# Contents

**Contributors**

**Preface**

1 **Scaling Up Machine Learning: Introduction**
   *Ron Bekkerman, Mikhail Bilenko, and John Langford*
   1.1 Machine Learning Basics 2
   1.2 Reasons for Scaling Up Machine Learning 3
   1.3 Key Concepts in Parallel and Distributed Computing 6
   1.4 Platform Choices and Trade-Offs 7
   1.5 Thinking about Performance 9
   1.6 Organization of the Book 10
   1.7 Bibliographic Notes 17
   References 19

2 **MapReduce and Its Application to Massively Parallel Learning of Decision Tree Ensembles**
   *Biswanath Panda, Joshua S. Herbach, Sugato Basu, and Roberto J. Bayardo*
   2.1 Preliminaries 24
   2.2 Example of PLANET 30
   2.3 Technical Details 33
   2.4 Learning Ensembles 38
   2.5 Engineering Issues 39
   2.6 Experiments 41
   2.7 Related Work 44
   2.8 Conclusions 46
   Acknowledgments 47
   References 47

© in this web service Cambridge University Press  
www.cambridge.org
3 Large-Scale Machine Learning Using DryadLINQ  49
Mihai Budiu, Dennis Fetterly, Michael Isard, Frank McSherry, and Yuan Yu
3.1 Manipulating Datasets with LINQ  49
3.2 \(k\)-Means in LINQ  52
3.3 Running LINQ on a Cluster with DryadLINQ  53
3.4 Lessons Learned  65
References  67

4 IBM Parallel Machine Learning Toolbox  69
Edwin Pednault, Elad Yom-Tov, and Amol Ghoting
4.1 Data-Parallel Associative-Commutative Computation  70
4.2 API and Control Layer  71
4.3 API Extensions for Distributed-State Algorithms  76
4.4 Control Layer Implementation and Optimizations  77
4.5 Parallel Kernel \(k\)-Means  79
4.6 Parallel Decision Tree  80
4.7 Parallel Frequent Pattern Mining  83
4.8 Summary  86
References  87

5 Uniformly Fine-Grained Data-Parallel Computing for Machine Learning Algorithms  89
Meichun Hsu, Ren Wu, and Bin Zhang
5.1 Overview of a GP-GPU  91
5.2 Uniformly Fine-Grained Data-Parallel Computing on a GPU  93
5.3 The \(k\)-Means Clustering Algorithm  97
5.4 The \(k\)-Means Regression Clustering Algorithm  99
5.5 Implementations and Performance Comparisons  102
5.6 Conclusions  105
References  105

Part Two  Supervised and Unsupervised Learning Algorithms

6 PSVM: Parallel Support Vector Machines with Incomplete Cholesky Factorization  109
Edward Y. Chang, Hongjie Bai, Kaihua Zhu, Hao Wang, Jian Li, and Zhihuan Qiu
6.1 Interior Point Method with Incomplete Cholesky Factorization  112
6.2 PSVM Algorithm  114
6.3 Experiments  121
6.4 Conclusion  125
Acknowledgments  125
References  125

