“Python Machine Learning”
12 assignments: from chapter 1 to 6.

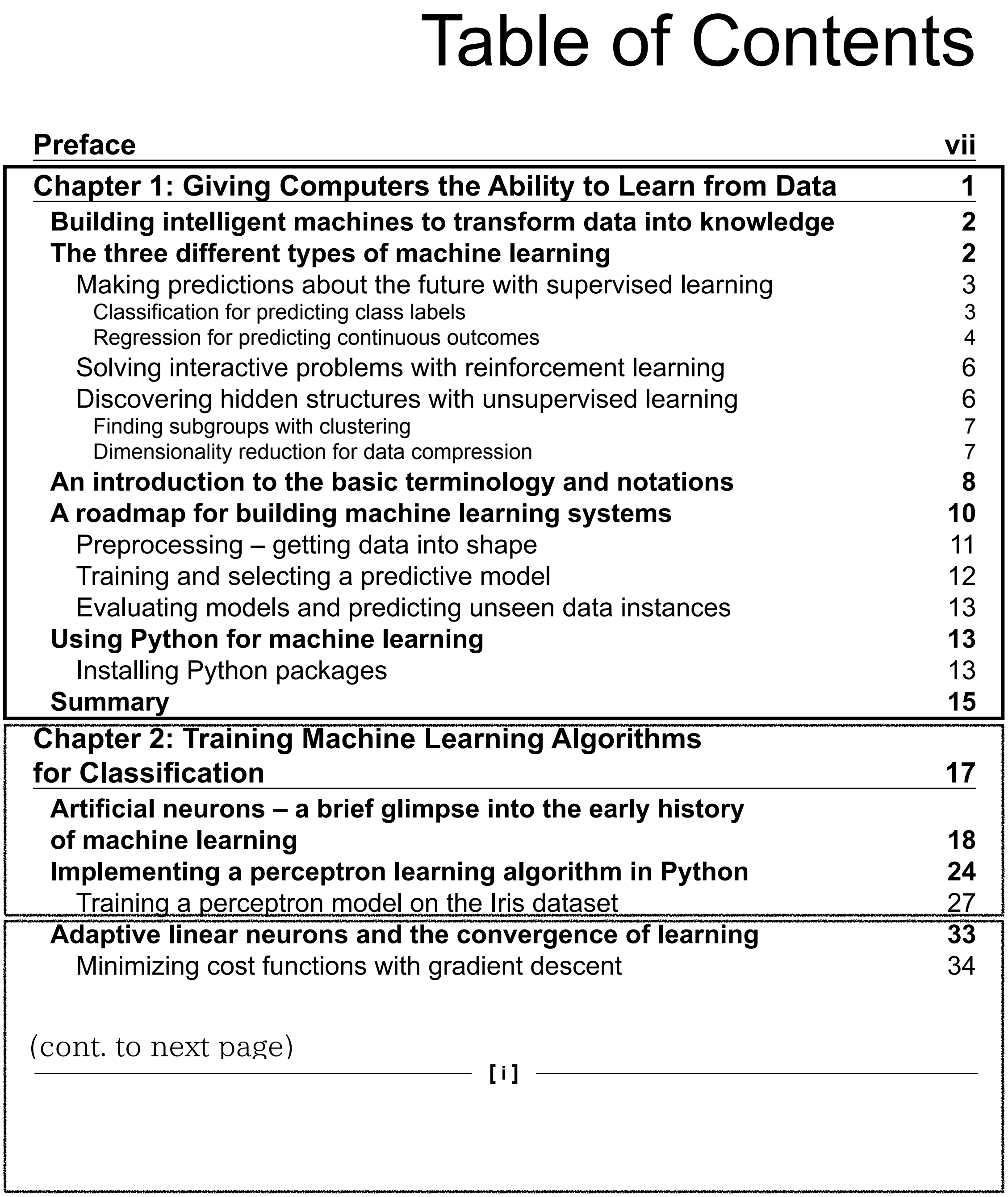


Table of Contents

Preface vii

Chapter 1: Giving Computers the Ability to Learn from Data 1
- Building intelligent machines to transform data into knowledge 2
- The three different types of machine learning 2
  - Making predictions about the future with supervised learning 3
    - Classification for predicting class labels 3
    - Regression for predicting continuous outcomes 4
  - Solving interactive problems with reinforcement learning 6
  - Discovering hidden structures with unsupervised learning 6
    - Finding subgroups with clustering 7
    - Dimensionality reduction for data compression 7
- An introduction to the basic terminology and notations 8
- A roadmap for building machine learning systems 10
  - Preprocessing – getting data into shape 11
  - Training and selecting a predictive model 12
  - Evaluating models and predicting unseen data instances 13
- Using Python for machine learning 13
  - Installing Python packages 13
- Summary 15

