Richard M. Golden

Statistical Machine Learning: A Unified Framework
## Contents

Symbols vii  
Algorithm index xv  
Preface xvii  

### I Inference and Learning Machines 1  
1 A Statistical Machine Learning Framework 3  
1.1 Machine Learning Environments 4  
1.1.1 Feature Vectors 4  
1.1.2 Stationary Statistical Environments 6  
1.1.3 Strategies for Teaching Machine Learning Algorithms 7  
1.1.4 Prior Knowledge 8  
1.1.4.1 Feature Representations Dictate Event Similarities 8  
1.1.4.2 Similar Inputs Predict Similar Responses 8  
1.1.4.3 Many Free Parameter Values are Zero 9  
1.1.4.4 Different Feature Detectors Share Parameters 9  
1.2 Empirical Risk Minimization Framework 11  
1.2.1 Objective Functions 11  
1.2.2 Regularization Terms 13  
1.2.3 Optimization Methods 14  
1.3 Theory-Based System Analysis and Design 17  
1.3.1 Stage 1: System Specification 17  
1.3.2 Stage 2: Theoretical Analyses 17  
1.3.3 Stage 3: Physical Implementation 18  
1.3.4 Stage 4: System Behavior Evaluation 18  
1.4 Supervised Learning Machines 21  
1.4.1 Discrepancy Functions 21  
1.4.2 Basis Functions and Hidden Units 24  
1.5 Unsupervised Learning Machines 33  
1.6 Reinforcement Learning Machines 46  
1.6.1 Reinforcement Learning in Stationary Environments 47  
1.6.2 Value Function Reinforcement Learning 52  
1.6.3 Policy Gradient Reinforcement Learning 54  
1.7 Further Readings 59  

2 Set Theory for Concept Modeling 63  
2.1 Set Theory and Logic 65  
2.2 Relations 67  
2.2.1 Types of Relations 67  
2.2.2 Directed Graphs 68
## Contents

2.2.3 Undirected Graphs ................................................. 69
2.3 Functions .................................................................. 71
2.4 Metric Spaces ............................................................. 73
2.5 Further Readings .......................................................... 78

3 Formal Machine Learning Algorithms ............................. 79
3.1 Environment Models .................................................. 79
  3.1.1 Time Models ......................................................... 79
  3.1.2 Event Environments ................................................ 80
3.2 Machine Models .......................................................... 82
  3.2.1 Dynamical Systems ................................................ 82
  3.2.2 Iterated Maps ....................................................... 83
  3.2.3 Vector Fields ....................................................... 86
3.3 Intelligent Machine Models ......................................... 88
3.4 Further Readings ........................................................ 92

II Deterministic Learning Machines .................................. 95

4 Linear Algebra for Machine Learning ............................. 97
4.1 Matrix Notation and Operators .................................... 97
4.2 Linear Subspace Projection Theorems ............................ 105
4.3 Linear System Solution Theorems .................................. 111
4.4 Further Readings ....................................................... 115

5 Vector Calculus for Machine Learning ............................ 117
5.1 Convergence and Continuity ....................................... 117
  5.1.1 Deterministic Convergence ..................................... 117
  5.1.2 Continuous Functions ............................................ 122
5.2 Vector Derivatives ..................................................... 126
  5.2.1 Vector Derivative Definitions .................................. 126
  5.2.2 Theorems for Computing Matrix Derivatives ............... 128
  5.2.3 Backpropagation of Derivatives in Feedforward Networks 130
  5.2.4 Example Derivative Calculations .............................. 132
5.3 Objective Function Analysis ........................................ 139
  5.3.1 Taylor Series Expansions ....................................... 139
  5.3.2 Gradient Descent Type Algorithms ........................... 140
  5.3.3 Critical Point Classification .................................... 143
    5.3.3.1 Identifying Critical Points ............................... 143
    5.3.3.2 Identifying Local Minimizers ............................ 145
    5.3.3.3 Identifying Global Minimizers .......................... 146
  5.3.4 Lagrange Multipliers ............................................ 152
5.4 Further Readings ..................................................... 164

6 Convergence of Time-Invariant Dynamical Systems ............ 167
6.1 Dynamical System Existence Theorems ......................... 168
6.2 Invariant Sets .......................................................... 170
6.3 Lyapunov Convergence Theorems ................................ 173
  6.3.1 Lyapunov Functions .............................................. 173
  6.3.2 Invariant Set Theorems ......................................... 175
    6.3.2.1 Convergence in Finite State Spaces .................... 175
    6.3.2.2 Convergence in Continuous State Spaces ............... 177
Contents