7 Massive SVM Parallelization Using Hardware Accelerators  127
Igor Durdanovic, Eric Cosatto, Hans Peter Graf, Srihari Cadambi, Venkata Jakula, Srimat Chakradhar, and Abhinandan Majumdar
7.1 Problem Formulation  128
7.2 Implementation of the SMO Algorithm  131
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3</td>
<td>Micro Parallelization: Related Work</td>
<td>132</td>
</tr>
<tr>
<td>7.4</td>
<td>Previous Parallelizations on Multicore Systems</td>
<td>133</td>
</tr>
<tr>
<td>7.5</td>
<td>Micro Parallelization: Revisited</td>
<td>136</td>
</tr>
<tr>
<td>7.6</td>
<td>Massively Parallel Hardware Accelerator</td>
<td>137</td>
</tr>
<tr>
<td>7.7</td>
<td>Results</td>
<td>145</td>
</tr>
<tr>
<td>7.8</td>
<td>Conclusion</td>
<td>146</td>
</tr>
<tr>
<td>References</td>
<td>146</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Large-Scale Learning to Rank Using Boosted Decision Trees</td>
<td>148</td>
</tr>
<tr>
<td>8.1</td>
<td>Related Work</td>
<td>149</td>
</tr>
<tr>
<td>8.2</td>
<td>LambdaMART</td>
<td>151</td>
</tr>
<tr>
<td>8.3</td>
<td>Approaches to Distributing LambdaMART</td>
<td>153</td>
</tr>
<tr>
<td>8.4</td>
<td>Experiments</td>
<td>158</td>
</tr>
<tr>
<td>8.5</td>
<td>Conclusions and Future Work</td>
<td>168</td>
</tr>
<tr>
<td>8.6</td>
<td>Acknowledgments</td>
<td>169</td>
</tr>
<tr>
<td>References</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>The Transform Regression Algorithm</td>
<td>170</td>
</tr>
<tr>
<td>9.1</td>
<td>Classification, Regression, and Loss Functions</td>
<td>171</td>
</tr>
<tr>
<td>9.2</td>
<td>Background</td>
<td>172</td>
</tr>
<tr>
<td>9.3</td>
<td>Motivation and Algorithm Description</td>
<td>173</td>
</tr>
<tr>
<td>9.4</td>
<td>TReg Expansion: Initialization and Termination</td>
<td>177</td>
</tr>
<tr>
<td>9.5</td>
<td>Model Accuracy Results</td>
<td>184</td>
</tr>
<tr>
<td>9.6</td>
<td>Parallel Performance Results</td>
<td>186</td>
</tr>
<tr>
<td>9.7</td>
<td>Summary</td>
<td>188</td>
</tr>
<tr>
<td>References</td>
<td>189</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Parallel Belief Propagation in Factor Graphs</td>
<td>190</td>
</tr>
<tr>
<td>10.1</td>
<td>Belief Propagation in Factor Graphs</td>
<td>191</td>
</tr>
<tr>
<td>10.2</td>
<td>Shared Memory Parallel Belief Propagation</td>
<td>195</td>
</tr>
<tr>
<td>10.3</td>
<td>Multicore Performance Comparison</td>
<td>209</td>
</tr>
<tr>
<td>10.4</td>
<td>Parallel Belief Propagation on Clusters</td>
<td>210</td>
</tr>
<tr>
<td>10.5</td>
<td>Conclusion</td>
<td>214</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>214</td>
<td></td>
</tr>
<tr>
<td>References</td>
<td>214</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Distributed Gibbs Sampling for Latent Variable Models</td>
<td>217</td>
</tr>
<tr>
<td>11.1</td>
<td>Latent Variable Models</td>
<td>217</td>
</tr>
<tr>
<td>11.2</td>
<td>Distributed Inference Algorithms</td>
<td>220</td>
</tr>
<tr>
<td>11.3</td>
<td>Experimental Analysis of Distributed Topic Modeling</td>
<td>224</td>
</tr>
<tr>
<td>11.4</td>
<td>Practical Guidelines for Implementation</td>
<td>229</td>
</tr>
<tr>
<td>11.5</td>
<td>A Foray into Distributed Inference for Bayesian Networks</td>
<td>231</td>
</tr>
<tr>
<td>11.6</td>
<td>Conclusion</td>
<td>236</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>237</td>
<td></td>
</tr>
<tr>
<td>References</td>
<td>237</td>
<td></td>
</tr>
</tbody>
</table>
## CONTENTS

### 12 Large-Scale Spectral Clustering with MapReduce and MPI

*Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and Edward Y. Chang*

12.1 Spectral Clustering
12.2 Spectral Clustering Using a Sparse Similarity Matrix
12.3 Parallel Spectral Clustering (PSC) Using a Sparse Similarity Matrix
12.4 Experiments
12.5 Conclusions

