Bayesian Speech and Language Processing

With this comprehensive guide you will learn how to apply Bayesian machine learning techniques systematically to solve various problems in speech and language processing.

A range of statistical models is detailed, from hidden Markov models to Gaussian mixture models, *n*-gram models, and latent topic models, along with applications including automatic speech recognition, speaker verification, and information retrieval. Approximate Bayesian inferences based on MAP, Evidence, Asymptotic, VB, and MCMC approximations are provided as well as full derivations of calculations, useful notations, formulas, and rules.

The authors address the difficulties of straightforward applications and provide detailed examples and case studies to demonstrate how you can successfully use practical Bayesian inference methods to improve the performance of information systems.

This is an invaluable resource for students, researchers, and industry practitioners working in machine learning, signal processing, and speech and language processing.

Shinji Watanabe received his Ph.D. from Waseda University in 2006. He has been a research scientist at NTT Communication Science Laboratories, a visiting scholar at Georgia Institute of Technology and a senior principal member at Mitsubishi Electric Research Laboratories (MERL), as well as having been an associate editor of the *IEEE Transactions on Audio Speech and Language Processing*, and an elected member of the IEEE Speech and Language Processing Technical Committee. He has published more than 100 papers in journals and conferences, and received several awards including the Best Paper Award from IEICE in 2003.

Jen-Tzung Chien is with the Department of Electrical and Computer Engineering and the Department of Computer Science at the National Chiao Tung University, Taiwan, where he is now the University Chair Professor. He received the Distinguished Research Award from the Ministry of Science and Technology, Taiwan, and the Best Paper Award of the 2011 IEEE Automatic Speech Recognition and Understanding Workshop. He serves currently as an elected member of the IEEE Machine Learning for Signal Processing Technical Committee.

"This book provides an overview of a wide range of fundamental theories of Bayesian learning, inference, and prediction for uncertainty modeling in speech and language processing. The uncertainty modeling is crucial in increasing the robustness of practical systems based on statistical modeling under real environment, such as automatic speech recognition systems under noise, and question answering systems based on limited size of training data. This is the most advanced and comprehensive book for learning fundamental Bayesian approaches and practical techniques."

Sadaoki Furui, Tokyo Institute of Technology

Bayesian Speech and Language Processing

SHINJI WATANABE
Mitsubishi Electric Research Laboratories

JEN-TZUNG CHIEN
National Chiao Tung University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

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/9781107055575

First published 2015

Printed in the United Kingdom by Clays, St Ives plc

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data
Watanabe, Shinji (Communications engineer) author.
Bayesian speech and language processing / Shinji Watanabe, Mitsubishi Electric Research Laboratories; Jen-Tzung Chien, National Chiao Tung University.
pages cm
ISBN 978-1-107-05557-5 (hardback)
1. Language and languages – Study and teaching – Statistical methods. 2. Bayesian statistical decision theory. I. Title.
P53.815.W38 2015
410.1′51–dc23

2014050265

ISBN 978-1-107-05557-5 Hardback

Contents

Preface *page* xi
Notation and abbreviations xiii

Part I General discussion 1

1 Introduction 3
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

2 Bayesian approach 13
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

3 Statistical models in speech and language processing 53
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

Part II Approximate inference 135

4 Maximum a-posteriori approximation 137
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

5 Evidence approximation 184
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

6 Asymptotic approximation 211
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

7 Variational Bayes 242
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

8 Markov chain Monte Carlo 337
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

Appendix A Basic formulas 388

Appendix B Vector and matrix formulas 390

Appendix C Probabilistic distribution functions 392

References 405
Index 422

Preface

In general, speech and language processing involves extensive knowledge of statistical models. The acoustic model using hidden Markov models and the language model using n-grams are mainly introduced here. Both acoustic and language models are important parts of modern speech recognition systems where the learned models from real-world data are full of complexity, ambiguity, and uncertainty. The uncertainty modeling is crucial to tackle the lack of robustness for speech and language processing.

This book addresses fundamental theories of Bayesian learning, inference, and prediction for the uncertainty modeling. Uniquely, compared with standard textbooks for dealing with the fundamental Bayesian approaches, this book focuses on the practical methods of the approaches to make them applicable to actual speech and language problems. We (the authors) have been studying these topics for a long time with a strong belief that the Bayesian approaches could solve "robustness" issues in speech and language processing, which are the most difficult problem and most serious shortcoming of real systems based on speech and language processing. In our experience, the most difficult issue in applying Bayesian approaches is how to appropriately choose a specific technique among the many Bayesian techniques proposed in statistics and machine learning so far. One of our answers to this question is to provide the approximated Bayesian inference methods rather than focusing on covering the whole Bayesian techniques. We categorize the Bayesian approaches into five categories: the maximum a-posteriori estimation; evidence approximation; asymptotic approximation; variational Bayes; and Markov chain Monte Carlo. We also describe the speech and language processing applications within this categorization so that readers can appropriately choose the approximated Bayesian techniques for their problems.

