum of weights. Therefore T' was not a minimal spanning tree. Thus, by contradiction, we have proven that T was optimal.

Section 10.10

40. PROBLEM NOT YET SOLVED

41. As given in Problem 35, the change in J_e due to the transfer of one point is

$$J_e(2) = J_e(1) - \frac{n_1 n_2}{n_1 + n_2}\|\mathbf{m}_1 - \mathbf{m}_2\|^2.$$

We calculate the expected value of $J_e(1)$ as:

$$\begin{aligned}
\mathcal{E}(J_e(1)) &= \mathcal{E}\left(\sum_{\mathbf{x}\in\mathcal{D}} \|\mathbf{x} - \mathbf{m}\|^2\right) \\
&= \sum_{i=1}^{d} \mathcal{E}\left(\sum_{\mathbf{x}\in\mathcal{D}} (x_i - m_i)^2\right) \\
&= \sum_{i=1}^{d} (n-1)\sigma^2,
\end{aligned}$$

where we have used the standard results

$$S^2 = \sum_i (\mathbf{x}_i - \mathbf{x})^2 \sim \sigma^2 \chi^2_{n-1}$$

and

$$\begin{aligned}
\mathcal{E}(S^2) &= \sigma^2(n-1) \\
&= (n-1)d\sigma^2 \\
&\simeq nd\sigma^2 \text{ for } n \text{ large.}
\end{aligned}$$

We take the expectation of both sides and find

$$\begin{aligned}
\mathcal{E}(J_e(2)) &= \mathcal{E}(J_e(1)) - \mathcal{E}\left[\frac{n_1 n_2}{n_1 + n_2}\|\mathbf{m}_1 - \mathbf{m}_2\|^2\right] \\
&\simeq nd\sigma^2 - \mathcal{E}\left[\frac{n_1 n_2}{n_1 + n_2}\|\mathbf{m}_1 - \mathbf{m}_2\|^2\right].
\end{aligned}$$

We now consider only sub-optimal hyperplanes through the sample mean. For n large, this restriction is equivalent to considering only hyperplanes through the population mean $\boldsymbol{\mu}$ and this, in turn, implies $n_1/n \to 1/2$ and $n_2/n \to 1/2$, by the weak Law of Large Numbers.

Consider points $\mathbf{x}$ distributed symmetrically in a hypersphere about $\boldsymbol{\mu}$. For such a configuration, the distribution of $J_e(2)$ will not depend on the choice of hyperplane; in particular, we can consider the hyperplane that divides $\mathcal{D}$ into

$$\begin{aligned}
\mathcal{D}_1 &= \{\mathbf{x} : x_1 > \mu_1\} \\
\mathcal{D}_2 &= \{\mathbf{x} : x_1 < \mu_1\}.
\end{aligned}$$

Thus we have

$$\begin{aligned}\mathcal{E}\left[\frac{n_1 n_2}{n_1+n_2}\|\mathbf{m}_1-\mathbf{m}_2\|^2\right] &= \sum_{i=1}^{d}\mathcal{E}\left[\frac{n_1 n_2}{n_1+n_2}(m_{1i}-m_{2i})^2\right] \\ &\simeq \mathcal{E}\left[\frac{n_1 n_2}{n_1+n_2}(m_{11}-m_{21})^2\right],\end{aligned}$$

as $\mathbf{x}$ comes from a symmetric normal distribution of the form $N(\boldsymbol{\mu},\sigma^2\mathbf{I})$. Thus we have $m_{1i}-m_{2i}\simeq 0$ for $i\neq 1$ independent of m_{11}, m_{21} by the Strong Law of Large Numbers. Now we have in the limit of large n

$$\begin{aligned}m_{11} &\rightarrow \mathcal{E}(x_1|x_1>\mu_1) \\ m_{21} &\rightarrow \mathcal{E}(x_1|x_1<\mu_1).\end{aligned}$$

Without loss of generality we let $\mu_1=0$. Then we have $p(x_1)\sim N(0,\sigma^2)$ and thus

$$\begin{aligned}\mathcal{E}(x_1|x_1>0) &= \sqrt{\frac{2}{\pi}}\sigma \\ \mathcal{E}(x_1|x_1<0) &= -\sqrt{\frac{2}{\pi}}\sigma.\end{aligned}$$

Therefore, we have $m_{11}-m_{21}\rightarrow 2\sqrt{\frac{2}{\pi}}\sigma$ for n large. We also have

$$\begin{aligned}\mathcal{E}(J_e(2)) &\simeq nd\sigma^2-\frac{(n/2)(n/2)}{(n/2)+(n/2)}\left(2\sqrt{\frac{2}{\pi}}\sigma\right)^2 \\ &= nd\sigma^2-\frac{n}{4}4\times\frac{2}{\pi}\sigma^2 \\ &= n\left(d-\frac{2}{\pi}\right)\sigma^2.\end{aligned}$$

Thus $J_e(2)$ is approximately independent of

$$J_e(1)-J_e(2) = \frac{n_1 n_2}{n_1+n_2}\|\mathbf{m}_1-\mathbf{m}_2\|^2.$$

We write

$$J_e(1)=J_e(2)+[J_e(1)-J_e(2)]$$

and take the variance of both sides and find

$$\mathrm{Var}[J_e(1)] = \mathrm{Var}[J_e(2)]+\mathrm{Var}[\Delta],$$

where

$$\Delta=\frac{n_1 n_2}{n_1+n_2}\|\mathbf{m}_1-\mathbf{m}_2\|^2.$$

Problem not yet solved

Section 10.11

42. Consider a simple greedy algorithm for creating a spanning tree based on the Euclidean distance between points.

(a) The following is known as Kruskal's minimal spanning tree algorithm.

Algorithm 0 (Kruskal's minimal spanning tree)

```
1    begin initialize i ← 0
2      T ← {}
3      do i ← i + 1
4          c_i ← x_i
5      until i = n
6      e_ij ← ||x_i − x_j||; sort in non − decreasing order
7      E ← ordered set of e_ij
8      do for each e_ij in E in non − decreasing order
9          if x_i and x_j belong to disjoint clusters
10             then Append e_ij to T; Append c_j to c_i
11             until all e_ij considered
12     return T
13   end
```

(b) The space complexity is $n(n-1)$ if we pre-compute all point-to-point distances. If we compute all distances on demand, we only need to store the spanning treee T and the clusters, which are both at most of length n.

(c) The time complexity is dominated by the initial sorting step, which takes $O(n^2 \log n)$ time. We have to examine each edge in the **for** loop and that means we have $O(n^2)$ iterations where each iteration can be done in $O(\log n)$ time assuming we store the clusters as disjoint-set-forests where searches and unions can be done in logarithmic time. Thus the total time complexity of the algorithm is $O(n^2 \log n)$.

Section 10.12

43. Consider the XOR problem and whether it can be implemented by an ART network as illustrated in the text.

(a) Learning the XOR problem means that the ART network is supposed to learn the following input-output relation. We augment all the input vectors and normalize them to unit weight wo that they lie on the unit sphere in R^3. Then we have

input			output	
0	0	1	1	$(*)$
$1/\sqrt{2}$	0	$1/\sqrt{2}$	0	$(**)$
0	$1/\sqrt{2}$	$1/\sqrt{2}$	0	$(***)$
$1/\sqrt{3}$	$1/\sqrt{3}$	$1/\sqrt{3}$	1	$(****)$

Associate the weight vector $\mathbf{w}^0$ with cluster 0, and $\mathbf{w}^1$ with cluster 1, and the subscripts with the different dimension. Then we get the following conditions:

$$\begin{aligned} x_3^0 < w_3^1 \qquad & (*) \\ w_1^0 + x_3^0 > w_1^1 + w_3^1 \text{ or } (w_1^0 - w_1^1) > (w_3^1 - w_3^0) \qquad & (**) \\ w_2^0 + w_3^0 > w_2^1 + w_3^1 \text{ or } (w_2^0 - w_2^1) > (w_3^1 - w_3^0) \qquad & (***) \\ w_1^0 + w_2^0 + w_3^0 < w_1^1 + w_2^1 + w_3^1. \qquad & (****) \end{aligned}$$

From $(***)$ we see

$$(w_1^0 - w_1^1) + (w_2^0 - w_2^1) < (w_3^1 - w_3^0).$$

We combine this with $(*)$, $(**)$ and $(***)$ and find

$$0 < 2(w_3^1 - w_3^0) < (w_1^0 - w_1^1) + (w_2^0 - w_2^1) < (w_3^1 - w_3^0).$$

This is a contradition. Thus there is no set of weights that the ART network could converge to that would implement that XOR solution.

(b) Given a weight vector $\mathbf{w}$ for cluster C_1, the vigilance parameter ρ and two inputs $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $\mathbf{w}^t\mathbf{x}_1 > \rho$ and $\mathbf{w}^t x_2 < \rho$, but also $(\mathbf{w}+\mu\mathbf{x})^t x_2 > \rho\|\mathbf{w}+\mu\mathbf{x}_1\|$. Since the top-down feedback tries to push the input vector to be closer to $\mathbf{w}$, we can use the angle between $\mathbf{w}$ and $\mathbf{x}_1$ in this case to make our point, since we cannot forecast the value of y. If we present $\mathbf{x}_1$ before $\mathbf{x}_2$, then $\mathbf{x}_1$ will be classified as belonging to cluster C_1. The weight vector $\mathbf{w}$ is now slightly adjusted and now also $\mathbf{x}_2$ is close enough to $\mathbf{w}$ to be classified as belonging to cluster C_1, thus we will not create another cluster. In constrast, if we present $\mathbf{x}_2$ before $\mathbf{x}_1$, the angle between $\mathbf{w}$ and $\mathbf{x}_2$ is smaller than ρ, thus we will introduce a new cluster C_2 containing $\mathbf{x}_2$. Due to the fixed value of ρ but weights that are changing in dependence on the input, the number of clusters depends on the order the samples are presented.

(c) In a stationary environment the ART network is able to classify inputs robustly because the feedback connections will drive the network into a stable state even when the input is corrupted by noise. In a non-stationary environment the feedback mechanism will delay the adaptation to the changing input vectors because the feedback connections will interpret the changing inputs as noisy versions of the original input and will try to force the input to stay the same. Without the feedback the adaptation of the weights would faster account for changing sample configurations.

Section 10.13

44. Let $\mathbf{e}$ be a vector of unit length and $\mathbf{x}$ the input vector.

(a) We can write the variance as

$$\begin{aligned}
\sigma^2 &= \mathcal{E}_x[a^2] \\
&= \mathcal{E}_x[(\mathbf{x}^t\mathbf{e})^t(\mathbf{x}^t\mathbf{e})] \\
&= \mathcal{E}_x[\mathbf{e}^t\mathbf{x}\mathbf{x}^t\mathbf{e}] \\
&= \mathbf{e}^t\mathcal{E}_x[\mathbf{x}\mathbf{x}^t]\mathbf{e} \\
&= \mathbf{e}^t\mathbf{\Sigma}\mathbf{e},
\end{aligned}$$

since $\mathbf{e}$ is fixed and thus can be taken out of the expectation operator.

(b) We use the Taylor expansion:

$$\begin{aligned}
\sigma^2(\mathbf{e}+\delta\mathbf{e}) &= (\mathbf{e}+\delta\mathbf{e})^t\mathbf{\Sigma}(\mathbf{e}+\delta\mathbf{e}) \\
&= \mathbf{e}^t\mathbf{\Sigma}\mathbf{e} + 2\delta\mathbf{e}^t\mathbf{\Sigma}\mathbf{e} + O(\|\delta\mathbf{e}\|^3).
\end{aligned}$$

If we disregard the cubic term, we see that

$$(\mathbf{e}+\delta\mathbf{e})^t\mathbf{\Sigma}(\mathbf{e}+\delta\mathbf{e}) = \mathbf{e}^t\mathbf{\Sigma}\mathbf{e}$$

implies $(\delta\mathbf{e})^t\mathbf{\Sigma}\mathbf{e} = 0$.

(c) Since $\delta\mathbf{e}$ is perpendicular to $\mathbf{e}$, we have $(\delta\mathbf{e})^t\mathbf{e} = 0$. We can now combine this with the condition from part (b) and get the following equation: $(\delta\mathbf{e})^t\mathbf{\Sigma}\mathbf{e} - \lambda(\delta\mathbf{e})^t\mathbf{e} = 0$. This can be rewritten $(\delta\mathbf{e})^t(\mathbf{\Sigma}\mathbf{e} - \lambda\mathbf{e}) = 0$. For this equation to hold it is sufficient that $\mathbf{\Sigma}\mathbf{e} = \lambda\mathbf{e}$ for some λ. It is also a necessary condition since $(\delta\mathbf{e})^t\mathbf{\Sigma}\mathbf{e}$ must vanish for all $\delta\mathbf{e}$ perpendicular to $\mathbf{e}$, thus $\mathbf{\Sigma}\mathbf{e}$ is itself perpendicular to $\delta\mathbf{e}$ or in other words parallel to $\mathbf{e}$. In other words, we have $\mathbf{\Sigma}\mathbf{e} = \lambda\mathbf{e}$.

(d) We denote the d eigenvalues of the covariance matrix $\mathbf{\Sigma}$ by λ_i and the corresponding eigenvectors by $\mathbf{e}_i$. Since $\mathbf{\Sigma}$ is symmetric and positive semi-definite, we know that all eigenvalues are real and greater than or equal to zero. Since the eigenvectors $\mathbf{e}_i$ can be chosen orthonormal, they span a basis of the d-dimensional space and we can write each vector $\mathbf{x}$ as

$$\sum_{i=1}^{d}(\mathbf{e}_i^t\mathbf{x})\mathbf{e}_i.$$

We define the error $E_k(\mathbf{x})$ as the sum-squared error between the vector $\mathbf{x}$ and its projection onto a k-dimensional subspace, that is

$$\begin{aligned} E_k(\mathbf{x}) &= \left[\mathbf{x} - \sum_{i=1}^{k}(\mathbf{e}_i^t\mathbf{x})\mathbf{e}\right]^t\left[\mathbf{x} - \sum_{i=1}^{k}(\mathbf{e}_i^t\mathbf{x})\mathbf{e}\right] \\ &= \left[\sum_{i=k+1}^{d}(\mathbf{e}_i^t\mathbf{x})\mathbf{e}_i\right]^t\left[\sum_{i=k+1}^{d}(\mathbf{e}_i^t\mathbf{x})\mathbf{e}_i\right] \\ &= \sum_{k=k+1}^{d}(\mathbf{e}_i^t\mathbf{x})^2, \end{aligned}$$

because $\mathbf{e}^t\mathbf{e} = \delta_{ij}$, where δ_{ij} is the Kronecker delta which has value 1 if $i = j$ and 0 otherwise. If we now take the expected value of the error over all $\mathbf{x}$, we get using the definitions in part (a):

$$\begin{aligned} \mathcal{E}_x[\mathcal{E}_k(\mathbf{x})] &= \sum_{i=k+1}^{d}\mathcal{E}_x[(\mathbf{x}^t\mathbf{e}_i)^t(\mathbf{x}^t\mathbf{e}_i)] \\ &= \sum_{i=k+1}^{d}\mathbf{e}_i^t\mathbf{\Sigma}_i\mathbf{e}_i \end{aligned}$$

since all $\mathbf{e}_i$ are eigenvectors of the covariance matrix $\mathbf{\Sigma}$, we find

$$\sum_{i=k+1}^{d}\lambda_i\mathbf{e}_i^t\mathbf{e}_i = \sum_{k=k+1}^{d}\lambda_i.$$

This then gives us the expected error. Thus to minimize the squared-error criterion, the k-dimensional subspace should be spanned by the k largest eigenvectors, wince the error is the sum of the remaining $d-k$ eigenvalues of the covariance matrix.

45. PROBLEM NOT YET SOLVED
46. PROBLEM NOT YET SOLVED
47. PROBLEM NOT YET SOLVED
48. PROBLEM NOT YET SOLVED
49. PROBLEM NOT YET SOLVED

Section 10.14

50. Consider the three points $\mathbf{x}_1 = \binom{1}{0}, \mathbf{x}_2 = \binom{0}{0}$, and $\mathbf{x}_3 = \binom{0}{1}$. Then we have the distances

$$\begin{aligned}
\delta_{12} = ||\mathbf{x}_1 - \mathbf{x}_2|| &= \sqrt{(1-0)^2 + (0-0)^2} = 1 \\
\delta_{13} = ||\mathbf{x}_1 - \mathbf{x}_3|| &= \sqrt{(1-0)^2 + (0-1)^2} = \sqrt{2} \\
\delta_{23} = ||\mathbf{x}_2 - \mathbf{x}_3|| &= \sqrt{(0-0)^2 + (0-1)^2} = 1.
\end{aligned}$$

We assume for definiteness, and without loss of generality, that the transformed points obey $0 = y_1 < y_2 < y_3$. We define

$$\begin{aligned}
y_2 &= \Delta_1^2 > 0 \\
y_3 &= \Delta_1^2 + \Delta_2^2 > \Delta_1^2 = y_2,
\end{aligned}$$

and therefore we have

$$\begin{aligned}
d_{12} &= y_2 - y_1 = \Delta_1^2 - 0 = \Delta_1^2 \\
d_{13} &= y_3 - y_1 = \Delta_1^2 + \Delta_2^2 - 0 = \Delta_1^2 + \Delta_2^2 \\
d_{23} &= y_3 - y_2 = \Delta_1^2 + \Delta_2^2 - \Delta_1^2 = \Delta_2^2.
\end{aligned}$$

(a) From the definition in Eq. 107 in the text we have

$$\begin{aligned}
J_{ee} &= \frac{\sum_{i<j} (d_{ij} - \delta_{ij})^2}{\sum_{i<j} \delta_{ij}^2} \\
&= \frac{(\Delta_1^2 - 1)^2 + (\Delta_1^2 + \Delta_2^2 - \sqrt{2})^2 + (\Delta_2^2 - 1)^2}{1 + 1 + 2}.
\end{aligned}$$

Therefore, minimizing J_{ee} is equivalent to minimizing

$$f = (\Delta_1^2 - 1)^2 + (\Delta_2^2 - 1)^2 + (\Delta_1^2 + \Delta_2^2 - \sqrt{2})^2.$$

In order to minimize f, we must calculate its derivatives:

$$\begin{aligned}
\frac{\partial f}{\partial \Delta_1} &= 2(\Delta_1^2 - 1)2\Delta_1 + 2(\Delta_1^2 + \Delta_2^2 - \sqrt{2})2\Delta_1 \\
\frac{\partial f}{\partial \Delta_2} &= 2(\Delta_2 - 1)2\Delta_2 + 2(\Delta_1^2 + \Delta_2^2 - \sqrt{2})2\Delta_2.
\end{aligned}$$

We set these derivatives to zero and find

$$\begin{aligned}
\left[(\Delta_1^2 - 1) + (\Delta_1^2 + \Delta_2^2 - \sqrt{2})\right] 4\Delta_1 = 0 \\
\left[(\Delta_2^2 - 1) + (\Delta_1^2 + \Delta_2^2 - \sqrt{2})\right] 4\Delta_2 = 0
\end{aligned}$$

Because $\Delta_1, \Delta_2 \neq 0$, we have

$$\Delta_1^2 - 1 = -(\Delta_1^2 + \Delta_2^2 - \sqrt{2}) = \Delta_2^2 - 1,$$

and thus $\Delta_1^2 = \Delta_2^2$. We have, moreover,

$$\Delta_1^2 - 1 + (\Delta_1^2 + \Delta_2^2 - \sqrt{2}) = 0$$

and thus $3\Delta_1^2 = 1 + \sqrt{2}$, which implies

$$\begin{aligned} \Delta_1^2 = \Delta_2^2 &= \frac{1+\sqrt{2}}{3} \\ y_2 = \Delta_1^2 &= \frac{1+\sqrt{2}}{3} \\ y_3 = \Delta_1^2 + \Delta_2^2 &= 2\Delta_1^2 = 2y_2. \end{aligned}$$

(b) From definition in Eq. 108 in the text we have

$$\begin{aligned} J_{ff} &= \sum_{i<j} \left(\frac{d_{ij} - \delta_{ij}}{\delta_{ij}} \right)^2 \\ &= \left(\frac{\Delta_1^2 - 1}{1} \right)^2 + \left(\frac{\Delta_1^2 + \Delta_2^2 - \sqrt{2}}{\sqrt{2}} \right)^2 + \left(\frac{\Delta_2^2 - 1}{1} \right)^2 \\ &= (\Delta_1^2 - 1)^2 + \frac{1}{2}(\Delta_1^2 + \Delta_2^2 - \sqrt{2})^2 + (\Delta_2^2 - 1)^2, \end{aligned}$$

and thus the derivatives

$$\begin{aligned} \frac{\partial J_{ff}}{\partial \Delta_1} &= 2(\Delta_1^2 - 1)2\Delta_1 + \frac{1}{2}2(\Delta_1^2 + \Delta_2^2 - \sqrt{2})2\Delta_1 \\ \frac{\partial J_{ff}}{\partial \Delta_2} &= 2(\Delta_2^2 - 1)2\Delta_2 + \frac{1}{2}2(\Delta_1^2 + \Delta_2^2 - \sqrt{2})2\Delta_2. \end{aligned}$$

We set $\partial J_{ff}/\partial \Delta_i = 0$ and obtain (for $\Delta_1, \Delta_2 \neq 0$)

$$\begin{aligned} \Delta_1^2 - 1 &= -\frac{1}{2}(\Delta_1^2 + \delta_2^2 - \sqrt{2}) \\ &= \Delta_2^2 - 1 \end{aligned}$$

and thus $\Delta_1^2 = \Delta_2^2$. This result implies

$$\Delta_1^2 - 1 + \frac{1}{2}(\Delta_1^2 + \Delta_2^2 - \sqrt{2}) = 0,$$

and thus

$$\begin{aligned} \Delta_1^2 - 1 + \frac{1}{2}2\Delta_1^2 - \frac{1}{\sqrt{2}} &= 0 \\ 2\Delta_1^2 = 1 + \frac{1}{\sqrt{2}} &= \frac{\sqrt{2}+1}{\sqrt{2}} = \frac{2+\sqrt{2}}{2}. \end{aligned}$$

We solve for Δ_1^2 and find

$$\Delta_1^2 = \frac{2+\sqrt{2}}{4}$$

Using the above relations, we find y_2 and y_3 to be

$$y_2 = \Delta_1^2 = \frac{2+\sqrt{2}}{4}$$
$$y_3 = \Delta_1^2 + \Delta_2^2 = 2\Delta_1^2 = 2y_2.$$

Computer Exercises

Section 10.4

1. Computer exercise not yet solved
2. Computer exercise not yet solved
3. Computer exercise not yet solved
4. Computer exercise not yet solved
5. Computer exercise not yet solved

Section 10.5

6. Computer exercise not yet solved
7. Computer exercise not yet solved

Section 10.6

8. Computer exercise not yet solved

Section 10.7

9. Computer exercise not yet solved

Section 10.8

10. Computer exercise not yet solved

Section 10.9

11. Computer exercise not yet solved
12. Computer exercise not yet solved

Section 10.11

13. Computer exercise not yet solved

Section 10.12

14. Computer exercise not yet solved

Section 10.13

15. Computer exercise not yet solved
16. Computer exercise not yet solved

Section 10.14

17. Computer exercise not yet solved

Sample final exams and solutions

EXAM 1, three hours, 100 points

1. (10 points total) Consider two categories, each of which is described by a d-dimensional Gaussian having the same (but arbitrary) covariance, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, arbitrary means, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and arbitrary priors, $P(\omega_1)$ and $P(\omega_2)$. Show that the minimum-error decision boundary is a hyperplane, described by $\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0$. Express $\mathbf{w}$ and $\mathbf{x}_0$ in terms of the variables given.

2. (20 points total) Consider a one-dimensional two-category classification problem with equal priors, $P(\omega_1) = P(\omega_2) = 1/2$, where the densities have the form

$$p(x|\omega_i) = \begin{cases} 0 & x < 0 \\ \frac{2}{w_i}\left(1 - \frac{x}{w_i}\right) & 0 \le x \le w_i \\ 0 & \text{otherwise,} \end{cases}$$

where the w_i for $i = 1, 2$, are positive but unknown parameters.

(a) Confirm that the distributions are normalized.

(b) The following data were collected: $\mathcal{D}_1 = \{2, 5\}$ and $\mathcal{D}_2 = \{3, 9\}$ for ω_1 and ω_2, respectively. Find the maximum-likelihood values $\hat{w}_1$ and $\hat{w}_2$.

(c) Given your answer to part (b), determine the decision boundary x^* for minimum classification error. Be sure to state which category is to right (higher) values than x^*, and which to the left (lower) values than x^*.

(d) What is the expected error of your classifier in part (b)?

3. (10 points total) Consider a two-dimensional, three-category pattern classification problem, with equal priors $P(\omega_1) = P(\omega_2) = P(\omega_3) = 1/3$. We define the "disk distribution" $D(\boldsymbol{\mu}, r)$ to be uniform inside a circular disc centered on $\boldsymbol{\mu}$ having radius r and elsewhere 0. Suppose we model each distribution as a "disk" $D(\boldsymbol{\mu}_i, r_i)$ and after training our classifier find the following parameters:

$$\omega_1 : \qquad \boldsymbol{\mu}_1 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \quad r_1 = 2$$

$$\omega_2: \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 4 \\ 1 \end{pmatrix}, \quad r_2 = 1$$

$$\omega_3: \qquad \boldsymbol{\mu}_3 = \begin{pmatrix} 5 \\ 4 \end{pmatrix}, \quad r_3 = 3$$

(a) (**1 pt**) Use this information to classify the point $\mathbf{x} = \binom{6}{2}$ with minimum probability of error.

(b) (**1 pt**) Use this information to classify the point $\mathbf{x} = \binom{3}{3}$.

(c) (**8 pts**) Use this information to classify the point $\mathbf{x} = \binom{*}{0.5}$, where $*$ denotes a missing feature.

4. (10 points total) It is easy to see that the nearest-neighbor error rate P can equal the Bayes rate P^* if $P^* = 0$ (the best possibility) or if $P^* = (c-1)/c$ (the worst possibility). One might ask whether or not there are problems for which $P = P^*$ where P^* is between these extremes.

(a) Show that the Bayes rate for the one-dimensional case where $P(\omega_i) = 1/c$ and

$$P(x|\omega_i) = \begin{cases} 1 & 0 \le x \le \frac{cr}{c-1} \\ 1 & i \le x \le i+1-\frac{cr}{c-1} \\ 0 & \text{elsewhere} \end{cases}$$

is $P^* = r$.

(b) Show that for this case the nearest-neighbor rate is $P = P^*$.

5. (10 points total) Consider a d-n_H-c three-layer backpropagation network, trained on the traditional sum-squared error criterion, where the inputs have been "standardized" to have mean zero and unit variance in each feature. All non-linear units are sigmoidal, with transfer function

$$f(net) = a\tanh(b\ net) = \frac{2a}{1+e^{-2b\ net}} - a$$

where $a = 1.716$ and $b = 2/3$, and net is the appropriate net activation at a unit, as described in the text.

(a) State why when initializing weights in the network we never set them all to have value zero.

(b) Instead, to speed learning we initialize weights in a range $-\tilde{w} \le w \le +\tilde{w}$. For the input-to-hidden weights, what should be this range? That is, what is a good value of $\tilde{w}$?

(c) Explain what your answer to part (b) achieves. That is, explain as specifically as possible the motivation that leads to your answer in part (b).

6. (15 points total) Consider training a tree-based classifier with the following eight points of the form $\binom{x_1}{x_2}$ in two categories

ω_1	ω_2
$\binom{0}{4}$, $\binom{1}{2}$, $\binom{6}{0}$, $\binom{7}{3}$	$\binom{3}{5}$, $\binom{4}{7}$, $\binom{5}{6}$, $\binom{8}{1}$

using an entropy or information impurity and queries of the form "Is $x_i \leq \theta$?" (or "Is $x_i \geq \theta$?").

(a) What is the information impurity at the root node (that is, before any splitting)?

(b) What is the optimal query at the root node?

(c) What is the information impurity at each of the immediate descendent nodes (that is, the two nodes at the next level)?

(d) Combine the information impurities of these nodes to determine the impurity at this level. How much is the information impurity reduced by your decision in part (b)?

