

Contents

Preface	vii
1 Introduction and Preview	1
1.1 Multivariate Analysis	1
1.2 Data Mining	3
1.2.1 From EDA to Data Mining	3
1.2.2 What Is Data Mining?	5
1.2.3 Knowledge Discovery	8
1.3 Machine Learning	9
1.3.1 How Does a Machine Learn?	9
1.3.2 Prediction Accuracy	10
1.3.3 Generalization	11
1.3.4 Generalization Error	12
1.3.5 Overfitting	13
1.4 Overview of Chapters	14
Bibliographical Notes	16
2 Data and Databases	17
2.1 Introduction	17

2.2	Examples	18
2.2.1	Example: DNA Microarray Data	18
2.2.2	Example: Mixtures of Polyaromatic Hydrocarbons	19
2.2.3	Example: Face Recognition	22
2.3	Databases	24
2.3.1	Data Types	25
2.3.2	Trends in Data Storage	26
2.3.3	Databases on the Internet	27
2.4	Database Management	29
2.4.1	Elements of Database Systems	29
2.4.2	Structured Query Language (SQL)	30
2.4.3	OLTP Databases	32
2.4.4	Integrating Distributed Databases	32
2.4.5	Data Warehousing	33
2.4.6	Decision Support Systems and OLAP	35
2.4.7	Statistical Packages and DBMSs	36
2.5	Data Quality Problems	36
2.5.1	Data Inconsistencies	37
2.5.2	Outliers	38
2.5.3	Missing Data	39
2.5.4	More Variables than Observations	40
2.6	The Curse of Dimensionality	41
	Bibliographical Notes	42
	Exercises	43
3	Random Vectors and Matrices	45
3.1	Introduction	45
3.2	Vectors and Matrices	45
3.2.1	Notation	45
3.2.2	Basic Matrix Operations	46
3.2.3	Vectoring and Kronecker Products	47
3.2.4	Eigenanalysis for Square Matrices	48
3.2.5	Functions of Matrices	49
3.2.6	Singular-Value Decomposition	50
3.2.7	Generalized Inverses	50
3.2.8	Matrix Norms	51

3.2.9	Condition Numbers for Matrices	52
3.2.10	Eigenvalue Inequalities	52
3.2.11	Matrix Calculus	53
3.3	Random Vectors	56
3.3.1	Multivariate Moments	57
3.3.2	Multivariate Gaussian Distribution	59
3.3.3	Conditional Gaussian Distributions	61
3.4	Random Matrices	62
3.4.1	Wishart Distribution	63
3.5	Maximum Likelihood Estimation for the Gaussian	65
3.5.1	Joint Distribution of Sample Mean and Sample Covariance Matrix	67
3.5.2	Admissibility	68
3.5.3	James–Stein Estimator of the Mean Vector	69
	Bibliographical Notes	72
	Exercises	72
4	Nonparametric Density Estimation	75
4.1	Introduction	75
4.1.1	Example: Coronary Heart Disease	76
4.2	Statistical Properties of Density Estimators	77
4.2.1	Unbiasedness	77
4.2.2	Consistency	78
4.2.3	Bona Fide Density Estimators	79
4.3	The Histogram	80
4.3.1	The Histogram as an ML Estimator	81
4.3.2	Asymptotics	82
4.3.3	Estimating Bin Width	84
4.3.4	Multivariate Histograms	85
4.4	Maximum Penalized Likelihood	87
4.5	Kernel Density Estimation	88
4.5.1	Choice of Kernel	89
4.5.2	Asymptotics	91
4.5.3	Example: 1872 Hidalgo Postage Stamps of Mexico	93
4.5.4	Estimating the Window Width	95
4.6	Projection Pursuit Density Estimation	100

