

CONTENTS

3.3.1	Crude Monte Carlo	191
3.3.2	Bootstrap Method	191
3.3.3	Variance Reduction	191
3.4	Monte Carlo for Optimization	202
3.4.1	Simulated Annealing	207
3.4.2	Cross-Entropy Method	215
3.4.3	Splitting for Optimization	215
3.4.4	Noisy Optimization	215
	Exercises	216
	Reproducing Kernel Hilbert Spaces	xiii
	Construction of Reproducing Kernels	44
	Reproducing Kernels via Feature Mapping	xvii
4.1	Introduction	225
4.2	Risk and Loss in Unsupervised Learning	225
4.3	Expectation-Maximization (EM) Algorithm	225
4.4	Empirical Distribution and Density Estimation	225
4.5	Clustering via Mixture Models	225
4.5.1	Mixture Models	225
4.5.2	EM Algorithm for Mixture Models	225
4.6	Clustering via Vector Quantization	225
4.6.1	k-Means	225
4.6.2	Clustering via Centroids	225
4.7	Hierarchical Clustering	225
4.8	Principal Component Analysis (PCA)	225
4.8.1	Motivation: Principal Axes of an Ellipsoid	225
4.8.2	PCA and Singular Value Decomposition (SVD)	225
	Exercises	225
1	Importing, Summarizing, and Visualizing Data	1
1.1	Introduction	1
1.2	Structuring Features According to Type	3
1.3	Summary Tables	6
1.4	Summary Statistics	7
1.5	Visualizing Data	8
1.5.1	Plotting Qualitative Variables	9
1.5.2	Plotting Quantitative Variables	9
1.5.3	Data Visualization in a Bivariate Setting	12
	Exercises	15
2	Statistical Learning	19
2.1	Introduction	19
2.2	Supervised and Unsupervised Learning	20
2.3	Training and Test Loss	23
2.4	Tradeoffs in Statistical Learning	31
2.5	Estimating Risk	35
2.5.1	In-Sample Risk	35
2.5.2	Cross-Validation	37
2.6	Modeling Data	40
2.7	Multivariate Normal Models	44
2.8	Normal Linear Models	46
2.9	Bayesian Learning	47
	Exercises	58
3	Monte Carlo Methods	67
3.1	Introduction	67
3.2	Monte Carlo Sampling	68
3.2.1	Generating Random Numbers	68
3.2.2	Simulating Random Variables	69
3.2.3	Simulating Random Vectors and Processes	74
3.2.4	Resampling	76
3.2.5	Markov Chain Monte Carlo	78
3.3	Monte Carlo Estimation	85

3.3.1	Crude Monte Carlo	85
3.3.2	Bootstrap Method	88
3.3.3	Variance Reduction	92
3.4	Monte Carlo for Optimization	96
3.4.1	Simulated Annealing	96
3.4.2	Cross-Entropy Method	100
3.4.3	Splitting for Optimization	103
3.4.4	Noisy Optimization	105
	Exercises	113
4	Unsupervised Learning	121
4.1	Introduction	121
4.2	Risk and Loss in Unsupervised Learning	122
4.3	Expectation–Maximization (EM) Algorithm	128
4.4	Empirical Distribution and Density Estimation	131
4.5	Clustering via Mixture Models	135
4.5.1	Mixture Models	135
4.5.2	EM Algorithm for Mixture Models	137
4.6	Clustering via Vector Quantization	142
4.6.1	K-Means	144
4.6.2	Clustering via Continuous Multiextremal Optimization	146
4.7	Hierarchical Clustering	147
4.8	Principal Component Analysis (PCA)	153
4.8.1	Motivation: Principal Axes of an Ellipsoid	153
4.8.2	PCA and Singular Value Decomposition (SVD)	155
	Exercises	160
5	Regression	167
5.1	Introduction	167
5.2	Linear Regression	169
5.3	Analysis via Linear Models	171
5.3.1	Parameter Estimation	171
5.3.2	Model Selection and Prediction	172
5.3.3	Cross-Validation and Predictive Residual Sum of Squares	173
5.3.4	In-Sample Risk and Akaike Information Criterion	175
5.3.5	Categorical Features	177
5.3.6	Nested Models	180
5.3.7	Coefficient of Determination	181
5.4	Inference for Normal Linear Models	182
5.4.1	Comparing Two Normal Linear Models	183
5.4.2	Confidence and Prediction Intervals	186
5.5	Nonlinear Regression Models	188
5.6	Linear Models in Python	191
5.6.1	Modeling	191
5.6.2	Analysis	193
5.6.3	Analysis of Variance (ANOVA)	195