(e) Continue splitting to create the full tree with "pure" leaf nodes. Show your final tree, being sure to indicate the queries and the labels on the leaf nodes.

7. (10 points total) We define the "20% trimmed mean" of a sample to be the mean of the data with the top 20% of the data and the bottom 20% of the data removed. Consider the following six points, $\mathcal{D} = \{0, 4, 5, 9, 14, 15\}$.

(a) Calculate the jackknife estimate of the 20% trimmed mean of $\mathcal{D}$.

(b) State the formula for the variance of the jackknife estimate.

(c) Calculate the variance of the jackknife estimate of the 20% trimmed mean of $\mathcal{D}$.

8. (8 points total) In multi-dimensional scaling, we take points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with inter-point distances δ_{ij} in a high-dimensional space and map them to points $\mathbf{y}_1 \ldots, \mathbf{y}_n$ in a low-dimensional space, having inter-point distances d_{ij}. One measure or criterion of quality of such a mapping is

$$J_{ee} = \frac{\sum_{i<j}^{n} (d_{ij} - \delta_{ij})^2}{\sum_{i<j}^{n} \delta_{ij}^2}.$$

(a) Suppose we had a non-optimal mapping (configuration) and wanted to adjust the position of one of the points $\mathbf{y}_k$ so as to reduce the J_{ee} criterion. Take the derivative $\boldsymbol{\nabla}_{\mathbf{y}_k} J_{ee}$ to show which direction $\mathbf{y}_k$ should be moved.

(b) Write pseudocode for an iterative procedure for full multi-dimensional scaling, using your result from part (a).

9. (7 points total) Short answer.

(a) In self-organizing feature maps (Kohonen maps, topologically correct maps), why is it important to employ a "window function" $\Lambda(|\mathbf{y} - \mathbf{y}^*|)$? What does $\mathbf{y}^*$ represent in this context?

(b) In backpropagation using sigmoids described in Problem 5 (above), why do we train with teaching values ± 1 rather than ± 1.716, the limits of the output units?

(c) Of all classifiers that can be described by a parameter vector $\boldsymbol{\theta}$, must the classifier with maximum-likelihood $\hat{\boldsymbol{\theta}}$ have the smallest error? Explain or give a simple example.

(d) State in just a few sentences the "Occam's razor" principle, and informally what it implies or counsels in pattern recognition.

(e) When creating a three-component classifier system for a c-category problem through standard boosting, we train the first component classifier C_1 on a subset of the data. We then select another subset data for training the second component classifier C_2. How do we select this next set of points for training C_2? Why this way, and not for instance randomly?

(f) Summarize briefly the No Free Lunch Theorem, referring specifically to the use of "off training set" data.

(g) State how cross validation is used in the training of a general classifier.

Important formulas

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{k=1}^{c} P(\omega_k)p(\mathbf{x}|\omega_k)}$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\begin{aligned}\mathbf{\Sigma} &= \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] \\ \mathbf{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^t \\ \mathbf{\Sigma} &= \int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})d\mathbf{x}\end{aligned}$$

$$P(\omega_i|\mathbf{x}_g) = \frac{\int P(\omega_i|\mathbf{x}_g,\mathbf{x}_b)p(\mathbf{x}_g,\mathbf{x}_b)d\mathbf{x}_b}{p(\mathbf{x}_g)}$$

$$l(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta})$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}$$

$$Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i) = \mathcal{E}_{\mathcal{D}_b}\left[\ln p(\mathcal{D}_g,\mathcal{D}_b;\ \boldsymbol{\theta})|\mathcal{D}_g;\ \boldsymbol{\theta}^i\right]$$

$$\lim_{n\to\infty} P_n(e|\mathbf{x}) = 1 - \sum_{i=1}^{c} P^2(\omega_i|\mathbf{x})$$

$$P^* \leq P \leq P^*\left(2 - \frac{c}{c-1}P^*\right)$$

$$\begin{aligned}J_p(\mathbf{a}) &= \sum_{\mathbf{y}\in\mathcal{Y}}(-\mathbf{a}^t\mathbf{y}) \\ \mathbf{a}(k+1) &= \mathbf{a}(k) + \eta(k)\sum_{\mathbf{y}\in\mathcal{Y}_k}\mathbf{y}\end{aligned}$$

$$\begin{aligned}\Delta w_{kj} &= \eta\delta_k y_j = \eta(t_k - z_k)f'(net_k)y_j \\ \Delta w_{ji} &= \eta\delta_j x_i = \eta\left[\sum_{k=1}^{c} w_{kj}\delta_k\right] f'(net_j)x_i\end{aligned}$$

$$f(net) = a\,\tanh[b\;net] = \frac{2a}{1+e^{-2b\;net}} - a$$

$$i(N) = -\sum_j P(\omega_j)\log_2 P(\omega_j)$$

$$\mathcal{E}_{\mathcal{D}}\left[(g(\mathbf{x};\;\mathcal{D}) - F(\mathbf{x}))^2\right] = (\mathcal{E}_{\mathcal{D}}\left[g(\mathbf{x};\;\mathcal{D}) - F(\mathbf{x})\right])^2 + \mathcal{E}_{\mathcal{D}}\left[g(\mathbf{x};\;\mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\;\mathcal{D})]\right)^2]$$

$$\begin{aligned}\hat{\theta}_{(i)} &= \hat{\theta}(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) \\ \hat{\theta}_{(\cdot)} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}\end{aligned}$$

SOLUTION EXAM 1, three hours

1. (10 points total) Consider two categories, each of which is described by a d-dimensional Gaussian having the same (but arbitrary) covariance, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, arbitrary means, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and arbitrary priors, $P(\omega_1)$ and $P(\omega_2)$. Show that the minimum-error decision boundary is a hyperplane, described by $\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0$. Express $\mathbf{w}$ and $\mathbf{x}_0$ in terms of the variables given.

Solution Under the stated conditions, the Gaussian densities are written

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right],$$

and the appropriate discriminant function here is $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$. In this case, then, we have

$$g_i(\mathbf{x}) = \underbrace{-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Sigma}|}_{\textit{same for both categories}} -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i).$$

We expand and cancel the quadratic term $\mathbf{x}\mathbf{\Sigma}^{-1}\mathbf{x}$, which is the same for both categories (and hence can be eliminated), regroup, and find

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0}$$

where

$$\mathbf{w}_i = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i$$

and

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i),$$

where we used the symmetry of $\mathbf{\Sigma}^{-1}$ to rewrite terms such as $\mathbf{x}^t\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ by $\boldsymbol{\mu}^t\mathbf{\Sigma}^{-1}\mathbf{x}$ and $[\mathbf{\Sigma}^{-1}]^t$ by $\mathbf{\Sigma}^{-1}$.

We now seek the decision boundary, that is, where $g_1(\mathbf{x}) - g_2(\mathbf{x}) = 0$, and this implies the $\mathbf{x}$-dependent term is merely the difference between the two weights $\mathbf{w}_i$, that is, our equation is $\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0$ where

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Now we seek $\mathbf{x}_0$. The $\mathbf{x}$-independent term is

$$\mathbf{w}^t\mathbf{x}_0 = -\frac{1}{2}\boldsymbol{\mu}_1^t\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \ln P(\omega_1) + \frac{1}{2}\boldsymbol{\mu}_2^t\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 - \ln P(\omega_2).$$

Thus we have

$$\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\mathbf{x}_0 = -\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \ln\frac{P(\omega_1)}{P(\omega_2)}.$$

We left-multiply both sides by $-(1/2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t$, then divide both sides by the scalar $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and find

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{\ln P(\omega_1)/P(\omega_2)}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

(Note, this is effectively the derivation of Eqs. 59–65 on pages 39–40 in the text.)

2. (20 points total) Consider a one-dimensional two-category classification problem with equal priors, $P(\omega_1) = P(\omega_2) = 1/2$, where the densities have the form

$$p(x|\omega_i) = \begin{cases} 0 & x < 0 \\ \frac{2}{w_i}\left(1 - \frac{x}{w_i}\right) & 0 \le x \le w_i \\ 0 & \text{otherwise,} \end{cases}$$

where the w_i for $i = 1, 2$, are positive but unknown parameters.

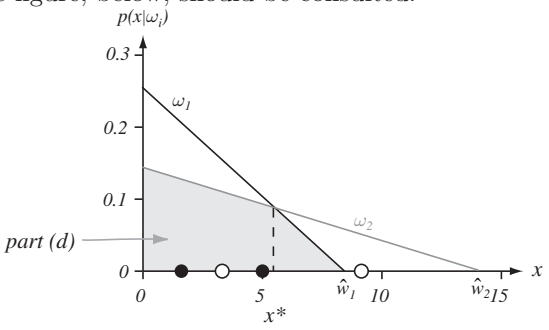
(a) Confirm that the distributions are normalized.

(b) The following data were collected: $\mathcal{D}_1 = \{2, 5\}$ and $\mathcal{D}_2 = \{3, 9\}$ for ω_1 and ω_2, respectively. Find the maximum-likelihood values $\hat{w}_1$ and $\hat{w}_2$.

(c) Given your answer to part (b), determine the decision boundary x^* for minimum classification error. Be sure to state which category is to right (higher) values than x^*, and which to the left (lower) values than x^*.

(d) What is the expected error of your classifier in part (b)?

Solution The figure, below, should be consulted.



(a) We can ignore subscripts as we check normalization

$$\int_0^w \frac{2}{w}\left(1 - \frac{x}{w}\right) dx = \left[\frac{2x}{w} - \frac{x^2}{w^2}\right]_0^w = [(2-1) - (0-0)] = 1.$$

(b) The likelihood is

$$\begin{aligned} p(\mathcal{D}|w) &= p(x_1|w)p(x_2|w) = \frac{2}{w}\left(1 - \frac{x_1}{w}\right)\frac{2}{w}\left(1 - \frac{x_2}{w}\right) \\ &= \frac{4}{w^2}\left[1 - \frac{x_1 + x_2}{w} + \frac{x_1 x_2}{w^2}\right]. \end{aligned}$$

To find the maximum-likelihood solution, we compute

$$\begin{aligned} \frac{\partial p(\mathcal{D}|w)}{\partial w} &= 4\left[-2w^{-3} + 3(x_1 + x_2)w^{-4} - 4x_1x_2w^{-5}\right] \\ &= -4w^{-5}\left[2w^2 - 3(x_1 + x_2)w + 4x_1x_2\right] \\ &= 0. \end{aligned}$$

Clearly, we are not interested in the solutions $w = \infty$. Thus we use the quadratic equation to find the solution:

$$\hat{w} = \frac{3(x_1 + x_2) \pm \sqrt{9(x_1 + x_2)^2 - 4(2)(4x_1x_2)}}{4}.$$

For category ω_1, we substitute $x_1 = 2$ and $x_2 = 5$ and find

$$\begin{aligned} \hat{w}_1 &= \frac{21 \pm \sqrt{9 \cdot 49 - 320}}{4} \\ &= \frac{21 \pm \sqrt{121}}{4} \\ &= 8 \text{ or } 2.5. \end{aligned}$$

Clearly, the solution 2.5 is invalid, since it is smaller than one of the points in $\mathcal{D}_1$; thus $\hat{w}_1 = 8$.

Likewise, for ω_2, we substitute $x_1 = 3$ and $x_2 = 9$ and find

$$\begin{aligned} \hat{w}_2 &= \frac{36 \pm \sqrt{9 \cdot 12^2 - 864}}{4} \\ &= \frac{36 \pm \sqrt{432}}{4} \\ &= 14.2 \text{ or } 3.1. \end{aligned}$$

Clearly, the solution 3.1 is invalid, since it is smaller than one of the points in $\mathcal{D}_2$; thus $\hat{w}_1 = 14.2$.

(c) We seek the value of x such that the posteriors are equal, that is, where

$$\begin{aligned} \frac{2}{\hat{w}_1}\left(1 - \frac{x}{\hat{w}_1}\right) &= \frac{2}{\hat{w}_2}\left(1 - \frac{x}{\hat{w}_2}\right) \\ \frac{2}{8}\left(1 - \frac{x}{8}\right) &= \frac{2}{14.2}\left(1 - \frac{x}{14.2}\right) \end{aligned}$$

which has solution

$$x^* = \frac{6.2}{\left(\frac{14.2}{8} - \frac{8}{14.2}\right)} = 5.1,$$

with $\mathcal{R}_1$ being points less than x^* and $\mathcal{R}_2$ being points higher than x^*.

(d) The probability of error in this optimal case is

$$\begin{aligned} P(e) &= \int \min[P(\omega_1)p(x|\omega_1), P(\omega_2)p(x|\omega_2)]dx \\ &= \int_0^{x^*=5.1} \frac{1}{2}\frac{2}{14.2}\left(1 - \frac{x}{14.2}\right)dx + \int_{x^*=5.1}^{\hat{w}_1=8} \frac{1}{2}\frac{2}{8}\left(1 - \frac{x}{8}\right)dx \\ &= \frac{1}{14.2}x\Big|_0^{5.1} - \frac{x^2}{2(14.2)^2}\Big|_0^{5.1} + \frac{1}{8}x\Big|_{5.1}^{8} - \frac{x^2}{2(8)^2}\Big|_{5.1}^{8} \\ &= 0.360. \end{aligned}$$

A more geometrically inspired approach is to use the area of a triangle, $A = (base \cdot height)/2$ and the figure. If you followed this approach, though, you had to remember to divide your shaded area by the total area under the two straight lines, and account for priors.

3. (10 points total) Consider a two-dimensional, three-category pattern classification problem, with equal priors $P(\omega_1) = P(\omega_2) = P(\omega_3) = 1/3$. We define the "disk distribution" $D(\boldsymbol{\mu}, r)$ to be uniform inside a circular disc centered on $\boldsymbol{\mu}$ having radius r and elsewhere 0. Suppose we model each distribution as such a "disk" $D(\boldsymbol{\mu}_i, r_i)$ and after training our classifier find the following parameters:

$$\omega_1 : \quad \boldsymbol{\mu}_1 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \quad r_1 = 2$$

$$\omega_2 : \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 4 \\ 1 \end{pmatrix}, \quad r_2 = 1$$

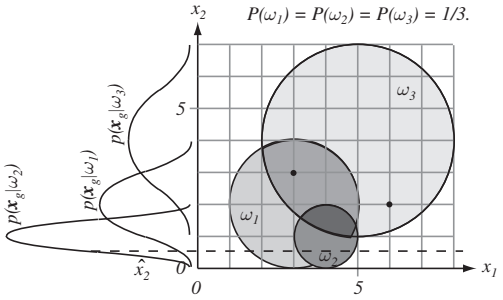
$$\omega_3 : \quad \boldsymbol{\mu}_3 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \quad r_3 = 3$$

(a) (**1 pt**) Use this information to classify the point $\mathbf{x} = \binom{6}{2}$ with minimum probability of error.

(b) (**1 pt**) Use this information to classify the point $\mathbf{x} = \binom{3}{3}$.

(c) (**8 pts**) Use this information to classify the point $\mathbf{x} = \binom{*}{0.5}$, where $*$ denotes a missing feature.

Solution



(a) As is clear from the figure below, $\binom{6}{2}$ should be classified as ω_3.

(b) As can be seen from the figure, $\binom{3}{3}$ is in the non-zero range of ω_1 and ω_3. Because the density $p(\mathbf{x}|\omega_1) > p(\mathbf{x}|\omega_3)$ at that position, however, $\binom{3}{3}$ should be classified as ω_1.

(c) Clearly, the deficient point cannot be in ω_3. We must marginalize over the bad feature, x_1, to find the class-conditional densities given the good feature, x_2. But notice that the dashed line goes through a greater percentage of the diameter of the ω_2 disk than the ω_1 disk. In short, if we marginalize the $p(\mathbf{x}|\omega_2)$ distribution over x_1 with $x_2 = 0.5$, we get a larger value than if we perform the same marginalization for $p(\mathbf{x}|\omega_1)$, as graphed in the figure. Thus we should classify $\binom{*}{0.5}$ as ω_2.

4. (10 points total) It is easy to see that the nearest-neighbor error rate P can equal the Bayes rate P^* if $P^* = 0$ (the best possibility) or if $P^* = (c-1)/c$ (the worst possibility). One might ask whether or not there are problems for which $P = P^*$ where P^* is between these extremes.

(a) Show that the Bayes rate for the one-dimensional case where $P(\omega_i) = 1/c$ and

$$P(x|\omega_i) = \begin{cases} 1 & 0 \le x \le \frac{cr}{c-1} \\ 1 & i \le x \le i+1-\frac{cr}{c-1} \\ 0 & \text{elsewhere} \end{cases}$$

is $P^* = r$.

(b) Show that for this case the nearest-neighbor rate is $P = P^*$.

Solution It is indeed possible to have the nearest-neighbor error rate P equal to the Bayes error rate P^* for non-trivial distributions.

(a) Consider uniform priors over c categories, that is, $P(\omega_i) = 1/c$, and one-dimensional distributions given in the problem statement. The evidence is

$$p(x) = \sum_{i=1}^{c} p(x|\omega_i)P(\omega_i) = \begin{cases} 1 & 0 \le x \le \frac{cr}{c-1} \\ 1/c & i \le x \le (i+1)-\frac{cr}{c-1} \\ 0 & \text{elsewhere.} \end{cases}$$

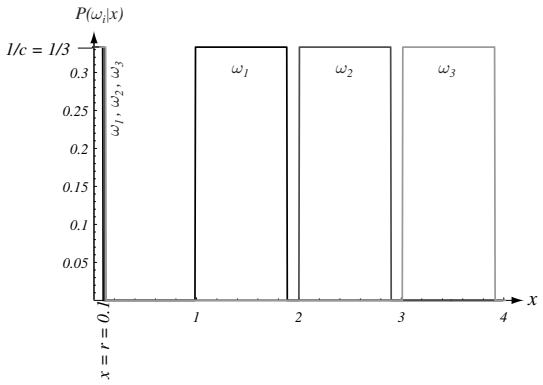
Note that this automatically imposes the restriction

$$0 \le \frac{cr}{c-1} \le 1.$$

Because the $P(\omega_i)$'s are constant, we have $P(\omega_i|x) \propto p(x|\omega_i)$ and thus

$$P(\omega_i|x) = \begin{cases} \frac{P(\omega_i)}{p(x)} = \frac{1/c}{p(x)} & 0 \le x \le \frac{cr}{c-1} \\ \left.\begin{array}{ll} 0 & \text{if } i \ne j \\ 1 & \text{if } i = j \end{array}\right\} & j \le x \le j+1-\frac{cr}{c-1} \\ 0 & \text{otherwise,} \end{cases}$$

as shown in the figure for $c = 3$ and $r = 0.1$. The conditional Bayesian



probability of error at a point x is

$$\begin{aligned} P^*(e|x) &= 1 - P(\omega_{max}|x) \\ &= \begin{cases} 1 - \frac{1/c}{p(x)} & \text{if} \quad 0 \le x \le \frac{cr}{c-1} \\ 1 - 1 = 0 & \text{if} \quad i \le x \le i+1-\frac{cr}{c-1}, \end{cases} \end{aligned}$$

and to calculate the full Bayes probability of error, we integrate as

$$\begin{aligned} P^* &= \int P^*(e|x)p(x)dx \\ &= \int_0^{cr/(c-1)} \left[1 - \frac{1/c}{p(x)}\right] p(x)dx \\ &= \left(1 - \frac{1}{c}\right)\frac{cr}{c-1} = r. \end{aligned}$$

(b) The nearest-neighbor error rate is

$$\begin{aligned} P &= \int \left[1 - \sum_{i=1}^{c} P^2(\omega_i|x)\right] p(x)dx \\ &= \int_0^{cr/(c-1)} \left[1 - \frac{c(\frac{1}{c})^2}{p^2(x)}\right] p(x)dx + \underbrace{\sum_{j=1}^{c} \int_j^{j+1-\frac{cr}{c-1}} [1-1]\, p(x)dx}_{0} \\ &= \int_0^{cr/(c-1)} \left(1 - \frac{1/c}{p^2(x)}\right) p(x)dx \\ &= \int_0^{cr/(c-1)} \left(1 - \frac{1}{c}\right) dx = \left(1 - \frac{1}{c}\right)\frac{cr}{c-1} = r. \end{aligned}$$

Thus we have demonstrated that $P^* = P = r$ in this nontrivial case. (Note: this is Problem 8 from Chapter 4 on your homework.)

5. (10 points total) Consider a d-n_H-c three-layer backpropagation network, trained on the traditional sum-squared error criterion, where the inputs have been "standardized" to have mean zero and unit variance in each feature. All non-linear units are sigmoidal, with transfer function

$$f(net) = a\tanh(b\ net) = \frac{2a}{1 + e^{-2b\ net}} - a$$

where $a = 1.716$ and $b = 2/3$, and net is the appropriate net activation at a unit, as described in the text.

(a) State why when initializing weights in the network we never set them all to have value zero.

(b) Instead, to speed learning we initialize weights in a range $-\tilde{w} \leq w \leq +\tilde{w}$. For the input-to-hidden weights, what should be this range? That is, what is a good value of $\tilde{w}$?

(c) Explain what your answer to part (b) achieves. That is, explain as specifically as possible the motivation that leads to your answer in part (b).

Solution

(a) If all the weights are zero, the net activation net_j in all hidden units is zero, and thus so is their output y_j. The weight update for the input-to-hidden weights would also be zero,

$$\Delta w_{kj} = \eta(t_k - z_k)f'(net_k)\underbrace{y_j}_{=0}.$$

Likewise, the weight update for the input-to-hidden weights would be zero

$$\Delta w_{ji} = \eta \underbrace{\left[\sum_{k=1}^{c} w_{kj}\delta_k\right]}_{=0} f'(net_j)x_i.$$

In short, if all the weights were zero, learning could not progress. Moreover, as described in the text, if all the weights are the *same* for a given pair of layers (even if these weights are not *zero*), then learning cannot progress.

(b) Suppose all the weights had value 1.0. In calculating its *net*, each hidden unit is summing d random variables with variance 1.0. The variance of the value of *net* would be, then, d. We want, however, for the net activation to be in roughly the linear range of $f(net)$, that is, $-1 < net < +1$. Thus we want our weights to be in the range $-1/\sqrt{d} < w < +1/\sqrt{d}$.

(c) We want to use the *linear* range of the transfer function, $f(net)$, so as to implement the simple linear model first. If the problem is complex, the training will change the weights and thus express the nonlinearities, if necessary.

6. (15 points total) Consider training a tree-based classifier with the following eight points of the form $\binom{x_1}{x_2}$ in two categories

ω_1	ω_2
$\binom{0}{4}$, $\binom{1}{2}$, $\binom{6}{0}$, $\binom{7}{3}$	$\binom{3}{5}$, $\binom{4}{7}$, $\binom{5}{6}$, $\binom{8}{1}$

using an entropy or information impurity and queries of the form "Is $x_i \leq \theta$?" (or "Is $x_i \geq \theta$?").

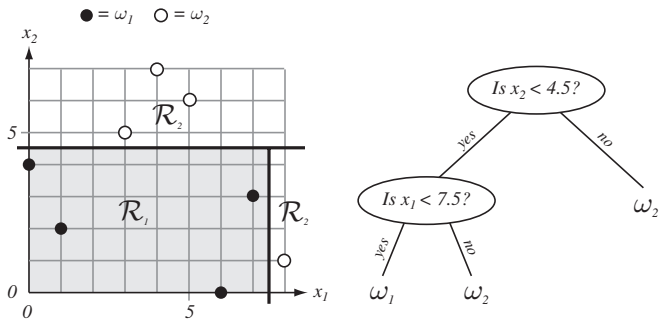
(a) What is the information impurity at the root node (that is, before any splitting)?

(b) What is the optimal query at the root node?

(c) What is the information impurity at each of the immediate descendent nodes (that is, the two nodes at the next level)?

(d) Combine the information impurities of these nodes to determine the impurity at this level. How much is the information impurity reduced by your decision in part (b)?

(e) Continue splitting to create the full tree with "pure" leaf nodes. Show your final tree, being sure to indicate the queries and the labels on the leaf nodes.



Solution

(a) The entropy impurity at the root node is

$$\begin{aligned} i(root) &= -\sum_{j=1}^{2} P(\omega_j)\log_2 P(\omega_j) \\ &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} \\ &= 1 \text{ bit.} \end{aligned}$$

(b) As is clear from the figure, the query "Is $x_2 < 4.5$" the optimal split at the root.

(c) The "left" node contains four ω_1 points and one ω_2 point. Its impurity is thus

$$-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.258 + 0.464 = 0.722 \text{ bits.}$$

The "right" node is "pure," containing three ω_2 points. It's impurity is $i(R) = -1\log_2 1 - 0\log_2 0 = 0$ bits.

(d) The impurity at the level beneath the root is just the sum of the impurities of the two nodes computed in part (b) weighted by the probability any pattern at the root goes to that node. Thus the impurity at the level is

$$\frac{5}{8}0.722 + \frac{3}{8}0 = 0.451 \text{ bits.}$$

The split at the root thus reduced the information impurity by $1.0 - 0.451 = 0.549$ bits.

(e) As can be seen in the figure, the query at the impure node should be "Is $x_1 < 7.5$."

7. (10 points total) We define the "20% trimmed mean" of a sample to be the mean of the data with the top 20% of the data and the bottom 20% of the data removed. Consider the following six points, $\mathcal{D} = \{0, 4, 5, 9, 14, 15\}$.

(a) Calculate the jackknife estimate of the 20% trimmed mean of $\mathcal{D}$.

(b) State the formula for the variance of the jackknife estimate.

(c) Calculate the variance of the jackknife estimate of the 20% trimmed mean of $\mathcal{D}$.

Solution The jackknife estimate of a statistic is merely the average of the leave-one-out estimates. In this case, there are five points, and thus

(a) We delete one point at a time, take the remaining data set, remove the top 20% and bottom 20% (that is, the highest point and the lowest point), and calculate the mean. We then average these trimmed means. As shown in the table, the jackknife estimate of the 20% trimmed mean is 8.33.

deleted point	trimmed set	mean of trimmed set
0	$\{5, 9, 14\}$	28/3 (9.33)
4	$\{5, 9, 14\}$	28/3 (9.33)
5	$\{4, 9, 14\}$	27/3 (9.00)
9	$\{4, 5, 14\}$	23/3 (7.66)
14	$\{4, 5, 9\}$	18/3 (6.00)
15	$\{4, 5, 9\}$	18/3 (6.00)
		$\frac{142/3}{6} = 7.89$

(b) The variance of the jackknife estimate is

$$\mathrm{Var}_{jack}[\theta] = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2,$$

though other forms are occassionally used here too.

(c) We perform the standard variance calculation, here consisting of the sum of six terms, each the square of the difference between the leave-one-out mean and the jackknife mean:

$$\begin{aligned}\mathrm{Var}_{jack}[\theta] &= \frac{5}{6}\big[(9.33-7.89)^2 + (9.33-7.89)^2 + (9.00-7.89)^2 \\ &\quad +(7.66-7.89)^2 + (6.00-7.89)^2 + (6.00-7.89)^2\big] \\ &= 0.833\,[2.07 + 2.07 + 1.23 + 0.53 + 3.57 + 3.57] \\ &= 10.48.\end{aligned}$$

8. (8 points total) In multi-dimensional scaling, we take points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with inter-point distances δ_{ij} in a high-dimensional space and map them to points $\mathbf{y}_1 \ldots, \mathbf{y}_n$ in a low-dimensional space, having inter-point distances d_{ij}. One measure or criterion of quality of such a mapping is

$$J_{ee} = \frac{\sum_{i<j}^{n}(d_{ij} - \delta_{ij})^2}{\sum_{i<j}^{n}\delta_{ij}^2}.$$

(a) Suppose we had a non-optimal mapping (configuration) and wanted to adjust the position of one of the points $\mathbf{y}_k$ so as to reduce the J_{ee} criterion. Take the derivative $\boldsymbol{\nabla}_{\mathbf{y}_k} J_{ee}$ to show which direction $\mathbf{y}_k$ should be moved.

(b) Write pseudocode for an iterative procedure for full multi-dimensional scaling, using your result from part (a).