6.4 Further Readings .................................................. 185

7 Batch Learning Algorithm Convergence .......................... 187
  7.1 Search Direction and Stepsize Choices ....................... 188
    7.1.1 Search Direction Selection ............................. 188
    7.1.2 Stepsize Selection ....................................... 189
  7.2 Descent Algorithm Convergence Analysis .................... 196
  7.3 Descent Strategies ............................................ 202
    7.3.1 Gradient and Steepest Descent ......................... 202
    7.3.2 Newton-Type Descent .................................... 204
      7.3.2.1 Newton-Raphson Algorithm ....................... 204
      7.3.2.2 Levenberg-Marquardt Algorithm .................. 205
    7.3.3 L-BFGS and Conjugate Gradient Descent Methods .... 207
  7.4 Further Readings ............................................... 210

III Stochastic Learning Machines .................................. 211

8 Random Vectors and Random Functions ......................... 213
  8.1 Probability Spaces ........................................... 214
    8.1.1 Sigma-Fields ........................................... 214
    8.1.2 Measures ................................................ 215
  8.2 Random Vectors ................................................ 218
    8.2.1 Measurable Functions ................................... 218
    8.2.2 Discrete, Continuous, and Mixed Random Vectors .... 221
  8.3 Existence of the Radon-Nikodým Density (Optional Reading) 225
    8.3.1 Lebesgue Integral ...................................... 225
    8.3.2 The Radon-Nikodým Probability Density Function .... 227
    8.3.3 Vector Support Specification Measures ............... 228
  8.4 Expectation Operations ....................................... 230
    8.4.1 Random Functions ....................................... 233
    8.4.2 Expectations of Random Functions .................... 234
    8.4.3 Conditional Expectation and Independence ........... 235
  8.5 Concentration Inequalities ................................. 237
  8.6 Further Readings ............................................. 239

9 Stochastic Sequences ............................................. 241
  9.1 Types of Stochastic Sequences ................................ 241
  9.2 Missing-Data Stochastic Sequences ......................... 244
  9.3 Stochastic Convergence ....................................... 247
    9.3.1 Convergence With Probability One ................. 248
    9.3.2 Convergence in Mean Square ......................... 251
    9.3.3 Convergence in Probability ......................... 252
    9.3.4 Convergence in Distribution ......................... 252
    9.3.5 Stochastic Convergence Relationships ............. 255
  9.4 Combining and Transforming Stochastic Sequences .......... 257
  9.5 Further Readings ............................................. 259
10 Probability Models of Data Generation 261
  10.1 Learnability of Probability Models ........................................... 261
  10.1.1 Probability Models ......................................................... 261
  10.1.2 Misspecified Probability Models ........................................... 262
  10.1.3 Parametric Probability Models ............................................ 263
  10.1.4 Missing-Data Probability Models ......................................... 266
  10.2 Gibbs Probability Models .................................................... 267
  10.3 Bayesian Networks ............................................................ 273
    10.3.1 Factoring a Chain ........................................................ 274
    10.3.2 Bayesian Network Factorization ....................................... 275
  10.4 Markov Random Fields ....................................................... 282
    10.4.1 The Markov Random Field Concept .................................... 283
    10.4.2 MRF Interpretation of Gibbs Distributions .......................... 286
  10.5 Further Readings ............................................................ 294

11 Monte Carlo Markov Chain Algorithm Convergence 297
  11.1 Monte Carlo Markov Chain (MCMC) Algorithms ............................ 298
    11.1.1 Countably Infinite First-Order Chains on Finite State Spaces ...... 298
    11.1.2 Convergence Analysis of Monte Carlo Markov Chains ................. 301
    11.1.3 Hybrid MCMC Algorithms .............................................. 303
    11.1.4 Finding Global Minimizers and Computing Expectations ............ 305
    11.1.5 Assessing and Improving MCMC Convergence Performance ........... 308
      11.1.5.1 Assessing Convergence When Estimating Expectations .......... 308
      11.1.5.2 Strategies for Addressing Convergence Challenges ............ 309
    11.2 MCMC Metropolis-Hastings (MH) Algorithms ............................. 312
      11.2.1 Metropolis-Hastings Algorithm Definition ........................ 312
      11.2.2 Convergence Analysis of Metropolis-Hastings Algorithms ........ 316
      11.2.3 Important Special Cases of the Metropolis-Hastings Algorithm .. 317
      11.2.4 Machine Learning Applications of MH-MCMC Methods ............. 318
    11.3 Further Readings .......................................................... 322

