

Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods

edited by

Joseph Keshet
Samy Bengio

Contents

Preface	11
1 Introduction	1
<i>Samy Bengio¹ and Joseph Keshet²</i>	
1.1 The Traditional Approach to Speech Processing	2
1.2 Potential Problems of the Probabilistic Approach	4
1.3 Support Vector Machines for Binary Classification	5
1.4 Outline	7
References	8
I Foundations	9
2 Theory and Practice of Support Vector Machines Optimization	11
<i>Shai Shalev-Shwartz and Nathan Srebro</i>	
2.1 Introduction	11
2.2 SVM and L_2 -Regularized Linear Prediction	12
2.2.1 Binary Classification and the Traditional SVM	12
2.2.2 More General Loss Functions	13
2.2.3 Examples	13
2.2.4 Kernels	14
2.2.5 Incorporating a Bias Term	15
2.3 Optimization Accuracy from a Machine Learning Perspective	16
2.4 Stochastic Gradient Descent	18
2.4.1 Subgradient calculus	20
2.4.2 Rate of convergence and stopping criteria	21
2.5 Dual Decomposition Methods	22
2.5.1 Duality	23
2.6 Summary	26
References	26
3 From Binary Classification to Categorical Prediction	29
<i>Koby Crammer</i>	
3.1 Multi Category Problems	29
3.2 Hypothesis Class	32

3.3	Loss Functions	33
3.3.1	Combinatorial Loss Functions	35
3.4	Hinge Loss Functions	37
3.5	A Generalized Perceptron Algorithm	38
3.6	A Generalized Passive-Aggressive Algorithm	41
3.6.1	Dual Formulation	42
3.7	A Batch Formulation	43
3.8	Concluding Remarks	45
3.9	Appendix	46
3.9.1	Derivation of the Dual of the Passive-Aggressive Algorithm	46
3.9.2	Derivation of the Dual of the Batch Formulation	49
	References	51

II Acoustic Modeling 53

4 A Large Margin Algorithm for Forced Alignment 55

Joseph Keshet¹, Shai Shalev-Shwartz², Yoram Singer³ and Dan Chazan⁴

4.1	Introduction	56
4.2	Problem Setting	56
4.3	Cost and Risk	57
4.4	A Large Margin Approach for Forced Alignment	58
4.5	An Iterative Algorithm	59
4.6	Efficient Evaluation of the Alignment Function	64
4.7	Base Alignment Functions	66
4.8	Experimental Results	68
4.9	Discussion	69
	References	70

5 A Kernel Wrapper for Phoneme Sequence Recognition 71

Joseph Keshet¹ and Dan Chazan²

5.1	Introduction	71
5.2	Problem Setting	72
5.3	Frame-based Phoneme Classifier	73
5.4	Kernel-based Iterative Algorithm for Phoneme Recognition	73
5.5	Non-Linear Feature Functions	77
5.5.1	Acoustic Modeling	77
5.5.2	Duration Modeling	79
5.5.3	Transition Modeling	80
5.6	Preliminary Experimental Results	80
5.7	Discussion: Can We Hope for Better Results?	81
	References	82

6 Augmented Statistical Models: using Dynamic Kernels for Acoustic Models 85

Mark J.F. Gales

6.1	Introduction	86
6.2	Temporal Correlation Modelling	87

6.3	Dynamic Kernels	89
6.3.1	Static and Dynamic Kernels	89
6.3.2	Generative Kernels	90
6.3.3	Simple Example	92
6.4	Augmented Statistical Models	93
6.4.1	Generative Augmented Models	94
6.4.2	Conditional Augmented Models	96
6.5	Experimental Results	97
6.6	Conclusions	99
	References	100
7	Large Margin Training of Continuous Density Hidden Markov Models	103
	<i>Fei Sha¹ and Lawrence K. Saul²</i>	
7.1	Introduction	104
7.2	Background	105
7.2.1	Maximum likelihood estimation	106
7.2.2	Conditional maximum likelihood	106
7.2.3	Minimum classification error	106
7.3	Large margin training	107
7.3.1	Discriminant function	107
7.3.2	Margin constraints and Hamming distances	108
7.3.3	Optimization	108
7.3.4	Related work	109
7.4	Experimental results	110
7.4.1	Large margin training	110
7.4.2	Comparison to CML and MCE	111
7.4.3	Other variants	111
7.5	Conclusion	114
	References	115
III	Language Modeling	117
8	A Survey of Discriminative Language Modeling Approaches for Large Vocabulary Continuous Speech Recognition	119
	<i>Brian Roark</i>	
8.1	Introduction	120
8.2	General Framework	121
8.2.1	Training Data and the GEN Function	122
8.2.2	Feature Mapping	125
8.2.3	Parameter Estimation	129
8.3	Further Developments	132
8.3.1	Novel Features	132
8.3.2	Novel Objectives	134
8.3.3	Domain Adaptation	134
8.4	Summary and Discussion	135

