

C.S. Wallace

Statistical and Inductive Inference by Minimum Message Length

With 22 Figures

 Springer

Contents

Preface	v
1. Inductive Inference	1
1.1 Introduction	1
1.2 Inductive Inference	5
1.3 The Demise of Theories	11
1.4 Approximate Theories	14
1.5 Explanation	14
1.5.1 Explanatory Power	16
1.5.2 The Explanation Message	16
1.6 Random Variables	19
1.7 Probability	21
1.8 Independence	22
1.9 Discrete Distributions	23
1.9.1 Example: The Binomial Distribution	24
1.10 Continuous Distributions	24
1.11 Expectation	25
1.12 Non-Bayesian Inference	28
1.12.1 Non-Bayesian Estimation	30
1.12.2 Non-Bayesian Model Selection	32
1.13 Bayesian Inference	35
1.14 Bayesian Decision Theory	40
1.15 The Origins of Priors	45
1.15.1 Previous Likelihood	46
1.15.2 Conjugate Priors	46
1.15.3 The Jeffreys Prior	48
1.15.4 Uninformative Priors	49
1.15.5 Maximum-Entropy Priors	51
1.15.6 Invariant Conjugate Priors	52
1.15.7 Summary	53
1.16 Summary of Statistical Critique	54

2. Information	57
2.1 Shannon Information	57
2.1.1 Binary Codes	59
2.1.2 Optimal Codes	63
2.1.3 Measurement of Information	66
2.1.4 The Construction of Optimal Codes	69
2.1.5 Coding Multi-Word Messages	72
2.1.6 Arithmetic Coding	73
2.1.7 Some Properties of Optimal Codes	76
2.1.8 Non-Binary Codes: The Nit	77
2.1.9 The Subjective Nature of Information	79
2.1.10 The Information Content of a Multinomial Distribution	81
2.1.11 Entropy	87
2.1.12 Codes for Infinite Sets	91
2.1.13 Unary and Punctuated Binary Codes	92
2.1.14 Optimal Codes for Integers	93
2.1.15 Feasible Codes for Infinite Sets	96
2.1.16 Universal Codes	98
2.2 Algorithmic Complexity	100
2.2.1 Turing Machines	101
2.2.2 Start and Stop Conditions	102
2.2.3 Dependence on the Choice of Turing Machine	103
2.2.4 Turing Probability Distributions	103
2.2.5 Universal Turing Machines	105
2.2.6 Algorithmic Complexity vs. Shannon Information	107
2.3 Information, Inference and Explanation	110
2.3.1 The Second Part of an Explanation	110
2.3.2 The First Part of an Explanation	112
2.3.3 Theory Description Codes as Priors	114
2.3.4 Universal Codes in Theory Descriptions	115
2.3.5 Relation to Bayesian Inference	116
2.3.6 Explanations and Algorithmic Complexity	118
2.3.7 The Second Part	120
2.3.8 The First Part	121
2.3.9 An Alternative Construction	123
2.3.10 Universal Turing Machines as Priors	124
2.3.11 Differences among UTMs	130
2.3.12 The Origins of Priors Revisited	133
2.3.13 The Evolution of Priors	135

3. Strict Minimum Message Length (SMML)	143
3.1 Problem Definition	144
3.1.1 The Set X of Possible Data	144
3.1.2 The Probabilistic Model of Data	146
3.1.3 Coding of the Data	146
3.1.4 The Set of Possible Inferences	147
3.1.5 Coding the Inference $\hat{\theta}$	148
3.1.6 Prior Probability Density	150
3.1.7 Meaning of the Assertion	152
3.2 The Strict Minimum Message Length Explanation for Discrete Data	153
3.2.1 Discrete Hypothesis Sets	156
3.2.2 Minimizing Relations for SMML	156
3.2.3 Binomial Example	157
3.2.4 Significance of $I_1 - I_0$	160
3.2.5 Non-Uniqueness of Θ^*	161
3.2.6 Sufficient Statistics	161
3.2.7 Binomial Example Using a Sufficient Statistic	163
3.2.8 An Exact Algorithm for the Binomial Problem	164
3.2.9 A Solution for the Trinomial Distribution	165
3.3 The SMML Explanation for Continuous Data	166
3.3.1 Mean of a Normal	169
3.3.2 A Boundary Rule for Growing Data Groups	171
3.3.3 Estimation of Normal Mean with Normal Prior	173
3.3.4 Mean of a Multivariate Normal	177
3.3.5 Summary of Multivariate Mean Estimator	183
3.3.6 Mean of a Uniform Distribution of Known Range	183
3.4 Some General Properties of SMML Estimators	187
3.4.1 Property 1: Data Representation Invariance	187
3.4.2 Property 2: Model Representation Invariance	188
3.4.3 Property 3: Generality	189
3.4.4 Property 4: Dependence on Sufficient Statistics	189
3.4.5 Property 5: Efficiency	189
3.4.6 Discrimination	192
3.4.7 Example: Discrimination of a Mean	193
3.5 Summary	195
4. Approximations to SMML	197
4.1 The "Ideal Group" (IG) Estimator	197
4.1.1 SMML-like codes	198
4.1.2 Ideal Data Groups	198
4.1.3 The Estimator	199
4.2 The Neyman-Scott Problem	200
4.3 The Ideal Group Estimator for Neyman-Scott	201
4.4 Other Estimators for Neyman-Scott	202

