

## Contents

	<i>Preface</i>	<i>page xi</i>
	<i>Notation and abbreviations</i>	<i>xiii</i>
	<b>Part I General discussion</b>	<b>1</b>
<b>1</b>	<b>Introduction</b>	<b>3</b>
	1.1 Machine learning and speech and language processing	3
	1.2 Bayesian approach	4
	1.3 History of Bayesian speech and language processing	8
	1.4 Applications	9
	1.5 Organization of this book	11
<b>2</b>	<b>Bayesian approach</b>	<b>13</b>
	2.1 Bayesian probabilities	13
	2.1.1 Sum and product rules	14
	2.1.2 Prior and posterior distributions	15
	2.1.3 Exponential family distributions	16
	2.1.4 Conjugate distributions	24
	2.1.5 Conditional independence	38
	2.2 Graphical model representation	40
	2.2.1 Directed graph	40
	2.2.2 Conditional independence in graphical model	40
	2.2.3 Observation, latent variable, non-probabilistic variable	42
	2.2.4 Generative process	44
	2.2.5 Undirected graph	44
	2.2.6 Inference on graphs	46
	2.3 Difference between ML and Bayes	47
	2.3.1 Use of prior knowledge	48
	2.3.2 Model selection	49
	2.3.3 Marginalization	50
	2.4 Summary	51

<b>3</b>	<b>Statistical models in speech and language processing</b>	<b>53</b>
3.1	Bayes decision for speech recognition	54
3.2	Hidden Markov model	59
3.2.1	Lexical unit for HMM	59
3.2.2	Likelihood function of HMM	60
3.2.3	Continuous density HMM	63
3.2.4	Gaussian mixture model	66
3.2.5	Graphical models and generative process of CDHMM	67
3.3	Forward–backward and Viterbi algorithms	70
3.3.1	Forward–backward algorithm	70
3.3.2	Viterbi algorithm	74
3.4	Maximum likelihood estimation and EM algorithm	76
3.4.1	Jensen’s inequality	77
3.4.2	Expectation step	79
3.4.3	Maximization step	86
3.5	Maximum likelihood linear regression for hidden Markov model	91
3.5.1	Linear regression for hidden Markov models	92
3.6	$n$ -gram with smoothing techniques	97
3.6.1	Class-based model smoothing	101
3.6.2	Jelinek–Mercer smoothing	101
3.6.3	Witten–Bell smoothing	103
3.6.4	Absolute discounting	104
3.6.5	Katz smoothing	106
3.6.6	Kneser–Ney smoothing	107
3.7	Latent semantic information	113
3.7.1	Latent semantic analysis	113
3.7.2	LSA language model	116
3.7.3	Probabilistic latent semantic analysis	119
3.7.4	PLSA language model	125
3.8	Revisit of automatic speech recognition with Bayesian manner	128
3.8.1	Training and test (unseen) data for ASR	128
3.8.2	Bayesian manner	129
3.8.3	Learning generative models	131
3.8.4	Sum rule for model	131
3.8.5	Sum rule for model parameters and latent variables	132
3.8.6	Factorization by product rule and conditional independence	132
3.8.7	Posterior distributions	133
3.8.8	Difficulties in speech and language applications	134
	<b>Part II Approximate inference</b>	<b>135</b>
<b>4</b>	<b>Maximum a-posteriori approximation</b>	<b>137</b>
4.1	MAP criterion for model parameters	138

4.2	MAP extension of EM algorithm	141
4.2.1	Auxiliary function	141
4.2.2	A recipe	143
4.3	Continuous density hidden Markov model	143
4.3.1	Likelihood function	144
4.3.2	Conjugate priors (full covariance case)	144
4.3.3	Conjugate priors (diagonal covariance case)	146
4.3.4	Expectation step	146
4.3.5	Maximization step	149
4.3.6	Sufficient statistics	158
4.3.7	Meaning of the MAP solution	160
4.4	Speaker adaptation	163
4.4.1	Speaker adaptation by a transformation of CDHMM	163
4.4.2	MAP-based speaker adaptation	165
4.5	Regularization in discriminative parameter estimation	166
4.5.1	Extended Baum–Welch algorithm	167
4.5.2	MAP interpretation of i-smoothing	169
4.6	Speaker recognition/verification	171
4.6.1	Universal background model	172
4.6.2	Gaussian super vector	173
4.7	$n$ -gram adaptation	174
4.7.1	MAP estimation of $n$ -gram parameters	175
4.7.2	Adaptation method	175
4.8	Adaptive topic model	176
4.8.1	MAP estimation for corrective training	177
4.8.2	Quasi-Bayes estimation for incremental learning	179
4.8.3	System performance	182
4.9	Summary	183
<b>5</b>	<b>Evidence approximation</b>	<b>184</b>
5.1	Evidence framework	185
5.1.1	Bayesian model comparison	185
5.1.2	Type-2 maximum likelihood estimation	187
5.1.3	Regularization in regression model	188
5.1.4	Evidence framework for HMM and SVM	190
5.2	Bayesian sensing HMMs	191
5.2.1	Basis representation	192
5.2.2	Model construction	192
5.2.3	Automatic relevance determination	193
5.2.4	Model inference	195
5.2.5	Evidence function or marginal likelihood	196
5.2.6	Maximum a-posteriori sensing weights	197
5.2.7	Optimal parameters and hyperparameters	197

