

# Table of Contents

<b>Preface</b>	<b>-i</b>
<b>Chapter 1: Getting Started with Python Machine Learning</b>	<b>7</b>
<b>Machine learning and Python - a dream team</b>	<b>8</b>
What the book will teach you - and what it will not	9
How to best read this book	10
What to do when you are stuck	12
Getting started	12
Introduction to NumPy, SciPy, Matplotlib, and TensorFlow	13
Installing Python	13
Chewing data efficiently with NumPy and intelligently with SciPy	14
Learning NumPy	14
Indexing	17
Handling nonexistent values	17
Comparing the runtime	18
Learning SciPy	19
Fundamentals of machine learning	20
Asking a question	21
Getting answers	21
Our first (tiny) application of machine learning	21
Reading in the data	22
Preprocessing and cleaning the data	22
Choosing the right model and learning algorithm	25
Before we build our first model	26
Starting with a simple straight line	26
Toward more complex models	28
Stepping back to go forward - another look at our data	31
Training and testing	34
Answering our initial question	36
<b>Summary</b>	<b>38</b>
<b>Chapter 2: Classifying with Real-World Examples</b>	<b>39</b>
<b>The Iris dataset</b>	<b>40</b>
Visualization is a good first step	40
Classifying with scikit-learn	42
Building our first classification model	42
<b>Evaluation - holding out data and cross-validation</b>	<b>43</b>
<b>How to measure and compare classifiers</b>	<b>47</b>
<b>A more complex dataset and the nearest-neighbor classifier</b>	<b>48</b>
Learning about the seeds dataset	48
Features and feature engineering	49
Nearest neighbor classification	50
Looking at the decision boundaries	51

<b>Which classifier to use</b>	54
<b>Summary</b>	55
<b>Chapter 3: Regression</b>	57
<b>Predicting house prices with regression</b>	57
<b>Multidimensional regression</b>	61
<b>Cross-validation for regression</b>	62
Penalized or regularized regression	63
L1 and L2 penalties	64
<b>Using Lasso or ElasticNet in scikit-learn</b>	66
Visualizing the Lasso path	66
P-greater-than-N scenarios	67
An example based on text documents	68
Setting hyperparameters in a principled way	70
<b>Regression with TensorFlow</b>	74
<b>Summary</b>	79
<b>Chapter 4: Classification I - Detecting Poor Answers</b>	81
<b>Sketching our roadmap</b>	82
<b>Learning to classify classy answers</b>	82
Tuning the instance	82
Tuning the classifier	82
<b>Fetching the data</b>	83
Slimming the data down to chewable chunks	84
Preselecting and processing attributes	84
Defining what a good answer is	86
<b>Creating our first classifier</b>	87
Engineering the features	87
Training the classifier	89
Measuring the classifier's performance	89
Designing more features	90
<b>Deciding howto improve the performance</b>	94
Bias, variance and their trade-off	95
Fixing high bias	95
Fixing high variance	95
High or low bias?	96
<b>Using logistic regression</b>	99
A bit of math with a smali example	99
Applying logistic regression to our post-classification problem	102
<b>Looking behind accuracy - precision and recall</b>	104
<b>Slimming the classifier</b>	107
<b>Ship it!</b>	108
<b>Classification using Tensorflow</b>	109
<b>Summary</b>	115

<b>Chapter 5: Dimensionality Reduction</b>	117
<b>Sketching our roadmap</b>	117
<b>Selecting features</b>	118
Detecting redundant features using filters	118
Correlation	119
Mutual information	122
Asking the model about the features using wrappers	127
Other feature selection methods	130
<b>Feature projection</b>	131
Principal component analysis	131
Sketching PCA	132
Applying PCA	131
Limitations of PCA and how LDA can help	135
<b>Multidimensional scaling</b>	138
<b>Autoencoders, or neural networks for dimensionality reduction</b>	144
<b>Summary</b>	145
<b>Chapter 6: Clustering - Finding Related Posts</b>	146
<b>Measuring the relatedness of posts</b>	146
How not to do it	146
How to do it	147
<b>Preprocessing - similarity measured as a similar number of common words</b>	147
Converting raw text into a bag of words	148
Counting words	149
Normalizing word count vectors	152
Removing less important words	153
Stemming	152
Installing and using NLTK	154
Extending the vectorizer with NLTK's stemmer	156
Stop words on steroids	156
Our achievements and goals	159
<b>Clustering</b>	159
K-means	163
Getting test data to evaluate our ideas	165
Clustering posts	165
<b>Solving our initial challenge</b>	168
Another look at noise	169
<b>Tweaking the parameters</b>	170
<b>Summary</b>	171
<b>Chapter 7: Recommendations</b>	172
<b>Rating predictions and recommendations</b>	173
<b>Splitting into training and testing</b>	169
<b>Normalizing the training data</b>	176
<b>A neighborhood approach to recommendations</b>	