References

### 13 Parallelizing Information-Theoretic Clustering Methods

*Ron Bekkerman and Martin Scholz*

13.1 Information-Theoretic Clustering
13.2 Parallel Clustering
13.3 Sequential Co-clustering
13.4 The DataLoom Algorithm
13.5 Implementation and Experimentation
13.6 Conclusion

References

### Part Three Alternative Learning Settings

### 14 Parallel Online Learning

*Daniel Hsu, Nikos Karampatziakis, John Langford, and Alex J. Smola*

14.1 Limits Due to Bandwidth and Latency
14.2 Parallelization Strategies
14.3 Delayed Update Analysis
14.4 Parallel Learning Algorithms
14.5 Global Update Rules
14.6 Experiments
14.7 Conclusion

References

### 15 Parallel Graph-Based Semi-Supervised Learning

*Jeff Bilmes and Amarnag Subramanya*

15.1 Scaling SSL to Large Datasets
15.2 Graph-Based SSL
15.3 Dataset: A 120-Million-Node Graph
15.4 Large-Scale Parallel Processing
15.5 Discussion

References

### 16 Distributed Transfer Learning via Cooperative Matrix Factorization

*Evan Xiang, Nathan Liu, and Qiang Yang*

16.1 Distributed Coalitional Learning
16.2 Extension of DisCo to Classification Tasks
CONTENTS

16.3 Conclusion 350
References 350

17 Parallel Large-Scale Feature Selection 352
Jeremy Kubica, Sameer Singh, and Daria Sorokina
17.1 Logistic Regression 353
17.2 Feature Selection 354
17.3 Parallelizing Feature Selection Algorithms 358
17.4 Experimental Results 363
17.5 Conclusions 368
References 368

Part Four Applications

18 Large-Scale Learning for Vision with GPUs 373
Adam Coates, Rajat Raina, and Andrew Y. Ng
18.1 A Standard Pipeline 374
18.2 Introduction to GPUs 377
18.3 A Standard Approach Scaled Up 380
18.4 Feature Learning with Deep Belief Networks 388
18.5 Conclusion 395
References 395

19 Large-Scale FPGA-Based Convolutional Networks 399
Clément Farabet, Yann LeCun, Koray Kavukcuoglu, Berin Martini, Polina Akselrod, Selcuk Talay, and Eugenio Culurciello
19.1 Learning Internal Representations 400
19.2 A Dedicated Digital Hardware Architecture 405
19.3 Summary 416
References 417

20 Mining Tree-Structured Data on Multicore Systems 420
Shirish Tatikonda and Srinivasan Parthasarathy
20.1 The Multicore Challenge 422
20.2 Background 423
20.3 Memory Optimizations 427
20.4 Adaptive Parallelization 431
20.5 Empirical Evaluation 437
20.6 Discussion 442
Acknowledgments 443
References 443

21 Scalable Parallelization of Automatic Speech Recognition 446
Jike Chong, Ekaterina Gonina, Kisun You, and Kurt Keutzer
21.1 Concurrency Identification 450
21.2 Software Architecture and Implementation Challenges 452
21.3 Multicore and Manycore Parallel Platforms 454
21.4 Multicore Infrastructure and Mapping 455
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>21.5 The Manycore Implementation</td>
<td>459</td>
</tr>
<tr>
<td>21.6 Implementation Profiling and Sensitivity Analysis</td>
<td>462</td>
</tr>
<tr>
<td>21.7 Application-Level Optimization</td>
<td>464</td>
</tr>
<tr>
<td>21.8 Conclusion and Key Lessons</td>
<td>467</td>
</tr>
<tr>
<td>References</td>
<td>468</td>
</tr>
</tbody>
</table>

*Subject Index*  
471