Solution

(a) We can break J_{ee} into terms that depend upon $\mathbf{y}_k$, and those that do not:

$$J_{ee} = \frac{1}{\sum_{i<j}^{n} \delta_{ij}^2} \left[\underbrace{\sum_{i\neq k, j\neq k}^{n} (d_{ij} - \delta_{ij})^2}_{does\ not\ depend\ on\ \mathbf{y}_k} + \underbrace{\sum_{i\neq k}^{n} (d_{ik} - \delta_{ik})^2}_{depends\ on\ \mathbf{y}_k} \right].$$

Thus when we take the derivative, we need only consider the second term. Thus we have

$$\begin{aligned} \nabla_{\mathbf{y}_k} J_{ee} &= \frac{1}{\sum_{i<j}^{n} \delta_{ij}^2} \frac{d}{d\mathbf{y}_k} \left[\sum_{i\neq k}^{n} (d_{ik} - \delta_{ik})^2 \right] \\ &= \frac{1}{\sum_{i<j}^{n} \delta_{ij}^2} \left[2 \sum_{i\neq k}^{n} (d_{ik} - \delta_{ik}) \frac{d}{d\mathbf{y}_k} d_{ik} \right]. \end{aligned}$$

We note that

$$d_{ik} = \sqrt{(\mathbf{y}_k - \mathbf{y}_i)^t (\mathbf{y}_k - \mathbf{y}_i)}$$

and thus

$$\frac{d}{d\mathbf{y}_k} d_{ik} = \frac{1}{2} \frac{2(\mathbf{y}_k - \mathbf{y}_i)}{d_{ik}} = \frac{\mathbf{y}_k - \mathbf{y}_i}{d_{ik}}.$$

We put all these results together and find

$$\nabla_{\mathbf{y}_k} J_{ee} = \frac{2}{\sum_{i<j}^{n} \delta_{ij}^2} \sum_{i\neq k}^{n} (d_{ki} - \delta_{ki}) \frac{\mathbf{y}_k - \mathbf{y}_i}{d_{ki}}.$$

(b) The multi-dimensional scaling algorithm can be written:

1 **Initialize** $\eta \leftarrow$ learning rate, $\mathbf{y}_i, i = 1, \ldots, n$
2 **Do** Randomly select a single $\mathbf{y}_k$
3 Compute $\nabla_{\mathbf{y}_k} J_{ee}$
4 $\mathbf{y}_k \leftarrow \mathbf{y}_k - \eta \nabla_{\mathbf{y}_k} J_{ee}$
5 **Until** No change in any $\mathbf{y}_k$
6 **Return** $\mathbf{y}_i, i = 1, \ldots n$
7 **End**

9. (7 points total) Short answer.

(a) In self-organizing feature maps (Kohonen maps, topologically correct maps), why is it important to employ a "window function" $\Lambda(|\mathbf{y} - \mathbf{y}^*|)$? What does $\mathbf{y}^*$ represent in this context?

Solution In this context, $\mathbf{y}^*$ is the output unit (in the low-dimensional space) currently most active. The learning rule updates its weight by adding to it the current input vector. In this way, subsequent presentations of the input pattern will lead to an even larger activation in $\mathbf{y}^*$. The window function $\Lambda(|\mathbf{y} - \mathbf{y}^*|)$ is large for $|\mathbf{y} - \mathbf{y}^*|$ small, and decreases as $|\mathbf{y} - \mathbf{y}^*|$ increases. Thus, units *near* $\mathbf{y}^*$ will be updated by adding to them the current input vector, though with a slightly smaller magnitude. In this way, after such learning neighborhoods in the input space map to neighborhoods in the output space — we have a "topologically correct" mapping.

(b) In backpropagation using sigmoids described in Problem 5 (above), why do we train with teaching values ± 1 rather than ± 1.716, the limits of the output units?

Solution If we used teaching values ± 1.716, learning would be unacceptably long, as weight magnitudes increase and increase so as to ensure net activations are $\pm\infty$. Using teaching values of ± 1 keeps weights in bounds and speeds learning, since no weights are driven to very large magnitude to obtain an output of 1.716.

(c) Of all classifiers that can be described by a parameter vector $\boldsymbol{\theta}$, must the classifier with maximum-likelihood $\hat{\boldsymbol{\theta}}$ have the smallest error? Explain or give a simple example.

Solution No. If the candidate model space does not include the true model (e.g., distribution), then even the classifier trained by maximum-likelihood methods need not be the best in this candidate model space. For instance, suppose in a one-dimensional, two-category classification problem the true distributions each consists of two spikes, one sufficiently small but a distant "outlier." Suppose we fit these distributions with Gaussians. It is then possible to "switch" the positions of the estimated means and yield a classifier with error = 100%, even though a classifier with error = 0% is in the model space. (This was illustrated in Problem 8 in Chapter 3.)

(d) State in just a few sentences the "Occam's razor" principle, and informally what it implies or counsels in pattern recognition.

Solution William of Occam (1285–1349) stated that "Entities should not be multiplied beyond necessity." (Actually, since he was writing in Latin, he stated "Entia non sunt multiplicanda praeter necessitatem.") In pattern classification, this has come to be interpreted as counselling the use of "simpler" models rather than complex ones, fewer parameters rather than more, and "smoother" generalizers rather than those that are less smooth. The mathematical descendents of this philosophical principle of parsimony appear in minimum-description-length principles, having numerous manifestations in learning, for instance regularization, pruning, and overfitting avoidance.

(e) When creating a three-component classifier system for a c-category problem through standard boosting, we train the first component classifier C_1 on a subset of the data. We then select another subset data for training the second component classifier C_2. How do we select this next set of points for training C_2? Why this way, and not for instance randomly?

Solution We seek a data set that contains information not already learned by C_1, that is, is not well represented by C_1. A data set which C_1 classifies

with 0% accuracy is, paradoxically, *not* independent of C_1. Likewise, a data set that is classified with 50% accuracy (in the general c-category case) is also not independent of C_1. Thus we seek a data set that C_1 correctly classifies $1/c$ of the patterns. A simple algorithm is to test all candidate patterns and place them in two sets: ones correctly classified by C_1 (call it $\mathcal{D}^+$) and those incorrectly classified by C_1 (call it $\mathcal{D}^-$). Then, if our data set is to have n_2 patterns, choose n_2/c patterns randomly from $\mathcal{D}^+$ and $n_2(1-1/c)$ from $\mathcal{D}^-$.

(f) Summarize briefly the No Free Lunch Theorem, referring specifically to the use of "off training set" data.

Solution In the absense of any information about the classification problem (or target function to be learned), on average no classifier method or learning algorithm is better than any other method, including random guessing on points in the "off training set" data, that is, the points not used for training the classifier.

(g) State how cross-validation is used in the training of a general classifier.

Solution Given a training set $\mathcal{D}$, we randomly remove some portion of the set, for instance 10%, and keep this as a "validation set." We train the classifier on the remaining 90%, and monitor the error on the validation set as training proceeds. This validation set acts as a representative of future test patterns that the classifier will classify. We stop training at the first minimum of the validation error — a heuristic to avoid overfitting and improve generalization.

EXAM 2, three hours, 100 points

1. (10 points total) Let the components of the vector $\mathbf{x} = (x_1, \ldots, x_d)^t$ be ternary valued (1, 0, or -1) with

$$\begin{aligned} p_{ij} &= \Pr[x_i = \ \ 1|\omega_j] \\ q_{ij} &= \Pr[x_i = \ \ 0|\omega_j] \\ r_{ij} &= \Pr[x_i = -1|\omega_j]. \end{aligned}$$

and with the components of x_i being statistically independent for all $\mathbf{x}$ in ω_j. Show that a minimum probability of error decision rule can be derived that involves discriminant functions $g_j(\mathbf{x})$ that are quadratic functions of the components x_i.

2. (15 points total) Consider a one-dimensional two-category classification problem with equal priors, $P(\omega_1) = P(\omega_2) = 1/2$, where the densities have the form

$$p(x|\omega_i) = \begin{cases} 0 & x < 0 \\ \theta_i e^{-\theta_i x} & x \geq 0, \end{cases}$$

where the θ_i for $i = 1, 2$, are positive but unknown parameters.

(a) (**1 pt**) Confirm that the distributions are normalized.

(b) (**8 pts**) The following data were collected: $\mathcal{D}_1 = \{1, 5\}$ and $\mathcal{D}_2 = \{3, 9\}$ for ω_1 and ω_2, respectively. Find the maximum-likelihood values $\hat{\theta}_1$ and $\hat{\theta}_2$.

(c) (**3 pts**) Given your answer to part (b), determine the decision boundary x^* for minimum classification error. Be sure to state which category is to right (higher) values than x^*, and which to the left (lower) values than x^*.

(d) (**3 pts**) What is the expected error of your classifier in part (c)?

3. (10 points total) Consider the application of the k-means clustering algorithm to the one-dimensional data set $\mathcal{D} = \{0, 1, 5, 8, 14, 16\}$ for $c = 3$ clusters.

(a) (**3 pt**) Start with the three cluster means: $m_1(0) = 2$, $m_2(0) = 6$ and $m_3(0) = 9$. What are the values of the means at the next iteration?

(b) (**5 pt**) What are the final cluster means, after convergence of the algorithm?

(c) (**2 pt**) For your final clusterer, to which cluster does the point $x = 3$ belong? To which cluster does $x = 11$ belong?

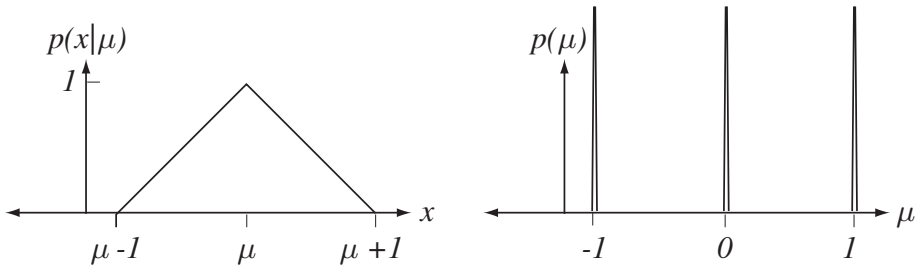
4. (10 points total) The task is to use Bayesian methods to estimate a one-dimensional probability density. The fundamental density function is a normalized triangle distribution $T(\mu, 1)$ with center at μ with half-width equal 1, defined by

$$p(x|\mu) \sim T(\mu, 1) = \begin{cases} 1 - |x - \mu| & |x - \mu| \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

as shown on the left figure. The prior information on the parameter μ is that it is equally likely to come from any of the three discrete values $\mu = -1, 0$ or 1. Stated mathematically, the prior consists of three delta functions, i.e.,

$$p(\mu) = \frac{1}{3}[\delta(x - 1) + \delta(x) + \delta(x + 1)],$$

as shown on the figure at the right. (Recall that the delta function has negligible width and unit integral.)



(a) (**2 pt**) Plot the "estimated density" before any data are collected (which we denote by $\mathcal{D}^0 = \{\}$). That is, plot $p(x|\mathcal{D}^0)$. Here and below, be sure to label and mark your axes and ensure normalization of your final estimated density.

(b) (**4 pts**) The single point $x = 0.25$ was sampled, and thus $\mathcal{D}^1 = \{0.25\}$. Plot the estimated density $p(x|\mathcal{D}^1)$.

(c) (**4 pts**) Next the point $x = 0.75$ was sampled, and thus the data set is $\mathcal{D}^2 = \{0.25, 0.75\}$. Plot the estimated density $p(x|\mathcal{D}^2)$.

5. (5 points total) Construct a cluster dendrogram for the one-dimensional data $\mathcal{D} = \{2, 3, 5, 10, 13\}$ using the distance measure $D_{max}(\mathcal{D}_i, \mathcal{D}_j)$.

6. (10 points total) Consider the use of traditional boosting for building a classifier for a two-category problem with n training points.

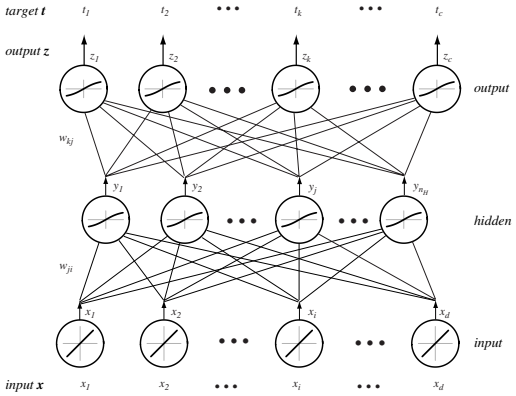
(a) (**8 pts**) Write pseudocode for traditional boosting, leading to three component classifiers.

(b) (**2 pts**) How are the resulting three component classifiers used to classify a text pattern?

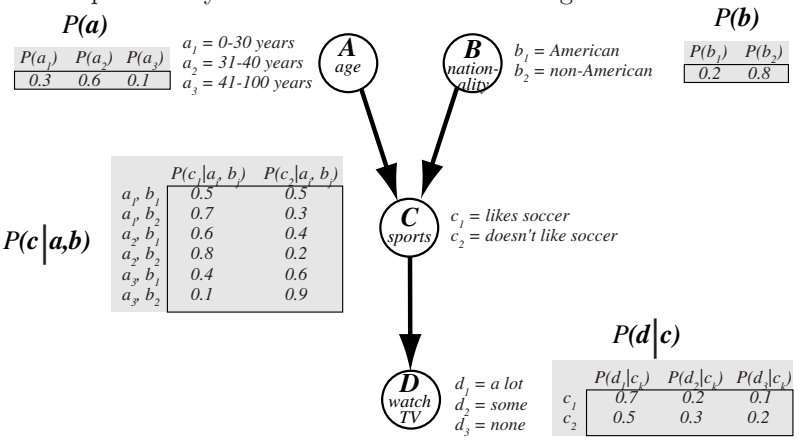
7. (5 points total) Consider a standard three-layer neural net as shown. Suppose the network is to be trained using the novel criterion function

$$J = \frac{1}{4}\sum_{k=1}^{c}(t_k - z_k)^4.$$

Derive the learning rule Δw_{kj} for the hidden-to-output weights.



8. (10 points total) It is World Cup Soccer season and a researcher has developed a Bayes belief net that expresses the dependencies among the age of a person (**A**), his or her nationality (**B**), whether he or she likes soccer (**C**), and how much he or she watches sports TV during the World Cup season (**D**). Use the conditional probability tables to answer the following.



(a) (**2 pts**) What is the probability we find a non-American who is younger than 30 who likes soccer and watches a lot of sports TV?

(b) (**4 pts**) Suppose we find someone who is an American between 31-40 years of age who watches "some" sports TV. What is the probability that this person likes soccer?

(c) (**4 pts**) Suppose we find someone over 40 years of age who never watches sports TV. What is the probability that this person likes soccer?

9. (15 points total) This problem conserns the construction of a binary decision tree for three categories from the following two-dimensional data:

ω_1	ω_2	ω_3
$\binom{1}{1}, \binom{2}{8}, \binom{3}{5}$	$\binom{4}{7}, \binom{6}{2}, \binom{7}{6}$	$\binom{5}{10}, \binom{7}{4}, \binom{7}{9}$

(a) (**3 pts**) What is the information impurity at the root node, i.e., before any splitting?

(b) (**4 pts**) The query at the root node is: "Is $x_1 > 3.5$?" What is the information impurity at the next level?

(c) (**6 pts**) Continue to grow your tree fully. Show the final tree and all queries.

(d) (**2 pts**) Use your tree to classify the point $\binom{7}{2}$.

10. (10 points total) Short answer (**1 pt each**).

(a) Explain using a diagram and a few sentences the technique of "learning with hints" and why it can improve a neural net pattern classifier.

(b) Use the Boltzmann factor to explain why at a sufficiently high "temperature" T, all configurations in a Boltzmann network are equally probable.

(c) Explain with a simple figure the "crossover" operation in genetic algorithms.

(d) What is a "surrogate split" and when is one used?

(e) If the cost for any fundamental string operation is 1.0, state the edit distance between `bookkeeper` and `beekeepers`.

(f) Suppose the Bayes error rate for a $c = 3$ category classification problem is 5%. What are the bounds on the error rate of a nearest-neighbor classifier trained with an "infinitely large" training set?

(g) What do we mean by the "language induced by grammar G"?

(h) In hidden Markov models, what does the term a_{ij} refer to? What does b_{jk} refer to?

Important formulas

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{k=1}^{c} P(\omega_k)p(\mathbf{x}|\omega_k)}$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\begin{aligned}\mathbf{\Sigma} &= \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] \\ \mathbf{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^t \\ \mathbf{\Sigma} &= \int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})d\mathbf{x}\end{aligned}$$

$$P(\omega_i|\mathbf{x}_g) = \frac{\int P(\omega_i|\mathbf{x}_g,\mathbf{x}_b)p(\mathbf{x}_g,\mathbf{x}_b)d\mathbf{x}_b}{p(\mathbf{x}_g)}$$

$$l(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta})$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}$$

$$Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i) = \mathcal{E}_{\mathcal{D}_b}\left[\ln p(\mathcal{D}_g,\mathcal{D}_b;\ \boldsymbol{\theta})|\mathcal{D}_g;\ \boldsymbol{\theta}^i\right]$$

$$d' = \frac{|\mu_2-\mu_1|}{\sigma}$$

$$\lim_{n\to\infty} P_n(e|\mathbf{x}) = 1 - \sum_{i=1}^{c} P^2(\omega_i|\mathbf{x})$$

$$P^* \leq P \leq P^*\left(2 - \frac{c}{c-1}P^*\right)$$

$$\begin{aligned} J_p(\mathbf{a}) &= \sum_{\mathbf{y}\in\mathcal{Y}}(-\mathbf{a}^t\mathbf{y}) \\ \mathbf{a}(k+1) &= \mathbf{a}(k)+\eta(k)\sum_{\mathbf{y}\in\mathcal{Y}_k}\mathbf{y} \end{aligned}$$

$$\begin{aligned} \Delta w_{kj} &= \eta\delta_k y_j = \eta(t_k - z_k)f'(net_k)y_j \\ \Delta w_{ji} &= \eta\delta_j x_i = \eta\left[\sum_{k=1}^{c} w_{kj}\delta_k\right]f'(net_j)x_i \end{aligned}$$

$$f(net) = a\,\tanh[b\;net] = a\left[\frac{e^{+b\;net}-e^{-b\;net}}{e^{+b\;net}+e^{-b\;net}}\right]$$

$$P(\gamma) = \frac{e^{-E_\gamma/T}}{Z}$$

$$Z = \sum_{\gamma'} e^{-E_{\gamma'}/T}$$

$$\Delta w_{ij} = \frac{\eta}{T}\left[\underbrace{\mathcal{E}_Q[s_i s_j]_{\alpha^i\;\alpha^o\;clamped}}_{learning} - \underbrace{\mathcal{E}[s_i s_j]_{\alpha^i\;clamped}}_{unlearning}\right]$$

$$i(N) = -\sum_{j=1}^{c} P(\omega_j)\log_2 P(\omega_j)$$

$$\mathcal{E}_{\mathcal{D}}\left[(g(\mathbf{x};\;\mathcal{D})-F(\mathbf{x}))^2\right] = \left(\mathcal{E}_{\mathcal{D}}\left[g(\mathbf{x};\;\mathcal{D})-F(\mathbf{x})\right]\right)^2 + \mathcal{E}_{\mathcal{D}}\left[g(\mathbf{x};\;\mathcal{D})-\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\;\mathcal{D})]\right]^2$$

$$\begin{aligned} \hat{\theta}_{(i)} &= \hat{\theta}(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) \\ \hat{\theta}_{(\cdot)} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \end{aligned}$$

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$$

$$\mathbf{S}_W = \sum_{i=1}^{c}\mathbf{S}_i$$

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$$

$$\mathbf{S}_T = \sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t$$

$$J_e = \mathrm{tr}[\mathbf{S}_W]$$

$$J_d = \|\mathbf{S}_W\|$$

$$\begin{aligned}
d_{min}(\mathcal{D}_i, \mathcal{D}_j) &= \min_{\substack{\mathbf{x}\in\mathcal{D} \\ \mathbf{x}\in\mathcal{D}'}} \|\mathbf{x} - \mathbf{x}'\| \\
d_{max}(\mathcal{D}_i, \mathcal{D}_j) &= \max_{\substack{\mathbf{x}\in\mathcal{D} \\ \mathbf{x}\in\mathcal{D}'}} \|\mathbf{x} - \mathbf{x}'\| \\
d_{avg}(\mathcal{D}_i, \mathcal{D}_j) &= \frac{1}{n_i n_j} \sum_{\mathbf{x}\in\mathcal{D}} \sum_{\mathbf{x}\in\mathcal{D}'} \|\mathbf{x} - \mathbf{x}'\| \\
d_{min}(\mathcal{D}_i, \mathcal{D}_j) &= \|\mathbf{m}_i - \mathbf{m}_j\|
\end{aligned}$$

$$\begin{aligned}
J_{ee} &= \frac{\sum_{i<j} (d_{ij} - \delta_{ij})^2}{\sum_{i<j} \delta_{ij}^2} \\
J_{ff} &= \sum_{i<j} \left(\frac{d_{ij} - \delta_{ij}}{\delta_{ij}} \right)^2 \\
J_{ef} &= \frac{1}{\sum_{i<j} \delta_{ij}} \sum_{i<j} \frac{(d_{ij} - \delta_{ij})^2}{\delta_{ij}}
\end{aligned}$$

$$w_{ki}(t+1) = w_{ki}(t) + \eta(t)\Lambda(|\mathbf{y} - \mathbf{y}^*|)\boldsymbol{\phi}_i$$

EXAM 2 Solutions

1. (10 points total) Let the components of the vector $\mathbf{x} = (x_1, \ldots, x_d)^t$ be ternary valued (1, 0, or -1) with

$$\begin{aligned} p_{ij} &= \Pr[x_i = \ \ 1|\omega_j] \\ q_{ij} &= \Pr[x_i = \ \ 0|\omega_j] \\ r_{ij} &= \Pr[x_i = -1|\omega_j]. \end{aligned}$$

and with the components of x_i being statistically independent for all $\mathbf{x}$ in ω_j. Show that a minimum probability of error decision rule can be derived that involves discriminant functions $g_j(\mathbf{x})$ that are quadratic functions of the components x_i.

Solution (Note: This is problem 44 from Chapter 2.)
The minimum probability of error is achieved by the following decision rule:

$$\text{Choose } \omega_k \text{ if } g_k(\mathbf{x}) \geq g_j(\mathbf{x}) \text{ for all } j \neq k,$$

where here we will use the discriminant function

$$g_j(\mathbf{x}) = \ln\, p(\mathbf{x}|\omega_j) + \ln\, P(\omega_j).$$

The components of $\mathbf{x}$ are statistically independent for all $\mathbf{x}$ in ω_j, and therefore,

$$p(\mathbf{x}|\omega_j) = p((x_1, \ldots, x_d)^t|\omega_j) = \prod_{i=1}^{d} p(x_i|\omega_j),$$

where

$$\begin{aligned} p_{ij} &= \Pr[x_i = \ \ 1|\omega_j] \\ q_{ij} &= \Pr[x_i = \ \ 0|\omega_j] \\ r_{ij} &= \Pr[x_i = -1|\omega_j]. \end{aligned}$$

As in Sect. 2.9.1 in the text, we use exponents to "select" the proper probability, that is, exponents that have value 1.0 when x_i has the value corresponding to the particular probability and value 0.0 for the other values of x_i. For instance, for the p_{ij} term, we seek an exponent that has value 1.0 when $x_i = +1$ but is 0.0 when $x_i = 0$ and when $x_i = -1$. The simplest such exponent is $\frac{1}{2}x_i + \frac{1}{2}x_i^2$. For the q_{ij} term, the simplest exponent is $1 - x_i^2$, and so on. Thus we write the class-conditional probability for a single component x_i as:

$$p(x_i|\omega_j) = p_{ij}^{\frac{1}{2}x_i+\frac{1}{2}x_i^2} q_{ij}^{1-x_i^2} r_{ij}^{-\frac{1}{2}x_i+\frac{1}{2}x_i^2} \quad \begin{matrix} i=1,\ldots,d \\ j=1,\ldots,c \end{matrix}$$

and thus for the full vector $\mathbf{x}$ the conditional probability is

$$p(\mathbf{x}|\omega_j) = \prod_{i=1}^{d} p_{ij}^{\frac{1}{2}x_i+\frac{1}{2}x_i^2} q_{ij}^{1-x_i^2} r_{ij}^{-\frac{1}{2}x_i+\frac{1}{2}x_i^2}.$$

Thus the discriminant functions can be written as

$$\begin{aligned} g_j(\mathbf{x}) &= \ln p(\mathbf{x}|\omega_j) + \ln P(\omega_j) \\ &= \sum_{i=1}^{d} \left[\left(\frac{1}{2}x_i + \frac{1}{2}x_i^2 \right) \ln p_{ij} + (1-x_i^2) \ln q_{ij} + \left(-\frac{1}{2}x_i + \frac{1}{2}x_i^2 \ln r_{ij} \right) \right] + \ln P(\omega_j) \\ &= \sum_{i=1}^{d} x_i^2 \ln \frac{\sqrt{p_{ij} r_{ij}}}{q_{ij}} + \frac{1}{2} \sum_{i=1}^{d} x_i \ln \frac{p_{ij}}{r_{ij}} + \sum_{i+1}^{d} \ln q_{ij} + \ln P(\omega_j), \end{aligned}$$

which are quadratic functions of the components x_i.

2. (15 points total) Consider a one-dimensional two-category classification problem with equal priors, $P(\omega_1) = P(\omega_2) = 1/2$, where the densities have the form

$$p(x|\omega_i) = \begin{cases} 0 & x < 0 \\ \theta_i e^{-\theta_i x} & x \geq 0, \end{cases}$$

where the θ_i for $i = 1, 2$, are positive but unknown parameters.

(a) (**1 pt**) Confirm that the distributions are normalized.
Solution: We can drop the subscripts and perform the integral

$$\int_0^\infty \theta e^{-\theta x} dx = \left[-e^{-\theta x} \right]_0^\infty = -0 - (-1) = 1.$$

(b) (**8 pts**) The following data were collected: $\mathcal{D}_1 = \{1, 5\}$ and $\mathcal{D}_2 = \{3, 9\}$ for ω_1 and ω_2, respectively. Find the maximum-likelihood values $\hat{\theta}_1$ and $\hat{\theta}_2$.
Solution: We temporarily drop the subscript on θ, denote the two training points as x_1 and x_2, and write the likelihood $p(\mathcal{D}|\theta)$ for any category as

$$\begin{aligned} p(x_1|\theta)p(x_2|\theta) &= \theta e^{-\theta x_1} \theta e^{-\theta x_2} \\ &= \theta^2 e^{-\theta(x_1+x_2)}. \end{aligned}$$

We take the derivative of this likelihood and set it to zero:

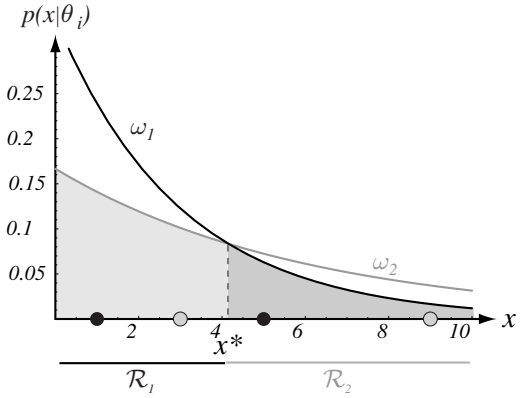
$$\begin{aligned} \frac{d}{d\theta} \left[\theta^2 e^{-\theta(x_1+x_2)} \right] &= 2\theta e^{-\theta(x_1+x_2)} + \theta^2(-(x_1+x_2)) e^{-\theta(x_1+x_2)} \\ &= \underbrace{\theta e^{-\theta(x_1+x_2)}}_{\neq 0} \underbrace{[2-(x_1+x_2)\theta]}_{=0} = 0, \end{aligned}$$

and thus $\hat{\theta} = 2/(x_1+x_2)$. For our data we have, then, $\hat{\theta}_1 = 2/(1+5) = 1/3$, and $\hat{\theta}_2 = 2/(3+9) = 1/6$.

