Machine Learning for Finance

Principles and practice for financial insiders

Jannes Klaas

Packt

BIRMINGHAM - MUMBAI

Table of Contents

Preface jx

Chapter 1: Neural Networks and Gradient-Based Optimization 1

Our journey in this book 2

What is machine learning? 4

Supervised learning 4

Unsupervised learning 5

Reinforcement learning 6

The unreasonable effectiveness of data 7

All models are wrong 7

Setting up your workspace 9

Using Kaggle kernels 9

Running notebooks locally 12

Installing TensorFlow 12

Installing Keras 13

Using data locally 13

Using the AWS deep learning AMI 13

Approximating functions 14

A forward pass 16

A logistic regressor 17

Python version of our logistic regressor 18

Optimizing model parameters 20

Measuring model loss 21

Gradient descent 22

Backpropagation 23

Parameter updates 25

Putting it all together 26

A deeper network 29

A brief introduction to Keras 33

Importing Keras 33

A two-layer model in Keras 33

Stacking layers 34

Compiling the model 35

Training the model 35

Keras and TensorFlow 36

Tensors and the computational graph 37

Exercises 39

Summary 39

Chapter 2: Applying Machine Learning to Structured Data 41
The data 42
Heuristic, feature-based, and E2E models 44
The machine learning software stack 46
The heuristic approach 46
Making predictions using the heuristic model 47
The F1 score 47
Evaluating with a confusion matrix 48
The feature engineering approach 50
A feature from intuition – fraudsters don't sleep 51
Expert insight – transfer, then cash out 53
Statistical quirks – errors in balances 53
Preparing the data for the Keras library 54
One-hot encoding 55
Entity embeddings 56
Tokenizing categories 57
Creating input models 57
Training the model 59
Creating predictive models with Keras 59
Extracting the target 59
Creating a test set 60
Creating a validation set 60
Oversampling the training data 61
Building the model 62
Creating a simple baseline 62
Building more complex models 65
A brief primer on tree-based methods 66
A simple decision tree 66
A random forest 67
XGBoost 69
E2E modeling 70
Exercises 70
Summary 70
Chapter 3: Utilizing Computer Vision 71
Convolutional Neural Networks 73
Filters on MNIST 73
Adding a second filter 75
Filters on color images 77
The building blocks of ConvNets in Keras 78
Conv2D 78
Kernel size 79
Stride size 79

Padding 80
Input shape 81
Simplified Conv2D notation 81
ReLU activation 81
MaxPooling2D 83
Flatten 84
Dense 84
Training MNIST 85
The model 85
Loading the data 87
Compiling and training 88
More bells and whistles for our neural network 90
Momentum 90
The Adam optimizer 91
Regularization 92
L2 regularization 93
L1 regularization 93
Regularization in Keras 94
Dropout 95
Batchnorm 97
Working with big image datasets 99
Working with pretrained models 101
Modifying VGG-16 104
Random image augmentation 105
Augmentation with ImageDataGenerator 105
The modularity tradeoff 107
Computer vision beyond classification 108
Facial recognition 109
Bounding box prediction 111
Exercises 113
Summary 113
Chapter 4: Understanding Time Series 115
Visualization and preparation in pandas 117
Aggregate global feature statistics 118
Examining the sample time series 122
Different kinds of stationarity 126
Why stationarity matters 127
Making a time series stationary 127
When to ignore stationarity issues 128
Fast Fourier transformations 129
Autocorrelation 131
Establishing a training and testing regime 134

A note on backtesting **136**
Median forecasting **137**
ARIMA **139**
Kalman filters **142**
Forecasting with neural networks **147**
Data preparation 147
Weekdays **148**
ConvID **153**
Dilated and causal convolution **154**
Simple RNN **156**
LSTM **157**
The carry 158
Recurrent dropout **160**
Bayesian deep learning **161**
Exercises **164**
Summary **165**
Chapter 5: Parsing Textual Data with Natural Language Processing **167**
An introductory guide to spaCy **168**
Named entity recognition **170**
Fine-tuning the NER 176
Part-of-speech (POS) tagging **178**
Rule-based matching **180**
Adding custom functions to matchers 182
Adding the matcher to the pipeline 184
Combining rule-based and learning-based systems 185
Regular expressions **185**
Using Python's regex module 186
Regex in pandas 187
When to use regexes and when not to 187
A text classification task **188**
Preparing the data **188**
Sanitizing characters 189
Lemmatization 190
Preparing the target 191
Preparing the training and test sets 191
Bag-of-words **191**
TF-IDF 193
Topic modeling **194**
Word embeddings **196**
Preprocessing for training with word vectors 196