<b>A regression approach to recommendations</b>	180
<b>Combining multiple methods</b>	182
<b>Basket analysis</b>	<b>184</b>
Obtaining useful predictions	186
Analyzing supermarket shopping baskets	186
<b>Association rule mining</b>	190
More advanced basket analysis	192
<b>Summary</b>	192
<b>Chapter 8: Artificial Neural Networks and Deep Learning</b>	195
<b>Using TensorFlow</b>	195
TensorFlow API	196
Graphs	197
Sessions	198
Useful operations	195
<b>Saving and restoring neural networks</b>	201
Training neural networks	203
Convolutional neural networks	204
Recurrent neural networks	214
<b>LSTM for predicting text</b>	214
<b>LSTM for image Processing</b>	219
<b>Summary</b>	221
<b>Chapter 9: Classification II - Sentiment Analysis</b>	223
<b>Sketching our roadmap</b>	223
<b>Fetching the Twitter data</b>	224
<b>Introducing the Nai've Bayes classifier</b>	224
Getting to know the Bayes theorem	225
Being nai've	224
Using Nai've Bayes to classify	227
Accounting for unseen words and other oddities	230
Accounting for arithmetic underflows	230
<b>Creating our first classifier and tuning it</b>	233
Solving an easy problem first	233
Using all classes	237
Tuning the classifier's parameters	239
<b>Cleaning tweets</b>	244
<b>Taking the word types into account</b>	246
Determining the word types	246
Successfully cheating using SentiWordNet	248
Our first estimator	250
Putting everything together	253
<b>Summary</b>	254
<b>Chapter 10: Topic Modeling</b>	255

<b>Latent Dirichlet allocation</b>	256
Building a topic model	256
Comparing documents by topic	262
Modeling the whole of Wikipedia	265
Choosing the number of topics	268
<b>Summary</b>	270
<b>Chapter 11: Classification III - Music Genre Classification</b>	271
<b>Sketching our roadmap</b>	271
<b>Fetching the music data</b>	272
Converting into WAV format	273
<b>Looking at music</b>	271
Decomposing music into sine-wave components	276
<b>Using FFT to build our first classifier</b>	278
Increasing experimentation agility	278
Training the classifier	281
Using a confusion matrix to measure accuracy in multiclass problems	
An alternative way to measure classifier performance using receiver-operator characteristics	284
<b>Improving classification performance with mel frequency cepstral coefficients</b>	286
<b>Music classification using Tensorflow</b>	291
<b>Summary</b>	298
<b>Chapter 12: Computer Vision</b>	301
<b>Introducing image processing</b>	301
Loading and displaying images	302
Thresholding	305
Gaussian blurring	305
Putting the center in focus	308
<b>Basic image classification</b>	310
<b>Computing features from images</b>	311
<b>Writing your own features</b>	314
<b>Using features to find similar images</b>	316
<b>Classifying a harder dataset</b>	318
<b>Local feature representations</b>	321
<b>Image generation with adversarial networks</b>	330
<b>Summary</b>	331
<b>Chapter 13: Reinforcement Learning</b>	331
<b>Types of reinforcement learning</b>	332
Policy and value network	333
Q-network	333
<b>Excelling at games</b>	333
A small example	

Using Tensorflow for the text game

Playing breakout

## **Summary**

## **Chapter 14: Bigger Data Learning about big data**

Using jug to break up your pipeline into tasks An introduction to tasks in jug **Looking under the hood**

**Using jug for data analysis Reusing partial results Using Amazon Web Services Creating your first virtual machines Installing Python packages on Amazon Linux Running jug on our cloud machine**

**Automating the generation of clusters with cfncluster**

## **Summary**

## **Appendix A: Where to Learn More About Machine Learning**

**Online courses**

**Books**

Blogs

Data sources Getting competitive

**Ali that was left out**

## **Summary**

## **Other Books You May Enjoy**

## **Index**