

Table of Contents

Preface.....	xi
--------------	----

Part I. Fundamentals of Unsupervised Learning

1. Unsupervised Learning in the Machine Learning Ecosystem.....	3
Basic Machine Learning Terminology	3
Rules-Based vs. Machine Learning	4
Supervised vs. Unsupervised	5
The Strengths and Weaknesses of Supervised Learning	5
The Strengths and Weaknesses of Unsupervised Learning	6
Using Unsupervised Learning to Improve Machine Learning Solutions	7
A Closer Look at Supervised Algorithms	10
Linear Methods	11
Neighborhood-Based Methods	12
Tree-Based Methods	13
Support Vector Machines	14
Neural Networks	15
A Closer Look at Unsupervised Algorithms	15
Dimensionality Reduction	15
Clustering	18
Feature Extraction	19
Unsupervised Deep Learning	20
Sequential Data Problems Using Unsupervised Learning	23
Reinforcement Learning Using Unsupervised Learning	23
Semisupervised Learning	24
Successful Applications of Unsupervised Learning	25
Anomaly Detection	25

Conclusion	26
2. End-to-End Machine Learning Project.	27
Environment Setup	27
Version Control: Git	27
Clone the Hands-On Unsupervised Learning Git Repository	28
Scientific Libraries: Anaconda Distribution of Python	28
Neural Networks: TensorFlow and Keras	28
Gradient Boosting, Version One: XGBoost	29
Gradient Boosting, Version Two: LightGBM	29
Clustering Algorithms	29
Interactive Computing Environment: Jupyter Notebook	30
Overview of the Data	30
Data Preparation	31
Data Acquisition	31
Data Exploration	32
Generate Feature Matrix and Labels Array	35
Feature Engineering and Feature Selection	36
Data Visualization	37
Model Preparation	38
Split into Training and Test Sets	38
Select Cost Function	39
Create k-Fold Cross-Validation Sets	39
Machine Learning Models (Part I)	40
Model #1: Logistic Regression	40
Evaluation Metrics	43
Confusion Matrix	44
Precision-Recall Curve	44
Receiver Operating Characteristic	46
Machine Learning Models (Part II)	48
Model #2: Random Forests	48
Model #3: Gradient Boosting Machine (XGBoost)	51
Model #4: Gradient Boosting Machine (LightGBM)	54
Evaluation of the Four Models Using the Test Set	57
Ensembles	61
Stacking	61
Final Model Selection	64
Production Pipeline	65
Conclusion	66

Part II. Unsupervised Learning Using Scikit-Learn

3. Dimensionality Reduction.....	69
The Motivation for Dimensionality Reduction	69
The MNIST Digits Database	70
Dimensionality Reduction Algorithms	74
Linear Projection vs. Manifold Learning	74
Principal Component Analysis	74
PCA, the Concept	74
PCA in Practice	75
Incremental PCA	80
Sparse PCA	81
Kernel PCA	82
Singular Value Decomposition	84
Random Projection	85
Gaussian Random Projection	85
Sparse Random Projection	86
Isomap	87
Multidimensional Scaling	89
Locally Linear Embedding	90
t-Distributed Stochastic Neighbor Embedding	91
Other Dimensionality Reduction Methods	92
Dictionary Learning	92
Independent Component Analysis	94
Conclusion	95
4. Anomaly Detection.....	97
Credit Card Fraud Detection	98
Prepare the Data	98
Define Anomaly Score Function	98
Define Evaluation Metrics	99
Define Plotting Function	101
Normal PCA Anomaly Detection	101
PCA Components Equal Number of Original Dimensions	102
Search for the Optimal Number of Principal Components	105
Sparse PCA Anomaly Detection	107
Kernel PCA Anomaly Detection	110
Gaussian Random Projection Anomaly Detection	112
Sparse Random Projection Anomaly Detection	114
Nonlinear Anomaly Detection	116
Dictionary Learning Anomaly Detection	116
ICA Anomaly Detection	118

Fraud Detection on the Test Set	120
Normal PCA Anomaly Detection on the Test Set	120
ICA Anomaly Detection on the Test Set	122
Dictionary Learning Anomaly Detection on the Test Set	124
Conclusion	125

5. Clustering..... 127

MNIST Digits Dataset	128
Data Preparation	128
Clustering Algorithms	129
k-Means	130
k-Means Inertia	130
Evaluating the Clustering Results	131
k-Means Accuracy	133
k-Means and the Number of Principal Components	134
k-Means on the Original Dataset	136
Hierarchical Clustering	138
Agglomerative Hierarchical Clustering	138
The Dendrogram	139
Evaluating the Clustering Results	141
DBSCAN	143
DBSCAN Algorithm	144
Applying DBSCAN to Our Dataset	144
HDBSCAN	146
Conclusion	147