Loading pretrained word vectors 199
Time series models with word vectors 202
Document similarity with word embeddings 204
A quick tour of the Keras functional API 205
Attention 209
Seq2seq models 213
Seq2seq architecture overview 213
The data 214
Encoding characters 216
Creating inference models 221
Making translations 222
Exercises 225
Summary 225
Chapter 6: Using Generative Models 227
Understanding autoencoders 228
Autoencoder for MNIST 229
Autoencoder for credit cards 232
Visualizing latent spaces with t-SNE 236
Variational autoencoders 242
MNIST example 243
Using the Lambda layer 244
Kullback-Leibler divergence 245
Creating a custom loss 247
Using a VAE to generate data 248
VAEs for an end-to-end fraud detection system 250
VAEs for time series 251
GANs 253
A MNIST GAN 255
Understanding GAN latent vectors 263
GAN training tricks 263
Using less data – active learning 266
Using labeling budgets efficiently 266
Leveraging machines for human labeling 267
Pseudo labeling for unlabeled data 268
Using generative models 269
SGANs for fraud detection 269
Exercises 276
Summary 276
Chapter 7: Reinforcement Learning for Financial Markets 277
Catch – a quick guide to reinforcement learning 278

Q-learning turns RL into supervised learning 281
Defining the Q-learning model 284
Training to play Catch 284
Markov processes and the bellman equation –
A more formal introduction to RL 287
The Bellman equation in economics 291
Advantage actor-critic models 292
Learning to balance 294
Learning to trade 306
Evolutionary strategies and genetic algorithms 310
Practical tips for RL engineering 312
Designing good reward functions 313
Careful, manual reward shaping 313
Inverse reinforcement learning 314
Learning from human preferences 314
Robust RL 315
Frontiers of RL 315
Multi-agent RL 316
Learning how to learn 317
Understanding the brain through RL 318
Exercises 319
Summary 319
Chapter 8: Privacy, Debugging, and Launching Your Products 321
Debugging data 322
How to find out whether your data is up to the task 322
What to do if you don't have enough data 324
Unit testing data 325
Keeping data private and complying with regulations 329
Preparing the data for training 332
Understanding which inputs led to which predictions 333
Debugging your model 335
Hyperparameter search with Hyperas 336
Efficient learning rate search 341
Learning rate scheduling 343
Monitoring training with TensorBoard 346
Exploding and vanishing gradients 349
Deployment 351
Launching fast 352
Understanding and monitoring metrics 352
Understanding where your data comes from 354
Performance tips 355

Using the right hardware for your problem 355
Making use of distributed training with TF estimators 355
Using optimized layers such as CuDNNLSTM 357
Optimizing your pipeline 357
Speeding up your code with Cython 360
Caching frequent requests 362
Exercises 363
Summary 363
Chapter 9: Fighting Bias 365
Sources of unfairness in machine learning 366
Legal perspectives 367
Observational fairness 368
Training to be fair 373
Causal learning 383
Obtaining causal models 384
Instrument variables 385
Non-linear causal models 386
Interpreting models to ensure fairness 388
Unfairness as complex system failure 395
Complex systems are intrinsically hazardous systems 395
Catastrophes are caused by multiple failures 396
Complex systems run in degraded mode 396
Human operators both cause and prevent accidents 396
Accident-free operation requires experience with failure 396
A checklist for developing fair models 397
What is the goal of the model developers? 397
Is the data biased? 397
Are errors biased? 397
How is feedback incorporated? 397
Can the model be interpreted? 398
What happens to models after deployment? 398
Exercises 398
Summary 398
Chapter 10: Bayesian Inference and Probabilistic Programming 401
An intuitive guide to Bayesian inference 402
Flat prior 403
<50% prior 406
Prior and posterior 407
Markov Chain Monte Carlo 409
Metropolis-Hastings MCMC 415
From probabilistic programming to deep probabilistic programming 419

Summary **420**
Farewell **421**
Further reading **422**
General data analysis 422
Sound science in machine learning 422
General machine learning 422
General deep learning 423
Reinforcement learning 423
Bayesian machine learning 423
Other Books You May Enjoy **425**
Index **429**