

Table of Contents

Preface	1
Chapter 1: A Gentle Introduction to Machine Learning	7
Introduction - classic and adaptive machines	7
Only learning matters	10
Supervised learning	11
Unsupervised learning	13
Reinforcement learning	15
Beyond machine learning - deep learning and bio-inspired adaptive systems	16
Machine learning and big data	18
Further reading	19
Summary	20
Chapter 2: Important Elements in Machine Learning	21
Data formats	21
Multiclass strategies	24
One-vs-all	24
One-vs-one	24
Learnability	25
Underfitting and overfitting	28
Error measures	29
PAC learning	31
Statistical learning approaches	33
MAP learning	35
Maximum-likelihood learning	35
Elements of information theory	40
References	43
Summary	43
Chapter 3: Feature Selection and Feature Engineering	45
scikit-learn toy datasets	45
Creating training and test sets	46
Managing categorical data	48
Managing missing features	51
Data scaling and normalization	52

Feature selection and filtering	55
Principal component analysis	57
Non-negative matrix factorization	63
Sparse PCA	65
Kernel PCA	66
Atom extraction and dictionary learning	69
References	71
Summary	71
Chapter 4: Linear Regression	73
<hr/>	
Linear models	73
A bidimensional example	74
Linear regression with scikit-learn and higher dimensionality	76
Regressor analytic expression	80
Ridge, Lasso, and ElasticNet	81
Robust regression with random sample consensus	87
Polynomial regression	88
Isotonic regression	92
References	94
Summary	94
Chapter 5: Logistic Regression	95
<hr/>	
Linear classification	96
Logistic regression	98
Implementation and optimizations	100
Stochastic gradient descent algorithms	104
Finding the optimal hyperparameters through grid search	108
Classification metrics	111
ROC curve	116
Summary	120
Chapter 6: Naive Bayes	121
<hr/>	
Bayes' theorem	121
Naive Bayes classifiers	123
Naive Bayes in scikit-learn	124
Bernoulli naive Bayes	124
Multinomial naive Bayes	127
Gaussian naive Bayes	129
References	132
Summary	133
Chapter 7: Support Vector Machines	135
<hr/>	

Linear support vector machines	135
scikit-learn implementation	140
Linear classification	140
Kernel-based classification	143
Radial Basis Function	144
Polynomial kernel	144
Sigmoid kernel	145
Custom kernels	145
Non-linear examples	145
Controlled support vector machines	151
Support vector regression	153
References	155
Summary	155
Chapter 8: Decision Trees and Ensemble Learning	157
Binary decision trees	158
Binary decisions	159
Impurity measures	161
Gini impurity index	162
Cross-entropy impurity index	162
Misclassification impurity index	163
Feature importance	163
Decision tree classification with scikit-learn	163
Ensemble learning	170
Random forests	170
Feature importance in random forests	173
AdaBoost	174
Gradient tree boosting	177
Voting classifier	179
References	183
Summary	183
Chapter 9: Clustering Fundamentals	185
Clustering basics	185
K-means	187
Finding the optimal number of clusters	192
Optimizing the inertia	192
Silhouette score	194
Calinski-Harabasz index	198
Cluster instability	200
DBSCAN	203
Spectral clustering	206
Evaluation methods based on the ground truth	208

Homogeneity	208
Completeness	209
Adjusted rand index	209
References	210
Summary	211
Chapter 10: Hierarchical Clustering	213
Hierarchical strategies	213
Agglomerative clustering	214
Dendrograms	217
Agglomerative clustering in scikit-learn	219
Connectivity constraints	223
References	225
Summary	226
Chapter 11: Introduction to Recommendation Systems	227
Naive user-based systems	227
User-based system implementation with scikit-learn	228
Content-based systems	231
Model-free (or memory-based) collaborative filtering	234
Model-based collaborative filtering	237
Singular Value Decomposition strategy	238
Alternating least squares strategy	240
Alternating least squares with Apache Spark MLlib	241
References	245
Summary	246
Chapter 12: Introduction to Natural Language Processing	247
NLTK and built-in corpora	247
Corpora examples	249
The bag-of-words strategy	250
Tokenizing	252
Sentence tokenizing	252
Word tokenizing	253
Stopword removal	254
Language detection	255
Stemming	256
Vectorizing	257
Count vectorizing	257
N-grams	259
Tf-idf vectorizing	260
A sample text classifier based on the Reuters corpus	262

References	264
Summary	264
Chapter 13: Topic Modeling and Sentiment Analysis in NLP	267
Topic modeling	267
Latent semantic analysis	268
Probabilistic latent semantic analysis	275
Latent Dirichlet Allocation	281
Sentiment analysis	288
VADER sentiment analysis with NLTK	292
References	293
Summary	293
Chapter 14: A Brief Introduction to Deep Learning and TensorFlow	295
Deep learning at a glance	295
Artificial neural networks	296
Deep architectures	300
Fully connected layers	300
Convolutional layers	301
Dropout layers	303
Recurrent neural networks	303
A brief introduction to TensorFlow	304
Computing gradients	306
Logistic regression	309
Classification with a multi-layer perceptron	313
Image convolution	317
A quick glimpse inside Keras	320
References	326
Summary	326
Chapter 15: Creating a Machine Learning Architecture	327
Machine learning architectures	327
Data collection	329
Normalization	330
Dimensionality reduction	330
Data augmentation	331
Data conversion	331
Modeling/Grid search/Cross-validation	332
Visualization	332
scikit-learn tools for machine learning architectures	332
Pipelines	333

Feature unions	337
References	338
Summary	338
Index	339