This book is part of our long-term cooperative efforts to promote the Bayesian approaches in speech and language processing. We have been pursuing this goal for more than ten years, and part of our efforts was to organize a tutorial lecture with this theme at the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in Kyoto, Japan, March 2012. The success of this tutorial lecture prompted the idea of writing a textbook with this theme. We strongly believe in the importance of the Bayesian approaches, and we sincerely encourage the researchers who work with Bayesian speech and language processing.

Acknowledgments

First we want to thank all of our colleagues and research friends, especially members of NTT Communication Science Laboratories, Mitsubishi Electric Research Laboratories (MERL), National Cheng Kung University, IBM T. J. Watson Research Center, and National Chiao Tung University (NCTU). Some of the studies in this book were actually conducted when the authors were working in these institutes. We also would like to thank many people for reading a draft and giving us valuable comments which greatly improved this book, including Tawara Naohiro, Yotaro Kubo, Seong-Jun Hahm, Yu Tsao, and all of the students from the Machine Learning Laboratory at NCTU. We are very grateful for support from Anthony Vetro, John R. Hershey, and Jonathan Le Roux at MERL, and Sin-Horng Chen, Hsueh-Ming Hang, Yu-Chee Tseng, and Li-Chun Wang at NCTU. The great efforts of the editors of Cambridge University Press, Phil Meyler, Sarah Marsh, and Heather Brolly, are also appreciated. Finally, we would like to thank our families for supporting our whole research lives.

Shinji Watanabe
Jen-Tzung Chien

Notation and abbreviations

General notation

This book observes the following general mathematical notation to avoid any confusion arising from notation:

$\mathbb{B} = \{\text{true}, \text{false}\}$

Set of boolean values

$\mathbb{Z}^+ = \{1, 2, \cdots\}$

Set of positive integers

$\mathbb{R}$

Set of real numbers

$\mathbb{R}_{>0}$

Set of positive real numbers

$\mathbb{R}^D$

Set of D dimensional real numbers

Σ^*

Set of all possible strings composed of letters

$\emptyset$

Empty set

a

Scalar variable

$\mathbf{a}$

Vector variable

$\mathbf{a} = \begin{bmatrix} a_1 & \cdots & a_N \end{bmatrix}^\mathsf{T} = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}$

Elements of a vector, which can be described with the square brackets $[\cdots]$, $^\mathsf{T}$ denotes the transpose operation

$\mathbf{A}$

Matrix variable

$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

Elements of a matrix, which can be described with the square brackets $[\cdots]$

$\mathbf{I}_D$

$D \times D$ identity matrix

$|\mathbf{A}|$

Determinant of square matrix

$\mathrm{tr}[\mathbf{A}]$

Trace of square matrix

$A, \mathcal{A}$

Set or sequential variable

$A = \{a_1, \cdots, a_N\} = \{a_n\}_{n=1}^N$

Elements in a set, which can be described with the curly brackets $\{\cdots\}$

$A = \{a_n\}$

Elements in a set, where the range of index n is omitted for simplicity

$a_{n:n'} = \{a_n, \cdots, a_{n'}\} \quad n' > n$

A set of sequential variables, which explicitly describes the range of elements from n to n' by using : in the subscript

$|A|$

The number of elements in a set A. For example $|\{a_n\}_{n=1}^N| = N$

$f(x)$ or f_x

Function of x

$p(x)$ or $q(x)$

Probabilistic distribution function of x

$\mathcal{F}[f]$

Functional of f. Note that a functional uses the square brackets $[\cdot]$ while a function uses the bracket $(\cdot)$.

$\mathbb{E}_{p(x|y)}[f(x)|y] = \int f(x)p(x|y)dx$

The expectation of $f(x)$ with respect to probability distribution $p(x|y)$

$\mathbb{E}_{(x)}[f(x)|y] = \int f(x)p(x|y)dx$ or $\mathbb{E}_{(x)}[f(x)] = \int f(x)p(x|y)dx$

Another form of the expectation of $f(x)$, where the subscript with the probability distribution and/or the conditional variable is omitted, when it is trivial.

$$\delta(a, a') = \begin{cases} 1 & a = a' \\ 0 & \text{Otherwise} \end{cases}$$

Kronecker delta function for discrete variables a and a'

$\delta(x - x')$

Dirac delta function for continuous variables x and x'

$A^{\mathrm{ML}}, A^{\mathrm{ML2}}, A^{\mathrm{MAP}}, A^{\mathrm{DT}}, \cdots$

The variables estimated by a specific criterion (e.g., Maximum Likelihood (ML)) are represented with the superscript of the abbreviation of the criterion.

