

Contents

Preface xv

1	Introduction	1
1.1	Data Science: Statistics, Probability, Calculus ... Python (or Perl) and Linux	2
1.2	Informatics and Data Analytics	3
1.3	FSA-Based Signal Acquisition and Bioinformatics	4
1.4	Feature Extraction and Language Analytics	7
1.5	Feature Extraction and Gene Structure Identification	8
1.5.1	HMMs for Analysis of Information Encoding Molecules	11
1.5.2	HMMs for Cheminformatics and Generic Signal Analysis	11
1.6	Theoretical Foundations for Learning	13
1.7	Classification and Clustering	13
1.8	Search	14
1.9	Stochastic Sequential Analysis (SSA) Protocol (Deep Learning Without NNs)	15
1.9.1	Stochastic Carrier Wave (SCW) Analysis – Nanoscope Signal Analysis	18
1.9.2	Nanoscope Cheminformatics – A Case Study for Device “Smartening”	19
1.10	Deep Learning using Neural Nets	20
1.11	Mathematical Specifics and Computational Implementations	21
2	Probabilistic Reasoning and Bioinformatics	23
2.1	Python Shell Scripting	23
2.1.1	Sample Size Complications	33
2.2	Counting, the Enumeration Problem, and Statistics	34
2.3	From Counts to Frequencies to Probabilities	35
2.4	Identifying Emergent/Convergent Statistics and Anomalous Statistics	35
2.5	Statistics, Conditional Probability, and Bayes’ Rule	37
2.5.1	The Calculus of Conditional Probabilities: The Cox Derivation	37

2.5.2	Bayes' Rule	38
2.5.3	Estimation Based on Maximal Conditional Probabilities	38
2.6	Emergent Distributions and Series	39
2.6.1	The Law of Large Numbers (LLN)	39
2.6.2	Distributions	39
2.6.3	Series	42
2.7	Exercises	42
3	Information Entropy and Statistical Measures	47
3.1	Shannon Entropy, Relative Entropy, Maxent, Mutual Information	48
3.1.1	The Khinchin Derivation	49
3.1.2	Maximum Entropy Principle	49
3.1.3	Relative Entropy and Its Uniqueness	51
3.1.4	Mutual Information	51
3.1.5	Information Measures Recap	52
3.2	Codon Discovery from Mutual Information Anomaly	58
3.3	ORF Discovery from Long-Tail Distribution Anomaly	66
3.3.1	<i>Ab initio</i> Learning with smORF's, Holistic Modeling, and Bootstrap Learning	69
3.4	Sequential Processes and Markov Models	72
3.4.1	Markov Chains	73
3.5	Exercises	75
4	<i>Ad Hoc</i>, <i>Ab Initio</i>, and Bootstrap Signal Acquisition Methods	77
4.1	Signal Acquisition, or Scanning, at Linear Order Time-Complexity	77
4.2	Genome Analytics: The Gene-Finder	80
4.3	Objective Performance Evaluation: Sensitivity and Specificity	93
4.4	Signal Analytics: The Time-Domain Finite State Automaton (tFSA)	93
4.4.1	tFSA Spike Detector	95
4.4.2	tFSA-Based Channel Signal Acquisition Methods with Stable Baseline	98
4.4.3	tFSA-Based Channel Signal Acquisition Methods Without Stable Baseline	103
4.5	Signal Statistics (Fast): Mean, Variance, and Boxcar Filter	107
4.5.1	Efficient Implementations for Statistical Tools ($O(L)$)	109
4.6	Signal Spectrum: Nyquist Criterion, Gabor Limit, Power Spectrum	110
4.6.1	Nyquist Sampling Theorem	110
4.6.2	Fourier Transforms, and Other Classic Transforms	110
4.6.3	Power Spectral Density	111
4.6.4	Power-Spectrum-Based Feature Extraction	111
4.6.5	Cross-Power Spectral Density	112
4.6.6	AM/FM/PM Communications Protocol	112
4.7	Exercises	112

