

Contents

19	Introduction to Machine Learning	1121
19.1	Preliminaries	1121
19.2	Statistical Learning	1122
19.2.1	Types of Learning	1122
19.2.2	DS and ML Tasks	1123
19.3	Association Rules Mining	1126
19.3.1	Overview	1126
19.3.2	Generating Rules	1131
19.3.3	The <i>A Priori</i> Algorithm	1133
19.3.4	Validation	1135
19.3.5	Case Study: Medical Data	1135
19.3.6	Toy Example: Titanic Data	1137
19.4	Classification & Regression	1138
19.4.1	Overview	1138
19.4.2	Classification Algorithms	1141
19.4.3	Decision Trees	1143
19.4.4	Performance Evaluation	1146
19.4.5	Case Study: Tax Audits	1149
19.4.6	Toy Example: Kyphosis Data	1154
19.5	Clustering	1156
19.5.1	Overview	1156
19.5.2	Clustering Algorithms	1159
19.5.3	<i>k</i> -Means	1160
19.5.4	Clustering Validation	1163
19.5.5	Case Study: Livehoods	1164
19.5.6	Toy Example: Iris Data	1167
19.6	Issues & Challenges	1169
19.6.1	Bad Data	1169
19.6.2	Overfitting/Underfitting	1170
19.6.3	Transferability	1172
19.6.4	Myths and Mistakes	1173
19.7	R Examples	1173
19.7.1	ARM: Titanic Data	1173
19.7.2	Classification: Kyphosis	1178
19.7.3	Clustering: Iris	1187
19.8	Exercises	1194
	Chapter References	1198
20	Regression and Value Estimation	1201
20.1	Statistical Learning	1201
20.1.1	Supervised Framework	1201
20.1.2	Systematic Component	1203
20.1.3	Model Evaluation	1208
20.1.4	Bias-Variance Trade-Off	1209

20.2	Regression Modeling	1212
20.2.1	Formalism	1214
20.2.2	Least Squares Properties	1218
20.2.3	Generalizations of OLS	1224
20.2.4	Shrinkage Methods	1225
20.3	Resampling Methods	1230
20.3.1	Cross-Validation	1231
20.3.2	Bootstrap	1238
20.3.3	Jackknife	1240
20.4	Model Selection	1242
20.4.1	Best Subset Selection	1243
20.4.2	Stepwise Selection	1243
20.4.3	Optimal Models	1244
20.5	Nonlinear Modeling	1251
20.5.1	Basis Function Models	1252
20.5.2	Splines	1259
20.5.3	GAMs	1273
20.6	Example: Algae Blooms	1275
20.6.1	Value Estimation Models	1275
20.6.2	Model Evaluation	1287
20.6.3	Model Predictions	1296
20.7	Exercises	1299
	Chapter References	1300
21	Focus on Classification and Supervised Learning	1301
21.1	Overview	1301
21.1.1	Formalism	1301
21.1.2	Model Evaluation	1303
21.1.3	Bias-Variance Trade-Off	1303
21.2	Simple Classifiers	1306
21.2.1	Logistic Regression	1310
21.2.2	Discriminant Analysis	1315
21.2.3	ROC Curve	1322
21.3	Rare Occurrences	1325
21.4	Other Approaches	1327
21.4.1	Tree-Based Methods	1327
21.4.2	Support Vector Machines	1342
21.4.3	Artificial Neural Networks	1358
21.4.4	Naive Bayes Classifiers	1383
21.5	Ensemble Learning	1391
21.5.1	Bagging	1392
21.5.2	Random Forests	1396
21.5.3	Boosting	1398
21.6	Exercises	1410
	Chapter References	1411
22	Focus on Clustering	1413
22.1	Overview	1413
22.1.1	Unsupervised Learning	1413
22.1.2	Clustering Framework	1414
22.1.3	Philosophical Approach	1417