4.6.1	The PPDE Paradigm	100
4.6.2	Projection Indexes	102
4.7	Assessing Multimodality	103
	Bibliographical Notes	103
	Exercises	104
5	Model Assessment and Selection in Multiple Regression	107
5.1	Introduction	107
5.2	The Regression Function and Least Squares	108
5.2.1	Random- X Case	109
5.2.2	Fixed- X Case	111
5.2.3	Example: Bodyfat Data	116
5.3	Prediction Accuracy and Model Assessment	117
5.3.1	Random- X Case	119
5.3.2	Fixed- X Case	119
5.4	Estimating Prediction Error	120
5.4.1	Apparent Error Rate	120
5.4.2	Cross-Validation	121
5.4.3	Bootstrap	122
5.5	Instability of LS Estimates	127
5.6	Biased Regression Methods	129
5.6.1	Example: PET Yarns and NIR Spectra	129
5.6.2	Principal Components Regression	131
5.6.3	Partial Least-Squares Regression	133
5.6.4	Ridge Regression	136
5.7	Variable Selection	142
5.7.1	Stepwise Methods	144
5.7.2	All Possible Subsets	146
5.7.3	Criticisms of Variable Selection Methods	147
5.8	Regularized Regression	148
5.9	Least-Angle Regression	152
5.9.1	The Forwards-Stagewise Algorithm	152
5.9.2	The LARS Algorithm	153
	Bibliographical Notes	154
	Exercises	155

6	Multivariate Regression	159
6.1	Introduction	159
6.2	The Fixed- X Case	160
6.2.1	Classical Multivariate Regression Model	161
6.2.2	Example: Norwegian Paper Quality	166
6.2.3	Separate and Multivariate Ridge Regressions	167
6.2.4	Linear Constraints on the Regression Coefficients	168
6.3	The Random- X Case	175
6.3.1	Classical Multivariate Regression Model	175
6.3.2	Multivariate Reduced-Rank Regression	176
6.3.3	Example: Chemical Composition of Tobacco	183
6.3.4	Assessing the Effective Dimensionality	185
6.3.5	Example: Mixtures of Polyaromatic Hydrocarbons	188
6.4	Software Packages	189
	Bibliographical Notes	189
	Exercises	191
7	Linear Dimensionality Reduction	195
7.1	Introduction	195
7.2	Principal Component Analysis	196
7.2.1	Example: The Nutritional Value of Food	196
7.2.2	Population Principal Components	199
7.2.3	Least-Squares Optimality of PCA	199
7.2.4	PCA as a Variance-Maximization Technique	202
7.2.5	Sample Principal Components	203
7.2.6	How Many Principal Components to Retain?	205
7.2.7	Graphical Displays	209
7.2.8	Example: Face Recognition Using Eigenfaces	209
7.2.9	Invariance and Scaling	210
7.2.10	Example: Pen-Based Handwritten Digit Recognition	211
7.2.11	Functional PCA	212
7.2.12	What Can Be Gained from Using PCA?	215
7.3	Canonical Variate and Correlation Analysis	215
7.3.1	Canonical Variates and Canonical Correlations	215
7.3.2	Example: COMBO-17 Galaxy Photometric Catalogue	216

7.3.3	Least-Squares Optimality of CVA	219
7.3.4	Relationship of CVA to RRR	222
7.3.5	CVA as a Correlation-Maximization Technique	223
7.3.6	Sample Estimates	226
7.3.7	Invariance	227
7.3.8	How Many Pairs of Canonical Variates to Retain?	228
7.4	Projection Pursuit	228
7.4.1	Projection Indexes	229
7.4.2	Optimizing the Projection Index	232
7.5	Visualizing Projections Using Dynamic Graphics	232
7.6	Software Packages	233
	Bibliographical Notes	233
	Exercises	234
8	Linear Discriminant Analysis	237
8.1	Introduction	237
8.1.1	Example: Wisconsin Diagnostic Breast Cancer Data	238
8.2	Classes and Features	240
8.3	Binary Classification	241
8.3.1	Bayes's Rule Classifier	241
8.3.2	Gaussian Linear Discriminant Analysis	242
8.3.3	LDA via Multiple Regression	247
8.3.4	Variable Selection	249
8.3.5	Logistic Discrimination	250
8.3.6	Gaussian LDA or Logistic Discrimination?	256
8.3.7	Quadratic Discriminant Analysis	257
8.4	Examples of Binary Misclassification Rates	258
8.5	Multiclass LDA	260
8.5.1	Bayes's Rule Classifier	261
8.5.2	Multiclass Logistic Discrimination	265
8.5.3	LDA via Reduced-Rank Regression	266
8.6	Example: Gilgaied Soil	271
8.7	Examples of Multiclass Misclassification Rates	272
8.8	Software Packages	277
	Bibliographical Notes	277
	Exercises	278

