

THE PROBABILITY COMPANION for Engineering and Computer Science



Adam Prügel-Bennett

CAMBRIDGE

Copyrighted material

The Probability Companion for Engineering and Computer Science

Adam Prügel-Bennett
University of Southampton



Copyrighted material

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781108480536

DOI: 10.1017/9781108635349

© Adam Prügel-Bennett 2019

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2019

Printed in the United Kingdom by TJ International, Padstow, Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Prügel-Bennett, Adam, 1963– author.

Title: The probability companion for engineering and computer science / Adam
Prügel-Bennett, University of Southampton.

Description: Cambridge, United Kingdom ; New York, NY, USA : University
Printing House, 2019. | Includes bibliographical references and index.

Identifiers: LCCN 2019015914 | ISBN 9781108480536 (hardback : alk. paper) |
ISBN 9781108727709 (paperback : alk. paper)

Subjects: LCSH: Engineering–Statistical methods. | Computer
science–Statistical methods. | Probabilities.

Classification: LCC TA340 .P84 2019 | DDC 519.2–dc23

LC record available at <https://lcn.loc.gov/2019015914>

ISBN 978-1-108-48053-6 Hardback

ISBN 978-1-108-72770-9 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of
URLs for external or third-party internet websites referred to in this publication
and does not guarantee that any content on such websites is, or will remain,
accurate or appropriate.

Copyrighted material

Bayes

Contents

8.1	Bayesian Statistics	188
8.2	Performing Bayesian Inference	197
8.2.1	Conjugate Priors	198
8.2.2	Uninformative Priors	205
8.2.3	Model Selection	211
8.3	Bayes with Complex Likelihoods	214
8.3.1	Hierarchical Models	214
8.3.2	MAP Solution	215
8.3.3	Variational Approximation	220
8.4	Latent Variable Models	227
8.4.1	Hidden Markov Models	235
8.4.2	Variational Auto-Encoders	240
8.5	Machine Learning	243
8.5.1	Naive Bayes Classifier	244
8.5.2	Graphical Models	245
8.A	Bertrand's Paradox	256

If you are trying to infer properties about real systems where there is uncertainty in the data and you have a good model of the data-generating process, then Bayesian methods are likely to be the best you can do. This field is now vast and continues to grow. In this chapter, we give an overview of the approach and explain why it has caused both interest and controversy. We then discuss some of the technical details including conjugate distributions, uninformative priors, and model selection. We show some of the tools for approximating the posterior and finish by showing some of the Bayesian-based machine learning tools.

Bayesian inference underlies a lot of modern technology. Its contribution can range from providing a state-of-the-art machine learning tool, such as Gaussian processes, to being the cornerstone of a field such as in statistical machine translation where the translation problem is posed entirely in the language of probabilistic inference. The basis of Bayesian inference is a simple mathematical identity, but its application requires a philosophical understanding of how to use it and a body of mathematical tools required to make the inference tractable.

8.1 Bayesian Statistics

Bayes' rule provides a very simple way to perform probabilistic inference. It was proposed by the Reverend Thomas Bayes in an article published posthumously in 1763. It was invented independently, but slightly later by Laplace. Given its long history and the authority of Laplace it is perhaps surprising that it was shunned in the first half of the twentieth century as 'subjective probability'. This arose because it requires putting a prior probability on unseen events – this was seen as subjective rather than objective, which no self-respecting scientist could tolerate. During this period, being Bayesian meant being unscientific and few people defended the Bayesian approach (the economist John Maynard Keynes and the mathematical physicist Harold Jeffreys were notable exceptions). Only in the latter half of the twentieth century did the Bayesian view reassert itself, most notably through the advocacy of Ed Jaynes. Eventually though it was computers that won the day. Inference is a significant part of machine learning, and (computerised) Bayesian techniques beat many more traditional methods. The more ardent Bayesians often proclaim that the only right way of doing inference is using a Bayesian approach, although there remain grounds for being critical of many techniques that are claimed to be Bayesian. Nevertheless, the Bayesian approach can no longer be dismissed.