References	137
9 Large Margin Methods for Part of Speech Tagging	141
<i>Yasemin Altun</i>	
9.1 Introduction	141
9.2 Modeling Sequence Labeling	143
9.2.1 Feature Representation	143
9.2.2 Empirical Risk Minimization	145
9.2.3 Conditional Random Fields and Sequence Perceptron	145
9.3 Sequence Boosting	146
9.3.1 Objective Function	147
9.3.2 Optimization Method	147
9.4 Hidden Markov Support Vector Machines	151
9.4.1 Objective Function	151
9.4.2 Optimization Method	152
9.4.3 Algorithm	153
9.5 Experiments	155
9.5.1 Data and Features for Part of Speech Tagging	155
9.5.2 Results of Sequence AdaBoost	156
9.5.3 Results of HM-SVMs	157
9.6 Discussion	158
References	158
10 A Proposal of a Kernel-Based Algorithm for Large Vocabulary Continuous Speech Recognition	161
<i>Joseph Keshet</i>	
10.1 Introduction	162
10.2 Segmental Models and Hidden Markov Models	163
10.3 Kernel-Based Model	165
10.4 Large Margin Training	166
10.5 Implementations Details	168
10.5.1 Iterative Algorithm	168
10.5.2 Recognition Feature Functions	170
10.5.3 The Decoder	171
10.5.4 Complexity	172
10.6 Discussion	172
References	173
IV Applications	175
11 Discriminative Keyword Spotting	177
<i>David Grangier¹, Joseph Keshet² and Samy Bengio³</i>	
11.1 Introduction	178
11.2 Previous Work	179
11.3 Discriminative Keyword Spotting	182
11.3.1 Problem Setting	182

11.3.2	Loss Function and Model Parameterization	184
11.3.3	An Iterative Training Algorithm	186
11.3.4	Analysis	187
11.4	Experiments and Results	189
11.4.1	The TIMIT Experiments	190
11.4.2	The WSJ Experiments	192
11.5	Conclusions	194
	References	195
12	Kernel Based Text-Independent Speaker Verification	197
	<i>Johnny Mariéthoz¹, Yves Grandvalet¹ and Samy Bengio²</i>	
12.1	Introduction	198
12.2	Generative Approaches	199
12.2.1	Rationale	199
12.2.2	Gaussian Mixture Models	200
12.3	Discriminative Approaches	201
12.3.1	Support Vector Machines	202
12.3.2	Kernels	202
12.4	Benchmarking Methodology	203
12.4.1	Data Splitting for Speaker Verification	203
12.4.2	Performance Measures	204
12.4.3	NIST Data	205
12.4.4	Pre-Processing	205
12.5	Kernels for Speaker Verification	206
12.5.1	Mean Operator Sequence Kernels	206
12.5.2	Fisher Kernels	207
12.5.3	Beyond Fisher Kernels	212
12.6	Parameter Sharing	215
12.6.1	Nuisance Attribute Projection	215
12.6.2	Other Approaches	217
12.7	Is the Margin Useful for this Problem?	218
12.8	Comparing All Methods	219
12.9	Conclusion	221
	References	221
13	Spectral Clustering for Speech Separation	225
	<i>Francis R. Bach¹ and Michael I. Jordan²</i>	
13.1	Introduction	225
13.2	Spectral clustering and normalized cuts	227
13.2.1	Similarity matrices	227
13.2.2	Normalized cuts	228
13.2.3	Spectral relaxation	229
13.2.4	Rounding	230
13.2.5	Spectral clustering algorithms	232
13.2.6	Variational formulation for the normalized cut	233
13.3	Cost functions for learning the similarity matrix	234