5	Text Analytics	125
5.1	Words	125
5.1.1	Text Acquisition: Text Scraping and Associative Memory	125
5.1.2	Word Frequency Analysis: Machiavelli's Polysemy on <i>Fortuna</i> and <i>Virtu</i>	130
5.1.3	Word Frequency Analysis: Coleridge's Hidden Polysemy on <i>Logos</i>	139
5.1.4	Sentiment Analysis	143
5.2	Phrases – Short (Three Words)	145
5.2.1	Shakespearean Insult Generation – Phrase Generation	147
5.3	Phrases – Long (A Line or Sentence)	150
5.3.1	Iambic Phrase Analysis: Shakespeare	150
5.3.2	Natural Language Processing	152
5.3.3	Sentence and Story Generation: Tarot	152
5.4	Exercises	153
6	Analysis of Sequential Data Using HMMs	155
6.1	Hidden Markov Models (HMMs)	155
6.1.1	Background and Role in Stochastic Sequential Analysis (SSA)	155
6.1.2	When to Use a Hidden Markov Model (HMM)?	160
6.1.3	Hidden Markov Models (HMMs) – Standard Formulation and Terms	161
6.2	Graphical Models for Markov Models and Hidden Markov Models	162
6.2.1	Hidden Markov Models	162
6.2.2	Viterbi Path	163
6.2.3	Forward and Backward Probabilities	164
6.2.4	HMM: Maximum Likelihood discrimination	165
6.2.5	Expectation/Maximization (Baum–Welch)	166
6.3	Standard HMM Weaknesses and their GHMM Fixes	168
6.4	Generalized HMMs (GHMMs – “Gems”): Minor Viterbi Variants	171
6.4.1	The Generic HMM	171
6.4.2	pMM/SVM	171
6.4.3	EM and Feature Extraction via EVA Projection	172
6.4.4	Feature Extraction via Data Absorption (a.k.a. Emission Inversion)	174
6.4.5	Modified AdaBoost for Feature Selection and Data Fusion	176
6.5	HMM Implementation for Viterbi (in C and Perl)	179
6.6	Exercises	206
7	Generalized HMMs (GHMMs): Major Viterbi Variants	207
7.1	GHMMs: Maximal Clique for Viterbi and Baum–Welch	207
7.2	GHMMs: Full Duration Model	216
7.2.1	HMM with Duration (HMMD)	216
7.2.2	Hidden Semi-Markov Models (HSMM) with sid-information	220
7.2.3	HMM with Binned Duration (HMMBD)	224
7.3	GHMMs: Linear Memory Baum–Welch Algorithm	228
7.4	GHMMs: Distributable Viterbi and Baum–Welch Algorithms	230

7.4.1	Distributed HMM processing via “Viterbi-overlap-chunking” with GPU speedup	230
7.4.2	Relative Entropy and Viterbi Scoring	231
7.5	Martingales and the Feasibility of Statistical Learning (further details in Appendix)	232
7.6	Exercises	234
8	Neuromanifolds and the Uniqueness of Relative Entropy	235
8.1	Overview	235
8.2	Review of Differential Geometry	236
8.2.1	Differential Topology – Natural Manifold	236
8.2.2	Differential Geometry – Natural Geometric Structures	240
8.3	Amari’s Dually Flat Formulation	243
8.3.1	Generalization of Pythagorean Theorem	246
8.3.2	Projection Theorem and Relation Between Divergence and Link Formalism	246
8.4	Neuromanifolds	247
8.5	Exercises	250
9	Neural Net Learning and Loss Bounds Analysis	253
9.1	Brief Introduction to Neural Nets (NNs)	254
9.1.1	Single Neuron Discriminator	254
9.1.2	Neural Net with Back-Propagation	258
9.2	Variational Learning Formalism and Use in Loss Bounds Analysis	261
9.2.1	Variational Basis for Update Rule	261
9.2.2	Review and Generalization of GD Loss Bounds Analysis	262
9.2.3	Review of the EG Loss Bounds Analysis	266
9.3	The “ $\sinh^{-1}(\omega)$ ” link algorithm (SA)	266
9.3.1	Motivation for “ $\sinh^{-1}(\omega)$ ” link algorithm (SA)	266
9.3.2	Relation of sinh Link Algorithm to the Binary Exponentiated Gradient Algorithm	268
9.4	The Loss Bounds Analysis for $\sinh^{-1}(\omega)$	269
9.4.1	Loss Bounds Analysis Using the Taylor Series Approach	273
9.4.2	Loss Bounds Analysis Using Taylor Series for the sinh Link (SA) Algorithm	275
9.5	Exercises	277
10	Classification and Clustering	279
10.1	The SVM Classifier – An Overview	281
10.2	Introduction to Classification and Clustering	282
10.2.1	Sum of Squared Error (SSE) Scoring	286
10.2.2	K-Means Clustering (Unsupervised Learning)	286
10.2.3	k-Nearest Neighbors Classification (Supervised Learning)	292
10.2.4	The Perceptron Recap (See Chapter 9 for Details)	295