9 Recursive Partitioning and Tree-Based Methods	281
9.1 Introduction	281
9.2 Classification Trees	282
9.2.1 Example: Cleveland Heart-Disease Data	284
9.2.2 Tree-Growing Procedure	285
9.2.3 Splitting Strategies	285
9.2.4 Example: Pima Indians Diabetes Study	292
9.2.5 Estimating the Misclassification Rate	294
9.2.6 Pruning the Tree	295
9.2.7 Choosing the Best Pruned Subtree	298
9.2.8 Example: Vehicle Silhouettes	302
9.3 Regression Trees	303
9.3.1 The Terminal-Node Value	305
9.3.2 Splitting Strategy	305
9.3.3 Pruning the Tree	306
9.3.4 Selecting the Best Pruned Subtree	306
9.3.5 Example: 1992 Major League Baseball Salaries	307
9.4 Extensions and Adjustments	309
9.4.1 Multivariate Responses	309
9.4.2 Survival Trees	310
9.4.3 MARS	311
9.4.4 Missing Data	312
9.5 Software Packages	313
Bibliographical Notes	313
Exercises	313
10 Artificial Neural Networks	315
10.1 Introduction	315
10.2 The Brain as a Neural Network	316
10.3 The McCulloch–Pitts Neuron	318
10.4 Hebbian Learning Theory	320
10.5 Single-Layer Perceptrons	321
10.5.1 Feedforward Single-Layer Networks	322
10.5.2 Activation Functions	323
10.5.3 Rosenblatt’s Single-Unit Perceptron	325

10.5.4	The Perceptron Learning Rule	326
10.5.5	Perceptron Convergence Theorem	326
10.5.6	Limitations of the Perceptron	328
10.6	Artificial Intelligence and Expert Systems	329
10.7	Multilayer Perceptrons	330
10.7.1	Network Architecture	331
10.7.2	A Single Hidden Layer	332
10.7.3	ANNs Can Approximate Continuous Functions	333
10.7.4	More than One Hidden Layer	334
10.7.5	Optimality Criteria	335
10.7.6	The Backpropagation of Errors Algorithm	336
10.7.7	Convergence and Stopping	340
10.8	Network Design Considerations	341
10.8.1	Learning Modes	341
10.8.2	Input Scaling	342
10.8.3	How Many Hidden Nodes and Layers?	343
10.8.4	Initializing the Weights	343
10.8.5	Overfitting and Network Pruning	343
10.9	Example: Detecting Hidden Messages in Digital Images	344
10.10	Examples of Fitting Neural Networks	347
10.11	Related Statistical Methods	348
10.11.1	Projection Pursuit Regression	349
10.11.2	Generalized Additive Models	350
10.12	Bayesian Learning for ANN Models	352
10.12.1	Laplace's Method	353
10.12.2	Markov Chain Monte Carlo Methods	361
10.13	Software Packages	364
	Bibliographical Notes	364
	Exercises	366
11	Support Vector Machines	369
11.1	Introduction	369
11.2	Linear Support Vector Machines	370
11.2.1	The Linearly Separable Case	371
11.2.2	The Linearly Nonseparable Case	376
11.3	Nonlinear Support Vector Machines	378

11.3.1	Nonlinear Transformations	379
11.3.2	The “Kernel Trick”	379
11.3.3	Kernels and Their Properties	380
11.3.4	Examples of Kernels	380
11.3.5	Optimizing in Feature Space	384
11.3.6	Grid Search for Parameters	385
11.3.7	Example: E-mail or Spam?	385
11.3.8	Binary Classification Examples	387
11.3.9	SVM as a Regularization Method	387
11.4	Multiclass Support Vector Machines	390
11.4.1	Multiclass SVM as a Series of Binary Problems	390
11.4.2	A True Multiclass SVM	391
11.5	Support Vector Regression	397
11.5.1	ϵ -Insensitive Loss Functions	398
11.5.2	Optimization for Linear ϵ -Insensitive Loss	398
11.5.3	Extensions	401
11.6	Optimization Algorithms for SVMs	401
11.7	Software Packages	403
	Bibliographical Notes	404
	Exercises	404
12	Cluster Analysis	407
12.1	Introduction	407
12.1.1	What Is a Cluster?	408
12.1.2	Example: Old Faithful Geyser Eruptions	409
12.2	Clustering Tasks	409
12.3	Hierarchical Clustering	411
12.3.1	Dendrogram	412
12.3.2	Dissimilarity	412
12.3.3	Agglomerative Nesting (agnes)	414
12.3.4	A Worked Example	414
12.3.5	Divisive Analysis (diana)	420
12.3.6	Example: Primate Scapular Shapes	420
12.4	Nonhierarchical or Partitioning Methods	422
12.4.1	K -Means Clustering (kmeans)	423
12.4.2	Partitioning Around Medoids (pam)	424