13.3.1	Distance between partitions	234
13.3.2	Cost functions as upper bounds	235
13.3.3	Functions of eigensubspaces	235
13.3.4	Empirical comparisons between cost functions	238
13.4	Algorithms for learning the similarity matrix	240
13.4.1	Learning algorithm	240
13.4.2	Related work	240
13.4.3	Testing algorithm	241
13.4.4	Handling very large similarity matrices	241
13.4.5	Simulations on toy examples	243
13.5	Speech separation as spectrogram segmentation	243
13.5.1	Spectrogram	245
13.5.2	Normalization and subsampling	246
13.5.3	Generating training samples	246
13.5.4	Features and grouping cues for speech separation	246
13.6	Spectral clustering for speech separation	248
13.6.1	Basis similarity matrices	248
13.6.2	Combination of similarity matrices	248
13.6.3	Approximations of similarity matrices	248
13.6.4	Experiments	249
13.7	Conclusions	252
	References	253

List of Contributors

Yasemin Altun

Dept. Schölkopf,
Max Planck Institute for Biological Cybernetics
yasemin.altun@tuebingen.mpg.de

Francis Bach

INRIA - Willow project,
Département d'Informatique,
Ecole Normale Supérieure
francis.bach@mines.org

Samy Bengio

Google Research Labs,
Google Inc.
bengio@google.com

Dan Chazan

Dept. of Electrical Engineering,
The Technion Institute of Technology
dan_chazan@yahoo.com

Koby Crammer

Dept. of Computer and Information Science,
University of Pennsylvania
crammer@cis.upenn.edu

Mark Gales

Dept. of Engineering,
University of Cambridge
mjfg@eng.cam.ac.uk

Yves Grandvalet

Heudiasyc,
Université de Technologie de Compiègne
yves.grandvalet@utc.fr

David Grangier

Dept. of Machine Learning,
NEC Laboratories America, Inc.
dgrangier@nec-labs.com

Michael I. Jordan

Computer Science Div. and Dept. of Statistics,
University of California at Berkeley
jordan@eecs.berkeley.edu

Joseph Keshet

Idiap Research Institute,
Martigny, Switzerland
jkeshet@idiap.ch

Johnny Mariéthoz

Idiap Research Institute,
Martigny, Switzerland
marietho@idiap.ch

Lawrence Saul

Dept. of Computer Science and Engineering,
University of California at San Diego
saul@cs.ucsd.edu

Brian Roark

Dept. of Computer Science and Electrical Eng.,
OGI School of Science and Engineering
roark@cslu.ogi.edu

Fei Sha

Computer Science Dept.,
University of Southern California
feisha@usc.edu

Shai Shalev-Shwartz

Toyota Technological Institute at Chicago
shai@tti-c.org

Yoram Singer

Google Research Labs,
Google Inc.
singer@google.com

Nathan Srebro

Toyota Technological Institute at Chicago
nati@uchicago.edu

Foreword

Preface

This is the first book dedicated to uniting research related to speech and speaker recognition based on the recent advances in large margin and kernel methods. The first part of the book presents theoretical and practical foundations of large margin and kernel methods, from support vector machines to large margin methods for structured learning. The second part of the book is dedicated to acoustic modeling of continuous speech recognizers, where the grounds for practical large margin sequence learning are set. The third part introduces large margin methods for discriminative language modeling. The last part of the book is dedicated to the application of keyword spotting, speaker verification and spectral clustering.

The book is an important reference to researchers and practitioners in the field of modern speech and speaker recognition. The purpose of the book is twofold; first, to set the theoretical foundation of large margin and kernel methods relevant to speech recognition domain; second, to propose a practical guide on implementation of these methods to the speech recognition domain. The reader is presumed to have basic knowledge of large margin and kernel methods and of basic algorithms in speech and speaker recognition.

August 2008

Joseph Keshet
Samy Bengio