Basic notation used for speech and language processing

We also list the notation specific for speech and language processing. This book tries to maintain consistency by using the same notation, while it also tries to use commonly used notation in each application. Therefore, some of the same characters are used to denote different variables, since this book needs to introduce many variables.

Common notation

Θ

Set of model parameters

M

Model variable including types of models, structure, hyperparameters, etc.

Ψ

Set of hyperparameters

$Q(\cdot|\cdot)$

Auxiliary function used in the EM algorithm

$\mathbf{H}$

Hessian matrix

Acoustic modeling

$T \in \mathbb{Z}^+$

Number of speech frames

$t \in \{1, \cdots, T\}$

Speech frame index

$\mathbf{o}_t \in \mathbb{R}^D$

D dimensional feature vector at time t

$\mathbf{O} = \{\mathbf{o}_t | t = 1, \cdots, T\}$

Sequence of T feature vectors

$J \in \mathbb{Z}^+$

Number of unique HMM states in an HMM

$s_t \in \{1, \cdots, J\}$

HMM state at time t

$S = \{s_t | t = 1, \cdots, T\}$

Sequence of HMM states for T speech frames

$K \in \mathbb{Z}^+$

Number of unique mixture components in a GMM

$v_t \in \{1, \cdots, K\}$

Latent mixture variable at time t

$V = \{v_t | t = 1, \cdots, T\}$

Sequence of latent mixture variables for T speech frames

$\alpha_t(j) \in [0, 1]$

Forward probability of the partial observations $\{\mathbf{o}_1, \cdots, \mathbf{o}_t\}$ until time t and state j at time t

$\beta_t(j) \in [0, 1]$

Backward probability of the partial observations $\{\mathbf{o}_{t+1}, \cdots, \mathbf{o}_T\}$ from $t + 1$ to the end given state j at time t

$\delta_t(j) \in [0, 1]$

The highest probability along a single path, at time t which accounts for previous observations $\{\mathbf{o}_1, \cdots, \mathbf{o}_t\}$ and ends in state j at time t

$\xi_t(i, j) \in [0, 1]$

Posterior probability of staying state i at time t and state j at time $t + 1$

$\gamma_t(j, k) \in [0, 1]$

Posterior probability of staying at state j and mixture component k at time t

$\pi_j \in [0, 1]$

Initial state probability of state j at time $t = 1$

$a_{ij} \in [0, 1]$

State transition probability from state $s_{t-1} = i$ to state $s_t = j$

$\omega_{jk} \in [0, 1]$

Gaussian mixture weight at component k of state j

$\boldsymbol{\mu}_{jk} \in \mathbb{R}^D$

Gaussian mean vector at component k of state j

$\boldsymbol{\Sigma}_{jk} \in \mathbb{R}^{D \times D}$

Gaussian covariance matrix at component k of state j. Symmetric matrix

$\mathbf{R}_{jk} \in \mathbb{R}^{D \times D}$

Gaussian precision matrix at component k of state j. Symmetric matrix, and the inverse of covariance matrix $\boldsymbol{\Sigma}_{jk}$

Language modeling

$w \in \Sigma^*$

Category (e.g., word in most cases, phoneme sometimes). The element is represented by a string in Σ^* (e.g., "I" and "apple" for words and /a/ and /k/ for phonemes) or a natural number in $\mathbb{Z}^+$ when the elements of categories are numbered.

$\mathcal{V} \subset \Sigma^*$

Vocabulary (dictionary), i.e., a set of distinct words, which is a subset of Σ^*

$|\mathcal{V}|$

Vocabulary size

$v \in \{1, \cdots, |\mathcal{V}|\}$

Ordered index number of distinct words in vocabulary $\mathcal{V}$

$w_{(v)} \in \mathcal{V}$

Word pointed by an ordered index v

$\{w_{(v)} | v = 1, \cdots, |\mathcal{V}|\} = \mathcal{V}$

A set of distinct words, which is equivalent to vocabulary $\mathcal{V}$

$J \in \mathbb{Z}^+$

Number of categories in a chunk (e.g., number of words in a sentence or number of phonemes or HMM states in a speech segment)

$i \in \{1, \cdots, J\}$

ith position of category (e.g., word or phoneme)