12.4.3	Fuzzy Analysis (fanny)	425
12.4.4	Silhouette Plot	426
12.4.5	Example: Landsat Satellite Image Data	428
12.5	Self-Organizing Maps (SOMs)	431
12.5.1	The SOM Algorithm	432
12.5.2	On-line Versions	433
12.5.3	Batch Version	434
12.5.4	Unified-Distance Matrix	435
12.5.5	Component Planes	437
12.6	Clustering Variables	439
12.6.1	Gene Clustering	439
12.6.2	Principal-Component Gene Shaving	440
12.6.3	Example: Colon Cancer Data	443
12.7	Block Clustering	443
12.8	Two-Way Clustering of Microarray Data	446
12.8.1	Biclustering	447
12.8.2	Plaid Models	449
12.8.3	Example: Leukemia (ALL/AML) Data	451
12.9	Clustering Based Upon Mixture Models	453
12.9.1	The EM Algorithm for Finite Mixtures	456
12.9.2	How Many Components?	459
12.10	Software Packages	459
	Bibliographical Notes	460
	Exercises	461
13	Multidimensional Scaling	
	and Distance Geometry	463
13.1	Introduction	463
13.1.1	Example: Airline Distances	464
13.2	Two Golden Oldies	468
13.2.1	Example: Perceptions of Color in Human Vision	468
13.2.2	Example: Confusion of Morse-Code Signals	469
13.3	Proximity Matrices	471
13.4	Comparing Protein Sequences	472
13.4.1	Optimal Sequence Alignment	472
13.4.2	Example: Two Hemoglobin Chains	475

13.5	String Matching	476
13.5.1	Edit Distance	476
13.5.2	Example: Employee Careers at Lloyds Bank	477
13.6	Classical Scaling and Distance Geometry	478
13.6.1	From Dissimilarities to Principal Coordinates	479
13.6.2	Assessing Dimensionality	480
13.6.3	Example: Airline Distances (Continued)	481
13.6.4	Example: Mapping the Protein Universe	484
13.7	Distance Scaling	486
13.8	Metric Distance Scaling	487
13.8.1	Metric Least-Squares Scaling	488
13.8.2	Sammon Mapping	488
13.8.3	Example: Lloyds Bank Employees	489
13.8.4	Bayesian MDS	489
13.9	Nonmetric Distance Scaling	492
13.9.1	Disparities	492
13.9.2	The Stress Function	497
13.9.3	Fitting Nonmetric Distance-Scaling Models	499
13.9.4	How Good Is an MDS Solution?	500
13.9.5	How Many Dimensions?	501
13.10	Software Packages	501
	Bibliographical Notes	502
	Exercises	503
14	Committee Machines	505
14.1	Introduction	505
14.2	Bagging	506
14.2.1	Bagging Tree-Based Classifiers	507
14.2.2	Bagging Regression-Tree Predictors	509
14.3	Boosting	511
14.3.1	ADABOOST: Boosting by Reweighting	512
14.3.2	Example: Aqueous Solubility in Drug Discovery	514
14.3.3	Convergence Issues and Overfitting	515
14.3.4	Classification Margins	518
14.3.5	ADABOOST and Maximal Margins	519
14.3.6	A Statistical Interpretation of ADABOOST	523