22.2	Simple Algorithms	1420
22.2.1	k -Means and Variants	1420
22.2.2	Hierarchical Clustering	1426
22.3	Clustering Evaluation	1432
22.3.1	Clustering Assessment	1432
22.3.2	Model Selection	1456
22.4	Advanced Methods	1461
22.4.1	Density-Based Clustering	1461
22.4.2	Spectral Clustering	1469
22.4.3	Probability Clustering	1481
22.4.4	Affinity Propagation	1491
22.4.5	Fuzzy Clustering	1496
22.4.6	Cluster Ensembles	1501
22.5	Exercises	1504
	Chapter References	1504
23	Feature Selection and Dimension Reduction	1507
23.1	Data Reduction for Insight	1507
23.1.1	NHL Game Reduction	1507
23.1.2	Meaning in Macbeth	1515
23.2	Dimension Reduction	1517
23.2.1	Sampling Observations	1517
23.2.2	Curse of Dimensionality	1518
23.2.3	PCA	1519
23.2.4	The Manifold Hypothesis	1523
23.3	Feature Selection	1531
23.3.1	Filter Methods	1532
23.3.2	Wrapper Methods	1540
23.3.3	Subset Selection Methods	1541
23.3.4	Regularization Methods	1541
23.3.5	SL & UL Feature Selection	1542
23.4	Advanced Topics	1542
23.4.1	SVD	1542
23.4.2	PC Regression & Partial LS	1546
23.4.3	Spectral Feature Selection	1548
23.4.4	UMAP	1565
23.5	Exercises	1571
	Chapter References	1571

List of Figures

19.1	<i>Amanita muscaria</i> in the wild	1124
19.2	Decision tree for the mushroom classification problem	1125
19.3	Decision path for <i>Amanita muscaria</i>	1126
19.4	Pruned supersets of an infrequent itemset in the <i>a priori</i> network of a dataset with 5 items	1133
19.5	Association rules for NHL playoff teams (1942-1967)	1134

19.6	COPD cluster in the Danish Medical Dataset	1137
19.7	Visualization of <i>Titanic</i> association rules	1138
19.8	A classification pipeline	1140
19.9	Illustrations of various classifiers – I	1142
19.10	Illustrations of various classifiers – II	1143
19.11	Picking the optimal information gain split	1144
19.12	Predicted and actual numerical responses	1149
19.13	Data sources for APGEN mining	1151
19.14	Feature selection process in APGEN mining	1152
19.15	Audit resource deployment efficiency	1153
19.16	Kyphosis decision tree visualization	1154
19.17	Pruning a decision tree	1155
19.18	Clusters and outliers in an artificial dataset	1156
19.19	Distance metrics between observations	1157
19.20	Cluster distances	1157
19.21	A clustering pipeline	1159
19.22	Illustration of hierarchical clustering, DBSCAN, and spectral clustering	1160
19.23	k -means cluster allocation and updated centres	1161
19.24	Cluster suggestions in an artificial dataset	1162
19.25	Illustration of the ambiguity of cluster number	1162
19.26	An illustration of ghost clustering with k -means	1163
19.27	Some livelihoods in metropolitan Pittsburgh, PA	1166
19.28	PCA plot of the iris dataset; one replicate of the optimal clustering results for the iris dataset	1168
19.29	Some clustering results on the iris dataset with k -means	1168
19.30	Optimal clustering results for the iris dataset	1169
19.31	Illustration of underfitting and overfitting for a classification task	1170
19.32	Underfitting and overfitting as a function of model complexity	1171
19.33	Schematic illustration of cross-fold validation	1171
20.1	Regression model for a Gapminder subset	1204
20.2	Regression model for a Gapminder subset, with vertical line	1204
20.3	Regression model for a Gapminder subset, with vertical line and neighbourhood	1205
20.4	The training/testing paradigm	1210
20.5	Illustration of the bias-variance trade-off	1210
20.6	Expected test error decomposition	1211
20.7	Predictor envelope for the Gapminder subset	1217
20.8	Ridge regression coefficients in a generic problem	1226
20.9	LASSO coefficients in a generic problem	1227
20.10	LASSO and RR level curves	1228
20.11	Various splines with a single knot	1260
20.12	Some hinge functions	1264
20.13	Overfit spline	1271
21.1	Illustration of the accuracy-interpretability trade-off for classifiers	1304
21.2	Illustration of k NN classifiers	1304
21.3	Classification based on OLS and k NN	1305
21.4	Illustration of LDA on a univariate dataset	1317
21.5	ROC schematics	1322
21.6	Illustration of undersampling	1325
21.7	Illustration of oversampling	1326
21.8	Stratification of predictor space	1329