$w_i \in \mathcal{V}$

Word at ith position

$W = \{w_i | i = 1, \cdots, J\}$

Word sequence from 1 to J

$w_{i-n+1}^{i} = \{w_{i-n+1} \cdots w_i\}$

Word sequence from $i - n + 1$ to i

$p(w_i | w_{i-n+1}^{i-1}) \in [0, 1]$

n-gram probability, which considers $n - 1$ order Markov model

$c(w_{i-n+1}^{i-1}) \in \mathbb{Z}^+$

Number of occurrences of word sequence w_{i-n+1}^{i-1} in a training corpus

$\lambda_{w_{i-n+1}^{i-1}}$

Interpolation weight for each w_{i-n+1}^{i-1}

$M \in \mathbb{Z}^+$

Number of documents

$m \in \{1, \cdots, M\}$

Document index

d_m

mth document, which would be represented by a string or positive integer

$c(w_{(v)}, d_m) \in \mathbb{Z}^+$

Number of co-occurrences of word $w_{(v)}$ in document d_m

$K \in \mathbb{Z}^+$

Number of unique latent topics

$z_i \in \{1, \cdots, K\}$

ith latent topic variable for word w_i

$Z = \{z_j | j = 1, \cdots, J\}$

Sequence of latent topic variables for J words

Abbreviations

AIC: Akaike Information Criterion (page 217)
AM: Acoustic Model (page 3)
ARD: Automatic Relevance Determination (page 194)
ASR: Automatic Speech Recognition (page 58)
BIC: Bayesian Information Criterion (page 8)
BNP: Bayesian Nonparametrics (pages 337, 345)
BPC: Bayesian Predictive Classification (page 218)
CDHMM: Continuous Density Hidden Markov Model (page 157)
CRP: Chinese Restaurant Process (page 350)
CSR: Continuous Speech Recognition (page 334)
DCLM: Dirichlet Class Language Model (page 326)

DHMM: Discrete Hidden Markov Model (page 62)
DNN: Deep Neural Network (page 224)
DP: Dirichlet Process (page 348)
EM: Expectation Maximization (page 9)
fMLLR: feature-space MLLR (page 204)
GMM: Gaussian Mixture Model (page 63)
HDP: Hierarchical Dirichlet Process (page 337)
HMM: Hidden Markov Model (page 59)
HPY: Hierarchical Pitman–Yor Process (page 383)
HPYLM: Hierarchical Pitman–Yor Language Model (page 384)
iid: Independently, identically distributed (page 216)
KL: Kullback–Leibler (page 79)
KN: Kneser–Ney (page 102)
LDA: Latent Dirichlet Allocation (page 318)
LM: Language Model (page 3)
LSA: Latent Semantic Analysis (page 113)
LVCSR: Large Vocabulary Continuous Speech Recognition (page 97)
MAP: Maximum A-Posteriori (page 7)
MAPLR: Maximum A-Posteriori Linear Regression (page 287)
MBR: Minimum Bayes Risk (page 56)
MCE: Minimum Classification Error (page 59)
MCMC: Markov Chain Monte Carlo (page 337)
MDL: Minimum Description Length (page 9)
MFCC: Mel-Frequency Cepstrum Coefficients (page 249)
MKN: Modified Kneser–Ney (page 111)
ML: Maximum Likelihood (page 77)
ML2: Type-2 Maximum Likelihood (page 188)
MLLR: Maximum Likelihood Linear Regression (page 200)
MLP: MultiLayer Perceptron (page 326)
MMI: Maximum Mutual Information (page 167)
MMSE: Minimum Mean Square Error (page 139)
MPE: Minimum Phone Error (page 167)
nCRP: nested Chinese Restaurant Process (page 360)
NDP: Nested Dirichlet Process (page 360)
NMF: Non-negative Matrix Factorization (page 124)
pdf: probability density function (page 63)
PLP: Perceptual Linear Prediction (page 54)
PLSA: Probabilistic Latent Semantic Analysis (page 113)
PY: Pitman–Yor Process (page 379)
QB: Quasi-Bayes (page 180)
RHS: Right-Hand Side (page 199)
RLS: Regularized Least-Squares (page 188)
RVM: Relevance Vector Machine (page 192)
SBL: Sparse Bayesian Learning (page 194)

SBP: Stick Breaking Process (page 348)
SMAP: Structural Maximum A-Posteriori (page 288)
SMAPLR: Structural Maximum A-Posteriori Linear Regression (page 288)
SVD: Singular Value Decomposition (page 114)
SVM: Support Vector Machine (page 188)
tf–idf: term frequency – inverse document frequency (page 113)
UBM: Universal Background Model (page 172)
VB: Variational Bayes (page 7)
VC: Vapnik–Chervonenkis (page 191)
VQ: Vector Quantization (page 62)
WB: Witten–Bell (page 102)
WER: Word Error Rate (page 56)
WFST: Weighted Finite State Transducer (page 60)
WSJ: Wall Street Journal (page 108)