14.3.7	Some Questions About ADABOOST	527
14.3.8	Gradient Boosting for Regression	530
14.3.9	Other Loss Functions	532
14.3.10	Regularization	533
14.3.11	Noisy Class Labels	535
14.4	Random Forests	536
14.4.1	Randomizing Tree Construction	536
14.4.2	Generalization Error	537
14.4.3	An Upper Bound on Generalization Error	538
14.4.4	Example: Diagnostic Classification of Four Childhood Tumors	541
14.4.5	Assessing Variable Importance	542
14.4.6	Proximities for Classical Scaling	544
14.4.7	Identifying Multivariate Outliers	545
14.4.8	Treating Unbalanced Classes	547
14.5	Software Packages	548
	Bibliographical Notes	548
	Exercises	549
15	Latent Variable Models for Blind Source Separation	551
15.1	Introduction	551
15.2	Blind Source Separation and the Cocktail-Party Problem	552
15.3	Independent Component Analysis	553
15.3.1	Applications of ICA	553
15.3.2	Example: Cutaneous Potential Recordings of a Pregnant Woman	554
15.3.3	Connection to Projection Pursuit	556
15.3.4	Centering and Sphering	557
15.3.5	The General ICA Problem	558
15.3.6	Linear Mixing: Noiseless ICA	560
15.3.7	Identifiability Aspects	560
15.3.8	Objective Functions	561
15.3.9	Nonpolynomial-Based Approximations	562
15.3.10	Mutual Information	564
15.3.11	The FastICA Algorithm	566

15.3.12	Example: Identifying Artifacts in MEG Recordings	569
15.3.13	Maximum-Likelihood ICA	572
15.3.14	Kernel ICA	575
15.4	Exploratory Factor Analysis	581
15.4.1	The Factor Analysis Model	582
15.4.2	Principal Components FA	583
15.4.3	Maximum-Likelihood FA	584
15.4.4	Example: Twenty-four Psychological Tests	587
15.4.5	Critiques of MLFA	588
15.4.6	Confirmatory Factor Analysis	590
15.5	Independent Factor Analysis	590
15.6	Software Packages	594
	Bibliographical Notes	594
	Exercises	595
16	Nonlinear Dimensionality Reduction and Manifold Learning	597
16.1	Introduction	597
16.2	Polynomial PCA	598
16.3	Principal Curves and Surfaces	600
16.3.1	Curves and Curvature	601
16.3.2	Principal Curves	603
16.3.3	Projection-Expectation Algorithm	604
16.3.4	Bias Reduction	605
16.3.5	Principal Surfaces	606
16.4	Multilayer Autoassociative Neural Networks	607
16.4.1	Main Features of the Network	607
16.4.2	Relationship to Principal Curves	608
16.5	Kernel PCA	609
16.5.1	PCA in Feature Space	610
16.5.2	Centering in Feature Space	612
16.5.3	Example: Food Nutrition (Continued)	612
16.5.4	Kernel PCA and Metric MDS	613
16.6	Nonlinear Manifold Learning	613
16.6.1	Manifolds	615

16.6.2	Data on Manifolds	616
16.6.3	ISOMAP	616
16.6.4	Local Linear Embedding	621
16.6.5	Laplacian Eigenmaps	625
16.6.6	Hessian Eigenmaps	626
16.6.7	Other Methods	628
16.6.8	Relationships to Kernel PCA	628
16.7	Software Packages	630
	Bibliographical Notes	630
	Exercises	631
17	Correspondence Analysis	633
17.1	Introduction	633
17.1.1	Example: Shoplifting in The Netherlands	634
17.2	Simple Correspondence Analysis	635
17.2.1	Two-Way Contingency Tables	635
17.2.2	Row and Column Dummy Variables	636
17.2.3	Example: Hair Color and Eye Color	638
17.2.4	Profiles, Masses, and Centroids	639
17.2.5	Chi-squared Distances	642
17.2.6	Total Inertia and Its Decomposition	644
17.2.7	Principal Coordinates for Row and Column Profiles	646
17.2.8	Graphical Displays	649
17.3	Square Asymmetric Contingency Tables	651
17.3.1	Example: Occupational Mobility in England	653
17.4	Multiple Correspondence Analysis	658
17.4.1	The Multivariate Indicator Matrix	658
17.4.2	The Burt Matrix	659
17.4.3	Equivalence and an Implication	660
17.4.4	Example: Satisfaction with Housing Conditions	660
17.4.5	A Weighted Least-Squares Approach	661
17.5	Software Packages	663
	Bibliographical Notes	663
	Exercises	663
	References	667

	Contents	xxv
Index of Examples		708
Author Index		710
Subject Index		721



<http://www.springer.com/978-0-387-78188-4>

Modern Multivariate Statistical Techniques
Regression, Classification, and Manifold Learning

Izenman, A.J.

2008, XXV, 733 p., Hardcover

ISBN: 978-0-387-78188-4