12 Adaptive Learning Algorithm Convergence 325
  12.1 Stochastic Approximation (SA) Theory ..................................... 326
    12.1.1 Passive versus Reactive Statistical Environments .................. 326
      12.1.1.1 Passive Learning Environments .................................. 326
      12.1.1.2 Reactive Learning Environments ................................ 326
    12.1.2 Average Downward Descent ............................................ 327
    12.1.3 Annealing Schedule .................................................... 328
    12.1.4 The Main Stochastic Approximation Theorem ........................ 329
    12.2 Learning in Passive Statistical Environments using SA ............... 336
      12.2.1 Implementing Different Optimization Strategies .................... 336
      12.2.2 Improving Generalization Performance .............................. 342
    12.3 Learning in Reactive Statistical Environments using SA .............. 348
      12.3.1 Policy Gradient Reinforcement Learning ................................ 348
      12.3.2 Stochastic Approximation Expectation Maximization .............. 350
      12.3.3 Markov Random Field Learning Algorithms .......................... 354
    12.4 Further Readings ........................................................ 357

IV Generalization Performance Evaluation 359
# Contents

13 Statistical Learning Objective Functions 361

13.1 Empirical Risk Function 363

13.2 Maximum Likelihood (ML) Estimation Methods 371

13.2.1 ML Estimation: Probability Theory Interpretation 371

13.2.2 ML Estimation: Information Theory Interpretation 375

13.2.2.1 Entropy: Asymptotic Correctly Specified Model Likelihood 376

13.2.2.2 Cross Entropy Minimization: ML Estimation 378

13.2.3 Pseudolikelihood Empirical Risk Function 382

13.2.4 Missing Data Likelihood Empirical Risk Function 384

13.3 Maximum A Posteriori (MAP) Estimation Methods 387

13.3.1 Parameter Priors and Hyperparameters 387

13.3.2 Maximum A Posteriori (MAP) Risk Function 388

13.3.3 Bayes Risk Interpretation of MAP Estimation 391

13.4 Further Readings 393

14 Simulation Methods for Evaluating Generalization 395

14.1 Sampling Distribution Concepts 398

14.1.1 K-Fold Cross-Validation 398

14.1.2 Sampling Distribution Estimation with Unlimited Data 399

14.2 Bootstrap Methods for Sampling Distribution Simulation 401

14.2.1 Bootstrap Approximation of Sampling Distribution 404

14.2.2 Monte Carlo Bootstrap Sampling Distribution Estimation 404

14.3 Further Readings 411

15 Analytic Formulas for Evaluating Generalization 413

15.1 Assumptions for Asymptotic Analysis 413

15.2 Theoretical Sampling Distribution Analysis 419

15.3 Confidence Regions 428

15.4 Hypothesis Testing for Model Comparison Decisions 433

15.5 Further Readings 438

16 Model Selection and Evaluation 439

16.1 Cross Validation Risk Model Selection Criteria 440

16.2 Bayesian Model Selection Criteria 449

16.2.1 Bayesian Model Selection Problem 449

16.2.2 Laplace Approximation for Multidimensional Integration 450

16.2.3 Generalized Bayesian Information Criterion 452

16.3 Model Misspecification Detection Model Selection Criteria 457

16.3.1 Nested Models Method for Assessing Model Misspecification 457

16.3.2 Information Matrix Discrepancy Model Selection Criteria 457

16.4 Further Readings 462

Bibliography 465

Subject index 467
Preface

Objectives

Statistical Machine Learning is a multidisciplinary field that integrates topics from the fields of Machine learning, Mathematical Statistics, and Numerical Optimization Theory. It is concerned with the problem of the development and evaluation of machines capable of inference and learning within an environment characterized by statistical uncertainty. The recent rapid growth in the variety and complexity of new machine learning architectures requires the development of improved methods for communicating relevant technological tools for supporting statistical machine learning algorithm analysis and design. The main objective of this textbook is to provide students, engineers, and scientists with practical established tools from mathematical statistics and nonlinear optimization theory to support the analysis and design of both existing and new state-of-the-art machine learning algorithms.