4.5	Maximum Likelihood for Neyman-Scott	203
4.5.1	Marginal Maximum Likelihood	203
4.6	Kullback-Leibler Distance	204
4.7	Minimum Expected K-L Distance (MEKL)	205
4.8	Minimum Expected K-L Distance for Neyman-Scott	206
4.9	Blurred Images	208
4.10	Dowe's Approximation IID to the Message Length	209
4.10.1	Random Coding of Estimates	210
4.10.2	Choosing a Region in Θ	211
4.11	Partitions of the Hypothesis Space	213
4.12	The Meaning of Uncertainty Regions	215
4.12.1	Uncertainty via Limited Precision	216
4.12.2	Uncertainty via Dowe's IID Construction	216
4.12.3	What Uncertainty Is Described by a Region?	216
4.13	Summary	218
5.	MML: Quadratic Approximations to SMML	221
5.1	The MML Coding Scheme	222
5.1.1	Assumptions of the Quadratic MML Scheme	226
5.1.2	A Trap for the Unwary	227
5.2	Properties of the MML Estimator	228
5.2.1	An Alternative Expression for Fisher Information	228
5.2.2	Data Invariance and Sufficiency	229
5.2.3	Model Invariance	229
5.2.4	Efficiency	230
5.2.5	Multiple Parameters	232
5.2.6	MML Multi-Parameter Properties	234
5.2.7	The MML Message Length Formulae	235
5.2.8	Standard Formulae	235
5.2.9	Small-Sample Message Length	235
5.2.10	Curved-Prior Message Length	236
5.2.11	Singularities in the Prior	237
5.2.12	Large- D Message Length	237
5.2.13	Approximation Based on I_0	237
5.2.14	Precision of Estimate Spacing	238
5.3	Empirical Fisher Information	240
5.3.1	Formula 11A for Many Parameters	240
5.3.2	Irregular Likelihood Functions	242
5.3.3	Transformation of Empirical Fisher Information	243
5.3.4	A Safer? Empirical Approximation to Fisher Information	244
5.4	A Binomial Example	246
5.4.1	The Multinomial Distribution	247
5.4.2	Irregularities in the Binomial and Multinomial Distributions	248

5.5	Limitations	249
5.6	The Normal Distribution	250
5.6.1	Extension to the Neyman-Scott Problem	252
5.7	Negative Binomial Distribution	253
5.8	The Likelihood Principle	254
6.	MML Details in Some Interesting Cases	257
6.1	Geometric Constants	257
6.2	Conjugate Priors for the Normal Distribution	258
6.2.1	Conjugate Priors for the Multivariate Normal Distribution	261
6.3	Normal Distribution with Perturbed Data	264
6.4	Normal Distribution with Coarse Data	265
6.5	von Mises-Fisher Distribution	266
6.5.1	Circular von Mises-Fisher distribution	267
6.5.2	Spherical von Mises-Fisher Distribution	268
6.6	Poisson Distribution	269
6.7	Linear Regression and Function Approximation	270
6.7.1	Linear Regression	270
6.7.2	Function Approximation	272
6.8	Mixture Models	275
6.8.1	ML Mixture Estimation: The EM Algorithm	276
6.8.2	A Message Format for Mixtures	279
6.8.3	A Coding Trick	281
6.8.4	Imprecise Assertion of Discrete Parameters	284
6.8.5	The Code Length of Imprecise Discrete Estimates	286
6.8.6	A Surrogate Class Label "Estimate"	288
6.8.7	The Fisher Information for Mixtures	290
6.8.8	The Fisher Information with Class Labels	291
6.8.9	Summary of the Classified Model	293
6.8.10	Classified vs. Unclassified Models	295
6.9	A "Latent Factor" Model	297
6.9.1	Multiple Latent Factors	300
7.	Structural Models	305
7.1	Inference of a Regular Grammar	305
7.1.1	A Mealey Machine Representation	305
7.1.2	Probabilistic FSMs	307
7.1.3	An Assertion Code for PFSMs	308
7.1.4	A Less Redundant FSM Code	309
7.1.5	Transparency and Redundancy	310
7.1.6	Coding Transitions	312
7.1.7	An Example	313
7.2	Classification Trees and Nets	314
7.2.1	A Decision Tree Explanation	315