(c) (**3 pts**) Given your answer to part (b), determine the decision boundary x^* for minimum classification error. Be sure to state which category is to right (higher) values than x^*, and which to the left (lower) values than x^*.
Solution: The decision boundary is at the point x^* where the two posteriors are equal, that is, where

$$P(\omega_1)\hat{\theta}_1 e^{-\hat{\theta}_1 x^*} = P(\omega_2)\hat{\theta}_2 e^{-\hat{\theta}_2 x^*},$$

or $1/3e^{-1/3x^*} = 1/6e^{-1/6x^*}$. We multiply each side by 6 and take the natural logarithm to find $\ln(2) - x^*/3 = -x^*/6$, or $x^* = 6\ln(2) \simeq 4.159$ with $\mathcal{R}_1$ corresponding to points less than x^*, and $\mathcal{R}_2$ points greater than x^*.

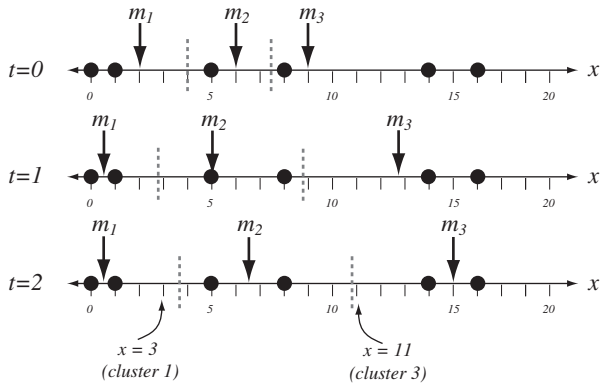


(d) (**3 pts**) What is the expected error of your classifier in part (c)?
Solution: The probability of error (in this case, the Bayes error rate) is

$$\begin{aligned}
P^* &= \int_0^\infty \text{Min}\left[P(\omega_1)\frac{1}{3}e^{x/3}, P(\omega_2)\frac{1}{6}e^{x/6}\right] dx \\
&= 0.5\left\{\int_0^{x^*} \frac{1}{6}e^{-x/6}dx + \int_{x^*}^\infty \frac{1}{3}e^{-x/3}dx\right\} \\
&= 0.5\left\{-\frac{1}{6}6e^{-x/6}\Big|_0^{x^*} + -\frac{1}{3}3e^{-x/3}\Big|_{x^*}^\infty\right\} \\
&= 0.5\left\{-e^{-x^*/6} + 1 - 0 + e^{-x^*/3}\right\} = 0.375.
\end{aligned}$$

3. (10 points total) Consider the application of the k-means clustering algorithm to the one-dimensional data set $\mathcal{D} = \{0, 1, 5, 8, 14, 16\}$ for $c = 3$ clusters.

(a) (**3 pt**) Start with the three cluster means: $m_1(0) = 2$, $m_2(0) = 6$ and $m_3(0) = 9$. What are the values of the means at the next iteration?
Solution: The top of the figure shows the initial state (i.e., at iteration $t = 0$). The dashed lines indicate the midpoints between adjacent means, and thus the cluster boundaries. The two points $x = 0$ and $x = 1$ are in the first cluster, and thus the mean for the cluster 1 in the next iteration is $m_1(1) = (0 + 1)/2 = 0.5$, as shown. Likewise, initially cluster 2 contains the single point $x = 5$, and thus the mean for cluster 2 on the next iteration is $m_2(1) = 5/1 = 5$. In the same manner, the mean for cluster 3 on the next iteration is $m_3(1) = (8 + 14 + 16)/3 = 12.67$.



(b) (**5 pt**) What are the final cluster means, after convergence of the algorithm?
Solution: After one more step of the algorithm, as above, we find $m_1(2) = 0.5$, $m_2(2) = 6.5$ and $m_3(2) = 15$.

(c) (**2 pt**) For your final clusterer, to which cluster does the point $x = 3$ belong? To which cluster does $x = 11$ belong?
Solution: As shown in the bottom figure, $x = 3$ is in cluster 1, and $x = 11$ is in cluster 3.

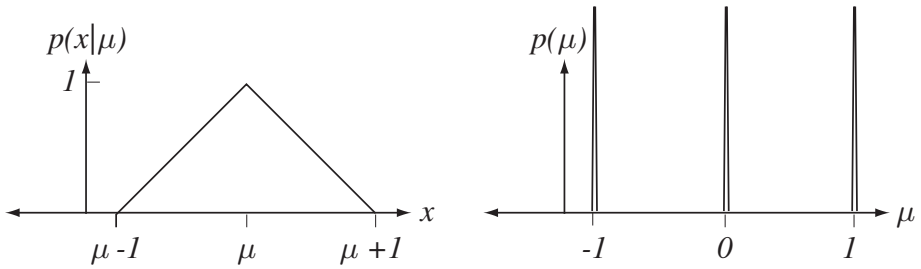
4. (10 points total) The task is to use Bayesian methods to estimate a one-dimensional probability density. The fundamental density function is a normalized triangle distribution $T(\mu, 1)$ with center at μ with half-width equal 1, defined by

$$p(x|\mu) \sim T(\mu, 1) = \begin{cases} 1 - |x - \mu| & |x - \mu| \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

as shown on the left figure. The prior information on the parameter μ is that it is equally likely to come from any of the three discrete values $\mu = -1, 0$ or 1. Stated mathematically, the prior consists of three delta functions, i.e.,

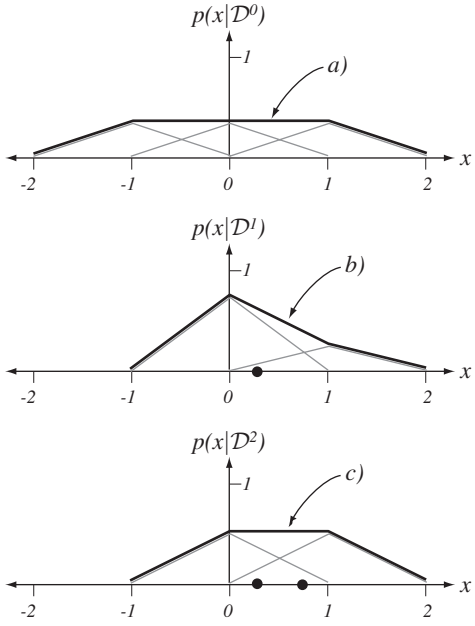
$$p(\mu) = \frac{1}{3}[\delta(x - 1) + \delta(x) + \delta(x + 1)],$$

as shown on the figure at the right. (Recall that the delta function has negligible width and unit integral.)



(a) (**2 pt**) Plot the "estimated density" before any data are collected (which we denote by $\mathcal{D}^0 = \{\}$). That is, plot $p(x|\mathcal{D}^0)$. Here and below, be sure to label and mark your axes and ensure normalization of your final estimated density.

Solution: In the absense of data, the estimate density is merely the sum of three triangle densities (of amiplitude 1/3 to ensure normalization), as shown in the top figure.



(b) (**4 pts**) The single point $x = 0.25$ was sampled, and thus $\mathcal{D}^1 = \{0.25\}$. Plot the estimated density $p(x|\mathcal{D}^1)$.

Solution: Bayesian density estimation

$$\begin{aligned} p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\ d\mu \\ &\propto \int p(x|\mu)p(\mathcal{D}|\mu)p(\mu)\ d\mu \end{aligned}$$

where here we do not worry about constants of proportionality as we shall normalize densities at the end. Because the prior consists of three delta functions, our final estimated density consists of the sum of three triangle distributions, centered on $x = -1, 0$ and 1 — the only challenge is to

determine the relative weighting of these triangle distributions. For the single point $\mathcal{D} = \{0.25\}$, clearly the weighting of the lefthand triangle (i.e., the one centered on $x = -1$) is zero because $p(\mathcal{D}^1|\mu = -1) = 0$. The relative weighting of the middle triangle (i.e., the one centered on $x = 0$) is 0.75 because $p(\mathcal{D}^1|\mu = 0) = 0.75$. Likewise, the weighting of the righthand triangle (i.e., centered on $x = 1$) is 0.25 because $p(\mathcal{D}^1|\mu = 1) = 0.25$. The priors are the same for all triangles, and thus our final density is:

$$p(x|\mathcal{D}^1) \sim 0.75T(0,1) + 0.25T(1,1),$$

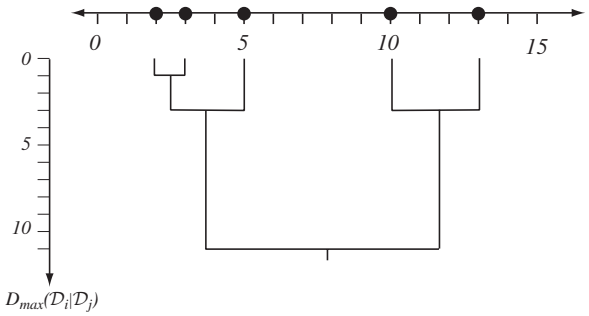
as shown in the figure.

(c) (**4 pts**) Next the point $x = 0.75$ was sampled, and thus the data set is $\mathcal{D}^2 = \{0.25, 0.75\}$. Plot the estimated density $p(x|\mathcal{D}^2)$.
Solution: As in part (b), there will be no contribution from the lefthand triangle because $p(\mathcal{D}^2|\mu = -1) = 0$. Because the points are placed symmetrically between the two other triangle centers, the contributions of these triangles must be equal, that is,

$$p(x|\mathcal{D}^2) \sim 0.5T(0,1) + 0.5T(1,1),$$

as shown in the figure.

5. (5 points total) Construct a cluster dendrogram for the one-dimensional data $\mathcal{D} = \{2, 3, 5, 10, 13\}$ using the distance measure $D_{max}(\mathcal{D}_i, \mathcal{D}_j)$.
Solution: The closest two points are 2 and 3, with distance $D_{max} = 2$, and thus they are merged. At $D_{max} = 3$, the point 5 is merged, since the distance is 3. Likewise, the points 10 and 13 are merged at this level. The two clusters $\{2, 3, 5\}$ and $\{10, 13\}$ are merged at $D_{max} = 11$, the Euclidean separation between 2 and 13.



6. (10 points total) Consider the use of traditional boosting for building a classifier for a two-category problem with n training points.

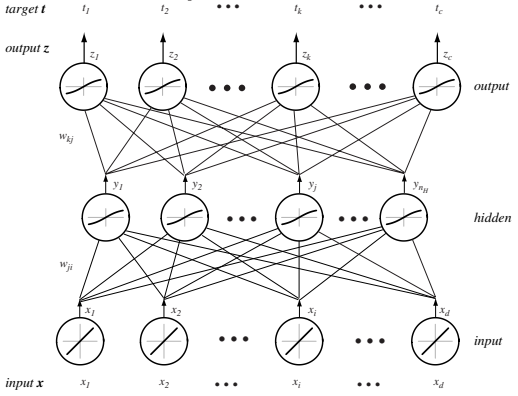
(a) (**8 pts**) Write pseudocode for traditional boosting, leading to three component classifiers.
Solution:
Algorithm[Boosting]
`Begin INITIALIZE` $\mathcal{D} = \{\mathbf{x}_1 \dots \mathbf{x}_n\}$
`Train` classifier C_1 on $\mathcal{D}_1$, i.e., $\sim n/3$ patterns chosen from $\mathcal{D}$
`Select` $\mathcal{D}_2$, i.e., roughly $n/3$ patterns that are "most informative"
`Train` classifier C_2 on $\mathcal{D}_2$
`Train` classifier C_3 on $\mathcal{D}_3$, i.e., all remaining patterns
`End`

(b) (**2 pts**) How are the resulting three component classifiers used to classify a text pattern?
Solution: Majority vote.

7. (5 points total) Consider a standard three-layer neural net as shown. Suppose the network is to be trained using the novel criterion function

$$J = \frac{1}{4}\sum_{k=1}^{c}(t_k - z_k)^4.$$

Derive the learning rule Δw_{kj} for the hidden-to-output weights.



Solution: The derivation is nearly the same as the one in the text. The chain rule for differentiation gives us:

$$\frac{\partial J}{\partial w_{kj}} = \frac{\partial J}{\partial net_k}\frac{\partial net_k}{\partial w_{kj}},$$

and because

$$J = \frac{1}{4}\sum_{k=1}^{c}(t_k - z_k)^4,$$

we have

$$\begin{aligned}\frac{\partial J}{\partial net_k} &= \frac{\partial J}{\partial z_k}\frac{\partial z_k}{\partial net_k} \\ &= -(t_k - z_k)^3 f'(net_k).\end{aligned}$$

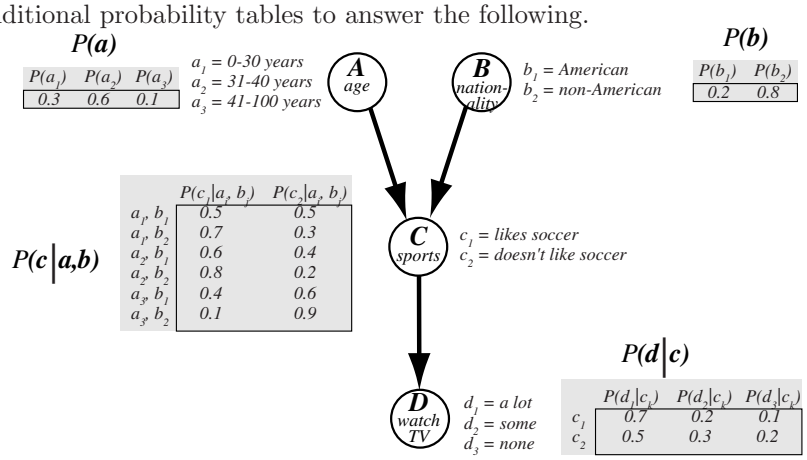
Recall, too, that

$$\frac{\partial net_k}{\partial w_{kj}} = y_j.$$

We put these together, and denote the learning rate by η to find the learning rule

$$\Delta w_{kj} = \eta(t_k - z_k)^3 f'(net_k) y_j.$$

8. (10 points total) It is World Cup Soccer season and a researcher has developed a Bayes belief net that expresses the dependencies among the age of a person (**A**), his or her nationality (**B**), whether he or she likes soccer (**C**), and how much he or she watches sports TV during the World Cup season (**D**). Use the conditional probability tables to answer the following.



(a) (**2 pts**) What is the probability we find a non-American who is younger than 30 who likes soccer and watches a lot of sports TV?
Solution: We use the conditional probabilities and compute

$$\begin{aligned} P(a_1, b_2, c_1, d_1) &= P(a_1)P(b_2)P(c_1|a_1, b_2)P(d_1|c_1) \\ &= 0.3 \cdot 0.8 \cdot 0.7 \cdot 0.7 = 0.1176. \end{aligned}$$

(b) (**4 pts**) Suppose we find someone who is an American between 31-40 years of age who watches "some" sports TV. What is the probability that this person likes soccer?
Solution: There are several ways to approach this, but perhaps the simplest is to compute

$$\begin{aligned} P(c_1|a_2, b_1, d_2) &= \frac{P(a_2, b_1, c_1, d_2)}{P(a_2, b_1, d_2)} = \alpha P(a_2, b_1, c_1, d_2) \\ &= \alpha\ P(a_2)P(b_1)P(c_1|a_2, b_1)P(d_2|c_1) \\ &= \alpha \cdot 0.6 \cdot 0.2 \cdot 0.6 \cdot 0.2 = \alpha\ 0.0144. \end{aligned}$$

Likewise we have

$$\begin{aligned} P(c_2|a_2, b_1, d_2) &= \alpha P(a_2, b_1, c_2, d_2) \\ &= \alpha\ P(a_2)P(b_1)P(c_c|a_2, b_1)P(d_2|c_c) \\ &= \alpha \cdot 0.6 \cdot 0.2 \cdot 0.4 \cdot 0.3 = \alpha\ 0.0144. \end{aligned}$$

Since $P(c_1|a_2, b_1, d_2) + P(c_2|a_2, b_1, d_2) = 1$, we fix the normalization and find

$$P(c_1|a_2, b_1, d_2) = \frac{0.0144}{0.0144 + 0.0144} = 0.5.$$

(c) (**4 pts**) Suppose we find someone over 40 years of age who never watches sports TV. What is the probability that this person likes soccer?
Solution: Again, there are several ways to calculate this probability, but this time we proceed as:

$$\begin{aligned} P(c_1|a_3, d_3) &\propto \sum_{\mathbf{b}} P(a_3, \mathbf{b}, c_1, d_3) \\ &= \alpha P(a_3)P(d_3|c_1) \sum_{\mathbf{b}} P(\mathbf{b})P(c_1|a_3, \mathbf{b}) \\ &= \alpha P(a_3)P(d_3|c_1) \left[P(b_1)P(c_1|a_3, b_1) + P(b_2)P(c_1|a_3, b_2)\right] \\ &= \alpha 0.1 \cdot 0.1 \left[0.2 \cdot 0.4 + 0.8 \cdot 0.1\right] = \alpha\ 0.0016. \end{aligned}$$

Likewise, we have

$$\begin{aligned} P(c_2|a_3, d_3) &\propto \sum_{\mathbf{b}} P(a_3, \mathbf{b}, c_2, d_3) \\ &= \alpha P(a_3)P(d_3|c_2) \sum_{\mathbf{b}} P(\mathbf{b})P(c_2|a_3, \mathbf{b}) \\ &= \alpha P(a_3)P(d_3|c_2) \left[P(b_1)P(c_2|a_3, b_1) + P(b_2)P(c_2|a_3, b_2)\right] \\ &= \alpha 0.1 \cdot 0.2 \left[0.2 \cdot 0.6 + 0.8 \cdot 0.9\right] = \alpha\ 0.0168. \end{aligned}$$

Since $P(c_1|a_3, d_3) + P(c_2|a_3, d_3) = 1$ we normalize our results to find

$$P(c_1|a_3, d_3) = \frac{0.0016}{0.0016 + 0.0168} = 0.087.$$

9. (15 points total) This problem conserns the construction of a binary decision tree for three categories from the following two-dimensional data:

ω_1	ω_2	ω_3
$\binom{1}{1}$, $\binom{2}{8}$, $\binom{3}{5}$	$\binom{4}{7}$, $\binom{6}{2}$, $\binom{7}{6}$	$\binom{5}{10}$, $\binom{7}{4}$, $\binom{7}{9}$

(a) (**3 pts**) What is the information impurity at the root node, i.e., before any splitting?
Solution: The information impurity at the root (before splitting) is

$$\begin{aligned} i(N) &= -\sum_{j=1}^{c} P(\omega_j) \log_2 P(\omega_j) \\ &= -3 \left[\frac{1}{3} \log_2 \frac{1}{3}\right] = \log_2 3 = 1.585 \text{ bits}. \end{aligned}$$

(b) (**4 pts**) The query at the root node is: "Is $x_1 > 3.5$?" What is the information impurity at the next level?

Solution: The impurity at the right node is 0, as all its points are in a single category, ω_1. The impurity at the left node, N_L, is:

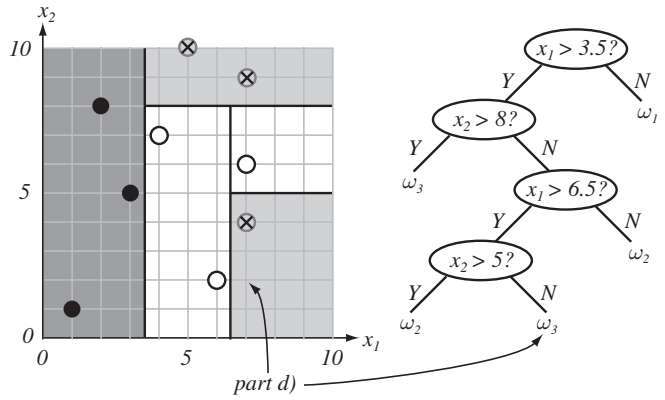
$$i(N_L) = -\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right] = \log_2 2 = 1.0 \text{ bit.}$$

The impurity at this level is the weighted sum of the impurities at the nodes, i.e.,

$$P_L \cdot 1.0 + (1 - P_L)0.0 = \frac{6}{9} \cdot 1.0 + \frac{3}{9}0.0 = 0.66 \text{ bits,}$$

where P_L and $P_R = 1 - P_L$ are the probabilities patterns go to the "left" and "right," respectively. Note that the reduction in impurity — 1.585 - 0.66 — is less than 1.0 bits, as it must from the answer to a single yes-no question.

(c) (**6 pts**) Continue to grow your tree fully. If two candidate splits are equally good, prefer the one based on x_1 (rather than x_2). Show the final tree and all queries.
Solution: See figure.

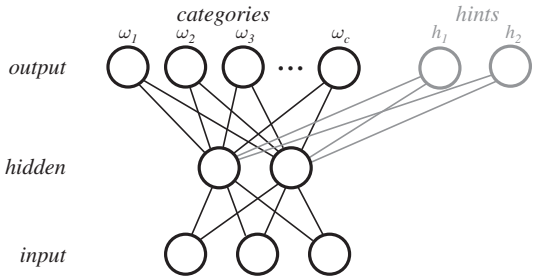


(d) (**2 pts**) Use your tree to classify the point $\binom{7}{2}$.
Solution: From the tree or the plot, it is clear that $\binom{7}{2}$ is classified as ω_3.

10. (10 points total) Short answer (**1 pt each**).

(a) Explain using a diagram and a few sentences the technique of "learning with hints" and why it can improve a neural net pattern classifier.
Solution: If a neural network is being trained on a particular classification problem, we can add to the network output units that correspond to a subsidiary, but related problem. We train the network on both problems simultaneously. We discard the hint output units after training. Such learning with hints can improve classification on the central task because the hidden units develop better features, relevant to the primary classificaation task.

(b) Use the Boltzmann factor to explain why at a sufficiently high "temperature" T, all configurations in a Boltzmann network are equally probable.

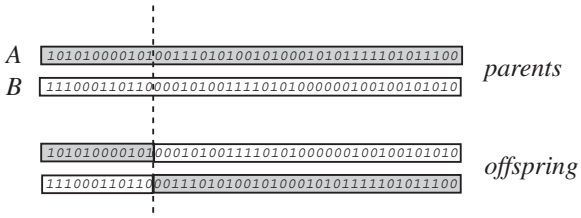


Solution: The probability of any configuration γ in a Boltzmann network is given by

$$P(\gamma) = \frac{e^{-E_\gamma/T}}{Z},$$

where the numerator is the Boltzmann factor and Z is the partition function — basically a normalization. If $T \gg E_\gamma$, then the exponent is nearly 0 regardless of γ, and the Boltzmann factor is nearly 1.0 for every state. Thus the probability of each configuration is roughly the same.

(c) Explain with a simple figure the "crossover" operation in genetic algorithms.
Solution: In genetic algorithms, we represent classifiers by a string of bits. The crossover operation employs a randomly chosen place along the string, then takes the left part from sequence A and splices it to the right side of sequence B, and vice versa, much in analogy with sexual reproduction in biology. This operation is fundamentally different from random variation and occassionally leads to particularly good or "fit" classifiers.



(d) What is a "surrogate split" and when is one used?
Solution: In addition to the traditional or "primary split" in a node in a decision tree classifier, we may provide "surrogate" splits, to be used whenever a test pattern is missing the feature queried by the primary split. The surrogate split is based on a feature *other* than the one used by the primary split, and is chosen to best approximate the action of the primary split, i.e., maximize the "predictive association" with the primary split.

(e) If the cost for any fundamental string operation is 1.0, state the edit distance between `bookkeeper` and `beekeepers`.
Solution: The edit distance is 4, as shown in the table.

x	`bookkeeper`	source string	
	`b`<u>`e`</u>`okkeeper`	substitute `e` for `o`	1
	`be`<u>`e`</u>`kkeeper`	substitute `e` for `o`	2
	`beekeeper`	delete `k`	3
y	`beekeepers`<u></u>	insert `s`	4

(f) Suppose the Bayes error rate for a $c = 3$ category classification problem is 5%. What are the bounds on the error rate of a nearest-neighbor classifier trained with an "infinitely large" training set?
Solution: In the limit of an infinitely large training set, the nearest-neighbor classifier has a probability of error P bounded as

$$P^* \leq P \leq P^* \left(2 - \frac{c}{c-1} P^*\right)$$

where P^* is the Bayes error rate. For the case $P^* = 0.05$ and $c = 3$, these limits are then

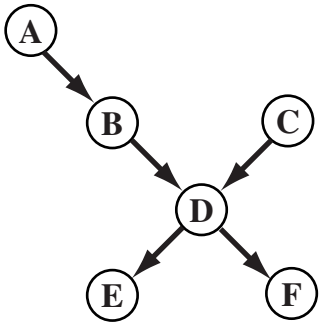
$$0.05 \leq P \leq 0.05 \left(2 - \frac{3}{2} 0.05\right) = 0.09625.$$

(g) What do we mean by the "language induced by grammar G"?
Solution: The complete set of strings (possibly infinite in number) that can be derived from the source symbol S by application of the production rules in all possible sequences.

(h) In hidden Markov models, what does the term a_{ij} refer to? What does b_{jk} refer to?
Solution: The $a_{ij} = P(\omega_j(t+1)|\omega_i(t))$ denote the transition probability from hidden state ω_i to hidden state ω_j in each time step. The $b_{jk} = P(v_k(t)|\omega_j(t))$ denote the transition probability that hidden state ω_j will emit visible symbol v_k in each time step.

EXAM 3, three hours, 100 points

1. (15 points total) A Bayes belief net consisting of six nodes and its associated conditional probability tables are shown below.



$P(a_1)$	$P(a_2)$	$P(a_3)$
0.5	0.3	0.2

$P(c_1)$	$P(c_2)$	$P(c_3)$
0.2	0.4	0.4

	$P(b_1\|a_i)$	$P(b_2\|a_i)$
a_1	0.4	0.6
a_2	0.3	0.7
a_3	0.5	0.5

	$P(d_1\|b_i,c_j)$	$P(d_2\|b_i,c_j)$
b_1,c_1	0.3	0.7
b_1,c_2	0.5	0.5
b_1,c_3	0.9	0.1
b_2,c_1	1.0	0.0
b_2,c_2	0.4	0.6
b_2,c_3	0.7	0.3

	$P(e_1\|d_i)$	$P(e_2\|d_i)$
d_1	0.1	0.9
d_2	0.8	0.2

	$P(f_1\|d_i)$	$P(f_2\|d_i)$	$P(f_3\|d_i)$
d_1	0.1	0.5	0.4
d_2	0.8	0.0	0.2

(a) (2 pt) Compute the probability $P(a_3, b_2, c_3, d_1, e_2, f_1)$.

(b) (2 pt) Compute the probability $P(a_2, b_2, c_2, d_2, e_1, f_2)$.

(c) (5 pt) Suppose we know the net is in the following (partial) state of evidence $\mathbf{e}$: a_3, b_1, c_2. What is the probability $P(f_1|\mathbf{e})$? What is the probability $P(e_2|\mathbf{e})$?

(d) (6 pt) Suppose we know the net is in the following (partial) state of evidence $\mathbf{e}$: f_1, e_2, a_2. What is the probability $P(d_1|\mathbf{e})$? What is the probability $P(e_2|\mathbf{e})$?

2. (15 points total) Consider a one-dimensional two-category classification problem with unequal priors, $P(\omega_1) = 0.7$ and $P(\omega_2) = 0.3$, where the densities have the form form of a half Gaussian "centered" at 0, i.e.,

$$p(x|\omega_i) = \begin{cases} 0 & x < 0 \\ \theta_i e^{-x^2/(2\sigma_i^2)} & x \geq 0, \end{cases}$$

where the θ_i for $i = 1, 2$, are positive but unknown parameters.

(a) (**1 pt**) Find the normalization constant θ_i as a function of σ_i.

(b) (**8 pts**) The following data were collected: $\mathcal{D}_1 = \{1, 4\}$ and $\mathcal{D}_2 = \{2, 8\}$ for ω_1 and ω_2, respectively. Find the maximum-likelihood values $\hat{\sigma}_1$ and $\hat{\sigma}_2$.

(c) (**3 pts**) Given your answer to part (b), determine the decision boundary for minimum classification error. Be sure to state which category label applies to each range in x values.

(d) (**3 pts**) Recall that the standard error function is defind as $erf(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2/2} dz$. Write a formula for the expected error of your classifier in part (c) in terms of error functions.

