

Information Theory, Inference, and Learning Algorithms

David J.C. MacKay

Information Theory, Inference, and Learning Algorithms

David J.C. MacKay
mackay@mrao.cam.ac.uk

©1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003

©Cambridge University Press 2003

Version 6.8 (second printing) March 14, 2004

Please send feedback on this book via
<http://www.inference.phy.cam.ac.uk/mackay/itila/>

This book (version 6.0) was published by C.U.P. in September 2003. It will remain viewable on-screen on the above website, in postscript, djvu, and pdf formats.

In the second printing (version 6.6) minor typos were corrected, and the book design was slightly altered to modify the placement of section numbers.

(C.U.P. replace this page with their own page ii.)

Contents

	Preface	v
1	Introduction to Information Theory	3
2	Probability, Entropy, and Inference	22
3	More about Inference	48
I	Data Compression	65
4	The Source Coding Theorem	67
5	Symbol Codes	91
6	Stream Codes	110
7	Codes for Integers	132
II	Noisy-Channel Coding	137
8	Correlated Random Variables	138
9	Communication over a Noisy Channel	146
10	The Noisy-Channel Coding Theorem	162
11	Error-Correcting Codes and Real Channels	177
III	Further Topics in Information Theory	191
12	Hash Codes: Codes for Efficient Information Retrieval	193
13	Binary Codes	206
14	Very Good Linear Codes Exist	229
15	Further Exercises on Information Theory	233
16	Message Passing	241
17	Communication over Constrained Noiseless Channels	248
18	Crosswords and Codebreaking	260
19	Why have Sex? Information Acquisition and Evolution	269
IV	Probabilities and Inference	283
20	An Example Inference Task: Clustering	286
21	Exact Inference by Complete Enumeration	295
22	Maximum Likelihood and Clustering	302
23	Useful Probability Distributions	313
24	Exact Marginalization	321
25	Exact Marginalization in Trellises	326
26	Exact Marginalization in Graphs	336
27	Laplace's Method	343

28	Model Comparison and Occam's Razor	345
29	Monte Carlo Methods	359
30	Efficient Monte Carlo Methods	389
31	Ising Models	402
32	Exact Monte Carlo Sampling	415
33	Variational Methods	424
34	Independent Component Analysis and Latent Variable Modelling	439
35	Random Inference Topics	447
36	Decision Theory	453
37	Bayesian Inference and Sampling Theory	459
V	Neural networks	469
38	Introduction to Neural Networks	470
39	The Single Neuron as a Classifier	473
40	Capacity of a Single Neuron	485
41	Learning as Inference	494
42	Hopfield Networks	507
43	Boltzmann Machines	524
44	Supervised Learning in Multilayer Networks	529
45	Gaussian Processes	537
46	Deconvolution	551
VI	Sparse Graph Codes	557
47	Low-Density Parity-Check Codes	559
48	Convolutional Codes and Turbo Codes	576
49	Repeat-Accumulate Codes	584
50	Digital Fountain Codes	591
VII	Appendices	599
A	Notation	600
B	Some Physics	603
C	Some Mathematics	607
	Bibliography	615
	Index	622

Preface

This book is aimed at senior undergraduates and graduate students in Engineering, Science, Mathematics, and Computing. It expects familiarity with calculus, probability theory, and linear algebra as taught in a first- or second-year undergraduate course on mathematics for scientists and engineers.

Conventional courses on information theory cover not only the beautiful *theoretical* ideas of Shannon, but also *practical* solutions to communication problems. This book goes further, bringing in Bayesian data modelling, Monte Carlo methods, variational methods, clustering algorithms, and neural networks.

Why unify information theory and machine learning? Because they are two sides of the same coin. In the 1960s, a single field, cybernetics, was populated by information theorists, computer scientists, and neuroscientists, all studying common problems. Information theory and machine learning still belong together. Brains are the ultimate compression and communication systems. And the state-of-the-art algorithms for both data compression and error-correcting codes use the same tools as machine learning.

How to use this book

The essential dependencies between chapters are indicated in the figure on the next page. An arrow from one chapter to another indicates that the second chapter requires some of the first.

Within Parts I, II, IV, and V of this book, chapters on advanced or optional topics are towards the end. All chapters of Part III are optional on a first reading, except perhaps for Chapter 16 (Message Passing).

The same system sometimes applies within a chapter: the final sections often deal with advanced topics that can be skipped on a first reading. For example in two key chapters – Chapter 4 (The Source Coding Theorem) and Chapter 10 (The Noisy-Channel Coding Theorem) – the first-time reader should detour at section 4.5 and section 10.4 respectively.

Pages vii–x show a few ways to use this book. First, I give the roadmap for a course that I teach in Cambridge: ‘Information theory, pattern recognition, and neural networks’. The book is also intended as a textbook for traditional courses in information theory. The second roadmap shows the chapters for an introductory information theory course and the third for a course aimed at an understanding of state-of-the-art error-correcting codes. The fourth roadmap shows how to use the text in a conventional course on machine learning.