5.6.4	Confidence and Prediction Intervals	198
5.6.5	Model Validation	198
5.6.6	Variable Selection	199
5.7	Generalized Linear Models	204
	Exercises	207
6	Regularization and Kernel Methods	215
6.1	Introduction	215
6.2	Regularization	216
6.3	Reproducing Kernel Hilbert Spaces	222
6.4	Construction of Reproducing Kernels	224
6.4.1	Reproducing Kernels via Feature Mapping	224
6.4.2	Kernels from Characteristic Functions	225
6.4.3	Reproducing Kernels Using Orthonormal Features	227
6.4.4	Kernels from Kernels	229
6.5	Representer Theorem	230
6.6	Smoothing Cubic Splines	235
6.7	Gaussian Process Regression	238
6.8	Kernel PCA	242
	Exercises	245
7	Classification	251
7.1	Introduction	251
7.2	Classification Metrics	253
7.3	Classification via Bayes' Rule	257
7.4	Linear and Quadratic Discriminant Analysis	259
7.5	Logistic Regression and Softmax Classification	266
7.6	K -Nearest Neighbors Classification	268
7.7	Support Vector Machine	269
7.8	Classification with Scikit-Learn	277
	Exercises	279
8	Decision Trees and Ensemble Methods	287
8.1	Introduction	287
8.2	Top-Down Construction of Decision Trees	289
8.2.1	Regional Prediction Functions	290
8.2.2	Splitting Rules	291
8.2.3	Termination Criterion	292
8.2.4	Basic Implementation	294
8.3	Additional Considerations	298
8.3.1	Binary Versus Non-Binary Trees	298
8.3.2	Data Preprocessing	298
8.3.3	Alternative Splitting Rules	298
8.3.4	Categorical Variables	299
8.3.5	Missing Values	299
8.4	Controlling the Tree Shape	300
8.4.1	Cost-Complexity Pruning	303

8.4.2	Advantages and Limitations of Decision Trees	304
8.5	Bootstrap Aggregation	305
8.6	Random Forests	309
8.7	Boosting	313
	Exercises	321
9	Deep Learning	323
9.1	Introduction	323
9.2	Feed-Forward Neural Networks	326
9.3	Back-Propagation	330
9.4	Methods for Training	334
9.4.1	Steepest Descent	334
9.4.2	Levenberg–Marquardt Method	335
9.4.3	Limited-Memory BFGS Method	336
9.4.4	Adaptive Gradient Methods	338
9.5	Examples in Python	340
9.5.1	Simple Polynomial Regression	340
9.5.2	Image Classification	344
	Exercises	349
A	Linear Algebra and Functional Analysis	355
A.1	Vector Spaces, Bases, and Matrices	355
A.2	Inner Product	360
A.3	Complex Vectors and Matrices	361
A.4	Orthogonal Projections	362
A.5	Eigenvalues and Eigenvectors	363
A.5.1	Left- and Right-Eigenvectors	364
A.6	Matrix Decompositions	368
A.6.1	(P)LU Decomposition	368
A.6.2	Woodbury Identity	370
A.6.3	Cholesky Decomposition	373
A.6.4	QR Decomposition and the Gram–Schmidt Procedure	375
A.6.5	Singular Value Decomposition	376
A.6.6	Solving Structured Matrix Equations	379
A.7	Functional Analysis	384
A.8	Fourier Transforms	390
A.8.1	Discrete Fourier Transform	392
A.8.2	Fast Fourier Transform	394
B	Multivariate Differentiation and Optimization	397
B.1	Multivariate Differentiation	397
B.1.1	Taylor Expansion	400
B.1.2	Chain Rule	400
B.2	Optimization Theory	402
B.2.1	Convexity and Optimization	403
B.2.2	Lagrangian Method	406
B.2.3	Duality	407

B.3	Numerical Root-Finding and Minimization	408
B.3.1	Newton-Like Methods	409
B.3.2	Quasi-Newton Methods	411
B.3.3	Normal Approximation Method	413
B.3.4	Nonlinear Least Squares	414
B.4	Constrained Minimization via Penalty Functions	415
C	Probability and Statistics	421
C.1	Random Experiments and Probability Spaces	421
C.2	Random Variables and Probability Distributions	422
C.3	Expectation	426
C.4	Joint Distributions	427
C.5	Conditioning and Independence	428
C.5.1	Conditional Probability	428
C.5.2	Independence	428
C.5.3	Expectation and Covariance	429
C.5.4	Conditional Density and Conditional Expectation	430
C.6	Functions of Random Variables	431
C.7	Multivariate Normal Distribution	434
C.8	Convergence of Random Variables	439
C.9	Law of Large Numbers and Central Limit Theorem	445
C.10	Markov Chains	451
C.11	Statistics	453
C.12	Estimation	454
C.12.1	Method of Moments	455
C.12.2	Maximum Likelihood Method	456
C.13	Confidence Intervals	457
C.14	Hypothesis Testing	458
D	Python Primer	463
D.1	Getting Started	463
D.2	Python Objects	465
D.3	Types and Operators	466
D.4	Functions and Methods	468
D.5	Modules	469
D.6	Flow Control	471
D.7	Iteration	472
D.8	Classes	473
D.9	Files	475
D.10	NumPy	478
D.10.1	Creating and Shaping Arrays	478
D.10.2	Slicing	480
D.10.3	Array Operations	480
D.10.4	Random Numbers	482
D.11	Matplotlib	483
D.11.1	Creating a Basic Plot	483

D.12 Pandas	485
D.12.1 Series and DataFrame	485
D.12.2 Manipulating Data Frames	487
D.12.3 Extracting Information	488
D.12.4 Plotting	490
D.13 Scikit-learn	490
D.13.1 Partitioning the Data	491
D.13.2 Standardization	491
D.13.3 Fitting and Prediction	492
D.13.4 Testing the Model	492
D.14 System Calls, URL Access, and Speed-Up	493
Bibliography	495
Index	503