

Contents

1	Introduction to Machine Learning	17
1.1	A simple supervised model: Nearest Neighbor	18
1.1.1	Tuning Hyperparameters with Cross-Validation	25
1.2	Preprocessing	30
1.2.1	Scaling Data	31
1.2.2	Forcing Data to be Gaussian: an Introduction to Power Transformations	35
1.2.3	Dealing with Categorical Variables	38
1.2.4	Handling with Missing Values	41
1.3	Methods for Dealing with Imbalanced Data	42
1.3.1	Random Oversampling of the Majority Class	45
1.3.2	Random Undersampling of the Majority Class	46
1.3.3	Oversampling using Synthetic Data: SMOTE	47
1.4	Reducing Dimensionality: Principal Component Analysis	48
1.4.1	PCA as dimensionality reduction	49
1.4.2	Feature extraction	52
1.4.3	Nonlinear Manifold Algorithm: t-SNE	55
2	Linear Models for Machine Learning	59
2.1	Linear Regression	60
2.2	Shrinkage Methods	62
2.2.1	Ridge Regression	62
2.2.2	Lasso Regression	67
2.2.3	Elastic Net	69
2.3	Robust Regression	70
2.3.1	Huber Regression	71
2.3.2	RANSAC	74

2.4	Logistic Regression	76
2.4.1	Why Logistic Regression is Linear?	77
2.4.2	Logistic Regression Predictions (Raw Model Output) vs Probabilities (Sigmoid Output)	78
2.4.3	Logistic Regression in Python	79
2.4.4	Model Performance Evaluation	80
2.4.5	Regularization	84
2.5	Linear Support Vector Machine	86
2.6	Beyond Linearity: Kernelized Models	91
2.6.1	Into the Hood of the Kernel Trick	94
2.6.2	Practical Classification Example: Face Recognition	95
3	Beyond Linearity: Ensemble Methods for ML	101
3.1	Introduction	101
3.2	Ensemble Methods	102
3.2.1	Bootstrap Aggregation	106
3.2.2	Out-of-Bag Estimation	108
3.3	Random Forests	109
3.3.1	Random Forests Classifier	109
3.3.2	Random Forests Regressor	112
3.4	Boosting Methods	113
3.4.1	AdaBoost	113
3.4.2	Gradient Boosting	114
3.4.3	Extreme Gradient Boosting (XGBoost)	117
3.4.4	CatBoost	124
4	An Introduction to Modern ML Techniques	133
4.1	Introduction to Natural language Processing	133
4.1.1	Preprocessing with Text Data	134
4.1.2	Numerical Representation of Documents: the Bag-of-Words	139
4.1.3	Practical Example: Sentiment Analysis with IMDb Reviews Dataset	142
4.1.4	Term Frequency-Inverse Document Frequency	144
4.1.5	Bag-of-Words with More Than One Word (n-Grams)	145
4.1.6	Beyond Bag-of-words: Word Embeddings	150
4.2	Introduction to Deep Learning	157
4.2.1	Dealing with Complex Data into a Neural Network	161

<i>CONTENTS</i>	15
4.2.2 Multiclass classification	163
Appendices	167
A A crash course in Python	169
A.1 Building Blocks in Python	169
A.1.1 Variables	169
A.1.2 Methods	170
A.2 Data Structure in Python	172
A.2.1 List and Tuples	172
A.2.2 Sets	173
A.2.3 Dictionaries	174
A.3 Loops in Python	174
A.3.1 The For Loop	174
A.3.2 The While Loop	175
A.4 Advanced Data Structure in Python	176
A.4.1 List comprehensions	176
A.4.2 Lambda Functions	177
A.5 Advanced Concepts on Functions	178
A.5.1 The magic of Wildcards into Function's arguments	178
A.5.2 Local vs Global Scope in Functions	182
A.6 Introduction to Object-Oriented Programming	183
A.6.1 Objects, Classes and Attributes	184
A.6.2 Subclasses and Inheritance	185
B Mathematics behind the skip-gram model	189