3. (10 points total) Consider the application of the k-means clustering algorithm to the two-dimensional data set $\mathcal{D} = \{\binom{-5}{3}, \binom{-3}{1}, \binom{-2}{6}, \binom{-1}{-7}, \binom{4}{-3}, \binom{6}{-1}\}$ for $c = 3$ clusters.

(a) (**3 pt**) Start with the three cluster means: $\mathbf{m}_1(0) = \binom{-7}{4}$, $\mathbf{m}_2(0) = \binom{7}{4}$, and $\mathbf{m}_3(0) = \binom{2}{-5}$. What are the values of the means at the next iteration?

(b) (**5 pt**) What are the final cluster means, after convergence of the algorithm?

(c) (**2 pt**) For your final clusterer, to which cluster does the point $\mathbf{x} = \binom{3}{3}$ belong? To which cluster does $\mathbf{x} = \binom{-3}{-4}$ belong?

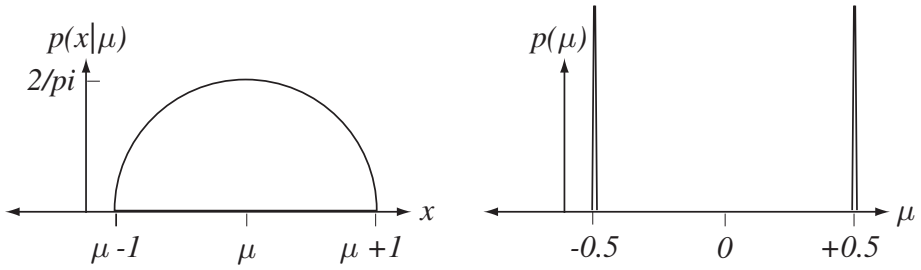
4. (15 points total) The task is to use Bayesian methods to estimate a one-dimensional probability density. The fundamental density function is a normalized "half circle" distribution $HC(\mu, 1)$ with center at μ with half-width equal 1, defined by

$$p(x|\mu) \sim HC(\mu, 1) = \begin{cases} \frac{2}{\pi}\sqrt{1-(x-\mu)^2} & |x-\mu| \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

as shown on the left figure. The prior information on the parameter μ is that it is equally likely to come from either of the two discrete values $\mu = -0.5$ or $+0.5$. Stated mathematically, the prior consists of two delta functions, i.e.,

$$p(\mu) = \frac{1}{2}[\delta(\mu - 0.5) + \delta(\mu + 0.5)],$$

as shown on the figure at the right. (Recall that the delta function has negligible width and unit integral.)



(a) (**3 pt**) Plot (sketch) the "estimated density" before any data are collected (which we denote by $\mathcal{D}^0 = \{\}$). That is, plot $p(x|\mathcal{D}^0)$. Here and below, be sure to label and mark your axes and ensure normalization of your final estimated density.

(b) (**4 pts**) The single point $x = 0.25$ was sampled, and thus $\mathcal{D}^1 = \{0.25\}$. Plot the density $p(x|\mathcal{D}^1)$ estimated by Bayesian methods.

(c) (**5 pts**) Next the point $x = 0.25$ was sampled, and thus the data set is $\mathcal{D}^2 = \{0.25, 0.25\}$. Plot the estimated density $p(x|\mathcal{D}^2)$.

(d) (**3 pts**) Suppose a very large number of points were selected and they were all 0.25, i.e., $\mathcal{D} = \{0.25, 0.25, \ldots, 0.25\}$. Plot the estimated density $p(x|\mathcal{D})$. (You don't need to do explicit calculations for this part.)

5. (5 points total) Construct a cluster dendrogram for the one-dimensional data $\mathcal{D} = \{5, 6, 9, 11, 18, 22\}$ using the distance measure $D_{avg}(\mathcal{D}_i, \mathcal{D}_j)$.

6. (5 points total) Consider learning a grammar from sentences.

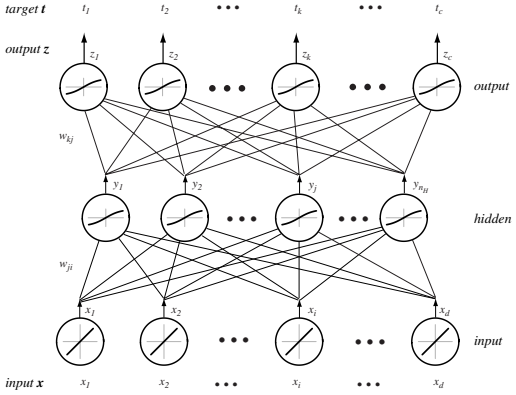
(a) (**8 pts**) Write pseudocode for simple grammatical inference. Define your terms.

(b) (**2 pts**) Define $\mathcal{D}^+$ and $\mathcal{D}^-$ and why your algorithm needs both.

7. (5 points total) Consider a standard three-layer neural net as shown. Suppose the network is to be trained using the novel criterion function

$$J = \frac{1}{6}\sum_{k=1}^{c}(t_k - z_k)^6.$$

Derive the learning rule Δw_{kj} for the hidden-to-output weights.



8. (5 points total) Prove that the single best representative pattern $\mathbf{x}_0$ for a data set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in the sum-squared-error criterion

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n} \|\mathbf{x}_0 - \mathbf{x}_k\|^2$$

is the sample mean $\mathbf{m} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$.

9. (15 points total) This problem concerns the construction of a binary decision tree for two categories from the following two-dimensional data using queries of the form "Is $x_i > x_i^*$?" for $i = 1, 2$ and the information impurity.

ω_1	ω_2
$\binom{1}{5}$, $\binom{2}{9}$, $\binom{4}{10}$, $\binom{5}{7}$, $\binom{8}{6}$	$\binom{3}{8}$, $\binom{6}{4}$, $\binom{7}{2}$, $\binom{9}{3}$

(a) (**2 pts**) What is the information impurity at the root node, i.e., before any splitting?

(b) (**3 pts**) What should be the query at the root node?

(c) (**3 pts**) How much has the impurity been reduced by the query at the root?

(d) (**3 pts**) Continue constructing your tree fully. (Whenever two candidate queries lead to the samee reduction in impurity, prefer the query that uses the x_1 feature.) Use your tree to classify $\mathbf{x} = \binom{6}{6}$ and $\mathbf{x} = \binom{3}{4}$.

(e) (**2 pts**) Suppose your tree is to be able to classify deficient patterns. What should be the first (and only) surrogate split at the root node?

10. (10 points total) Short answer (**1 pt each**).

(a) What are the four major components of a grammar G? What do we mean by the language induced by grammar G, i.e., $\mathcal{L}(G)$?

(b) Use the Boltzmann factor to explain why at a sufficiently high "temperature" T, all configurations in a Boltzmann network are equally probable.

(c) Use an equation and a few sentences to explain the minimum description length (MDL) principle.

(d) Use an equation and a few sentences to explain what is the discriminability in signal detection theory.

(e) If the cost for any fundamental string operation is 1.0, state the edit distance between `streets` and `scrams`.

(f) Suppose the Bayes error rate for a $c = 5$ category classification problem is 1%. What are the upper and lower bounds on the error rate of a nearest-neighbor classifier trained with an "infinitely large" training set?

(g) Use a formula and a sentence to explain learning with momentum in back-propagation.

(h) What is the evaluation problem in hidden Markov models?

Important formulas

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{k=1}^{c} P(\omega_k)p(\mathbf{x}|\omega_k)}$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\begin{aligned}\mathbf{\Sigma} &= \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] \\ \mathbf{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^t \\ \mathbf{\Sigma} &= \int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})d\mathbf{x}\end{aligned}$$

$$P(\omega_i|\mathbf{x}_g) = \frac{\int P(\omega_i|\mathbf{x}_g,\mathbf{x}_b)p(\mathbf{x}_g,\mathbf{x}_b)d\mathbf{x}_b}{p(\mathbf{x}_g)}$$

$$l(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta})$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}$$

$$Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^i) = \mathcal{E}_{\mathcal{D}_b}\left[\ln p(\mathcal{D}_g,\mathcal{D}_b;\ \boldsymbol{\theta})|\mathcal{D}_g;\ \boldsymbol{\theta}^i\right]$$

$$d' = \frac{|\mu_2-\mu_1|}{\sigma}$$

$$\lim_{n\to\infty} P_n(e|\mathbf{x}) = 1 - \sum_{i=1}^{c} P^2(\omega_i|\mathbf{x})$$

$$P^* \leq P \leq P^*\left(2 - \frac{c}{c-1}P^*\right)$$

$$\begin{aligned} J_p(\mathbf{a}) &= \sum_{\mathbf{y}\in\mathcal{Y}}(-\mathbf{a}^t\mathbf{y}) \\ \mathbf{a}(k+1) &= \mathbf{a}(k)+\eta(k)\sum_{\mathbf{y}\in\mathcal{Y}_k}\mathbf{y} \end{aligned}$$

$$\begin{aligned} \Delta w_{kj} &= \eta\delta_k y_j = \eta(t_k - z_k)f'(net_k)y_j \\ \Delta w_{ji} &= \eta\delta_j x_i = \eta\left[\sum_{k=1}^{c} w_{kj}\delta_k\right]f'(net_j)x_i \end{aligned}$$

$$f(net) = a\,\tanh[b\;net] = a\left[\frac{e^{+b\;net}-e^{-b\;net}}{e^{+b\;net}+e^{-b\;net}}\right]$$

$$P(\gamma) = \frac{e^{-E_\gamma/T}}{Z}$$

$$Z = \sum_{\gamma'} e^{-E_{\gamma'}/T}$$

$$\Delta w_{ij} = \frac{\eta}{T}\left[\underbrace{\mathcal{E}_Q[s_i s_j]_{\alpha^i\;\alpha^o\;clamped}}_{learning} - \underbrace{\mathcal{E}[s_i s_j]_{\alpha^i\;clamped}}_{unlearning}\right]$$

$$i(N) = -\sum_{j=1}^{c} P(\omega_j)\log_2 P(\omega_j)$$

$$\mathcal{E}_{\mathcal{D}}\left[(g(\mathbf{x};\;\mathcal{D}) - F(\mathbf{x}))^2\right] = \left(\mathcal{E}_{\mathcal{D}}\left[g(\mathbf{x};\;\mathcal{D}) - F(\mathbf{x})\right]\right)^2 + \mathcal{E}_{\mathcal{D}}\left[g(\mathbf{x};\;\mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\;\mathcal{D})]\right]^2$$

$$\begin{aligned} \hat{\theta}_{(i)} &= \hat{\theta}(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) \\ \hat{\theta}_{(\cdot)} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \end{aligned}$$

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$$

$$\mathbf{S}_W = \sum_{i=1}^{c}\mathbf{S}_i$$

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t$$

$$\mathbf{S}_T = \sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t$$

$$J_e = \mathrm{tr}[\mathbf{S}_W]$$

$$J_d = \|\mathbf{S}_W\|$$

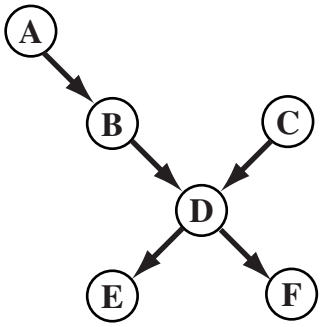
$$\begin{aligned}
d_{min}(\mathcal{D}_i, \mathcal{D}_j) &= \min_{\substack{\mathbf{x}\in\mathcal{D} \\ \mathbf{x}\in\mathcal{D}'}} \|\mathbf{x} - \mathbf{x}'\| \\
d_{max}(\mathcal{D}_i, \mathcal{D}_j) &= \max_{\substack{\mathbf{x}\in\mathcal{D} \\ \mathbf{x}\in\mathcal{D}'}} \|\mathbf{x} - \mathbf{x}'\| \\
d_{avg}(\mathcal{D}_i, \mathcal{D}_j) &= \frac{1}{n_i n_j} \sum_{\mathbf{x}\in\mathcal{D}} \sum_{\mathbf{x}\in\mathcal{D}'} \|\mathbf{x} - \mathbf{x}'\| \\
d_{min}(\mathcal{D}_i, \mathcal{D}_j) &= \|\mathbf{m}_i - \mathbf{m}_j\|
\end{aligned}$$

$$\begin{aligned}
J_{ee} &= \frac{\sum_{i<j} (d_{ij} - \delta_{ij})^2}{\sum_{i<j} \delta_{ij}^2} \\
J_{ff} &= \sum_{i<j} \left(\frac{d_{ij} - \delta_{ij}}{\delta_{ij}} \right)^2 \\
J_{ef} &= \frac{1}{\sum_{i<j} \delta_{ij}} \sum_{i<j} \frac{(d_{ij} - \delta_{ij})^2}{\delta_{ij}}
\end{aligned}$$

$$w_{ki}(t+1) = w_{ki}(t) + \eta(t)\Lambda(|\mathbf{y} - \mathbf{y}^*|)\boldsymbol{\phi}_i$$

EXAM 3 Solutions

1. (15 points total) A Bayes belief net consisting of six nodes and its associated conditional probability tables are shown below.



$P(a_1)$	$P(a_2)$	$P(a_3)$
0.5	0.3	0.2

$P(c_1)$	$P(c_2)$	$P(c_3)$
0.2	0.4	0.4

	$P(b_1\|a_i)$	$P(b_2\|a_i)$
a_1	0.4	0.6
a_2	0.3	0.7
a_3	0.5	0.5

	$P(d_1\|b_i,c_j)$	$P(d_2\|b_i,c_j)$
b_1,c_1	0.3	0.7
b_1,c_2	0.5	0.5
b_1,c_3	0.9	0.1
b_2,c_1	1.0	0.0
b_2,c_2	0.4	0.6
b_2,c_3	0.7	0.3

	$P(e_1\|d_i)$	$P(e_2\|d_i)$
d_1	0.1	0.9
d_2	0.8	0.2

	$P(f_1\|d_i)$	$P(f_2\|d_i)$	$P(f_3\|d_i)$
d_1	0.1	0.5	0.4
d_2	0.8	0.0	0.2

(a) (2 pt) Compute the probability $P(a_3, b_2, c_3, d_1, e_2, f_1)$.
Solution

$$\begin{aligned} P(a_3, b_2, c_3, d_1, e_2, f_1) &= P(a_3)P(b_2|a_3)P(c_3)P(d_1|b_2, c_3)P(e_2|d_1)P(f_1|d_1) \\ &= 0.2 \cdot 0.5 \cdot 0.4 \cdot 0.7 \cdot 0.9 \cdot 0.1 = 0.0025. \end{aligned}$$

(b) (2 pt) Compute the probability $P(a_2, b_2, c_2, d_2, e_1, f_2)$.
Solution

$$\begin{aligned} P(a_2, b_2, c_2, d_2, e_1, f_2) &= P(a_2)P(b_2|a_2)P(c_2)P(d_2|b_2, c_2)P(e_1|d_2)P(f_2|d_2) \\ &= 0.3 \cdot 0.7 \cdot 0.4 \cdot 0.6 \cdot 0.8 \cdot 0.0 = 0.0. \end{aligned}$$

(c) (5 pt) Suppose we know the net is in the following (partial) state of evidence $\mathbf{e}$: a_3, b_1, c_2. What is the probability $P(f_1|\mathbf{e})$? What is the probability $P(c_2|\mathbf{e})$?
Solution

$$\begin{aligned} P(f_1|\mathbf{e}) &= P(d_1|b_1, c_2)P(f_1|d_1) + P(d_2|b_1, c_2)P(f_1|d_2) \\ &= 0.5 \cdot 0.1 + 0.5 \cdot 0.8 = 0.45. \\ P(e_2|\mathbf{e}) &= P(d_1|b_1, c_2)P(e_2|d_1) + P(d_2|b_1, c_2)P(e_2|d_2) \\ &= 0.5 \cdot 0.9 + 0.5 \cdot 0.2 = 0.55. \end{aligned}$$

(d) (6 pt) Suppose we know the net is in the following (partial) state of evidence **e**: f_1, e_2, a_2. What is the probability $P(d_1|\mathbf{e})$? What is the probability $P(e_2|\mathbf{e})$?
Solution

$$
\begin{aligned}
P(d_1|\mathbf{e}) &= \frac{\sum_{b,c} P(b,c,d_1,a_2,e_2,f_1)}{\sum_{b,c,d} P(b,c,d,a_2,e_2,f_1)} \\
&= \frac{\sum_{b,c} P(a_2)P(b|a_2)P(c)P(d_1|b,c)P(e_2|d_1)P(f_1|d_1)}{\sum_{b,c,d} P(a_2)P(b|a_2)P(c)P(d|b,c)P(e_2|d)P(f_1|d)} \\
&= \frac{P(e_2|d_1)P(f_1|d_1)\sum_{b,c} P(b|a_2)P(c)P(d_1|b,c)}{\sum_{b,c,d} P(b|a_2)P(c)P(d|b,c)P(e_2|d)P(f_1|d)} \\
&= \frac{0.9\cdot 0.1[P(b_1|a_2)\sum_c P(c)P(d_1|b_1,c)) + P(b_2|a_2)\sum_c P(c)P(d_1|b_2,c)]}{P(e_2|d_1)P(f_1|d_1)\sum_{b,c} P(b|a_2)P(c)P(d_1|b,c) + P(e_2|d_2)P(f_1|d_2)\sum_{b,c} P(b|a_2)P(c)P(d_2|b,c)} \\
&= \frac{0.9\cdot 0.1[0.3(0.2\cdot 0.3 + 0.4\cdot 0.5 + 0.4\cdot 0.9) + 0.7(0.2\cdot 1.0 + 0.4\cdot 0.4 + 0.4\cdot 0.7)]}{P(e_2|d_1)P(f_1|d_1)\sum_{b,c} P(b|a_2)P(c)P(d_1|b,c) + P(e_2|d_2)P(f_1|d_2)\sum_{b,c} P(b|a_2)P(c)P(d_2|b,c)} \\
&= \frac{0.05706}{0.05706 + 0.2\cdot 0.8(0.3(0.2\cdot 0.7 + 0.4\cdot 0.5 + 0.4\cdot 0.1) + 0.7(0.2\cdot 0.0 + 0.4\cdot 0.6 + 0.4\cdot 0.3))} \\
&= \frac{0.05706}{0.05706 + 0.05856} \approx 0.4935.
\end{aligned}
$$

$$
\begin{aligned}
P(c_2|\mathbf{e}) &= \frac{\sum_{b,d} P(b,d,c_2,a_2,e_2,f_1)}{\sum_{b,c,d} P(b,c,d,a_2,e_2,f_1)} \\
&= \frac{\sum_{b,d} P(a_2)P(b|a_2)P(c_2)P(d|b,c_2)P(e_2|d)P(f_1|d)}{\sum_{b,c,d} P(a_2)P(b|a_2)P(c)P(d|b,c)P(e_2|d)P(f_1|d)} \\
&= \frac{P(c_2)\sum_b P(b|a_2)\sum_d P(d|b,c_2)P(e_2|d)P(f_1|d)}{\sum_c P(c)\sum_b P(b|a_2)\sum_d P(d|b,c)P(e_2|d)P(f_1|d)} \\
&= \frac{0.4(0.3(0.5\cdot 0.9\cdot 0.1 + 0.5\cdot 0.2\cdot 0.8) + 0.7(0.4\cdot 0.9\cdot 0.1 + 0.6\cdot 0.2\cdot 0.8))}{\sum_c P(c)\sum_b P(b|a_2)\sum_d P(d|b,c)P(e_2|d)P(f_1|d)} \\
&= .05196/\left(.05196 + P(c_1)\sum_b P(b|a_2)\sum_d P(d|b,c_1)P(e_2|d)P(f_1|d)\right. \\
&\quad \left. + P(c_3)\sum_b P(b|a_2)\sum_d P(d|b,c_3)P(e_2|d)P(f_1|d)\right) \\
&= .05196/\left(.05196 + 0.2(0.3(0.3\cdot 0.9\cdot 0.1 + 0.7\cdot 0.2\cdot 0.8) + 0.7(1.0\cdot 0.9\cdot 0.1 + 0.0\cdot 0.2\cdot 0.8))\right. \\
&\quad \left. + P(c_3)\sum_b P(b|a_2)\sum_d P(d|b,c_3)P(e_2|d)P(f_1|d)\right) \\
&= .05196/(.05196 + .02094 \\
&\quad + 0.4(0.3(0.9\cdot 0.9\cdot 0.1 + 0.1\cdot 0.2\cdot 0.8) + 0.7(0.7\cdot 0.9\cdot 0.1 + 0.3\cdot 0.2\cdot 0.8))) \\
&= \frac{.05196}{.05196 + .02094 + .04272} \approx 0.4494.
\end{aligned}
$$

2. (15 points total) Consider a one-dimensional two-category classification problem with unequal priors, $P(\omega_1) = 0.7$ and $P(\omega_2) = 0.3$, where the densities have the form form of a half Gaussian "centered" at 0, i.e.,

$$p(x|\omega_i) = \begin{cases} 0 & x < 0 \\ \theta_i e^{-x^2/(2\sigma_i^2)} & x \geq 0, \end{cases}$$

where the θ_i for $i = 1, 2$, are positive but unknown parameters.

(a) (**1 pt**) Find the normalization constant θ_i as a function of σ_i.
Solution
Because the normalization on a full one-dimensional Gaussian (as given on the equations on the exam) is $\frac{1}{\sqrt{2\pi}\sigma}$, the normalization on a "half" Gaussian must be twice as large, i.e.,

$$\theta_i = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma_i}.$$

(b) (**8 pts**) The following data were collected: $\mathcal{D}_1 = \{1, 4\}$ and $\mathcal{D}_2 = \{2, 8\}$ for ω_1 and ω_2, respectively. Find the maximum-likelihood values $\hat{\sigma}_1$ and $\hat{\sigma}_2$.
Solution
We drop the subscripts, denote the two training points x_1 and x_2, and compute the likelihood:

$$\begin{aligned} p(\mathcal{D}|\sigma) &= p(\{x_1, x_2\}|\sigma) = p(x_1|\sigma)p(x_2|\sigma) \\ &= \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}e^{-x_1^2/2\sigma^2} \cdot \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}e^{-x_2^2/2\sigma^2} \\ &= \frac{2}{\pi}\frac{1}{\sigma^2}e^{-(x_1^2+x_2^2)/2\sigma^2}. \end{aligned}$$

Clearly, because of the exponential we will want to work with the log-likelihood,

$$l \equiv \ln(2/\pi) - 2\ln\sigma - (x_1^2 + x_2^2)/2\sigma^2.$$

To find the maximum-likelihood solution, we first take the derivative

$$\frac{dl}{d\sigma} = \frac{-2}{\sigma} - \frac{(x_1^2 + x_2^2)}{2}\frac{-2}{\sigma^{-3}},$$

and set it to zero. Clearly we must ignore the solution $\hat{\sigma} = \infty$. The solution we seek is

$$\hat{\sigma} = \sqrt{\frac{x_1^2 + x_2^2}{2}},$$

which for the training data given yields

$$\begin{aligned} \hat{\sigma}_1 &= \sqrt{\frac{1^2 + 4^2}{2}} = \sqrt{\frac{17}{2}} \\ \hat{\sigma}_2 &= \sqrt{\frac{2^2 + 8^2}{2}} = \sqrt{34}. \end{aligned}$$

(c) (**3 pts**) Given your answer to part (b), determine the decision boundary for minimum classification error. Be sure to state which category label applies to each range in x values.

Solution

We set the posteriors equal, yielding

$$P(\omega_1)p(x^*|\omega_1) = P(\omega_2)p(x^*|\omega_2)$$

which gives the following equalities:

$$\begin{aligned} 0.7\sqrt{\frac{2}{\pi}}\frac{1}{\hat{\sigma}_1}e^{-x^{*2}/(2\hat{\sigma}_1^2)} &= 0.3\sqrt{\frac{2}{\pi}}\frac{1}{\hat{\sigma}_2}e^{-x^{*2}/(2\hat{\sigma}_2^2)} \\ \ln\left[\frac{0.7}{0.3}\frac{\hat{\sigma}_1}{\hat{\sigma}_2}\right] &= -\frac{x^{*2}}{2}\left[\frac{1}{\hat{\sigma}_2^2+\hat{\sigma}_1^2}\right] \end{aligned}$$

or

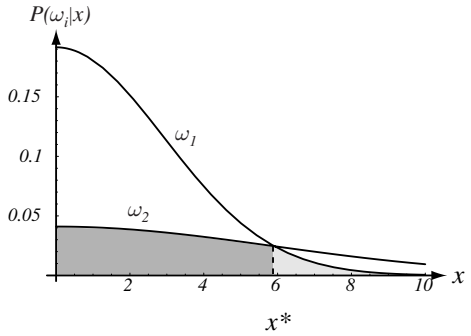
$$x^* = \sqrt{\frac{-2\ln\left[\frac{0.7}{0.3}\frac{\hat{\sigma}_1}{\hat{\sigma}_2}\right]}{\left[\frac{1}{\hat{\sigma}_2^2+\hat{\sigma}_1^2}\right]}} \simeq 5.91,$$

with $\mathcal{R}_1$ for points $0 \leq x < x^*$ and $\mathcal{R}_2$ for points $x > x^*$.

(d) (**3 pts**) Recall that the standard error function is defind as $\text{erf}(x) \equiv \frac{2}{\sqrt{\pi}}\int_0^x e^{-z^2/2}dz$. Write a formula for the expected error of your classifier in part (c) in terms of error functions.

Solution

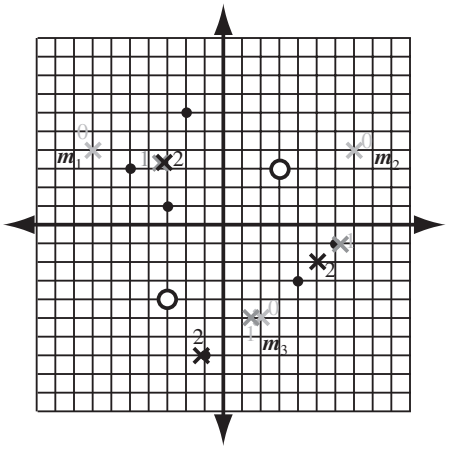
Consult the figure.



We sum the two shaded areas to find the probability of error,

$$\begin{aligned} P(error) &= \int \min[P(\omega_1)p(x|\omega_1), P(\omega_2)p(x|\omega_2)]dx \\ &= \int_0^{x^*} P(\omega_2)\sqrt{\frac{2}{\pi}}\frac{1}{\hat{\sigma}_2}e^{-x^2/(2\hat{\sigma}_2^2)}dx + \int_{x^*}^{\infty} P(\omega_1)\sqrt{\frac{2}{\pi}}\frac{1}{\hat{\sigma}_1}e^{-x^2/(2\hat{\sigma}_1^2)}dx \\ &= 0.3\frac{1}{\sqrt{2}}\frac{1}{\hat{\sigma}_2}\frac{2}{\sqrt{\pi}}\int_0^{x^*/\hat{\sigma}_2}\hat{\sigma}_2 e^{-z^2/2}dz + 0.7\frac{1}{\sqrt{2}}\frac{1}{\hat{\sigma}_1}\frac{2}{\sqrt{\pi}}\int_{x^*/\hat{\sigma}_1}^{\infty}\hat{\sigma}_1 e^{-z^2/2}dz \\ &= \frac{0.3}{\sqrt{2}}\text{erf}[x^*/\hat{\sigma}_2] + \frac{0.7}{\sqrt{2}}\left[1 - \text{erf}[x^*/\hat{\sigma}_1]\right] \simeq 0.237. \end{aligned}$$

3. (10 points total) Consider the application of the k-means clustering algorithm to the two-dimensional data set $\mathcal{D} = \{\binom{-5}{3}, \binom{-3}{1}, \binom{-2}{6}, \binom{-1}{-7}, \binom{4}{-3}, \binom{6}{-1}\}$ for $c = 3$ clusters.