7.2.2	Coding the Tree Structure	316
7.2.3	Coding the Class Distributions at the Leaves	317
7.2.4	Decision Graphs and Other Elaborations	318
7.3	A Binary Sequence Segmentation Problem	321
7.3.1	The Kearns <i>et al.</i> "MDL" Criterion	322
7.3.2	Correcting the Message Length	323
7.3.3	Results Using the MML Criterion	324
7.3.4	An SMML Approximation to the Sequence Problem	325
7.4	Learning Causal Nets	326
7.4.1	The Model Space	327
7.4.2	The Message Format	328
7.4.3	Equivalence Sets	329
7.4.4	Insignificant Effects	329
7.4.5	Partial Order Equivalence	330
7.4.6	Structural Equivalence	330
7.4.7	Explanation Length	331
7.4.8	Finding Good Models	331
7.4.9	Prior Constraints	335
7.4.10	Test Results	335
8.	The Feathers on the Arrow of Time	337
8.1	Closed Systems and Their States	339
8.2	Reversible Laws	340
8.3	Entropy as a Measure of Disorder	341
8.4	Why Entropy Will Increase	343
8.5	A Paradox?	344
8.6	Deducing the Past	345
8.6.1	Macroscopic Deduction	345
8.6.2	Deduction with Deterministic Laws, Exact View	346
8.6.3	Deduction with Deterministic Laws, Inexact View	347
8.6.4	Deduction with Non-deterministic Laws	348
8.6.5	Alternative Priors	350
8.6.6	A Tale of Two Clocks	353
8.7	Records and Memories	355
8.8	Induction of the Past (<i>A la recherche du temps perdu</i>)	356
8.8.1	Induction of the Past by Maximum Likelihood	357
8.8.2	Induction of the Past by MML	358
8.8.3	The Uses of Deduction	361
8.8.4	The Inexplicable	362
8.8.5	Induction of the Past with Deterministic Laws	363
8.9	Causal and Teleological Explanations	365
8.10	Reasons for Asymmetry	367
8.11	Summary: The Past Regained?	369
8.12	Gas Simulations	370

8.12.1	Realism of the Simulation	372
8.12.2	Backtracking to the Past	373
8.12.3	Diatomic Molecules	374
8.12.4	The Past of a Computer Process	375
8.13	Addendum: Why Entropy Will Increase (Additional Simulation Details)	376
8.13.1	Simulation of the Past	381
8.13.2	A Non-Adiabatic Experiment	382
9.	MML as a Descriptive Theory	385
9.1	The Grand Theories	386
9.2	Primitive Inductive Inferences	387
9.3	The Hypotheses of Natural Languages	388
9.3.1	The Efficiencies of Natural Languages	389
9.4	Some Inefficiencies of Natural Languages	390
9.5	Scientific Languages	391
9.6	The Practice of MML Induction	391
9.6.1	Human Induction	394
9.6.2	Evolutionary Induction	396
9.6.3	Experiment	397
10.	Related Work	401
10.1	Solomonoff	401
10.1.1	Prediction with Generalized Scoring	405
10.1.2	Is Prediction Inductive?	405
10.1.3	A Final Quibble	407
10.2	Rissanen, MDL and NML	408
10.2.1	Normalized Maximum Likelihood	410
10.2.2	Has NML Any Advantage over MML?	413
	Bibliography	417
	Index	421