Contents

Preface vii

List of Figures xiii

1 Introduction to Data Mining 1
1.1 Why data mining? 1
1.2 Data mining techniques 3
1.2.1 A brief overview 3
1.2.2 Data representation 6
1.3 General applications of data mining 10
1.3.1 Data mining for studying brain dynamics 11
1.3.2 Data mining in telecommunications 12
1.3.3 Mining market data 13
1.4 Data mining and optimization 14
1.4.1 The simulated annealing algorithm 17
1.5 Data mining and agriculture 19
1.6 General structure of the book 20

2 Statistical Based Approaches 23
2.1 Principal component analysis 23
2.2 Interpolation and regression 30
2.3 Applications 36
2.3.1 Checking chicken breast quality 37
2.3.2 Effects of energy use in agriculture 40
2.4 Experiments in MATLAB® 40
2.5 Exercises 44

3 Clustering by *k*-means 47
3.1 The basic *k*-means algorithm 47
3.2 Variants of the *k*-means algorithm 56
3.3 Vector quantization 62

3.4 Fuzzy c-means clustering 64
3.5 Applications 67
3.5.1 Prediction of wine fermentation problem 68
3.5.2 Grading method of apples 71
3.6 Experiments in MATLAB 73
3.7 Exercises 80

4 k-Nearest Neighbor Classification 83
4.1 A simple classification rule 83
4.2 Reducing the training set 85
4.3 Speeding k-NN up 88
4.4 Applications 89
4.4.1 Climate forecasting 91
4.4.2 Estimating soil water parameters 93
4.5 Experiments in MATLAB 96
4.6 Exercises 103

5 Artificial Neural Networks 107
5.1 Multilayer perceptron 107
5.2 Training a neural network 111
5.3 The pruning process 113
5.4 Applications 114
5.4.1 Pig cough recognition 116
5.4.2 Sorting apples by watercore 118
5.5 Software for neural networks 121
5.6 Exercises 122

6 Support Vector Machines 123
6.1 Linear classifiers 123
6.2 Nonlinear classifiers 126
6.3 Noise and outliers 129
6.4 Training SVMs 130
6.5 Applications 131
6.5.1 Recognition of bird species 133
6.5.2 Detection of meat and bone meal 135
6.6 MATLAB and LIBSVM 136
6.7 Exercises 139

7 Biclustering 143
7.1 Clustering in two dimensions 143
7.2 Consistent biclustering 148
7.3 Unsupervised and supervised biclustering 151
7.4 Applications 153
7.4.1 Biclustering microarray data 153
7.4.2 Biclustering in agriculture 155
7.5 Exercises 159

8 Validation . . . 161
8.1 Validating data mining techniques . . . 161
8.2 Test set method . . . 163
8.2.1 An example in MATLAB . . . 163
8.3 Leave-one-out method . . . 166
8.3.1 An example in MATLAB . . . 166
8.4 k-fold method . . . 168
8.4.1 An example in MATLAB . . . 170

9 Data Mining in a Parallel Environment . . . 173
9.1 Parallel computing . . . 173
9.2 A simple parallel algorithm . . . 176
9.3 Some data mining techniques in parallel . . . 177
9.3.1 k-means . . . 178
9.3.2 k-NN . . . 179
9.3.3 ANNs . . . 181
9.3.4 SVMs . . . 182
9.4 Parallel computing and agriculture . . . 184

10 Solutions to Exercises . . . 185
10.1 Problems of Chapter 2 . . . 185
10.2 Problems of Chapter 3 . . . 191
10.3 Problems of Chapter 4 . . . 200
10.4 Problems of Chapter 5 . . . 204
10.5 Problems of Chapter 6 . . . 211
10.6 Problems of Chapter 7 . . . 216

Appendix A: The MATLAB Environment . . . 219
A.1 Basic concepts . . . 219
A.2 Graphic functions . . . 224
A.3 Writing a MATLAB function . . . 228

Appendix B: An Application in C . . . 231
B.1 h-means in C . . . 231
B.2 Reading data from a file . . . 238
B.3 An example of main function . . . 241
B.4 Generating random data . . . 244
B.5 Running the applications . . . 247

References . . . 253

Glossary . . . 265

Index . . . 269