21.9	Generic recursive binary partition regression tree	1331
21.10	Different tree topologies with small changes in the training set	1334
21.11	Two-class artificial dataset and classification tree	1343
21.12	Separating hyperplane on a two-class artificial dataset	1343
21.13	Linearly separable subset of a two-class dataset with separating hyperplanes, maximal margin hyperplane, and support vectors.	1344
21.14	Non-linearly separable two-class datasets	1346
21.15	Hard margins and soft margins for linearly and non-linearly separable classifiers	1346
21.16	Toy classification problem	1352
21.17	Conceptual timeline of AI interest and optimism	1358
21.18	Artificial neural network topology	1360
21.19	Relationship between the network, layers, loss function, and optimizer	1361
21.20	A 3D time series data tensor	1362
21.21	A 4D image data tensor	1362
21.22	Signal propagating forward through an ANN	1365
21.23	SGD with one parameter	1368
22.1	Realizations of k -means on the 2011 Gapminder data	1425
22.2	Conceptual representation of AGNES and DIANA	1427
22.3	Cluster dendrogram for the hierarchical cluster structure of a dataset with 50 observations and 3 variables	1427
22.4	Conceptual notions of linkage	1429
22.5	Realizations of hierarchical clustering on the 2011 Gapminder data	1431
22.6	Artificial fruit image toy dataset	1438
22.7	Two clusters in a subset of the fruit image toy dataset	1438
22.8	Illustration of cluster quality measurements	1439
22.9	Schematics of instance/cluster properties and relationships	1440
22.10	Schematics of relative clustering validation	1445
22.11	Useful external quality metric considerations	1451
22.12	World regions in the Gapminder data	1454
22.13	Density path connection in a DBSCAN cluster	1464
22.14	Illustration of DBSCAN on an artificial dataset	1465
22.15	Realizations of DBSCAN on the (scaled) 2011 Gampinder data	1468
22.16	Schematics of spectral clustering	1470
22.17	Spectral clusters for the artificial dataset	1475
22.18	Comparing k -means and spiral clustering	1475
22.19	High-contrast image segmentation with spectral clustering	1477
22.20	Low-contrast image segmentation with spectral clustering	1478
22.21	Spectral clustering image segmentation of images at different resolutions	1479
22.22	Two realizations of spectral clustering, using the NJW algorithm	1479
22.23	Illustration of 3-medoids on an artificial dataset	1492
22.24	Illustration of affinity propagation	1493
22.25	Illustration of fuzzy c -means clustering	1498
22.26	FANNY clusters and silhouette profiles for 2011 Gampinder data	1500
23.1	Schematic diagram of data reduction – NHL game	1508
23.2	Play-by-play extract – NHL Game	1509
23.3	Advanced boxscore (I) – NHL Game	1510
23.4	Advanced boxscore (II) – NHL Game	1511
23.5	Advanced boxscore (III) – NHL Game	1512
23.6	Simple boxscore – NHL Game	1513

23.7	Visualizations – NHL Game	1514
23.8	Schematic diagram of data reduction – NHL Game	1515
23.9	Illustration of the curse of dimensionality	1518
23.10	Illustration of PCA on an artificial 2D dataset	1520
23.11	Selecting the number of principal component	1520
23.12	Degrees of freedom manifolds for faces and digits	1524
23.13	High-dimensional manifold unfolding	1525
23.14	Geodesic and Euclidean paths on the Earth	1525
23.15	Comparison of manifold learning methods on an artificial dataset.	1529
23.16	Sample of the MNIST dataset	1530
23.17	Manifold learning on a subset of MNIST	1530
23.18	Feature selection process for wrapper methods in classification problems	1540
23.19	SVD image reconstruction	1544

List of Tables

19.1	A general binary classifier	1147
19.2	Performance metrics for two (artificial) binary classifiers	1148
19.3	Performance metrics for multi-level classifiers	1148
19.4	Confusion matrices for audit evaluation	1153
19.5	Kyphosis decision tree performance evaluation	1155