5.2.8	Discriminative training	200
5.2.9	System performance	203
5.3	Hierarchical Dirichlet language model	205
5.3.1	$n$ -gram smoothing revisited	205
5.3.2	Dirichlet prior and posterior	206
5.3.3	Evidence function	207
5.3.4	Bayesian smoothed language model	208
5.3.5	Optimal hyperparameters	208
<b>6</b>	<b>Asymptotic approximation</b>	<b>211</b>
6.1	Laplace approximation	211
6.2	Bayesian information criterion	214
6.3	Bayesian predictive classification	218
6.3.1	Robust decision rule	218
6.3.2	Laplace approximation for BPC decision	220
6.3.3	BPC decision considering uncertainty of HMM means	222
6.4	Neural network acoustic modeling	224
6.4.1	Neural network modeling and learning	225
6.4.2	Bayesian neural networks and hidden Markov models	226
6.4.3	Laplace approximation for Bayesian neural networks	229
6.5	Decision tree clustering	230
6.5.1	Decision tree clustering using ML criterion	230
6.5.2	Decision tree clustering using BIC	235
6.6	Speaker clustering/segmentation	237
6.6.1	Speaker segmentation	237
6.6.2	Speaker clustering	239
6.7	Summary	240
<b>7</b>	<b>Variational Bayes</b>	<b>242</b>
7.1	Variational inference in general	242
7.1.1	Joint posterior distribution	243
7.1.2	Factorized posterior distribution	244
7.1.3	Variational method	246
7.2	Variational inference for classification problems	248
7.2.1	VB posterior distributions for model parameters	249
7.2.2	VB posterior distributions for latent variables	251
7.2.3	VB–EM algorithm	251
7.2.4	VB posterior distribution for model structure	252
7.3	Continuous density hidden Markov model	254
7.3.1	Generative model	254
7.3.2	Prior distribution	255
7.3.3	VB Baum–Welch algorithm	257
7.3.4	Variational lower bound	269
7.3.5	VB posterior for Bayesian predictive classification	274

7.3.6	Decision tree clustering	282
7.3.7	Determination of HMM topology	285
7.4	Structural Bayesian linear regression for hidden Markov model	287
7.4.1	Variational Bayesian linear regression	288
7.4.2	Generative model	289
7.4.3	Variational lower bound	289
7.4.4	Optimization of hyperparameters and model structure	303
7.4.5	Hyperparameter optimization	304
7.5	Variational Bayesian speaker verification	306
7.5.1	Generative model	307
7.5.2	Prior distributions	308
7.5.3	Variational posteriors	310
7.5.4	Variational lower bound	316
7.6	Latent Dirichlet allocation	318
7.6.1	Model construction	318
7.6.2	VB inference: lower bound	320
7.6.3	VB inference: variational parameters	321
7.6.4	VB inference: model parameters	323
7.7	Latent topic language model	324
7.7.1	LDA language model	324
7.7.2	Dirichlet class language model	326
7.7.3	Model construction	327
7.7.4	VB inference: lower bound	328
7.7.5	VB inference: parameter estimation	330
7.7.6	Cache Dirichlet class language model	332
7.7.7	System performance	334
7.8	Summary	335
<b>8</b>	<b>Markov chain Monte Carlo</b>	<b>337</b>
8.1	Sampling methods	338
8.1.1	Importance sampling	338
8.1.2	Markov chain	340
8.1.3	The Metropolis–Hastings algorithm	341
8.1.4	Gibbs sampling	343
8.1.5	Slice sampling	344
8.2	Bayesian nonparametrics	345
8.2.1	Modeling via exchangeability	346
8.2.2	Dirichlet process	348
8.2.3	DP: Stick-breaking construction	348
8.2.4	DP: Chinese restaurant process	349
8.2.5	Dirichlet process mixture model	351
8.2.6	Hierarchical Dirichlet process	352
8.2.7	HDP: Stick-breaking construction	353
8.2.8	HDP: Chinese restaurant franchise	355

---

8.2.9	MCMC inference by Chinese restaurant franchise	356
8.2.10	MCMC inference by direct assignment	358
8.2.11	Relation of HDP to other methods	360
8.3	Gibbs sampling-based speaker clustering	360
8.3.1	Generative model	361
8.3.2	GMM marginal likelihood for complete data	362
8.3.3	GMM Gibbs sampler	365
8.3.4	Generative process and graphical model of multi-scale GMM	367
8.3.5	Marginal likelihood for the complete data	368
8.3.6	Gibbs sampler	370
8.4	Nonparametric Bayesian HMMs to acoustic unit discovery	372
8.4.1	Generative model and generative process	373
8.4.2	Inference	375
8.5	Hierarchical Pitman–Yor language model	378
8.5.1	Pitman–Yor process	379
8.5.2	Language model smoothing revisited	380
8.5.3	Hierarchical Pitman–Yor language model	383
8.5.4	MCMC inference for HPYLM	385
8.6	Summary	387
<b>Appendix A</b>	<b>Basic formulas</b>	388
<b>Appendix B</b>	<b>Vector and matrix formulas</b>	390
<b>Appendix C</b>	<b>Probabilistic distribution functions</b>	392
	<i>References</i>	405
	<i>Index</i>	422