PREFACE

Classification analysis is a long-established technique that originated from the areas of statistics and pattern recognition, and is a broad rubric that subsumes *class discovery* and *class prediction*. Earlier forms of class discovery methods of statistical origin include natural grouping techniques such as hierarchical cluster analysis, *K*-means cluster analysis, and Gaussian mixture models employing the expectation–maximization algorithm. Conversely, pattern recognition-based class discovery techniques include self-organizing maps (Kohonen networks), neural gas, and fuzzy *K*-means cluster analysis. Class prediction methods that originated in statistics include linear discriminant analysis and logistic regression, while prototype learning, artificial neural networks, and swarm intelligence are more popular in pattern recognition. The main difference between the statistical and pattern recognition approaches is clear; the statistical methods tend to depend more on large sample Gaussian inference, while pattern recognition approaches tend to depend more on distribution-free heuristics employed in machine learning, evolutionary algorithms, or computational intelligence methods.

This book introduces the reader to a variety of statistical, machine learning, and computational intelligence classification algorithms that have been in use for several decades, as well as more recently developed algorithms based on fuzzy methods (soft computing), evolutionary algorithms, and swarm intelligence. Dimensional reduction and text mining for concept and document clustering are also covered to introduce the reader to information retrieval.

The layout of the book is divided into three parts. Part I, on class discovery, includes Chapters 1–9, which address various techniques for identifying the cluster structure of a dataset. Part II, on dimensional reduction, includes Chapters 10 and 11, which focus on linear and nonlinear approaches for reducing the dimensions of a dataset. Part III, on class prediction, covers Chapters 12–28, which present numerous approaches for predicting class membership of test objects after algorithm training is performed. There are also five appendixes, which cover probability, matrix algebra, mathematical functions, statistical primitives, and probability distributions. These are

followed by a glossary of the symbols and notation presented throughout the book.

Chapter 1 summarizes class discovery, novel diagnostic classes, comorbidity and disease overlap, outliers and heterogeneity, class prediction, rules of thumb, and descriptions of the nine microarray datasets used throughout this book. Chapter 2 introduces a crisp K-means cluster analysis algorithm, distance metrics, cluster validity to determine the optimal number of clusters, and cluster initialization. Chapter 3 describes use of fuzzification to develop membership functions, which enable the presentation of cluster weights for each microarray. Chapter 4 covers unsupervised cluster analysis using Kohonen networks [self-organizing maps (SOMs)] and the numerous uses of SOM for understanding cluster structure. The fundamental components of self-organizing maps such as neighborhood functions, best-matching units, component maps, and U matrices are also discussed. Chapter 5 introduces prototype learning and the exploration of the cluster structure of data through neural adaptive learning with prototypes. Chapter 6 covers agglomerative clustering, correlation and distance-based agglomeration, dendograms, and heatmaps. Chapter 7 addresses Gaussian mixture models and the expectation–maximization (EM) algorithm for clustering. Chapter 8 develops document and concept clustering via text mining. Methods discussed include stopping, stemming, hash tables, inverse document frequency, and concept vectors. Chapter 9 extends text mining with the use of N grams.

Chapter 10, which discusses linear dimensional reduction by eigenanalysis of the gene-by-gene or array-by-array correlation matrix, develops the concepts of principal component score coefficients, loadings, and principal component scores. Chapter 11 is presented as a means of distance metric learning and dimensional reduction from a nonlinear standpoint, and includes kernel principal component analysis (PCA), diffusion maps, Laplacian eigenmaps, local linear embedding, locality preserving projections, and Sammon mapping.

Chapter 12 presents various methods used for filtering genes in order to develop "optimal" gene lists. A review of variable types (continuous, nominal, ordinal, etc.) is provided, as well as several 2- and k-sample parametric and nonparametric statistical tests for identifying best-ranked genes. A sequential forward–reverse sequential floating method known as "greedy plus takeaway" (greedy PTA) is also introduced, which forms the basis for optimal gene sets used throughout the book. Chapter 13 reviews computational efficiency, confusion matrices and accuracy calculations, cross-validation, bootstrapping, ensemble classifier fusion, random oracles, sensitivity and specificity, receiver–operator characteristic (ROC) curves, and area under the curve (AUC). Chapter 14 presents a matrix algebra approach to multivariate regression in which dependent variables for

each class are binarized in the form $y = \pm 1$ and regressed on feature values in order to determine classification decision rules. Chapter 15 discusses decision tree classification, along with parent and daughter nodes, node splitting, terminal nodes, and stopping criteria. Chapter 16 discusses random forests as an extension of decision trees and also discusses feature importance scores, strength and correlation, proximity, unsupervised clustering, and outlier detection. Chapter 17 explores one of the most basic supervised classification methods, in which decision rules are determined from nearest neighbors and applied to test objects to predict class membership. Chapter 18 presents a probabilistic approach that uses Bayes' theorem to formulate class membership predictions. Chapter 19 discusses multivariate definitions, covariance matrices, quadratic discriminants, and Fisher's discriminant function. Chapter 20 presents supervised prototype learning to assign class labels to test objects based on the class of the nearest prototype. Chapter 21 describes binary and polytomous logistic regression. Maximum likelihood theory for the logistic regression model is first developed and then hinged to decision rules for test object class membership prediction. Chapter 22 focuses on linear separability, kernel mapping, hard and soft margins, gradient ascent, and least-squares support vector machines. Chapter 23 covers backpropagation learning, RPROP learning, cycles and epochs, training methods, training with K-means cluster analysis and principal components analysis, recursive feature elimination, gene list generation, and decision rules for class membership assignment. Chapter 24 addresses the use of kernels in linear regression for decision rule generation. Chapter 25 shows how to use metaheuristics such as genetic algorithms, covariance matrix self-adaptation, particle swarm optimization, and ant colony optimization for training a multilayer perceptron (neural network) for class prediction. Chapter 26 discusses the supervised form of neural gas, and describes a distance ranking approach for the neighborhood function. Chapter 27 covers gating probabilities, expert networks, and the EM algorithm for minimizing class prediction error. Chapter 28 covers random matrix theory and methods for filtering covariance (correlation) matrices.

The majority of chapters on class prediction in Part III present results of cross-validation based on fold values of 2, 5, and 10, and bootstrap bias, in which accuracy is shown as a function of increasing bootstrap sample size, multiclass receiver operator characteristic curves and area under the curve as a function of sample size, and class decision boundaries based on posttraining class prediction of arrays and 90,000 reference vectors of a trained 300×300 self-organizing map.

LEIF E. PETERSON