(a) (**3 pt**) Start with the three cluster means: $\mathbf{m}_1(0) = \binom{-7}{4}$, $\mathbf{m}_2(0) = \binom{7}{4}$, and $\mathbf{m}_3(0) = \binom{2}{-5}$. What are the values of the means at the next iteration?
Solution
The points nearest each mean are as follows:

$$\begin{aligned} \mathbf{m}_1 &: \quad \tbinom{-5}{3}, \tbinom{-3}{1}, \tbinom{-2}{6} \\ \mathbf{m}_2 &: \quad \tbinom{6}{-1} \\ \mathbf{m}_3 &: \quad \tbinom{-1}{-7}, \tbinom{4}{-3}. \end{aligned}$$

We compute the mean for each of these sets to find the new means, i.e.,

$$\mathbf{m}_1(1) = \begin{pmatrix} -3.33 \\ 3 \end{pmatrix}, \quad \mathbf{m}_2(1) = \begin{pmatrix} 6 \\ -1 \end{pmatrix}, \quad \mathbf{m}_3(1) = \begin{pmatrix} 1.5 \\ -5 \end{pmatrix}.$$

(b) (**5 pt**) What are the final cluster means, after convergence of the algorithm?
Solution
On the next iteration, $\mathbf{m}_1$ does not change, but the others do:

$$\mathbf{m}_1(2) = \begin{pmatrix} -3.33 \\ 3 \end{pmatrix} \quad \mathbf{m}_2(2) = \begin{pmatrix} 5 \\ -2 \end{pmatrix} \quad \mathbf{m}_3(2) = \begin{pmatrix} -1 \\ -7 \end{pmatrix},$$

which is the final state.

(c) (**2 pt**) For your final clusterer, to which cluster does the point $\mathbf{x} = \binom{3}{3}$ belong? To which cluster does $\mathbf{x} = \binom{-3}{-4}$ belong?
Solution
The point $\binom{3}{3}$ is in cluster 2; the point $\binom{-3}{-4}$ is in cluster 3.

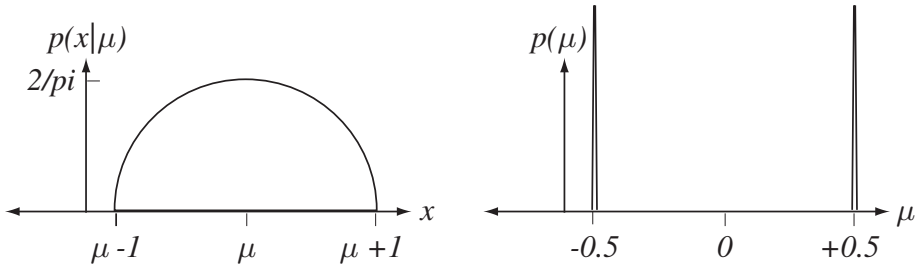
4. (15 points total) The task is to use Bayesian methods to estimate a one-dimensional probability density. The fundamental density function is a normalized "half circle" distribution $HC(\mu, 1)$ with center at μ with half-width equal 1, defined by

$$p(x|\mu) \sim HC(\mu, 1) = \begin{cases} \frac{2}{\pi}\sqrt{1-(x-\mu)^2} & |x-\mu| \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

as shown on the left figure. The prior information on the parameter μ is that it is equally likely to come from either of the two discrete values $\mu = -0.5$ or $+0.5$. Stated mathematically, the prior consists of two delta functions, i.e.,

$$p(\mu) = \frac{1}{2}[\delta(\mu - 0.5) + \delta(\mu + 0.5)],$$

as shown on the figure at the right. (Recall that the delta function has negligible width and unit integral.)

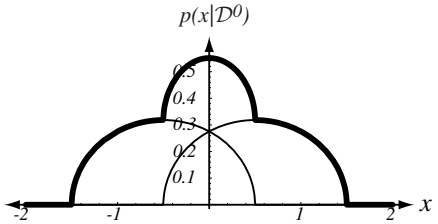


(a) (**3 pt**) Plot (sketch) the "estimated density" before any data are collected (which we denote by $\mathcal{D}^0 = \{\}$). That is, plot $p(x|\mathcal{D}^0)$. Here and below, be sure to label and mark your axes and ensure normalization of your final estimated density.

Solution In the absense of training data, our distribution is based merely on the prior values of the parameters. Written out in full, we have

$$\begin{aligned} p(x|\mathcal{D}^0) &= \int p(x|\mu)p(\mu)d\mu \\ &= \int HC(\mu, 1)0.5[\delta(\mu - 0.5) + \delta(\mu + 0.5)]d\mu \\ &= 0.5\left[HC(0.5, 1) + HC(-0.5, 1)\right], \end{aligned}$$

which is just the sum of the two disk distributions, as shown in the figure.



(b) (**4 pts**) The single point $x = 0.25$ was sampled, and thus $\mathcal{D}^1 = \{0.25\}$. Plot the density $p(x|\mathcal{D}^1)$ estimated by Bayesian methods.

Solution

We use recursive Bayesian learning, in particular $p(\mu|\mathcal{D}^1) \propto p(x_1|\mu)p(\mu|\mathcal{D}^0$, where we defer the normalization.

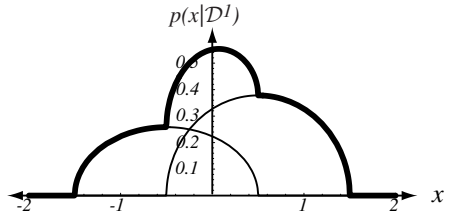
$$p(x|\mu) = \int p(x|\mu)p(x_1|\mu)\frac{1}{2}\left[\delta(\mu - 0.5) + \delta(\mu + 0.5)\right]$$

The ratio of contributions of the two components is given by the values of the component densities at $x_1 = 0.25$, i.e.,

$$\frac{\sqrt{1-(0.5-0.25)^2}}{\sqrt{1-(-.5-0.25)^2}} = \sqrt{\frac{1-1/16}{1-9/16}} = \sqrt{15/7} \equiv \alpha,$$

where we let α be this ratio. The ratio of contributions is then $\alpha/(1+\alpha)$ for the $HC(0.5, 1)$ component and $1/(1+\alpha)$ for the $HC(-0.5, 1)$ component, i.e.,

$$p(x|\mathcal{D}^1) = \frac{\alpha}{1+\alpha}HC(0.5, 1) + \frac{1}{1+\alpha}HC(-0.5, 1).$$

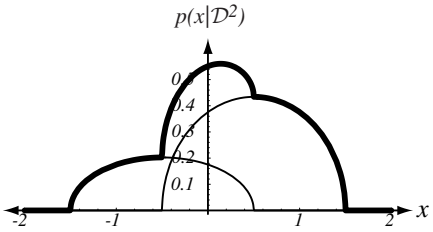


(c) (**5 pts**) Next the point $x = 0.25$ was sampled, and thus the data set is $\mathcal{D}^2 = \{0.25, 0.25\}$. Plot the estimated density $p(x|\mathcal{D}^2)$.

Solution

We use recursive Bayesian learning, in particular $p(\mu|\mathcal{D}^2) \propto p(x_2|\mu)p(\mu|\mathcal{D}^1$, where again we defer the normalization. As before, we now have $\alpha/(1+\alpha)$ for the $HC(0.5, 1)$ component and $1/(1+\alpha)$ for the $HC(-0.5, 1)$ component, i.e.,

$$p(x|\mathcal{D}^1) = \frac{\alpha^2}{1+\alpha^2}HC(0.5, 1) + \frac{1}{1+\alpha^2}HC(-0.5, 1).$$



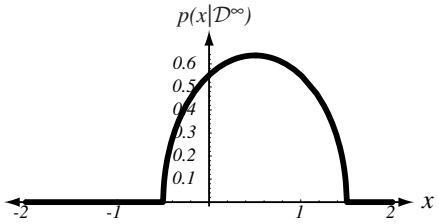
(d) (**3 pts**) Suppose a very large number of points were selected and they were all 0.25, i.e., $\mathcal{D} = \{0.25, 0.25, \ldots, 0.25\}$. Plot the estimated density $p(x|\mathcal{D})$. (You don't need to do explicit calculations for this part.)

Solution

Clearly, each subsequent piont $x = 0.5$ introduces another factor of α into the ratio of the two components. In the limit of large n, we have

$$\begin{aligned} \lim_{n\to\infty} \frac{\alpha^n}{1+\alpha^n} &\rightarrow 1; \\ \lim_{n\to\infty} \frac{1}{1+\alpha^n} &\rightarrow 0, \end{aligned}$$

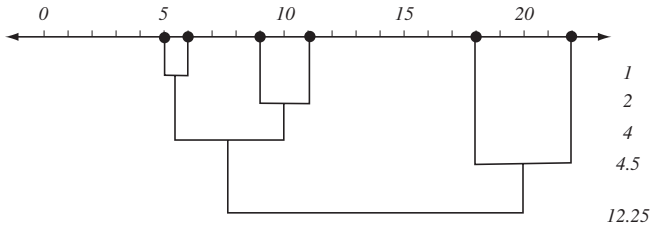
and thus the density consists solely of the component $HC(0.5, 1)$. Incidentally, this would be the appropriate maximum-likelihood solution too.



5. (5 points total) Construct a cluster dendrogram for the one-dimensional data $\mathcal{D} = \{5, 6, 9, 11, 18, 22\}$ using the distance measure $d_{avg}(\mathcal{D}_i, \mathcal{D}_j)$.

Solution

See figure. Throughout, we use $d_{avg}(\mathcal{D}_i, \mathcal{D}_j) = \frac{1}{n_i n_j} \sum_{\mathbf{x} \in \mathcal{D}} \sum_{\mathbf{x} \in \mathcal{D}'} \|\mathbf{x} - \mathbf{x}'\|$. The distances at each successive levels are: $1, 2, 4, 4.5, 12.25$.



6. (**5 points total**) Consider learning a grammar from sentences.

   (a) (**8 pts**) Write pseudocode for simple grammatical inference. Define your terms.

   Solution

   See Algorithm 5 in Chapter 8.

   (b) (**2 pts**) Define $\mathcal{D}^+$ and $\mathcal{D}^-$ and why your algorithm needs both.

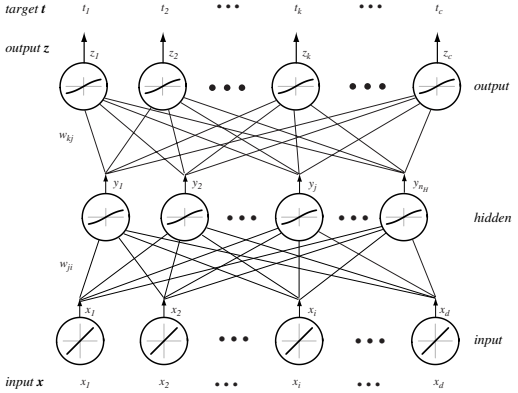
   Solution

   $\mathcal{D}^+$ is a set of sentences known to be derivable in the grammar; $\mathcal{D}^-$ is a set of sentences known to be *not* derivable in the grammar. One needs both to reduce the number of grammars consistent with training data and make learning tractable.

7. (5 points total) Consider a standard three-layer neural net as shown. Suppose the network is to be trained using the novel criterion function

$$J = \frac{1}{6} \sum_{k=1}^{c} (t_k - z_k)^6.$$

Derive the learning rule Δw_{kj} for the hidden-to-output weights.



Solution

We have $J = \frac{1}{6}\sum_{k=1}^{c}(t_k - z_k)^6$, and thus we have the derivative $\frac{\partial J}{\partial z_k} = -(t_k - z_k)^5$. We use the chain rule and find

$$\Delta w_{kj} = -\eta\frac{\partial J}{\partial w_{kj}} = -\eta\frac{\partial J}{\partial z_k}\frac{\partial z_k}{\partial \text{net}_k}\frac{\partial \text{net}_k}{\partial w_{kj}} = \eta(t_k - z_k)^5 f'(\text{net}_k)y_j.$$

8. (5 points total) Prove that the single best representative pattern $\mathbf{x}_0$ for a data set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in the sum-squared-error criterion

$$J_0(\mathbf{x}_0) = \sum_{k=1}^{n}\|\mathbf{x}_0 - \mathbf{x}_k\|^2$$

is the sample mean $\mathbf{m} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$.

Solution

Minimizing $\mathbf{x}_0$ under the sum-squared-error criterion $\sum_{k=1}^{n}\|\mathbf{x}_0 - \mathbf{x}_k\|^2$

$$0 = \frac{\partial}{\partial \mathbf{x}_0}\sum_{k=1}^{n}\|\mathbf{x}_0 - \mathbf{x}_k\|^2 = 2\sum_{k=1}^{n}(\mathbf{x}_0 - \mathbf{x}_k) = 2\left(n\cdot\mathbf{x}_0 - \sum_{k=1}^{n}\mathbf{x}_k\right)$$

and thus

$$\mathbf{x}_0 = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$$

9. (15 points total) This problem concerns the construction of a binary decision tree for two categories from the following two-dimensional data using queries of the form "Is $x_i > x_i^*$?" for $i = 1, 2$ and the information impurity.

ω_1	ω_2
$\binom{1}{5}$, $\binom{2}{9}$, $\binom{4}{10}$, $\binom{5}{7}$, $\binom{8}{6}$	$\binom{3}{8}$, $\binom{6}{4}$, $\binom{7}{2}$, $\binom{9}{3}$

(a) (**2 pts**) What is the information impurity at the root node, i.e., before any splitting?
Solution

$$i(N) = -\sum_{j=1}^{2}P(\omega_j)\log_2 P(\omega_j) = -\frac{5}{9}\log_2\frac{5}{9} + -\frac{4}{9}\log_2\frac{4}{9} \approx 0.9911$$

(b) (**3 pts**) What should be the query at the root node?
Solution

By inspection, the first splitting criterion should be $x_2 = 4.5$.

(c) (**3 pts**) How much has the impurity been reduced by the query at the root?
Solution
Trivially, $i(N_R) = 0$.

$$i(N_L) = -\frac{1}{6}\log_2\frac{1}{6} + -\frac{5}{6}\log_2\frac{5}{6} \approx 0.65$$

The reduction in extropy is

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R) \approx 0.99 - \frac{6}{9}0.65 - 0 \approx 0.56$$

.

(d) (**3 pts**) Continue constructing your tree fully. (Whenever two candidate queries lead to the same reduction in impurity, prefer the query that uses the x_1 feature.) Use your tree to classify $\mathbf{x} = \binom{6}{6}$ and $\mathbf{x} = \binom{3}{4}$.
Solution

We need to decide whether the second splitting criterion is $x_1 = 2.5$ or $x_1 = 3.5$. The entropy reduction for $x_1 = 2.5$ is

$$\Delta i(N) = \left(-\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6}\right) - \frac{4}{6}\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right) \approx 0.11.$$

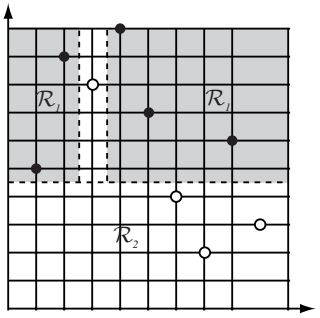
The entropy reduction for $x_1 = 3.5$ is

$$\Delta i(N) = \left(-\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6}\right) - \frac{3}{6}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) \approx 0.19.$$

So the second split should be $x_1 = 3.5$. The last split is then $x_1 = 2.5$. Thus the tree is

(e) (**2 pts**) Suppose your tree is to be able to classify deficient patterns. What should be the first (and only) surrogate split at the root node?
Solution

$\mathbf{x} = (6, 6)^t \Rightarrow \omega_1$; $\mathbf{x} = (3, 4)^t \Rightarrow \omega_2$. The surrogate split tries to achieve the same partitioning of the samples as the $x_2 > 4.5$ primary split. By inspection, it is $x_1 < 5.5$.



10. (10 points total) Short answer (**1 pt each**).

(a) What are the four major components of a grammar G? What do we mean by the language induced by grammar G, i.e., $\mathcal{L}(G)$?
Solution
The four major components of a grammar is symbols in the alphabet $\mathcal{A}$, variables $\mathcal{I}$, root symbol in set $\mathcal{S}$, and productions $\mathcal{P}$. The language induced by a grammar $\mathcal{G}$ is the set of all strings that can be generated by $\mathcal{G}$.

(b) Use the Boltzmann factor to explain why at a sufficiently high "temperature" T, all configurations in a Boltzmann network are equally probable.
Solution
At high temperature T, the probability of all states are equal.

$$\lim_{T\to\infty} \frac{P(\gamma)}{p(\gamma')} = \lim_{T\to\infty} e^{-(E_\gamma - E_{\gamma'})/T} \to 1.$$

(c) Use an equation and a few sentences to explain the minimum description length (MDL) principle.
Solution
The MDL principle states that one should seek a model h^* that minimizes the sum of the model's algorithmic complexity and the description of the training data with respect to that model. That is,

$$h^* = \arg\min_h K(h, D) = \arg\min_h K(h) + K(D \text{ using } h).$$

(d) Use an equation and a few sentences to explain what is the discriminability in signal detection theory.
Solution
For two normally distributed classes in one dimension (with same variance), *discriminability* is a measure of the ease of discriminating the two classes. It is defined as $d' = \frac{|\mu_2 - \mu_1|}{\sigma}$.

(e) If the cost for any fundamental string operation is 1.0, state the edit distance between `streets` and `scrams`.
Solution
The cost is 4. `streets` → `screets` → `scraets` → `scramts` → `scrams`.

(f) Suppose the Bayes error rate for a $c = 5$ category classification problem is 1%. What are the upper and lower bounds on the error rate of a nearest-neighbor classifier trained with an "infinitely large" training set?
Solution
Using the equation given with the exam, $P^* \leq P \leq P^*(2 - \frac{c}{c-1}P^*)$, we get $0.01 \leq P \leq 0.01(2 - \frac{5}{5-1}0.01) = 0.019875 \approx 0.02$.

(g) Use a formula and a sentence to explain learning with momentum in back-propagation.
Solution
Momentum in neural net learning is a heuristic that weight changes should tend to keep moving in the same direction. Let $\Delta\mathbf{w}(m) = \mathbf{w}(m) - \mathbf{w}(m-1)$

and $\mathbf{w}_{bp}(m)$ be the change in $\mathbf{w}(m)$ that would be called for by the back-propagation algorithm (i.e., $-\eta\frac{\partial J}{\partial \mathbf{w}}$). Then,

$$\mathbf{w}(m+1) = \mathbf{w}(m) + (1-\alpha)\Delta\mathbf{w}_{bp}(m) + \alpha\Delta\mathbf{w}(m-1).$$

(h) What is the evaluation problem in hidden Markov models?
Solution
The evaluation problem is determining the probability that a particular sequence of visible states was generated by an HMM model.

Worked examples

Below are worked examples, organized by book section, that may be of use to students.

[to be written]

Errata and additions in the text

Below are errata and minor alterations that improve the style or clarify the book. To see which printing of the book you have, look on the third page of the text itself, which faces the dedication page. At the bottom you will see:

Printed in the United States of America
10 9 8 7 6 5 4 3 2

The last number at the right gives the number of the printing; thus, as illustrated here, "2" means that this is the second printing.

First and second printings

Below, "*line +7*" means the seventh line from the top of the text body (not including figure or table captions, or other headlines unless otherwise indicated), and "*line -5*" means the fifth line from the bottom of the text body. In Algorithms, the numbering refers to the line numbers within the algorithm itself. Thus **Algorithm 4**, *line 5*" means line 5 in Algorithm 4, not the fifth line from the top of the page.

Front matter

page x *line +6*: Change "4.8 Reduced Coulomb" to "*4.8 Reduced Coulomb"

page xv *line -13*: Change "A.4.7 The Law of Total Probability and Bayes' Rule" to "A.4.7 The Law of Total Probability and Bayes Rule"

page xviii Take the last sentence under **Examples**, "In addition, in response to popular demand, a Solutions Manual has been prepared to help instructors who adopt this book for courses." and move it to be the final sentence under **Problems**, lower on the same page.

page xviii *lines -10– -11*: Change "and they are generally" to "and they are typically"

page xix *line -15*: Change "Ricoh Silicon Valley" to "Ricoh Innovations"

Chapter 1

page 1 *line +4*: Change "data and taking" to "data and making"

page 4 the only equation on the page: Change "$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$" to "$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$"

page 11 *line +21*: Change "us for practical, rather than" to "us for practical rather than"

page 14 *line -4* in the caption to Figure 1.8: Change "of the data impact both" to "of the data affect both"

page 19 *line +15*: Change "is achieved in humans" to "is performed by humans"

Chapter 2

page 21 *line -6* in the footnote: Change "should be written as $p_X(x|\omega)$" to "should be written as $p_x(x|\omega)$"

page 21 *line -4* in the footnote: Change "clear that $p_X(\cdot)$ and $p_Y(\cdot)$" to "clear that $p_x(\cdot)$ and $p_y(\cdot)$"

page 22 *second line after Eq. 3*: Change "probability (or *posterior*) probability" to "probability (or *posterior*)"

page 23 *second line after Eq. 7*: Change "By using Eq. 1, we can" to "By using Eq. 1 we can"

page 26 *first line after Eq. 17*: Change "and ω_2 otherwise." to "and otherwise decide ω_2."

page 28 Equation 23: Change "$(\lambda_{11}-\lambda_{22})-(\lambda_{21}-\lambda_{11})$" to "$(\lambda_{11}-\lambda_{22})+(\lambda_{21}-\lambda_{11})$"

page 28 *second line after Eq. 24*: Change "decision boundary gives" to "decision boundary then gives"

page 32 *second line after Eq. 33*: Change "expected values — by these" to "expected values by these"

page 36 *first equation after Eq. 49*: Change "Let us examine the discriminant" to "Let us examine this discriminant"

page 41 Figure 2.13, caption, *line +2*: Change "unequal variance." to "unequal variance, as shown in this case with $P(\omega_1) = P(\omega_2)$."

page 47 Equation 73: Change "for0" to "for 0" (i.e., add space)

page 47 *line - 10*: Change "substituting the results in Eq. 73" to "substituting this β into Eq. 73"

page 47 *line -2*: Change "This result is the so-called" to "This gives the so-called"

page 48 Example **2**, *line +3*: Change "4.11," to "4.06,"

page 48 Example **2**, *line +4*: Change "0.016382." to "0.0087."

page 50 The x-axis label on Fig. 2.20: Change "$P(x < x^*|x \in \omega_2)$" to "$P(x < x^*|x \in \omega_1)$"

pages 56 − 62 At the time of the release of the first printing, the problem of inference in Bayes belief nets *with loops* was not fully solved. Since that time, however, such a solution has emerged and for this reason Section 2.11 has been rewritten accordingly. This revision is posted on the Wiley site.

page 66 Problem **2**, part (b), *line +2*: Change "for arbitrary a_i and b_i." to "for arbitrary a_i and positive b_i."

page 66 Problem **3**, part (a) equation: End the equation with a period (full stop).

page 67 Problem **5**, part (d): Change "What is the minimax risk?" to "What is the minimax risk for part (c)?"

page 67 Problem **6**, part (2), *line +2*: Change "Determine the decision boundary" to "Determine the single-point decision boundary"

page 69 Move the title "**Section 2.4**" to the top of the page so that Problem **13** is now under **Section 2.4**.

page 71 Problem **20**, part (a), *line +1*: Change "we know only that a distribution is nonzero in" to "we know solely that a distribution is nonzero only in"

page 71 Problem **20**, part (b), *line +1*: Change "we know only that a distribution is nonzero for" to "we know solely that a distribution is nonzero only for"

page 71 Problem **23**, at the center of the typeset equation, change "and$\mathbf{\Sigma}$" to "and $\mathbf{\Sigma}$" (i.e., add space)

page 72 Problem **24**, *line +1*: Change "normal density for which $\sigma_{ij} = 0$" to "normal density with mean $\boldsymbol{\mu}$, $\sigma_{ij} = 0$"

page 73 Problem **34**, *line +6–7*: Change "assume the distributions" to "assume $P(\omega_1) = P(\omega_2) = 0.5$ and the distributions"

page 73 Problem **34**, part c), *line +4*: Change "Bayes error is 0.5." to "Bayes error is 0.25."

page 75 Problem **37**, *first equation*: Change "and$P(\omega_1)$" to "and $P(\omega_1)$" (i.e., add space)

page 75 Problem **37**, part (c) *equation*: Change "and$p(\mathbf{x}|\omega_2)$" to "and $p(\mathbf{x}|\omega_2)$" (i.e., add space)

page 75 Problem **39**, *line +1*: Change "Use the signal detection" to "Use signal detection"

page 75 Problem **39**, part (a), *line +1*: Change "and $P(x < x^*|x \in \omega_2)$, taken" to "and $P(x > x^*|x \in \omega_1)$, taken"

page 75 Problem **39**, part (b): Replace the last two sentences with "Estimate d' if $P(x > x^*|x \in \omega_1) = 0.8$ and $P(x > x^*|x \in \omega_2) = 0.3$. Repeat for $P(x > x^*|x \in \omega_1) = 0.7$ and $P(x > x^*|x \in \omega_1) = 0.4$."

page 75 Problem **39**, part (d): Replace the two equation lines with "**Case A:** $P(x > x^*|x \in \omega_1) = 0.8$, $P(x > x^*|x \in \omega_2) = 0.3$ or
Case B: $P(x > x^*|x \in \omega_1)$, $P(x > x^*|x \in \omega_2) = 0.7$."

page 76 Problem **41**, *first line after the equation*: Change "$(\mu_2 - \mu_1)/\delta_i$ " to "$(\mu_2 - \mu_1)/\delta$"

page 76 Problem **41**, part (b): Change "$d'_T = 1.0$." to "$d'_T = 1.0$ and 2.0."

page 76 Problem **41**, part (c): Change "$P(x > x^*|x \in \omega_1) = .2$." to "$P(x > x^*|x \in \omega_1) = .7$."

page 76 Problem **76**, part (e): Change "measure $P(x > x^*|x \in \omega_2) = .9$ and $(x > x^*|x \in \omega_1) = .3$." to "measure $P(x > x^*|x \in \omega_2) = .3$ and $P(x > x^*|x \in \omega_1) = .9$."

page 81 Computer exercise **6**, part (b), *line +1*: Change "Consider" to "Consider the normal distributions"

page 81 Computer exercise **6**, part (b), *equation*: Change "and$p(\mathbf{x}|\omega_2)$" to "and $p(\mathbf{x}|\omega_2)$" (i.e., add space)

page 81 Computer exercise **6**, part (b), *equation*: Move "with $P(\omega_1) = P(\omega_2) = 1/2$." out of the centered equation, and into the following line of *text*.

page 83 *first column*, entry for [21], *lines +3 – 4*: Change "Silverman edition, 1963." to "Silverman, 1963."

