Introduction

Chapter 1  The Two Essential Algorithms for Making Predictions
  Why Are These Two Algorithms So Useful?  1
  What Are Penalized Regression Methods?  2
  What Are Ensemble Methods?  7
  How to Decide Which Algorithm to Use  9
  The Process Steps for Building a Predictive Model  11
    Framing a Machine Learning Problem  13
    Feature Extraction and Feature Engineering  15
    Determining Performance of a Trained Model  17
  Chapter Contents and Dependencies  18
  Summary  20

Chapter 2  Understand the Problem by Understanding the Data
  The Anatomy of a New Problem  23
    Different Types of Attributes and Labels  24
    Drive Modeling Choices  26
    Things to Notice about Your New Data Set  27
  Classification Problems: Detecting Unexploded Mines Using Sonar  28
    Physical Characteristics of the Rocks Versus Mines Data Set  29
    Statistical Summaries of the Rocks versus Mines Data Set  32
    Visualization of Outliers Using Quantile-Quantile Plot  35
    Statistical Characterization of Categorical Attributes  37
    How to Use Python Pandas to Summarize the Rocks Versus Mines Data Set  37
Visualizing Properties of the Rocks versus Mines Data Set 40
  Visualizing with Parallel Coordinates Plots 40
  Visualizing Interrelationships between Attributes and Labels 42
  Visualizing Attribute and Label Correlations
    Using a Heat Map 49
  Summarizing the Process for Understanding Rocks versus Mines Data Set 50
Real-Valued Predictions with Factor Variables:
  How Old Is Your Abalone? 50
  Parallel Coordinates for Regression Problems—Visualize Variable Relationships for Abalone Problem 56
  How to Use Correlation Heat Map for Regression—Visualize Pair-Wise Correlations for the Abalone Problem 60
Real-Valued Predictions Using Real-Valued Attributes:
  Calculate How Your Wine Tastes 62
Multiclass Classification Problem: What Type of Glass Is That? 68

**Chapter 3**  Predictive Model Building: Balancing Performance, Complexity, and Big Data 75

The Basic Problem: Understanding Function Approximation 76
  Working with Training Data 76
  Assessing Performance of Predictive Models 78
Factors Driving Algorithm Choices and Performance—Complexity and Data
  Contrast Between a Simple Problem and a Complex Problem 80
  Contrast Between a Simple Model and a Complex Model 82
  Factors Driving Predictive Algorithm Performance 86
Choosing an Algorithm: Linear or Nonlinear? 87
Measuring the Performance of Predictive Models 88
  Performance Measures for Different Types of Problems 88
Simulating Performance of Deployed Models 99
Achieving Harmony Between Model and Data
  Choosing a Model to Balance Problem Complexity,
    Model Complexity, and Data Set Size 102
  Using Forward Stepwise Regression to Control Overfitting 103
  Evaluating and Understanding Your Predictive Model 108
Control Overfitting by Penalizing Regression Coefficients—Ridge Regression 110
Summary 119

**Chapter 4**  Penalized Linear Regression 121

Why Penalized Linear Regression Methods Are So Useful 122
  Extremely Fast Coefficient Estimation 122
  Variable Importance Information 122
  Extremely Fast Evaluation When Deployed 123
Contents

Reliable Performance 123
Sparse Solutions 123
Problem May Require Linear Model 124
When to Use Ensemble Methods 124

Penalized Linear Regression: Regulating Linear Regression for Optimum Performance 124
Training Linear Models: Minimizing Errors and More 126
Adding a Coefficient Penalty to the OLS Formulation 127
Other Useful Coefficient Penalties—Manhattan and ElasticNet 128
Why Lasso Penalty Leads to Sparse Coefficient Vectors 129
ElasticNet Penalty Includes Both Lasso and Ridge 131

Solving the Penalized Linear Regression Problem 132
Understanding Least Angle Regression and Its Relationship to Forward Stepwise Regression 132
How LARS Generates Hundreds of Models of Varying Complexity 136
Choosing the Best Model from The Hundreds LARS Generates 139
Using Glmnet: Very Fast and Very General 144
Comparison of the Mechanics of Glmnet and LARS Algorithms 145
Initializing and Iterating the Glmnet Algorithm 146

Extensions to Linear Regression with Numeric Input 151
Solving Classification Problems with Penalized Regression 151
Working with Classification Problems Having More Than Two Outcomes 155
Understanding Basis Expansion: Using Linear Methods on Nonlinear Problems 156
Incorporating Non-Numeric Attributes into Linear Methods 158

Summary 163

Chapter 5  Building Predictive Models Using Penalized Linear Methods 165
Python Packages for Penalized Linear Regression 166
Multivariable Regression: Predicting Wine Taste 167
Building and Testing a Model to Predict Wine Taste 168
Training on the Whole Data Set before Deployment 172
Basis Expansion: Improving Performance by Creating New Variables from Old Ones 178
Binary Classification: Using Penalized Linear Regression to Detect Unexploded Mines 181
Building a Rocks versus Mines Classifier for Deployment 191
Multiclass Classification: Classifying Crime Scene Glass Samples 204
Summary 209
Chapter 6  Ensemble Methods  211