6. Group Segmentation..... 149

Lending Club Data	149
Data Preparation	150
Transform String Format to Numerical Format	152
Impute Missing Values	152
Engineer Features	154
Select Final Set of Features and Perform Scaling	154
Designate Labels for Evaluation	155
Goodness of the Clusters	157
k-Means Application	158
Hierarchical Clustering Application	161
HDBSCAN Application	165
Conclusion	166

Part III. Unsupervised Learning Using TensorFlow and Keras

7. Autoencoders.....	169
Neural Networks	170
TensorFlow	171
Keras	172
Autoencoder: The Encoder and the Decoder	173
Undercomplete Autoencoders	173
Overcomplete Autoencoders	174
Dense vs. Sparse Autoencoders	175
Denoising Autoencoder	175
Variational Autoencoder	176
Conclusion	177
8. Hands-On Autoencoder.....	179
Data Preparation	179
The Components of an Autoencoder	182
Activation Functions	182
Our First Autoencoder	183
Loss Function	184
Optimizer	184
Training the Model	185
Evaluating on the Test Set	187
Two-Layer Undercomplete Autoencoder with Linear Activation Function	189
Increasing the Number of Nodes	193
Adding More Hidden Layers	195
Nonlinear Autoencoder	196
Overcomplete Autoencoder with Linear Activation	198
Overcomplete Autoencoder with Linear Activation and Dropout	201
Sparse Overcomplete Autoencoder with Linear Activation	203
Sparse Overcomplete Autoencoder with Linear Activation and Dropout	206
Working with Noisy Datasets	208
Denoising Autoencoder	208
Two-Layer Denoising Undercomplete Autoencoder with Linear Activation	208
Two-Layer Denoising Overcomplete Autoencoder with Linear Activation	211
Two-Layer Denoising Overcomplete Autoencoder with ReLu Activation	213
Conclusion	215
9. Semisupervised Learning.....	217
Data Preparation	217
Supervised Model	220
Unsupervised Model	222

Semisupervised Model	224
The Power of Supervised and Unsupervised	226
Conclusion	227

Part IV. Deep Unsupervised Learning Using TensorFlow and Keras

10. Recommender Systems Using Restricted Boltzmann Machines.....	231
Boltzmann Machines	231
Restricted Boltzmann Machines	232
Recommender Systems	233
Collaborative Filtering	233
The Netflix Prize	234
MovieLens Dataset	234
Data Preparation	234
Define the Cost Function: Mean Squared Error	238
Perform Baseline Experiments	239
Matrix Factorization	240
One Latent Factor	240
Three Latent Factors	242
Five Latent Factors	242
Collaborative Filtering Using RBMs	243
RBM Neural Network Architecture	243
Build the Components of the RBM Class	245
Train RBM Recommender System	247
Conclusion	249
11. Feature Detection Using Deep Belief Networks.....	251
Deep Belief Networks in Detail	251
MNIST Image Classification	252
Restricted Boltzmann Machines	254
Build the Components of the RBM Class	254
Generate Images Using the RBM Model	257
View the Intermediate Feature Detectors	257
Train the Three RBMs for the DBN	258
Examine Feature Detectors	260
View Generated Images	261
The Full DBN	265
How Training of a DBN Works	269
Train the DBN	270
How Unsupervised Learning Helps Supervised Learning	271
Generate Images to Build a Better Image Classifier	271

Image Classifier Using LightGBM	279
Supervised Only	279
Unsupervised and Supervised Solution	280
Conclusion	281
12. Generative Adversarial Networks.....	283
GANs, the Concept	283
The Power of GANs	284
Deep Convolutional GANs	284
Convolutional Neural Networks	285
DCGANs Revisited	289
Generator of the DCGAN	290
Discriminator of the DCGAN	292
Discriminator and Adversarial Models	293
DCGAN for the MNIST Dataset	293
MNIST DCGAN in Action	295
Synthetic Image Generation	296
Conclusion	297
13. Time Series Clustering.....	299
ECG Data	300
Approach to Time Series Clustering	300
k-Shape	300
Time Series Clustering Using k-Shape on ECGFiveDays	301
Data Preparation	301
Training and Evaluation	308
Time Series Clustering Using k-Shape on ECG5000	309
Data Preparation	309
Training and Evaluation	312
Time Series Clustering Using k-Means on ECG5000	314
Time Series Clustering Using Hierarchical DBSCAN on ECG5000	315
Comparing the Time Series Clustering Algorithms	316
Full Run with k-Shape	317
Full Run with k-Means	318
Full Run with HDBSCAN	319
Comparing All Three Time Series Clustering Approaches	320
Conclusion	322
14. Conclusion.....	323
Supervised Learning	324
Unsupervised Learning	324
Scikit-Learn	325

TensorFlow and Keras	325
Reinforcement Learning	326
Most Promising Areas of Unsupervised Learning Today	326
The Future of Unsupervised Learning	328
Final Words	329

Index.....	331
-------------------	------------