assign 1

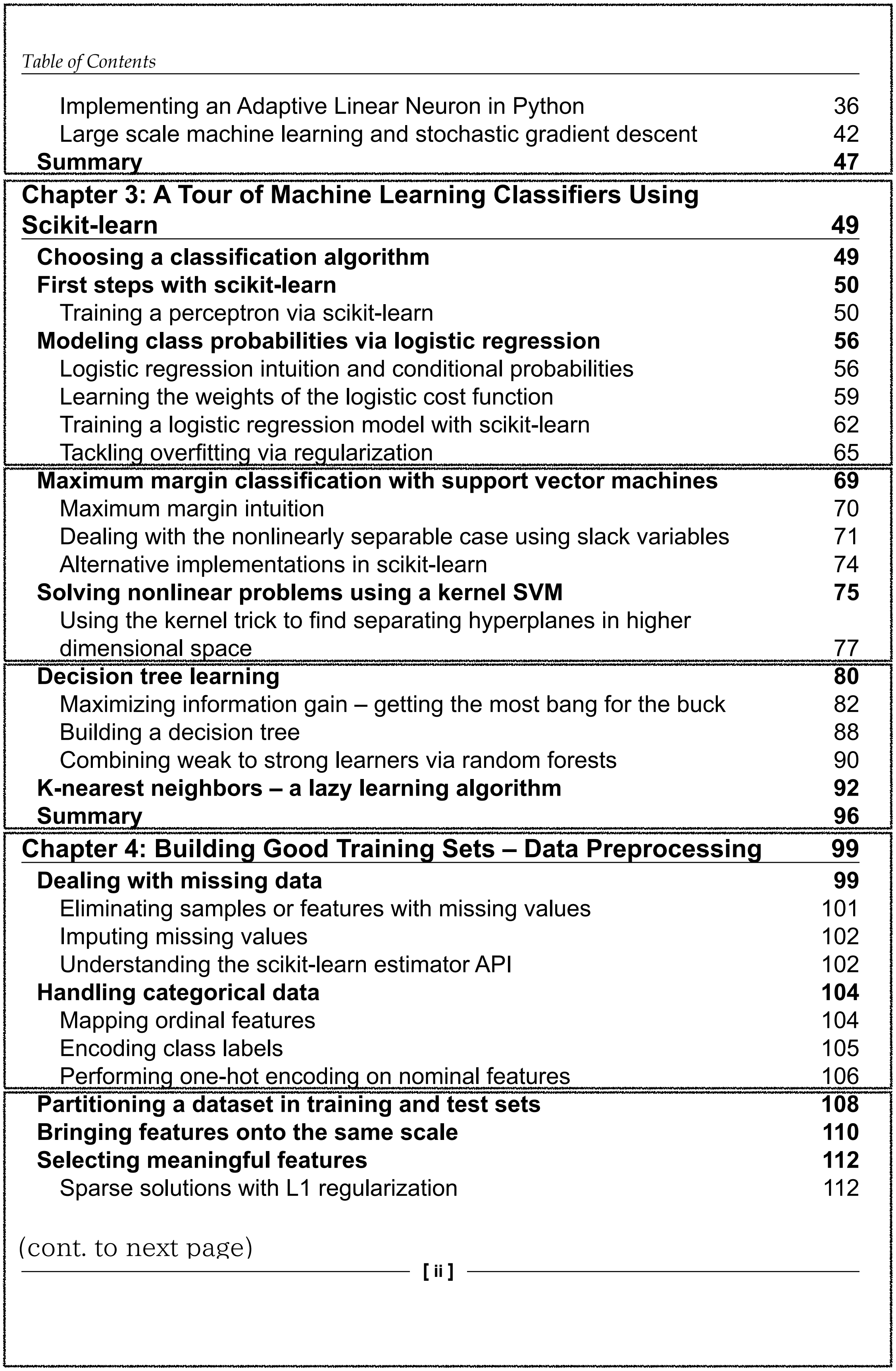
Chapter 2: Training Machine Learning Algorithms for Classification 17
- Artificial neurons – a brief glimpse into the early history of machine learning 18
- Implementing a perceptron learning algorithm in Python 24
  - Training a perceptron model on the Iris dataset 27