10.3	Lagrangian Optimization and Structural Risk Minimization (SRM)	296
10.3.1	Decision Boundary and SRM Construction Using Lagrangian	296
10.3.2	The Theory of Classification	301
10.3.3	The Mathematics of the Feasibility of Learning	303
10.3.4	Lagrangian Optimization	306
10.3.5	The Support Vector Machine (SVM) – Lagrangian with SRM	308
10.3.6	Kernel Construction Using Polarization	310
10.3.7	SVM Binary Classifier Derivation	312
10.4	SVM Binary Classifier Implementation	318
10.4.1	Sequential Minimal Optimization (SMO)	318
10.4.2	Alpha-Selection Variants	320
10.4.3	Chunking on Large Datasets: $O(N^2) \rightarrow n O(N^2/n^2) = O(N^2)/n$	320
10.4.4	Support Vector Reduction (SVR)	331
10.4.5	Code Examples (in OO Perl)	335
10.5	Kernel Selection and Tuning Metaheuristics	346
10.5.1	The “Stability” Kernels	346
10.5.2	Derivation of “Stability” Kernels	349
10.5.3	Entropic and Gaussian Kernels Relate to Unique, Minimally Structured, Information Divergence and Geometric Distance Measures	351
10.5.4	Automated Kernel Selection and Tuning	353
10.6	SVM Multiclass from Decision Tree with SVM Binary Classifiers	356
10.7	SVM Multiclass Classifier Derivation (Multiple Decision Surface)	359
10.7.1	Decomposition Method to Solve the Dual	361
10.7.2	SVM Speedup via Differentiating BSVs and SVs	362
10.8	SVM Clustering	364
10.8.1	SVM-External Clustering	365
10.8.2	Single-Convergence SVM-Clustering: Comparative Analysis	368
10.8.3	Stabilized, Single-Convergence Initialized, SVM-External Clustering	375
10.8.4	Stabilized, Multiple-Convergence, SVM-External Clustering	379
10.8.5	SVM-External Clustering – Algorithmic Variants	381
10.9	Exercises	385
11	Search Metaheuristics	389
11.1	Trajectory-Based Search Metaheuristics	389
11.1.1	Optimal-Fitness Configuration Trajectories – Fitness Function Known and Sufficiently Regular	390
11.1.2	Optimal-Fitness Configuration Trajectories – Fitness <i>Function</i> not Known	392
11.1.3	Fitness Configuration Trajectories with Nonoptimal Updates	397
11.2	Population-Based Search Metaheuristics	399