It is important to emphasize that this is a mathematics textbook intended for readers interested in a concise mathematically rigorous introduction to the statistical machine learning literature. For readers interested in non-mathematical introductions to the machine learning literature, many alternative options are available. For example, there are many useful software-oriented machine learning textbooks which support the rapid development and evaluation of a wide range of machine learning architectures (Geron, 2017; Muller and Guida, 2017, Bell, 2015, James et al., 2017). A student can use these software tools to rapidly create and evaluate a bewildering wide range of machine learning architectures. After an initial exposure to such tools, the student will want to obtain a deeper understanding of such systems in order to properly apply and properly evaluate such tools. To address this issue, there are now many excellent textbooks (e.g., Hastie el al., 2001; Bishop et al., Stork et al., Ripley et al.; Hastie and Tibshirani, 2016; Goodfellow and Bengio, 2016) which provide detailed discussions of a variety of machine learning architectures and principles by focusing attention on basic principles. Such textbooks specifically omit particular technical mathematical details under the assumption that students without the relevant technical background should not be distracted, while students with graduate level training in optimization theory and mathematical statistics can obtain such details elsewhere.

However, such mathematical technical details are essential for providing a principled methodology for supporting the communication, analysis, and design of novel nonlinear machine learning architectures. Thus, it is desirable to explicitly incorporate such details into self-contained concise discussions of machine learning applications. Technical mathematical details support improved methods for machine learning algorithm specification, validation, classification, and understanding. Such methods can provide important support for rapid machine learning algorithm development and deployment as well as novel insights into reusable modular software design architectures.
Book Overview

A distinguishing feature of this textbook is that a particular empirical risk minimization framework is introduced for the purpose of analyzing both the asymptotic behavior and generalization performance of commonly encountered machine learning algorithms. In particular, a small set of explicit theorems define a useful pedagogical framework for understanding machine learning algorithms. Explicit examples from the machine learning literature are provided to show students how to properly interpret the assumptions and conclusions of such theorems. Machine learning algorithms that do not conform to this unified framework are easily identified as exceptional cases.

Part 1 is concerned with introducing the concept of machine learning algorithms through examples and providing mathematical tools for specifying such algorithms. Chapter 1 informally shows, by example, that the large class of supervised, unsupervised, and reinforcement learning algorithms which are the focus of this textbook may be interpreted as nonlinear optimization algorithms. Chapter 3 provides a formal description of this large class of nonlinear optimization algorithms and shows how optimization may be semantically interpreted within a rational decision making framework.

Part 2 is concerned with characterizing the asymptotic behavior of deterministic learning machines. Chapter 6 provides sufficient conditions for characterizing the asymptotic behavior of discrete-time and continuous-time time-invariant dynamical systems. Chapter 7 provides sufficient conditions for ensuring a large class of deterministic batch learning algorithms converge to the critical points of the objective function for learning.

Part 3 is concerned with characterizing the asymptotic behavior of stochastic inference and stochastic learning machines. Chapter 11 develops the asymptotic convergence theory for Monte Carlo Markov Chains for the special case where the Markov chain is defined on a finite state space. Chapter 12 provides relevant asymptotic convergence analyses of adaptive learning algorithms for both passive and reactive learning environments.

Part 4 is concerned with the problem of characterizing the generalization performance of a machine learning algorithm. Chapter 13 discusses the analysis and design of semantically interpretable objective functions. Chapters 14, 15, and 16 show how both bootstrap simulation methods (Chapter 14) and asymptotic formulas (Chapters 15, 16) can be used to characterize the generalization performance of the class of machine learning algorithms considered here.

In addition, the book includes self-contained relevant introductions to real analysis (Chapter 2, 5), linear algebra (Chapter 4), measure theory (Chapter 8), and stochastic sequences (Chapter 9) to reduce the required mathematical prerequisites for the analyses presented here.

Targeted Audience

The textbook is written for a multidisciplinary audience with multidisciplinary objectives. It is assumed students taking a course based upon this book have taken lower-division coursework in linear algebra and calculus as well as an upper-division calculus-based probability theory course. Students with these mathematical prerequisites will find this textbook challenging but nevertheless accessible.