assign 2

- Adaptive linear neurons and the convergence of learning 33
  - Minimizing cost functions with gradient descent 34

assign 3

(cont. to next page)



assign 3

- Implementing an Adaptive Linear Neuron in Python 36
- Large scale machine learning and stochastic gradient descent 42
- **Summary 47**

assign 4

Chapter 3: A Tour of Machine Learning Classifiers Using Scikit-learn 49

- **Choosing a classification algorithm 49**
- **First steps with scikit-learn 50**
  - Training a perceptron via scikit-learn 50
- **Modeling class probabilities via logistic regression 56**
  - Logistic regression intuition and conditional probabilities 56
  - Learning the weights of the logistic cost function 59
  - Training a logistic regression model with scikit-learn 62
  - Tackling overfitting via regularization 65

assign 5

- **Maximum margin classification with support vector machines 69**
  - Maximum margin intuition 70
  - Dealing with the nonlinearly separable case using slack variables 71
  - Alternative implementations in scikit-learn 74
- **Solving nonlinear problems using a kernel SVM 75**
  - Using the kernel trick to find separating hyperplanes in higher dimensional space 77

assign 6

- **Decision tree learning 80**
  - Maximizing information gain – getting the most bang for the buck 82
  - Building a decision tree 88
  - Combining weak to strong learners via random forests 90
- **K-nearest neighbors – a lazy learning algorithm 92**
- **Summary 96**

assign 7

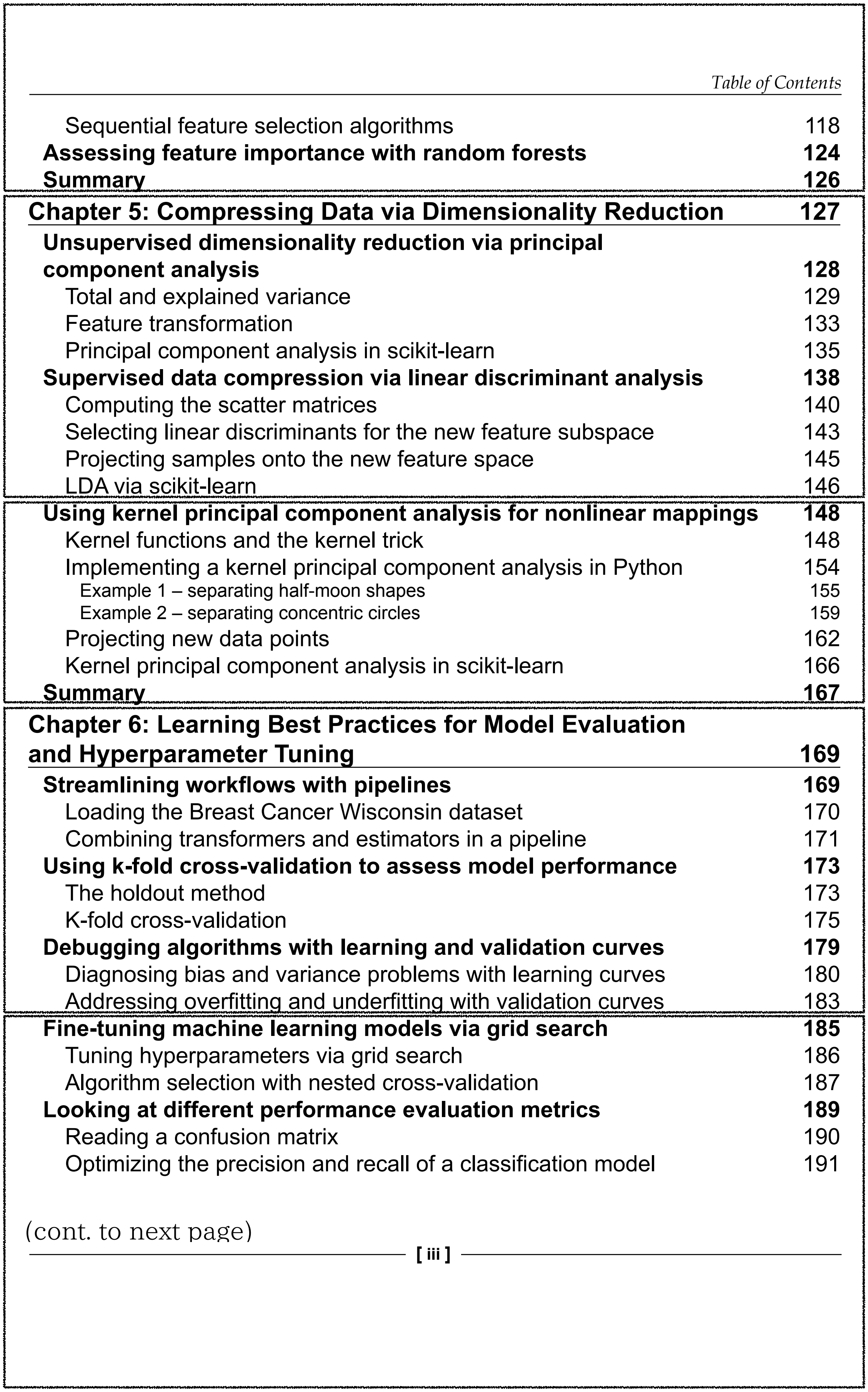
Chapter 4: Building Good Training Sets – Data Preprocessing 99