11.2.1	Population with Evolution	400
11.2.2	Population with Group Interaction – Swarm Intelligence	402
11.2.3	Population with Indirect Interaction via Artifact	403
11.3	Exercises	404
12	Stochastic Sequential Analysis (SSA)	407
12.1	HMM and FSA-Based Methods for Signal Acquisition and Feature Extraction	408
12.2	The Stochastic Sequential Analysis (SSA) Protocol	410
12.2.1	(Stage 1) Primitive Feature Identification	415
12.2.2	(Stage 2) Feature Identification and Feature Selection	416
12.2.3	(Stage 3) Classification	418
12.2.4	(Stage 4) Clustering	418
12.2.5	(All Stages) Database/Data-Warehouse System Specification	419
12.2.6	(All Stages) Server-Based Data Analysis System Specification	420
12.3	Channel Current Cheminformatics (CCC) Implementation of the Stochastic Sequential Analysis (SSA) Protocol	420
12.4	SCW for Detector Sensitivity Boosting	423
12.4.1	NTD with Multiple Channels (or High Noise)	424
12.4.2	Stochastic Carrier Wave	426
12.5	SSA for Deep Learning	430
12.6	Exercises	431
13	Deep Learning Tools – TensorFlow	433
13.1	Neural Nets Review	433
13.1.1	Summary of Single Neuron Discriminator	433
13.1.2	Summary of Neural Net Discriminator and Back-Propagation	433
13.2	TensorFlow from Google	435
13.2.1	Installation/Setup	436
13.2.2	Example: Character Recognition	437
13.2.3	Example: Language Translation	440
13.2.4	TensorBoard and the TensorFlow Profiler	441
13.2.5	Tensor Cores	444
13.3	Exercises	444
14	Nanopore Detection – A Case Study	445
14.1	Standard Apparatus	447
14.1.1	Standard Operational and Physiological Buffer Conditions	448
14.1.2	α -Hemolysin Channel Stability – Introduction of Chaotropes	448
14.2	Controlling Nanopore Noise Sources and Choice of Aperture	449
14.3	Length Resolution of Individual DNA Hairpins	451
14.4	Detection of Single Nucleotide Differences (Large Changes in Structure)	454
14.5	Blockade Mechanism for 9bphp	455

14.6	Conformational Kinetics on Model Biomolecules	459
14.7	Channel Current Cheminformatics	460
14.7.1	Power Spectra and Standard EE Signal Analysis	460
14.7.2	Channel Current Cheminformatics for Single-Biomolecule/Mixture Identifications	462
14.7.3	Channel Current Cheminformatics: Feature Extraction by HMM	464
14.7.4	Bandwidth Limitations	465
14.8	Channel-Based Detection Mechanisms	467
14.8.1	Partitioning and Translocation-Based ND Biosensing Methods	467
14.8.2	Transduction Versus Translation	468
14.8.3	Single-Molecule Versus Ensemble	469
14.8.4	Biosensing with High Sensitivity in Presence of Interference	470
14.8.5	Nanopore Transduction Detection Methods	471
14.9	The NTD Nanoscope	474
14.9.1	Nanopore Transduction Detection (NTD)	475
14.9.2	NTD: A Versatile Platform for Biosensing	479
14.9.3	NTD Platform	481
14.9.4	NTD Operation	484
14.9.5	Driven Modulations	487
14.9.6	Driven Modulations with Multichannel Augmentation	490
14.10	NTD Biosensing Methods	495
14.10.1	Model Biosensor Based on Streptavidin and Biotin	495
14.10.2	Model System Based on DNA Annealing	501
14.10.3	Y-Aptamer with Use of Chaotropes to Improve Signal Resolution	506
14.10.4	Pathogen Detection, miRNA Detection, and miRNA Haplotyping	508
14.10.5	SNP Detection	510
14.10.6	Aptamer-Based Detection	512
14.10.7	Antibody-Based Detection	512
14.11	Exercises	516

Appendix A: Python and Perl System Programming in Linux 519

A.1	Getting Linux and Python in a Flash (Drive)	519
A.2	Linux and the Command Shell	520
A.3	Perl Review: I/O, Primitives, String Handling, Regex	521

Appendix B: Physics 529

B.1	The Calculus of Variations	529
-----	----------------------------	-----

Appendix C: Math 531

C.1	Martingales	531
C.2	Hoeffding Inequality	537

References 541

Index 559