不适于直接生成表格，需要将文本转换为表格形式，如下所示：

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary Decision Trees</td>
<td>212</td>
</tr>
<tr>
<td>How a Binary Decision Tree Generates Predictions</td>
<td>213</td>
</tr>
<tr>
<td>How to Train a Binary Decision Tree</td>
<td>214</td>
</tr>
<tr>
<td>Tree Training Equals Split Point Selection</td>
<td>218</td>
</tr>
<tr>
<td>How Split Point Selection Affects Predictions</td>
<td>218</td>
</tr>
<tr>
<td>Algorithm for Selecting Split Points</td>
<td>219</td>
</tr>
<tr>
<td>Multivariable Tree Training—Which Attribute to Split?</td>
<td>219</td>
</tr>
<tr>
<td>Recursive Splitting for More Tree Depth</td>
<td>220</td>
</tr>
<tr>
<td>Overfitting Binary Trees</td>
<td>221</td>
</tr>
<tr>
<td>Measuring Overfit with Binary Trees</td>
<td>221</td>
</tr>
<tr>
<td>Balancing Binary Tree Complexity for Best Performance</td>
<td>222</td>
</tr>
<tr>
<td>Modifications for Classification and Categorical Features</td>
<td>225</td>
</tr>
<tr>
<td>Bootstrap Aggregation: “Bagging”</td>
<td>226</td>
</tr>
<tr>
<td>How Does the Bagging Algorithm Work?</td>
<td>226</td>
</tr>
<tr>
<td>Bagging Performance—Bias versus Variance</td>
<td>229</td>
</tr>
<tr>
<td>How Bagging Behaves on Multivariable Problem</td>
<td>231</td>
</tr>
<tr>
<td>Bagging Needs Tree Depth for Performance</td>
<td>235</td>
</tr>
<tr>
<td>Summary of Bagging</td>
<td>236</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>236</td>
</tr>
<tr>
<td>Basic Principle of Gradient Boosting Algorithm</td>
<td>237</td>
</tr>
<tr>
<td>Parameter Settings for Gradient Boosting</td>
<td>239</td>
</tr>
<tr>
<td>How Gradient Boosting Iterates Toward a Predictive Model</td>
<td>240</td>
</tr>
<tr>
<td>Getting the Best Performance from Gradient Boosting</td>
<td>240</td>
</tr>
<tr>
<td>Gradient Boosting on a Multivariable Problem</td>
<td>244</td>
</tr>
<tr>
<td>Summary for Gradient Boosting</td>
<td>247</td>
</tr>
<tr>
<td>Random Forest</td>
<td>247</td>
</tr>
<tr>
<td>Random Forests: Bagging Plus Random Attribute Subsets</td>
<td>250</td>
</tr>
<tr>
<td>Random Forests Performance Drivers</td>
<td>251</td>
</tr>
<tr>
<td>Random Forests Summary</td>
<td>252</td>
</tr>
<tr>
<td>Summary</td>
<td>252</td>
</tr>
</tbody>
</table>

Chapter 7  Building Ensemble Models with Python  255

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solving Regression Problems with Python</td>
<td>255</td>
</tr>
<tr>
<td>Ensemble Packages</td>
<td>255</td>
</tr>
<tr>
<td>Building a Random Forest Model to Predict Wine Taste</td>
<td>256</td>
</tr>
<tr>
<td>Constructing a RandomForestRegressor Object</td>
<td>256</td>
</tr>
<tr>
<td>Modeling Wine Taste with RandomForestRegressor</td>
<td>259</td>
</tr>
<tr>
<td>Visualizing the Performance of a RandomForest Forests Regression Model</td>
<td>262</td>
</tr>
<tr>
<td>Using Gradient Boosting to Predict Wine Taste</td>
<td>263</td>
</tr>
<tr>
<td>Using the Class Constructor for GradientBoostingRegressor</td>
<td>263</td>
</tr>
<tr>
<td>Using GradientBoostingRegressor to Implement a Regression Model</td>
<td>267</td>
</tr>
<tr>
<td>Assessing the Performance of a Gradient Boosting Model</td>
<td>269</td>
</tr>
</tbody>
</table>
Coding Bagging to Predict Wine Taste 270
Incorporating Non-Numeric Attributes in Python Ensemble Models 275
  Coding the Sex of Abalone for Input to Random Forest Regression in Python 275
  Assessing Performance and the Importance of Coded Variables 278
  Coding the Sex of Abalone for Gradient Boosting Regression in Python 278
  Assessing Performance and the Importance of Coded Variables with Gradient Boosting 282
Solving Binary Classification Problems with Python Ensemble Methods 284
  Detecting Unexploded Mines with Python Random Forest 285
  Constructing a Random Forests Model to Detect Unexploded Mines 287
  Determining the Performance of a Random Forests Classifier 291
  Detecting Unexploded Mines with Python Gradient Boosting 291
  Determining the Performance of a Gradient Boosting Classifier 298
Solving Multiclass Classification Problems with Python Ensemble Methods 302
  Classifying Glass with Random Forests 302
  Dealing with Class Imbalances 305
  Classifying Glass Using Gradient Boosting 307
  Assessing the Advantage of Using Random Forest Base Learners with Gradient Boosting 311
Comparing Algorithms 314
Summary 315
Index 319