- **Dealing with missing data 99**
  - Eliminating samples or features with missing values 101
  - Imputing missing values 102
  - Understanding the scikit-learn estimator API 102
- **Handling categorical data 104**
  - Mapping ordinal features 104
  - Encoding class labels 105
  - Performing one-hot encoding on nominal features 106

assign 8

- **Partitioning a dataset in training and test sets 108**
- **Bringing features onto the same scale 110**
- **Selecting meaningful features 112**
  - Sparse solutions with L1 regularization 112

(cont. to next page)



assign 8

Sequential feature selection algorithms 118

Assessing feature importance with random forests 124

Summary 126

assign 9

Chapter 5: Compressing Data via Dimensionality Reduction 127

Unsupervised dimensionality reduction via principal component analysis 128

Total and explained variance 129

Feature transformation 133

Principal component analysis in scikit-learn 135

Supervised data compression via linear discriminant analysis 138

Computing the scatter matrices 140

Selecting linear discriminants for the new feature subspace 143

Projecting samples onto the new feature space 145

LDA via scikit-learn 146

assign 10

Using kernel principal component analysis for nonlinear mappings 148

Kernel functions and the kernel trick 148

Implementing a kernel principal component analysis in Python 154

Example 1 – separating half-moon shapes 155

Example 2 – separating concentric circles 159

Projecting new data points 162

Kernel principal component analysis in scikit-learn 166

Summary 167

assign 11

Chapter 6: Learning Best Practices for Model Evaluation and Hyperparameter Tuning 169

Streamlining workflows with pipelines 169

Loading the Breast Cancer Wisconsin dataset 170

Combining transformers and estimators in a pipeline 171

Using k-fold cross-validation to assess model performance 173

The holdout method 173

K-fold cross-validation 175

Debugging algorithms with learning and validation curves 179

Diagnosing bias and variance problems with learning curves 180

Addressing overfitting and underfitting with validation curves 183

assign 12

Fine-tuning machine learning models via grid search 185

Tuning hyperparameters via grid search 186

Algorithm selection with nested cross-validation 187

Looking at different performance evaluation metrics 189

Reading a confusion matrix 190

Optimizing the precision and recall of a classification model 191

(cont. to next page)

assign 12

Plotting a receiver operating characteristic 193
The scoring metrics for multiclass classification 197
Summary 198

Chapter 7: Combining Different Models for Ensemble Learning 199

Learning with ensembles 199
Implementing a simple majority vote classifier 203
Combining different algorithms for classification with majority vote 210
Evaluating and tuning the ensemble classifier 213
Bagging – building an ensemble of classifiers from bootstrap samples 219
Leveraging weak learners via adaptive boosting 224
Summary 232

Chapter 8: Applying Machine Learning to Sentiment Analysis 233

Obtaining the IMDb movie review dataset 233
Introducing the bag-of-words model 236
Transforming words into feature vectors 236
Assessing word relevancy via term frequency-inverse document frequency 238
Cleaning text data 240
Processing documents into tokens 242
Training a logistic regression model for document classification 244
Working with bigger data – online algorithms and out-of-core learning 246
Summary 250

Chapter 9: Embedding a Machine Learning Model into a Web Application 251

Serializing fitted scikit-learn estimators 252
Setting up a SQLite database for data storage 255
Developing a web application with Flask 257
Our first Flask web application 258
Form validation and rendering 259
Turning the movie classifier into a web application 264
Deploying the web application to a public server 272
Updating the movie review classifier 274
Summary 276