

Contents

Acknowledgments, xv

SECTION I Overview

CHAPTER 1 ■ Across Statistical Modeling and Machine Learning on a Shoestring 小數財	3
1.1 ABOUT THIS BOOK	3
1.2 WHAT WILL THIS BOOK COVER?	4
1.2.1 Clustering	4
1.2.2 Regression	5
1.2.3 Classification	6
1.3 ORGANIZATION OF THIS BOOK	6
1.4 WHY ARE THERE MATHEMATICAL CALCULATIONS IN THE BOOK?	8
1.5 WHAT WON'T THIS BOOK COVER?	11
1.6 WHY IS THIS A BOOK?	12
REFERENCES AND FURTHER READING	14
CHAPTER 2 ■ Statistical Modeling	15
2.1 WHAT IS STATISTICAL MODELING?	15
2.2 PROBABILITY DISTRIBUTIONS ARE THE MODELS	18
2.3 AXIOMS OF PROBABILITY AND THEIR CONSEQUENCES: "RULES OF PROBABILITY"	23
2.4 HYPOTHESIS TESTING: WHAT YOU PROBABLY ALREADY KNOW ABOUT STATISTICS	26

2.5	TESTS WITH FEWER ASSUMPTIONS	30
2.5.1	Wilcoxon Rank-Sum Test, Also Known As the Mann-Whitney U Test (or Simply the WMW Test)	30
2.5.2	Kolmogorov-Smirnov Test (KS-Test)	31
2.6	CENTRAL LIMIT THEOREM	33
2.7	EXACT TESTS AND GENE SET ENRICHMENT ANALYSIS	33
2.8	PERMUTATION TESTS	36
2.9	SOME POPULAR DISTRIBUTIONS	38
2.9.1	The Uniform Distribution	38
2.9.2	The T-Distribution	39
2.9.3	The Exponential Distribution	39
2.9.4	The Chi-Squared Distribution	39
2.9.5	The Poisson Distribution	39
2.9.6	The Bernoulli Distribution	40
2.9.7	The Binomial Distribution	40
	EXERCISES	40
	REFERENCES AND FURTHER READING	41
CHAPTER 3 ■ Multiple Testing		43
3.1	THE BONFERRONI CORRECTION AND GENE SET ENRICHMENT ANALYSIS	43
3.2	MULTIPLE TESTING IN DIFFERENTIAL EXPRESSION ANALYSIS	46
3.3	FALSE DISCOVERY RATE	48
3.4	eQTLs: A VERY DIFFICULT MULTIPLE-TESTING PROBLEM	49
	EXERCISES	51
	REFERENCES AND FURTHER READING	52
CHAPTER 4 ■ Parameter Estimation and Multivariate Statistics		53
4.1	FITTING A MODEL TO DATA: OBJECTIVE FUNCTIONS AND PARAMETER ESTIMATION	53
4.2	MAXIMUM LIKELIHOOD ESTIMATION	54
4.3	LIKELIHOOD FOR GAUSSIAN DATA	55

10.7 TRAINING NAÏVE BAYES CLASSIFIERS	221
10.8 NAÏVE BAYES AND DATA INTEGRATION	222
EXERCISES	223
REFERENCES AND FURTHER READING	223
CHAPTER 11 ■ Nonlinear Classification	225
11.1 TWO APPROACHES TO CHOOSE NONLINEAR BOUNDARIES: DATA-GUIDED AND MULTIPLE SIMPLE UNITS	226
11.2 DISTANCE-BASED CLASSIFICATION WITH k -NEAREST NEIGHBORS	228
11.3 SVMs FOR NONLINEAR CLASSIFICATION	230
11.4 DECISION TREES	234
11.5 RANDOM FORESTS AND ENSEMBLE CLASSIFIERS: THE WISDOM OF THE CROWD	236
11.6 MULTICLASS CLASSIFICATION	237
EXERCISES	238
REFERENCES AND FURTHER READING	239
CHAPTER 12 ■ Evaluating Classifiers	241
12.1 CLASSIFICATION PERFORMANCE STATISTICS IN THE IDEAL CLASSIFICATION SETUP	241
12.2 MEASURES OF CLASSIFICATION PERFORMANCE	242
12.3 ROC CURVES AND PRECISION-RECALL PLOTS	245
12.4 EVALUATING CLASSIFIERS WHEN YOU DON'T HAVE ENOUGH DATA	248
12.5 LEAVE-ONE-OUT CROSS-VALIDATION	251
12.6 BETTER CLASSIFICATION METHODS VERSUS BETTER FEATURES	253
EXERCISES	254
REFERENCES AND FURTHER READING	255
INDEX, 257	