Chapter 3

page 88 Equation 9: Change "$\boldsymbol{\nabla_{\theta\mu}}$" to "$\boldsymbol{\nabla_{\mu}}$"

page 91 Ninth line after Eq. 22: Change "shall consider), the samples" to "shall consider) the samples"

page 99 Caption to first figure, change "starts our as a flat" to "starts out as a flat"

page 100 *line -5*: Change "are equivalent to" to "are more similar to"

page 100 *line -5 – -4*: Change "If there are much data" to "If there is much data"

page 102 *line -5*: Change "(Computer exercise 22)" to "(Problem 22)"

page 103 *line -2*: Change "choice of an prior" to "choice of a prior"

page 104 *line +6*: Change "if" to "only if"

page 104 *line +16*: Change "only if" to "if"

page 104 Equation 62: Make the usage and style of the summation sign ($\sum$) uniform in this equation. Specifically, in two places put the arguments *beneath* the summation sign, that is, change "$\sum_{\mathcal{D}\in\bar{\mathcal{D}}}$" to "$\sum\limits_{\mathcal{D}\in\bar{\mathcal{D}}}$"

page 105 first line after Eq 63: Change "to this kind of scaling." to "to such scaling."

page 111 *lines +9 – 10*: Change "constants c_0 and x_0 such that $|f(x)| \leq c_0|h(x)|$ for all" to "constants c and x_0 such that $|f(x)| \leq c|h(x)|$ for all"

page 111 *line +14*: Change "proper choice of c_0 and x_0." to "proper choice of c and x_0."

page 116 Equation 86: Change "$\lambda \mathbf{e}^t\mathbf{e}$" to "$\lambda(\mathbf{e}^t\mathbf{e} - 1)$"

page 125 **Algorithm 1**, *line 1*: Change "$i = 0$" to "$i \leftarrow 0$"

page 126 *first line of the equation at the middle of the page*: Change

$$Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^0) = \mathcal{E}_{x_{41}}[\ln p(\mathbf{x}_g, \mathbf{x}_b;\ \boldsymbol{\theta}|\boldsymbol{\theta}^0;\ \mathcal{D}_g)]$$

to

$$Q(\boldsymbol{\theta};\ \boldsymbol{\theta}^0) = \mathcal{E}_{x_{41}}[\ln p(\mathbf{x}_g, \mathbf{x}_b;\ \boldsymbol{\theta})|\boldsymbol{\theta}^0;\ \mathcal{D}_g]$$

page 128 *line +6*, (second line after the Example): Change "the EM algorithm, and they" to "the EM algorithm as they"

page 129 second line aboove Section 3.10.3: Change "while the ω_i are unobservable" to "while the ω_j are unobservable"

page 132 *line +3*: Change "b_{kj}, and thus" to "b_{jk}, and thus"

page 132 Equation 136: Replace by:

$$\alpha_j(t) = \begin{cases} 0 & t = 0 \text{ and } j \neq \text{initial state} \\ 1 & t = 0 \text{ and } j = \text{initial state} \\ \left[\sum_i \alpha_i(t-1)a_{ij}\right] b_{jk}v(t) & \text{otherwise.} \end{cases}$$

page 132 third line after Eq. 136: Change "Consequently, $\alpha_i(t)$ represents" to "Consequently, $\alpha_j(t)$ represents"

page 132 fourth line after Eq. 136: Change "hidden state ω_i" to "hidden state ω_j"

page 132 **Algorithm 2**, *line 1*: Delete "$\omega(1)$,"

page 132 **Algorithm 2**, *line 1*: Change "$t = 0$" to "$t \leftarrow 0$"

page 132 **Algorithm 2**, *line 1*: Change "$\alpha(0) = 1$" to "$\alpha_j(0)$"

page 132 **Algorithm 2**, *line 3*: Replace entire line by "$\alpha_j(t) \leftarrow b_{jk}v(t)\sum_{i=1}^{c} \alpha_i(t-1)a_{ij}$"

page 132 **Algorithm 3**: Somehow the line numbering became incorrect. Change the line number to be sequential, $1, 2, \ldots 6$.

page 132 **Algorithm 3**, *line 1*: Change "$\omega(t), t = T$" to "$\beta_j(T), t \leftarrow T$"

page 132 **Algorithm 3**, old *line 4*, renumbered to be *line 3*: Replace entire line by "$\beta_i(t) \leftarrow \sum_{j=1}^{c} \beta_j(t+1)a_{ij}b_{jk}v(t+1)$"

page 132 **Algorithm 3**, old *line 7*, renumbered *line 5*: Change "$P(V^T)$" to "$P(\mathbf{V}^T)$" (i.e., make the "V" bold)

page 133 Figure 3.10, change the label on the horizontal arrow from "a_{12}" to "a_{22}".

page 133 Figure 3.10, caption, *line +5*: Change "was in state $\omega_j(t = 2)$" to "was in state $\omega_i(t = 2)$"

page 133 Figure 3.10, caption, *line +6*: Change "is $\alpha_j(2)$ for $j = 1, 2$" to "is $\alpha_i(t)$ for $i = 1, 2$"

page 133 Figure 3.10, caption, *line -1*: Change "$\alpha_2(3) = b_{2k} \sum_{j=1}^{c} \alpha_j(2)a_{j2}$" to "$\alpha_2(3) = b_{2k} \sum_{i=1}^{c} \alpha_i(2)a_{i2}$"

page 133 *line -6*: Change "$\mathbf{V}^5 = \{v_3, v_1, v_3, v_2, v_0\}$" to "$\mathbf{V}^4 = \{v_1, v_3, v_2, v_0\}$"

page 133 *line -44*: Change "is shown above," to "is shown at the top of the figure"

page 134 Figure in Example 3, caption, *line +4*: Change "$\alpha_i(t)$ — the probability" to "$\alpha_j(t)$ — the probability"

page 134 Figure in Example 3, caption, *line +6*: Change "and $\alpha_i(0) = 0$ for $i \neq 1$." to "and $\alpha_j(0) = 0$ for $j \neq 1$."

page 134 Figure in Example 3, caption, *line +6*: Change "calculation of $\alpha_i(1)$." to "calculation of $\alpha_j(1)$."

page 134 Figure in Example 3, caption, *line -6*: Change "calculation of $\alpha_i(1)$" to "calculation of $\alpha_j(1)$"

page 134 igure in Example 3, caption, *line -4*: Change "contribution to $\alpha_i(1)$."to "contribution to $\alpha_j(1)$."

page 135 Algorithm 4: somehow the line numbering became incorrect. Change the line numbering to be sequential, 1, 2, 3, ..., 11. In the old line 4 (now renumbered 3): Change "$k = 0, \alpha_0 = 0$" to "$j \leftarrow -1$"

page 135 Algorithm 4 old *line 5* (now renumbered 4): Change "$k \leftarrow k + 1$" to "$j \leftarrow j + 1$"

page 135 Algorithm 4 old *line 7* (now renumbered 5): Change "$\alpha_k(t)$" to "$\alpha_j(t)$"

page 135 Algorithm 4, old *line 8* (now renumbered 6): Change "$k = c$" to "$j = c$"

page 135 Algorithm 4, old *line 11* (now renumbered 8): Change "*AppendTo* Path $\omega_{j'}$" to "*Append* $\omega_{j'}$ *to* Path"

page 135 *line -5*: Change "The red line" to "The black line"

page 135 *line -4*: Change "value of α_i at each step" to "value of α_j at each step"

page 137 Equation 138: Replace equation by

$$\beta_i(t) = \begin{cases} 0 & \omega_i(t) \neq \omega_0 \text{ and } t = T \\ 1 & \omega_i(t) = \omega_0 \text{ and } t = T \\ \sum_j \beta_j(t+1)a_{ij}b_{jk}v(t+1) & \text{otherwise.} \end{cases}$$

page 137 seventh line after Eq. 138: Change "$\beta_i(T-1) = \sum_j a_{ij}b_{ij}v(T)\beta_i(T)$." to "$\beta_i(T-1) = \sum_j a_{ij}b_{jk}v(T)\beta_j(T)$."

page 137 fourth line before Eq. 139: Change "probabilities a_{ij} and b_{ij}" to "probabilities a_{ij} and b_{jk}"

page 137 Equation 139: Change "b_{ij}" to "b_{jk}"

page 138 *line +3*: Change "whereas at step t it is" to "whereas the total expected number of any transitions from ω_i is"

page 138 first line after Eq. 140: Change "$\hat{b}_{ij}$" to "$\hat{b}_{jk}$"

page 138 Equation 141: Replace equation by:

$$\hat{b}_{jk} = \frac{\sum_{\substack{t=1 \\ v(t)=v_k}}^{T} \sum_{l} \gamma_{jl}(t)}{\sum_{t=1}^{T} \sum_{l} \gamma_{jl}(t)}$$

page 138 **Algorithm 5**, *line 1*: Change "criterion θ"to "criterion $\theta, z \leftarrow 0$"

page 143 Problem **11**, second and third lines after first equation: Change "$p_2(\mathbf{x})$ by a normal $p_1(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$" to "$p_1(\mathbf{x})$ by a normal $p_2(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$"

page 143 Problem **11**: Second equations: Change "$\mathcal{E}_2$" to "$\mathcal{E}_1$" in two places

page 143 Problem **11**, last line: Change "over the density $p_2(\mathbf{x})$" to "over the density $p_1(\mathbf{x})$"

page 147 Problem **22**, line between the two equations: Change "has a uniform" to "has a uniform distribution"

page 148 Problem **27**, part (a), line after the equation: Change "as given in Table 3.1." to "as in Table 3.1.1."

page 149 Problem **31**, *line +1*: Change "suppose a and b are constants" to "suppose a and b are positive constants"

page 149 Problem **32**, *line +1*: Change "where the n coefficients" to "at a point x, where the n coefficients"

page 150 Problem **34**, *line +4*: Change "the number of operations n" to "the maximum size n"

page 151 Problem **38**, *line +1*: Change "$p_x(\mathbf{x}|\omega_i)$" to "$p_{\mathbf{x}}(\mathbf{x}|\omega_i)$"

page 151 Problem **38 (b)** bottom equation on page: Change "$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^2$" to "$(\mu_1 - \mu_2)^2$"

page 152 *line +1*: Change "and" to "is maximized by"

page 153 Problem **43**, *line +4*: Change "and the d mean vectors." to "and the c mean vectors."

page 154 Problem **46**, (top equation): Change "0 otherwise." to "ϵ otherwise."

page 154 Problem **46**, *line +1* (after the equation): Change "missing feature values." to "missing feature values and ϵ is a very small positive constant that can be neglected when normalizing the density within the above bounds."

page 154 Problem **46**, part (b): Add "You may make some simplifying assumptions."

page 154 Problem **47**, first equation: Change "$e^{-\theta_1 x_1}$" to "e^{-x_1/θ_1}"

page 156 Computer exercise 2, after the equation: Change "calculate a density" to "calculate the density"

page 156 Computer exercise 2. After "x_2 feature of category ω_2." add "Assume your priors on the parameters are uniform throughout the range of the data."

page 157 Computer exercise **4**, *line -3*: Change "apply it to the $x_1 - x_2$ components" to "apply it to the x_1–x_2 components" (i.e., eliminate spaces and note that the dash is not subtraction sign, but an n-dash)

page 157 Computer exercise **6**, part (a), *line +2*: Change "in the Table above." to "in the table above."

page 159 Computer exercise **13** table, sample 4 under ω_1: Change "`AD`" to "`ADB`"

Chapter 4

page 172 Figure 4.9, caption, *line -1*: Change "where $\mathbf{I}$ is the $d \times d$ idenity matrix." to "where $\mathbf{I}$ is the d-by-d identity matrix."

page 173 **Algorithm 1**, *line 1*: Change "$j = 0$" to "$j \leftarrow 0$"

page 173 **Algorithm 2**, *line 1*: Change "test pattern," to "test pattern"

page 178 *line +3*: Change "Becuase" to "Because"

page 179 Equation 41: Change "$\int_{\mathbf{x}' \in \mathcal{S}}$" to "$\int\limits_{\mathbf{x}' \in \mathcal{S}}$" (i.e., place the limits underneath the integral sign)

page 184 *lines +2 – 3*: Keep "$k - i > i$" on the same line (i.e., do not put a line break in this equation)

page 186 **Algorithm 3**, *line 1*: Change "$j = 0, \mathcal{D} =$ data set, $n =$ # prototypes" to "$j \leftarrow 0, \mathcal{D} \leftarrow$ data set, $n \leftarrow$ # prototypes"

page 187 Figure 4.18, caption, *line -4*: Change "by a factor 1/3" to "by a factor $\alpha = 1/3$"

page 188 First margin note: Change "MINKOWSKI METRIC" to "MINKOWSKI METRIC" (i.e., capitalize the M in MINKOWSKI)

page 188 Second margin note: Change "MANHATTAN DISTANCE" to "MANHATTAN DISTANCE" (i.e., capitalize the M in MANHATTAN)

page 188 Third margin note: Change "TANIMOTO METRIC" to "TANIMOTO METRIC" (i.e., capitalize the T in TANIMOTO)

page 189 *line +5*: Change "the relative shift is a mere" to "the relative shift s is a mere"

page 192 Figure 4.23, caption, *lines -4– -2*: Eliminate the second-to-last sentence.

page 193 Equation 61: Replace the full equation with "$\mu_x(x) \cdot \mu_y(y)$"

page 194 *line +19*: Change "beief about memberships" to "belief about memberships"

page 195 *line +9*: Change in the title "**4.8 REDUCED COULOMB**" to "***4.8 REDUCED COULOMB**"

page 196 **Algorithm 4**, *line 1*: Change "$j = 0, n = \#$ patterns, $\epsilon =$ small parameter, $\lambda_m =$ max radius" to "$j \leftarrow 0, n \leftarrow \#$ patterns, $\epsilon \leftarrow$ small parameter, $\lambda_m \leftarrow$ max radius"

page 197 **Algorithm 5**, *line 1*: Change "$j = 0, k = 0, \mathbf{x} =$ test pattern, $\mathcal{D}_t = \{\}$" to "$j \leftarrow 0, k \leftarrow 0, \mathbf{x} \leftarrow$ test pattern, $\mathcal{D}_t \leftarrow \{\}$"

page 199 *line -7*: Change "prior knowledge." to "prior beliefs."

page 202 Problem 6c, first line: Change "increases" to "increase"

page 203 Problem **11**, part (d), *line +1*: Change "close to an edge" to "close to a face"

page 203 Problem **11**, part (d) *line +4*: Change "closer to an edge" to "closer to a face"

page 203 Problem **11**, part (d), *line +5*: Change "even though it is easier to calculate here" to "and happens to be easier to calculate"

page 204 Problem **13** *line +1*: Change "from the distributions" to "with priors $P(\omega_1) = P(\omega_2) = 0.5$ and the distributions"

page 205 Problem **21**, part (a), *line +1*: Change "As given in the text, take" to "Follow the treatment in the text and take"

page 206 Problem **23** part (d) *line +2*: Change "space, then the **b**" to "space, then in the **b**"

page 206 Problem **26**, *line +4*: Change "and find its nearest" to "and seek its nearest"

page 207 Problem **26** part (d), first line: change "Calculate the probability" to "Estimate through simulation the probability"

page 207 Problem **27**, part (a) equation: Replace current equation with "$D_{Tanimoto}(\mathcal{S}_1, \mathcal{S}_2) = \frac{n_1+n_2-2n_{12}}{n_1+n_2-n_{12}}$,"

page 207 Problem **29**, first equation: Change "δ_i" to "δ_1" in all three places

page 207 Problem **29**, second equation: Change "$\hat{C}(x, \mu_i$" to "$\hat{C}(x;\ \mu_i$"

page 207 Problem **29** second equation: Change "δ_i" to "δ_2" in all three places

page 207 Problem **29** *line -5*: Change "we have for the length $\delta_i = 5$" to "we have for the length $\delta_1 = 5$"

page 207 Problem **29** *line -4*: Change "and for lightness $\delta_j = 30$" to "and for the lightness $\delta_2 = 30$"

Chapter 5

page 218 *line +6*: Change "the problem to $c - 1$" to "the problem to c"

page 218 Equation 2: Change "$\mathbf{w}^t\mathbf{x}_i$" to "$\mathbf{w}_i^t\mathbf{x}$"

page 219 second line after the second (unnumbered) equation: Change "is given by $(g_i - g_j)/\|\mathbf{w}_i - \mathbf{w}_j\|$" to "is given by $(g_i(\mathbf{x}) - g_j(\mathbf{x}))/\|\mathbf{w}_i - \mathbf{w}_j\|$"

page 220 sentence before Eq. 5, Change "this in turn suggest" to "this in turn suggests"

pages 221–222 (across the page break): Change "mul-tilayer" to "multi-layer" (i.e., hyphenate as "multi-layer")

page 225 **Algorithm 1**, *line 1*: Change "$k = 0$" to "$k \leftarrow 0$"

page 228 **Algorithm 3**, *line 1*: Change "$k = 0$" to "$k \leftarrow 0$"

page 229 *line +5*: Change "We shall begin our examination" to "We begin our examination"

page 229 *line -3*: Change "Thus we shall denote" to "Thus we denote"

page 230 **Algorithm 4**, *line 1*: Change "$k = 0$" to "$k \leftarrow 0$"

page 230 fourth line before **Theorem 5.1**: Change "correction is clearly moving" to "correction is hence moving"

page 230 *line -2*: Change "From Eq. 20," to "From Eq. 20 we have"

page 232 first line after Eq. 24: Change "Because the squared distance" to "Because this squared distance"

page 233 **Algorithm 5**, *line 1*: Change "$k = 0$" to "$k \leftarrow 0$"

page 233 **Algorithm 6**, *line 1*: Change "$k = 0$" to "$k \leftarrow 0$"

page 234 *line +22* (counting from the end of the Algorithm): Change "that it will have little effect at all." to "that it will have little if any effect."

page 235 *lines +3 –4* (counting from the end of the Algorithm): Change "and this means the "gap," determined by these two vectors, can never increase in size for separable data." to "and this means that for separable data the "gap," determined by these two vectors, can never increase in size."

page 235 second and third lines after the section title **5.6 RELAXATION PROCEDURES**: Change "in so-called "relaxation procedures" to include" to "in so-called "relaxation procedures," to include"

page 235 first and second lines after Eq. 32: Change "misclassified by $\mathbf{a}$, as do J_p, J_q focus attention" to "misclassified by $\mathbf{a}$. Both J_p and J_q focus attention"

page 235 second line after Eq. 32: Change "Its chief difference is that its gradient" to "The chief difference is that the gradient of J_q is"

page 236 **Algorithm 8**, *line 1*: Change "$k = 0$" to "$k \leftarrow 0$"

page 236 **Algorithm 8**, *line 4*: Change "$j = 0$" to "$j \leftarrow 0$"

page 236 **Algorithm 9**, *line 1*: Change "$k = 0$" to "$k \leftarrow 0$"

page 238 *line -5*: Change "*procedures*, because" to "*procedures* because"

page 242 *line -2*: Change "We begin by writing Eq. 47" to "We begin by writing Eq. 45"

page 246 *lines +1 – 2*: Don't split the equation by the line break

page 246 first (unnumbered) equation, second line: Change "$(\mathbf{Ya}_k - \mathbf{b})$." to $(\mathbf{Ya}(k) - \mathbf{b})$."

page 246 margin note: Change "LMS RULE" to "LMS RULE" (i.e., capitalize "LMS")

page 246 Equation 61, second line: Change "$(b_k - \mathbf{a}(k)^t \mathbf{y}^k)$" to "$(b(k) - \mathbf{a}^t(k)\mathbf{y}^k)$"

page 246 **Algorithm 10**, *line 1*: Change "$k = 0$" to "$k \leftarrow 0$"

page 248 second line after Eq. 66: Change "obtain the MSE optimal" to "obtain the MSE-optimal"

page 250 Equation 79: Co-align vertically the ">" in the top equation with the "=" in the bottom equation

page 250 Equation 79: Change "$\mathbf{a}(k)$" to "$\mathbf{b}(k)$"

page 251 Algorithm 11, *line 5*: Change "$\mathbf{a}$" to "$\mathbf{b}$"

page 252 top equation: Change "$= \mathbf{0} =$" to "$= 0 =$" (i.e., de-bold the "$\mathbf{0}$")

page 253 *line -7*: Change "requires that $\mathbf{e}^+(k) = 0$ for" to "requires that $\mathbf{e}^+(k) = \mathbf{0}$ for"

page 254 line after Eq. 90: Change "constant, positive-definite" to "constant, symmetric, positive-definite"

page 255 First (unnumbered) equation on page: Change the first term in parentheses on the right-hand side from "$\eta^2 \mathbf{YRY}^t \mathbf{YRY}$" to "$\eta^2 \mathbf{YRY}^t \mathbf{YRY}^t$"

page 255 Eq. 92: Last term on the right-hand-side, change "$\eta^2 \mathbf{RY}^t \mathbf{R}$" to "$\eta^2 \mathbf{RY}^t \mathbf{YR}$"

page 257 Figure 5.18, caption, *line +2*: Change "form $\mathbf{Au}\boldsymbol{\beta}$" to "form $\mathbf{Au} = \boldsymbol{\beta}$"

page 262 fourth line after Eq. 105: Change "with the largest margin" to "with the *largest* margin" (i.e., italicize "largest")

page 263 fourth line after Eq. 107: Change "equation in Chapter 9," to "topic in Chapter 9,"

page 266 Equation 114: Change "$\mathbf{a}_i^t(k)\mathbf{y}^k \leq \mathbf{a}_j(k)^t\mathbf{y}^k$." to "$\mathbf{a}_i^t(k)\mathbf{y}^k \leq \mathbf{a}_j^t(k)\mathbf{y}^k$."

page 270 twelfth line under **BIBLIOGRAPHICAL AND HISTORICAL REMARKS**: Change "error-free case [7] and [11] and" to "error-free case [7,11] and"

page 270 *line -13*: Change "support vector machines" to "Support Vector Machines"

page 270 *line -8*: Change "support vector machines" to "Support Vector Machines"

page 271 Problem **2**, *line +3*: Change "if $0 \leq \lambda \leq 1$." to "for $0 \leq \lambda \leq 1$."

page 272 Problem **8**, *line +2*: Change "if $\mathbf{a}^t\mathbf{y}_i \geq 0$" to "if $\mathbf{a}^t\mathbf{y}_i \geq b$"

page 274 Problem **22**, second term on the right-hand side change "$(\mathbf{a}^t\mathbf{y} - (\lambda_{12} - \lambda_{22}))^2$" to "$(\mathbf{a}^t\mathbf{y} + (\lambda_{12} - \lambda_{22}))^2$"

page 275 Problem **27**, last line: Change "by Eq. 85." to "by Eq. 95."

page 277 *lines +2 – 3*: Change "satisfies $z_k\mathbf{a}^t\mathbf{y}_k = 0$" to "satisfies $z_k\mathbf{a}^t\mathbf{y}_k = 1$"

page 277 Problem **38**, *line -2*: Change "procedures Perceptron" to "procedures. Generalize the Perceptron"

page 278 Problem **1**, part (a): change "data in in order" to "data in order"

page 278 Computer exercise **2**, part (a): Change "Starting with $\mathbf{a} = 0$," to "Starting with $\mathbf{a} = \mathbf{0}$," (i.e., make bold the "0")

page 278 First heading after the table: Change "**Section 5.4**" to "**Section 5.5**"

page 278 Computer Exercise 1: Change "(Algorithm 1) and Newton's algorithm (Algorithm 2) applied" to "(Algorithm 1) and the Perceptron criterion (Eq. 16)"

page 278 Second heading after the table: Delete "**Section 5.5**"

page 279 *line +1*: Change "length is greater than the pocket" to "length is greater than with the pocket"

page 279 Computer exercise **4**, part (a), *line +2*: Change "and $\boldsymbol{\mu}_1 = 0$," to "and $\boldsymbol{\mu}_1 = \mathbf{0}$," (i.e., make bold the "0")

Chapter 6

page 286 Equation 5: Change "$z_k = f(net_k)$." to "$z_k = f(net_k)$,"

page 287 *line +2*: Change "all identical." to "all the same."

page 288 *line -3*: Change "depend on the" to "depends on the"

page 291 Two lines before Eq. 3: Change "hidden-to-output weights, w_{ij}" to "hidden-to-output weights, w_{kj}"

page 292 Equation 19, first line (inside brackets): Change "1/2" to "$\frac{1}{2}$" (i.e., typeset as a full fraction)

page 292 After Eq. 19, *line +3*: Change "activiation" to "activation"

page 293 Figure 6.5: Change "w_{ij}" to "w_{ji}"

page 294 Algorithm 2, *line 3*: Change "$\Delta w_{kj} \leftarrow$" to "$\Delta w_{kj} \leftarrow 0$"

page 295 *line -5*: Change "In addition to the use of the training set, here are" to "In addition to the use of the training set, there are"

page 299 *line -1*: Change "and can be linearly separable" to "and are linearly separable"

page 305 *line 10*: Change "ratio of such priors." to "ratio of such priors, though this need not ensure minimal error."

page 306 sixth and seventh line after Eq. 33: Change "in a sum squared error sense" to "in a sum-squared-error sense"

page 307 *lines +2 – 3*: Change "been found useful" to "found to be useful"

page 307 *line +6*: Change "as continuity of f and its derivative" to "as continuity of $f(\cdot)$ and its derivative"

page 308 *line +10*: Change "that is," to "or is an "odd" function, that is,"

page 308 *line +19*: Change "values that ensure $f'(0) \simeq 1$" to "values that ensure $f'(0) \simeq 0.5$"

page 315 first line after Eq. 38: Change "reducing the criterion" to "reducing the error"

page 316 Figure 6.19 caption, *line -1*: Change "trained network." to "trained network (red)."

page 318 *line +8*: Change "to compute is nonnegative" to "to compute, is nonnegative"

page 318 margin note: Change "MINKOWSKI ERROR" to "MINKOWSKI ERROR" (i.e., captialize "M" in "MINKOWSKI")

page 320 line between Eqs. 50 and 51: Change "The optimimum change" to "Therefore, the optimum change"

page 319 Equation 48: Change last entry from "$f'(net)y_{n_H}x_d$" to "$f'(net)f'(net_{n_H}x_d$"

page 322 Equation 56: Replace the current equation by

$$\beta_m = \frac{\nabla J^t(\mathbf{w}(m))\nabla J(\mathbf{w}(m))}{\nabla J^t(\mathbf{w}(m-1))\nabla J(\mathbf{w}(m-1))} \qquad (56)$$

page 322 Equation 57: Replace the current equation by

$$\beta_m = \frac{\nabla J^t(\mathbf{w}(m))[\nabla J(\mathbf{w}(m)) - \nabla J(\mathbf{w}(m-1))]}{\nabla J^t(\mathbf{w}(m-1))\nabla J(\mathbf{w}(m-1))} \qquad (57)$$

page 323 Fourth (unnumbered) equation on the page: Replace the left portion by

$$\beta_1 = \frac{\nabla J^t(\mathbf{w}(1))\nabla J(\mathbf{w}(1))}{\nabla J^t(\mathbf{w}(0))\nabla J(\mathbf{w}(0))} =$$

page 323 Caption to bottom figure, *line +2*: Change "shown in the contour plot," to "shown in the density plot,"

page 325 last line of the section **Special Bases**: Add "This is very closely related to model-dependent maximum-likelihood techniques we saw in Chapter 3."

page 326 first line after Eq. 63: Change "of the filter in analogy" to "of the filter, in analogy"

page 326 *line -5*: Add a red margin note "TIME DELAY NEURAL NETWORK"

page 330 three lines above Eq. 67: Change "write the error as the sum" to "write the new error as the sum"

page 332 Figure 6.28, caption, *line +1*: Change "a function of weights, $J(\mathbf{w})$" to "a function of weights, $J(\mathbf{w})$,"

page 337 Problem **8**, pagt (b) *line +2*: Change "if the sign if flipped" to "if the sign is flipped"

page 337 Problem **14**, part (c), *line +2*: Change "the 2×2 identity" to "the 2-by-2 identity"

page 339 Problem **22**, *line +4*: Add "Are the discriminant functions independent?"

page 341 Problem **31**, *lines +1 – 2*: Change "for a sum squared error criterion" to "for a sum-square-error criterion"

page 344 Computer exercise **2**, *line +2*: Change "backpropagation to (Algorithm 1)" to "backpropagation (Algorithm 1)"

page 345 Computer exercise **7**, *line +2*: Change "on a random problem." to "on a two-dimensional two-category problem with 2^k patterns chosen randomly from the unit square. Estimate k such that the expected error is 25% . Discuss your results.

page 346 Computer exercise **10**, part (c), *line +1*: Change "Use your network" to "Use your trained network"

page 347 Column 2, entry for [14], *line +5*: Change "volume 3. Morgan Kaufmann" to "volume 3, pages 853–859. Morgan Kaufmann"

page 348 column 2, entry for [43], *line +4*: Change "2000." to "2001."