So what is Bayesian inference? Suppose we have a number of mutually exclusive hypotheses. These might be, for example, *who is the murderer in Cluedo?* or *how many fish are there in the sea?* We label these hypotheses $\mathcal{H}_i$ for $i = 1, 2, \dots, n$. We also have some data, $\mathcal{D}$, pertaining to the hypotheses. Bayes' rule – see Equation (1.4) on page 7 – then tells us the probability of the hypotheses given the data

$$\mathbb{P}(\mathcal{H}_i|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{H}_i) \mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(\mathcal{D})}. \quad (8.1)$$

This is a mathematical identity so in itself it is rather uncontroversial. We call the probability on the left-hand side, $\mathbb{P}(\mathcal{H}_i|\mathcal{D})$, the *posterior* (as it is the probability *after* seeing the data). The probability $\mathbb{P}(\mathcal{H}_i)$ is known as the *prior* (being the probability of the hypothesis *before* seeing the data). The other terms are the *likelihood*, $\mathbb{P}(\mathcal{D}|\mathcal{H}_i)$, and the probability of the data, $\mathbb{P}(\mathcal{D})$, which is just equal to

$$\mathbb{P}(\mathcal{D}) = \sum_{i=1}^n \mathbb{P}(\mathcal{D}, \mathcal{H}_i) = \sum_{i=1}^n \mathbb{P}(\mathcal{D}|\mathcal{H}_i) \mathbb{P}(\mathcal{H}_i) \quad (8.2)$$

and guarantees that the posterior probability is normalised. We will see later in this subsection, however, that this term – sometimes referred to as the *evidence* – can play a very important role in model selection.

Bayes' rule also passes seamlessly to continuous variables, where instead of considering probabilities we consider probability densities. If we are trying to infer the value of a continuous parameter, x , describing the distribution of discrete data, Y , we can encode our uncertainty in terms of a probability density function $f(x)$. Then we can again use Bayes' rule

Historically the Bayesian approach was seen to lead to inconsistencies as you would get different results by assuming different prior distributions. One of the many contributions of Ed Jaynes was to show that by applying symmetry principles it is possible to unambiguously decide what is the correct prior given no information – see Appendix 8.A on Bertrand's paradox.

$$\mathbb{P}(x < X < x + \delta x | Y) = \frac{\mathbb{P}(Y|x) \mathbb{P}(x < X < x + \delta x)}{\mathbb{P}(Y)}$$

or dividing through by δx and taking the limit

$$\lim_{\delta x \rightarrow 0} \frac{\mathbb{P}(x < X < x + \delta x | Y)}{\delta x} = \lim_{\delta x \rightarrow 0} \frac{\mathbb{P}(Y|x) \mathbb{P}(x < X < x + \delta x)}{\mathbb{P}(Y) \delta x}$$

$$f(x|Y) = \frac{\mathbb{P}(Y|x) f(x)}{\mathbb{P}(Y)}.$$

Similarly, when our data is continuous, the probability that a data-point Y is in some ball, $B_\epsilon(y)$ of radius ϵ around the point y is

$$\mathbb{P}(Y \in B_\epsilon(y)) \approx f(y) |B_\epsilon(y)|$$



where $|B_\epsilon(y)|$ is the volume of the ball; the approximation becomes increasingly accurate as $\epsilon \rightarrow 0$. Putting this into Bayes' rule

$$f(x|Y=y) \approx \frac{\mathbb{P}(Y \in B_\epsilon(y)|x) f(x)}{\mathbb{P}(Y \in B_\epsilon(y))} \approx \frac{f(y|x) |B_\epsilon(y)| f(x)}{f(y) |B_\epsilon(y)|} = \frac{f(y|x) f(x)}{f(y)},$$

which becomes exact (assuming the density is sufficiently smooth) in the limit $\epsilon \rightarrow 0$. Similarly, if we were trying to infer a discrete parameter with continuous data we would get

$$\mathbb{P}(x|Y) = \frac{f(Y|x) \mathbb{P}(x)}{f(Y)}.$$

We see that the form of Bayes' rule does not vary whether we have probability masses or probability densities. It is easy to forget the difference between probability masses and probability densities when working in the Bayesian formalism, but remember that $f(x)$ and $f(x|Y)$ are not probabilities (although positive they can be greater than 1).

The likelihood. Before discussing the technicalities of performing a Bayesian calculation some comments are in order. Firstly, Bayes' rule allows us to reverse the conditional probabilities. That is, rather than having to compute the posterior, $\mathbb{P}(\mathcal{H}_i|\mathcal{D})$, directly we need to compute the likelihood, $\mathbb{P}(\mathcal{D}|\mathcal{H}_i)$. But the likelihood is often much easier to compute than the posterior. To understand why this is so let us consider collecting data from some instrument. For example, we might want to deduce the structure of a crystal from some scattering experiment. The scattering pattern is usually related to the Fourier transform of the crystal structure. In principle, then we could deduce the crystal structure by taking the inverse Fourier transform. Here we are having to solve an *inverse problem*.