Chapter 7

page 351 fourteenth line after Eq. 1: Change "of the magnets with the most stable configuration" to "of the magnets that is the most stable"

page 351 footnote, *line -1*: Change "in a range of problem domains." to "in many problem domains."

page 352 Figure 7.1, caption, *line +7*: Change "While our convention" to "While for neural nets our convention"

page 352 Figure 7.1, caption, *line +8*: Change "Boltzmann networks is" to "Boltzmann networks here is"

page 352 Caption to Figure 7.1, *line -1*: Change "$0 \leq \alpha \leq 2^{10}$" to "$0 \leq \alpha < 2^{10}$"

page 353 Figure 7.2, caption, *line +3*: Change "or "temperature" T to avoid" to "or "temperature" T, to avoid"

page 357 Figure 7.4, caption, *line +3*: Change "values e^{-E_γ}/T." to "values $e^{E_\gamma/T}$."

page 360 first line in Section 7.3: Change "will use modify the" to "will use the"

page 360 second line in Section 7.3: Change "to specifically identify" to "and specifically identify"

page 360 second line in Section 7.3: Change "and other units as outputs" to "and others as outputs"

page 361 fourth line after Eq. 6: Change "possible hidden states." to "possible hidden states consistent with α."

page 364 at the end of the body of the text: insert "One benefit of such stochastic learning is that if the final error seems to be unacceptably high, we can merely increase the temperature and anneal — we do not need to re-initialize the weights and re-start the full anneal."

page 365 seventh line after the subsection **Pattern Completion**: Change "components of a partial pattern" to "components of the partial pattern"

page 367 *line +1*: Change "Recall, at the end" to "As mentioned, at the end"

page 373 fourteenth line in Section 7.5: Change "repeated for subsequent" to "repeated for the subsequent"

page 373 fifth line above the subsection **Genetic Algorithms**: Change "In both cases, a key" to "In both cases a key"

page 374 *line +2*: Change "used in the algorithm. Below" to "used in the algorithm; below"

page 374 *line +3*: Change "P_{co} and P_{mut}, respectively, but first we present the general algorithm:" to "P_{co} and P_{mut}."

page 379 third line in the subsection **Representation**: Change "Here the syntactic" to "Whereas the syntactic"

page 381 third line under **BIBLIOGRAPHICAL AND HISTORICAL REMARKS**: Change "branch-and-bound, A^*" to "branch-and-bound and A^*

page 382 *line +28*: Change "been fabricated as described" to "been fabricated, as described"

page 383 Problem **3**, part (b), *line +1*: Change "The figure shows" to "That figure shows"

page 384 Problem **7**, part (c), *line +1*: Change "magnets, total" to "magnets, the total"

page 384 Problem **8**, part (b), equation: Change "$= P(s = +1)(+1) + P(s = -1)(-1)$." to "$= \Pr[s = +1](+1) + \Pr[s = -1](-1)$."

page 385 Problem **12**, *line +1*: Change "Train a six-unit Hopfield network with the following three patterns using the" to "Determine the weights in a six-unit Hopfield network trained with the following three patterns. Use the"

page 385 Problem **15** *line +1*: Change "not be in a set of" to "not be in a subset of"

page 387 Problem **24**: Change "crossover operator" to "crossover operator, and the multiplication operator, `*`, and the addition operator, `+`, can take two or more operands."

page 393 column 1, entry for [54], *line +3*: Change "*Evoultions-*" to "*Evolutions-*"

page 393 column 2, entry for [62], *line +2*: Change "neurobilogical system." to "neurobiological systems."

Chapter 8

page 403 line above **Section 8.3.4 Pruning**: Change "splitting is stopped." to "splitting should be stopped."

page 405 Table at top, x_2 entry in fifth row under ω_1, change ".48" to ".44"

page 405 caption to figure, *line -2*: Change "marked * were instead slightly lower (marked †)," to "marked * were instead slightly lower (marked †)," (i.e., change the color of the special symbols to red)

page 414 *line +16*: Change "⁄⁄⁄ `GACTG`" to "⁄⁄⁄`GACTG`" (i.e., eliminate space)

page 416 Algorithm 2, line 2: Change "$\mathcal{F}(\mathbf{x})$" to "$\mathcal{F}$"

page 416 Algorithm 2, line 3: Change "$\mathcal{G}(\mathbf{x})$" to "$\mathcal{G}$"

page 416 Algorithm 2, line 11: Change "$\mathcal{G}(0)$" to "1"

page 416 *line -1*: Change "$\mathcal{F}(\mathbf{x})$" to "$\mathcal{F}$"

page 417 *lines -9 – -4*, Replace last full paragraph by "Consider target string $\mathbf{x}$. Each location j (for $j < m$) defines a suffix of $\mathbf{x}$, that is, $\mathbf{x}[j+1, \ldots, m]$. The *good-suffix function* $\mathcal{G}(j)$ returns the starting location of the right-most instance of another occurrence of that suffix (if one exists). In the example in Fig. 8.8, $\mathbf{x}$ = `estimates` and thus $j = 8$ defines the suffix `s`. The right-most occurrence of another `s` is 2; therefore $\mathcal{G}(8) = 2$. Similarly $j = 7$ defines the suffix `es`. The right-most occurrence of another `es` is 2; therefore $\mathcal{G}(7) = 1$. No other suffix appears repeatedly within $\mathbf{x}$, and thus $\mathcal{G}$ is undefined for $j < 7$."

page 418 fifth line before **Algorithm 3**: Change "consider interchanges." to "consider the interchange operation."

page 418 fourth line before **Algorithm 3**: Change "be an $m \times n$ matrix" to "be an m-by-n matrix"

page 422 *line -3*: Change "specify how to transform" to "specifies how to transform"

page 426 *line +1*: Change "The grammar takes *digit6* and" to "This grammar takes *digits6* and"

page 428 *line -1*: Change "subsequent characters." to "subsequent characters, or instead starting and the last (right) character in the sentence."

page 429 caption to Figure 8.16, add after the last line: "Such finite-state machines are sometimes favored because of their clear interpretation and learning methods based on addition of nodes and links. In Section 8.7, though, we shall see general methods for grammatical learning that apply to a broader range of grammatical models."

page 438 Problem 5, *line +2*: Change "Eqs. 1 and 5." to "Eqs. 1 and 5 for the case of an arbitrary number of categories."

page 438 Problem **6** after the first set of equations: Replace the $i^*(\boldsymbol{\alpha})$ equation by "$i*(\alpha) = i(\alpha P^a(\omega_1) + (1-\alpha)P^b(\omega_1), ..., \alpha P^a(\omega_c) + (1-\alpha)P^b(\omega_c))$"

page 438 Problem **6** last line before part (a): Replace line by "then we have $i^* \geq \alpha i_a + (1-\alpha) i_b$."

page 445 Problem **36**, part (d), *line +1*: Change "Give a derivation" to "Attempt a derivation"

page 445 Problem **40**, part (d), *line +2*: Change "either grammar as" to "either grammar, or can be parsed in both grammars, as"

page 446 Table, sample 12: Change "D" to "E"

page 447 Computer exercise **3**, part b): Change "{C, D, J, L, M}" to "{C, E, J, L, M}"

Chapter 9

page 455 *line -9 – -10*: Change "a zero-one loss function, or, more generally, the cost for a general loss function $L(\cdot,\cdot)$" to "a zero-one or other loss function."

page 460 Table for rank $r = 3$, third row: Change "$\mathbf{x}_1\ OR\ \mathbf{x}_3\ OR\ \mathbf{x}_3$" to "$\mathbf{x}_1\ OR\ \mathbf{x}_3\ OR\ \mathbf{x}_4$"

page 462 *line +12*: Change "upon a specification method L," by "upon a specification method,"

page 462 *line +13*: Change "transmitted as y, denoted $L(y) = x$." to "transmitted as y and decoded given some fixed method L, denoted $L(y) = x$."

page 462 *line +15 – 16*: Change "denoted $\min_{|y|} \mathrm{L(y)} = \mathrm{x}$; this minimal...[[to end of paragraph]]" to "denoted $\min_{y:L(y)=x} |y|$."

page 462 *line +17 – 18*: Change "by analogy to entropy, where instead of a specification method L we consider" to "by analogy to communication, where instead of a fixed decoding method L, we consider"

page 462 *line +25*: Change "and so on. A universal description would" to "and so on. Such a description would"

page 462 *line +26*: Change "different binary strings." to "different binary strings x_1 and x_2."

page 462 third and second line above Eq. 7: Change "the *shortest* program y (where the length" to "the *shortest* program string y (where y's length"

page 462 Equation 7: Replace the entire equation by:

$$K(x) = \min_{y:U(y)=x} |y|,$$

page 463 *line -10*: Change "No Free Lunch Theorems." to "No Free Lunch Theorem."

page 472 Equation 23: Change "$\frac{1}{(n-1)}$" to "$\frac{1}{n(n-1)}$"

page 472 Equation 24: Change "$\sum_{j\neq i}$" to "$\sum_{j\neq i}^{n}$" (i.e., place the upper limit n over the summation sign)

page 472 *line -4*: Change "jackknife estimate of the variance" to "variance of the jackknife estimate"

page 479 *line +1*: Change "in line 4" to "in line 5"

page 485 *line -10*: Change "good estimates, because" to "good estimates because"

page 488 Caption to Fig. 9.13, *line +6*: Change "approximated as a k-dimensional" to "approximated as a p-dimensional"

page 488 *line +10* (i.e., second line of the second paragraph of text): Change "is k-dimensional and the" to "is p-dimensional and the"

page 496 Equation 54: Change "$P(r|\mathbf{x}, \boldsymbol{\eta}^0)$" to "$P(r|\mathbf{x}, \boldsymbol{\theta}_0^0)$"

page 497 Equation 58: Change $\boldsymbol{\mu}_r$ to $\boldsymbol{\theta}_r$ in two places

page 501 *line -15*: Change "and learning algorithm was first described" to "and learning algorithm were first described"

page 502 Problem **6**, last line: Change "this way" to "this sense"

page 504 Problem **20**, *line -2*: Change "denoted $p(g(\mathbf{x};\ \mathcal{D}))$ is a" to "denoted $p(g(\mathbf{x};\ \mathcal{D}))$, is a"

page 508 Problem **45**, *line +1*: Change "mixture of experts classifier" to "mixture-of-experts classifier"

page 512 *line -3*: Add "Now what is the training error measured using ω_A and ω_B?"

Chapter 10

page 524 The second to last equation: Increase the size of the final bracket, "$\big]$" to match its mate

page 525 line following Eq. 23: Change "results in Eq. 12" to "results into Eq. 12"

page 529 *line +5*: Change "$\hat{P}(w_j)$" to "$\hat{P}(\omega_j)$"

page 529 Algorithm 2: Change "**(Fuzzy k-Means Clustering)**" to "**(Fuzzy k-Means Clustering)**"

page 534 *line +2*: Change "overlap, thus" to "overlap; thus"

page 535 First equation: move the second, third, and fourth lines so as to co-align vertically the corresponding terms

page 536 *line +19*: Change "classification analog of Chapter 3" to "classification case in Chapter 3"

page 537 *line -20*: Change "distributed, these statistics" to "distributed these statistics"

page 571 *line +1 – +2*: Change "mi-crophones" to "micro-phones" (i.e., re-hyphenate)

page 578 Figure 10.31, caption, *line +2*: Change "space that leads maximally" to "space that maximally"

page 579 Figure 10.32, caption, *line -1*: Change "to this center region" to "to this central region"

page 580 *line +15*: Change "cluster centers being used to" to "*cluster centers* being used to" (i.e., italicize "cluster centers")

page 580 *line +16*: Change "with combined features being" to "with *combined features* being" (i.e., italicize "combined features")

page 582 four lines above **BIBLIOGRAPHICAL and HISTORICAL REMARKS**: Change "between points that, too, seeks to preserve neighborhoods" to "between points that preserves neighborhoods"

page 583 *lines +9 – 10*: Change "The classificatory foundations of biology, cladistics (from the Greek *klados*, branch) provide useful" to "Cladistics, the classifactory foundation of biology (from the Greek *klados*, branch), provides useful"

page 583 *line +18*: Change "analysis, and explained the very close" to "analysis, and in reference [36] explained the very close"

page 583 *line +19*: Change "information maximization in reference [36]." to "information maximization."

page 584 Problem **2**, equation: Change "$(1-|x-\mu_1|)/(2w_i)$" to "$(w_i-|x-\mu_i|)/w_i^2$"

page 584 Problem **4** a, right-hand side: Change "$\mathbf{x}_i$" to "x_j"

page 585 Problem **6**, *line +1*: Change "Consider a c component" to "Consider a c-component"

page 587 Problem **13**, *line +3*: Change "that for any observed x, all but one" to "that for any observed x all but one"

page 589 Problem **24** b, Change "the trace criterion" to "the determinant criterion"

page 591 Problem **41**, *line +1*: Change "null hypothesis in associated" to "null hypothesis associated"

page 592 Problem **44**, part (c), *line +3*: Change "$(\delta\mathbf{e})^t\mathbf{\Sigma}\lambda(\delta\mathbf{e})^t\mathbf{e} = 0$" to "$(\delta\mathbf{e})^t\mathbf{\Sigma}\mathbf{e} - \lambda(\delta\mathbf{e})^t\mathbf{e} = 0$"

page 592 Problem **46**, *line +1*: Change "principal componet analysis" to "principal component analysis"

page 593 Problem **48**, *line +3*: Change "linearity is given by" to "linearity is the one given by"

page 596 Problem **11**, *lines +3 – 4*: Change "to the date in the table above using the distance measure indicated" to "to the data in the table above using the distance measures indicated"

page 597 Problem **13**, *line +1*: Change "a basic competitive learning" to "a basic Competitive Learning"

page 597 Problem **13**, *lines +4 – 5*: Change "hy-persphere" to "hyper-sphere" (i.e., re-hyphenate "hypersphere")

Appendix

page 608 Section A.2.5 *line +5*: Change "In this case the absolute value of the determinant" to "In this case the determinant"

page 609 first line after Eq. 26: Change "the $i,\ \ j$ *cofactor* or" to "the i, j *cofactor* or" (i.e., eliminate the space in "i, j")

page 609 second line after Eq. 26: Change "is the $(d-1) \times (d-1)$ matrix" to "is the $(d-1)$-by-$(d-1)$ matrix"

page 609 fourth line after Eq. 26: Change "whose $i,\ j$ entry is the $j,\ i$ cofactor" to "whose i, j entry is the j, i cofactor" (i.e., eliminate the space in "$i,\ j$" and "$j,\ i$")

page 609 *line -2*: Add "The inverse of the product of two square matrices obeys $[\mathbf{MN}]^{-1} = \mathbf{N}^{-1}\mathbf{M}^{-1}$, as can be verified by multiplying on the right or the left by $\mathbf{MN}$."

page 615 *line -12* (i.e., just before Section A.4.7): Change "and $\frac{n_{11}/n}{(n_{01}+n_{11})/n}$ is approximately" to "and $(n_{01}+n_{11})/n$ is approximately"

page 624 Figure A.3, caption, *line +2*: Change "between $-\sqrt{2}u$ and $\sqrt{2}u$, that is" to "between $-\sqrt{2}u$ and $\sqrt{2}u$; that is"

page 629 *line +9*: Change "from a distribution" to "from a standarized Gaussian distribution"

page 631 *line +9*: Change "equally likely is" to "equally likely, is"

page 631 *line +12*: Change "outcome and $H = \log_2 2^3 = 3$" to "outcome and $H = -\sum_{i=0}^{7} \frac{1}{2^3} \log_2 2^3 = \log_2 2^3 = 3$"

page 631 Equation 118: Change "$\ln p(x)$" to "$\ln p(x)$" (i.e., reduce the space between "ln" and "$p(x)$")

page 631 first line after Eq. 119: Change "For this Dirac function" to "For this Dirac density"

page 632 Red margin note: Change "KULLBACK-LEIBLER DISTANCE"to "KULLBACK-LEIBLER DISTANCE" (i.e., capitalize the "L" in "LEIBLER")

page 632 *line -2*: Change is always larger than" to "is never smaller than"

page 634 *line +3*: Change "$f(x) \leq c_0 g(x)$ for all" to "$f(x) \leq c g(x)$ for all"

page 634 *line +7*: Change "proper choice of c_0 and x_0" to "proper choice of c and x_0"

Index

page 644 column 1, *line -14*: Insert "Gini impurity, 399, 401"

page 653 column 1, *line -3*: Change "multi-variate" to "multivariate"

Third and fourth printings

Chapter 2

page 53 Bottom equation (for w_0): Change "= 1.2" to "= -1.75"

page 58 In the figure, the entry for $P(c_3|x_1)$ should be changed from 0.1 to 0.2.

page 60 Equation 99: Change "$\mathbf{X}$" to "$\mathbf{x}$" in two places

page 73 Problem **34**, *line +6–7*: Change "assume the distributions" to "assume $P(\omega_1) = P(\omega_2) = 0.5$ and the distributions"

page 73 Problem **34**, part c), *line +4*: Change "Bayes error is 0.5." to "Bayes error is 0.25."

Chapter 3

page 99 Caption to first figure, change "starts our as a flat" to "starts out as a flat"

page 133 Figure 3.10, change the label on the horizontal arrow from "a_{12}" to "a_{22}"

page 143 Problem **11**, second and third lines after first equation: Change "$p_2(\mathbf{x})$ by a normal $p_1(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$" to "$p_1(\mathbf{x})$ by a normal $p_2(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$"

page 143 Problem **11**: Second equations: Change "$\mathcal{E}_2$" to "$\mathcal{E}_1$" in two places

page 143 Problem **11**, last line: Change "over the density $p_2(\mathbf{x})$" to "over the density $p_1(\mathbf{x})$"

page 149 Problem **31**, *line 1*: Change "suppose a and b are positive constants and" to "suppose a and b are constants greater than 1 and"

page 151 Problem **38 (b)** bottom equation on page: Change "$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^2$" to "$(\mu_1 - \mu_2)^2$"

page 156 Computer exercise 2, after the equation: Change "calculate a density" to "calculate the density"

page 156 Computer exercise 2. After "x_2 feature of category ω_2." add "Assume your priors on the parameters are uniform throughout the range of the data."

page 159 Computer exercise **13** table, sample 4 under ω_1: Change "`AD`" to "`ADB`"

Chapter 4

page 178 *line +3*: Change "Becuase" to "Because"

page 202 Problem **6** part (c), first line: Change "increases" to "increase"

page 207 Problem **26** part (d), first line: change "Calculate the probability" to "Estimate through simulation the probability"

page 207 Problem **26** delete part (e)

Chapter 5

page 220 sentence before Eq. 5, Change "this in turn suggest" to "this in turn suggests"

page 250 Equation 79: Change "$\mathbf{a}(k)$" to "$\mathbf{b}(k)$"

page 251 Algorithm 11, *line 5*: Change "$\mathbf{a}$" to "$\mathbf{b}$"

page 254 line after Eq. 90: Change "constant, positive-definite" to "constant, symmetric, positive-definite"

page 255 First (unnumbered) equation on page: Change the first term in parentheses on the right-hand side from "$\eta^2\mathbf{YRY}^t\mathbf{YRY}$" to "$\eta^2\mathbf{YRY}^t\mathbf{YRY}^t$"

page 255 Eq. 92: Last term on the right-hand-side, change "$\eta^2\mathbf{RY}^t\mathbf{R}$" to "$\eta^2\mathbf{RY}^t\mathbf{YR}$"

page 274 Problem **22**, second term on the right-hand side change "$(\mathbf{a}^t\mathbf{y} - (\lambda_{12} - \lambda_{22}))^2$" to "$(\mathbf{a}^t\mathbf{y} + (\lambda_{12} - \lambda_{22}))^2$"

page 275 Problem **27**, last line: Change "by Eq. 85." to "by Eq. 95."

page 278 Problem **1**, part (a), *line +1*: change "data in in order" to "data in order"

page 278 Problem **1**, part (a), *line +2*: change "use $\eta(k) = 0.1$." to "use $\eta(k) = 0.01$."

page 278 First heading after the table: Change "**Section 5.4**" to "**Section 5.5**"

page 278 Computer Exercise 1: Change "(Algorithm 1) and Newton's algorithm (Algorithm 2) applied" to "(Algorithm 1) and the Perceptron criterion (Eq. 16) applied"

page 278 Second heading after the table: Delete "**Section 5.5**"

page 279 *line +1*: Change "length is greater than the pocket" to "length is greater than with the pocket"

Chapter 6

page 291 Two lines before Eq. 3: Change "hidden-to-output weights, w_{ij}" to "hidden-to-output weights, w_{kj}"

page 292 After Eq. 19, *line +3*: Change "activiation" to "activation"

page 293 Figure 6.5: Change "w_{ij}" to "w_{ji}"

page 294 Algorithm 2, *line 3*: Change "$\Delta w_{kj} \leftarrow$" to "$\Delta w_{kj} \leftarrow 0$"

page 295 Second paragraph in Section 6.3.3: Change "indpendently selected" to "independently selected"

page 302 *line +3–4*: Change "weights merely leads" to "weights merely lead"

page 305 *line 10*: Change "ratio of such priors." to "ratio of such priors, though this need not ensure minimal error."

page 314 *line +6*: Change "weight changes are response" to "weight changes are a response"

page 330 *line -2– -1*: Change "While it is natural" to "It is natural"

page 337 Problem **8**, pagt (b) *line +2*: Change "if the sign if flipped" to "if the sign is flipped"

page 346 Problem **10**, part (c), *line +3*: Change "$\mathbf{x}_5 = (0, 0, 0)^t$" to "$\mathbf{x}_5 = (0, 0, 1)^t$"

Chapter 7

page 352 Caption to Figure 7.1, *line -1*: Change "$0 \leq \alpha \leq 2^{10}$" to "$0 \leq \alpha < 2^{10}$"

Chapter 8

page 405 Table at top, x_2 entry in fifth row under ω_1, change ".48" to ".44"

page 409 *line -7*: Change "queries involves" to "queries involve"

page 416 Algorithm 2, line 2: Change "$\mathcal{F}(\mathbf{x})$" to "$\mathcal{F}$"

page 416 Algorithm 2, line 3: Change "$\mathcal{G}(\mathbf{x})$" to "$\mathcal{G}$"

page 416 Algorithm 2, line 11: Change "$\mathcal{G}(0)$" to "1"

page 416 *line -1*: Change "$\mathcal{F}(\mathbf{x})$" to "$\mathcal{F}$"

page 417 *lines -9 – -4*, Replace last full paragraph by "Consider target string $\mathbf{x}$. Each location j (for $j < m$) defines a suffix of $\mathbf{x}$, that is, $\mathbf{x}[j+1, \ldots, m]$. The *good-suffix function* $\mathcal{G}(j)$ returns the starting location of the right-most instance of another occurrence of that suffix (if one exists). In the example in Fig. 8.8, $\mathbf{x}$ = `estimates` and thus $j = 8$ defines the suffix `s`. The right-most occurrence of another `s` is 2; therefore $\mathcal{G}(8) = 2$. Similarly $j = 7$ defines the suffix `es`. The right-most occurrence of another `es` is 2; therefore $\mathcal{G}(7) = 1$. No other suffix appears repeatedly within $\mathbf{x}$, and thus $\mathcal{G}$ is undefined for $j < 7$."

page 438 Problem **5**, *line +2*: Change "Eqs. 1 and 5." to "Eqs. 1 and 5 for the case of an arbitrary number of categories."

page 438 Problem **6** after the first set of equations: Replace the $i^*(\boldsymbol{\alpha})$ equation by "$i * (\alpha) = i(\alpha P^a(\omega_1) + (1-\alpha)P^b(\omega_1), ..., \alpha P^a(\omega_c) + (1-\alpha)P^b(\omega_c))$"

page 438 Problem **6** last line before part (a): Replace line by "then we have $i^* \geq \alpha i_a + (1-\alpha) i_b$."

page 446 Table, sample 12: Change "D" to "E"

page 447 Computer exercise **3**, part b): Change "{C, D, J, L, M}" to "{C, E, J, L, M}"

Chapter 9

page 460 Table for rank $r = 3$, third row: Change "$\mathbf{x}_1\ OR\ \mathbf{x}_3\ OR\ \mathbf{x}_3$" to "$\mathbf{x}_1\ OR\ \mathbf{x}_3\ OR\ \mathbf{x}_4$"

page 468 Five lines after Eq. 12: Change "regardless the *amount*" to "regardless of the *amount*"

page 474 Caption to Figure: Change "and jackknife estimate" to "and whose jackknife estimate"

page 497 Equation 58: Change "$\boldsymbol{\mu}_r$" to "$\boldsymbol{\theta}_r$" in two places

page 505 Take the heading "**Section 9.4**" and move it to the top of the page, i.e., between Problems **20** and **21**.

page 508 Problem **45**, *line +2*: Change "$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$" to "$N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$"

Chapter 10

page 526 *line -1* Change "In the absense" to "In the absence"

page 541 *line +9*: Change "this is a symmetric functions" to "this is a symmetric function"

page 549 Eq. 76: Change "$\mathbf{m}$" to "$\mathbf{m}_i$" on the righthand side.

page 573 Eq. 107: Put the lower limit underneath the summation sign, that is, change "$\sum_{i<j}$" to "$\sum_{i<j}$"

page 573 Eq. 109: Put the lower limit underneath the summation sign, that is, change "$\sum_{i<j}$" to "$\sum_{i<j}$"

page 574 First equation: Put the lower limit underneath the summation sign, that is, change "$\sum_{i<j}$" to "$\sum_{i<j}$"

page 574 Third equation: Put the lower limit underneath the summation sign, that is, change "$\sum_{i<j}$" to "$\sum_{i<j}$"

page 576 Eq. 112: Put the lower limit underneath the summation sign, that is, change "$\sum_{i<j}$" to "$\sum_{i<j}$"

page 558 Last full paragraph, *line +1*: Change "This result agrees with out statement" to "This result agrees with our statement"

page 584 Problem **4** part (a), right-hand side: Change "$\mathbf{x}_i$" to "x_j"

page 585 Problem **4** part (b): Change "$\hat{\boldsymbol{\theta}}_i$" to "$\hat{\boldsymbol{\theta}}$" in two places only on the right-hand side of the equation.

page 589 Problem **24** part (b), Change "the trace criterion" to "the determinant criterion"

Appendix

page 608 Section A.2.5 *line +5*: Change "In this case the absolute value of the determinant" to "In this case the determinant"

Fifth printing

Chapter 2

page 73 Problem **34**, *line +6–7*: Change "assume the distributions" to "assume $P(\omega_1) = P(\omega_2) = 0.5$ and the distributions"

page 73 Problem **34**, part c), *line +4*: Change "Bayes error is 0.5." to "Bayes error is 0.25."

Chapter 3

page 143 Problem **11**, second and third lines after first equation: Change "$p_2(\mathbf{x})$ by a normal $p_1(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$" to "$p_1(\mathbf{x})$ by a normal $p_2(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$"

page 143 Problem **11**: Second equations: Change "$\mathcal{E}_2$" to "$\mathcal{E}_1$" in two places

page 143 Problem **11**, last line: Change "over the density $p_2(\mathbf{x})$" to "over the density $p_1(\mathbf{x})$"

page 151 Problem **38 (b)** bottom equation on page: Change "$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^2$" to "$(\mu_1 - \mu_2)^2$"

page 159 Computer exercise **13** table, sample 4 under ω_1: Change "`AD`" to "`ADB`"

Chapter 5

page 250 Equation 79: Change "$\mathbf{a}(k)$" to "$\mathbf{b}(k)$"

page 251 Algorithm 11, *line 5*: Change "**a**" to "**b**"

page 275 Problem **27**, last line: Change "by Eq. 85." to "by Eq. 95."

Chapter 6

page 294 Algorithm 2, *line 3*: Change "$\Delta w_{kj} \leftarrow$" to "$\Delta w_{kj} \leftarrow 0$"

page 305 *line 10*: Change "ratio of such priors." to "ratio of such priors, though this need not ensure minimal error."

page 337 Problem **8**, pagt (b) *line +2*: Change "if the sign if flipped" to "if the sign is flipped"

Chapter 8

page 438 Problem **6** after the first set of equations: Replace the $i^*(\boldsymbol{\alpha})$ equation by "$i*(\alpha) = i(\alpha P^a(\omega_1) + (1-\alpha)P^b(\omega_1), ..., \alpha P^a(\omega_c) + (1-\alpha)P^b(\omega_c))$"

page 438 Problem **6** last line before part (a): Replace line by "then we have $i^* \geq \alpha i_a + (1-\alpha) i_b$."

page 446 Table, sample 12: Change "D" to "E"

page 447 Computer exercise **3**, part b): Change "{C, D, J, L, M}" to "{C, E, J, L, M}"

Chapter 9

page 460 Table for rank $r = 3$, third row: Change "$\mathbf{x}_1\ OR\ \mathbf{x}_3\ OR\ \mathbf{x}_3$" to "$\mathbf{x}_1\ OR\ \mathbf{x}_3\ OR\ \mathbf{x}_4$"

page 508 Problem **45**, *line +2*: Change "$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$" to "$N(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$"

Chapter 10

page 553 *line +13–14*: Change "each of which is an $O(d^2)$ calculation" to "each of which is an $O(d)$ calculation"

page 553 *line +17–18*: Change "the complexity is $O(n(n-1)(d^2+1)) = O(n^2d^2)$" to "the complexity is $O(n(n-1)(d+1)) = O(n^2d)$"

page 553 *line +21*: Change "complexity is thus $O(cn^2d^2)$" to "complexity is thus $O(cn^2d)$"