Feature Engineering for Machine Learning

Principles and Techniques for Data Scientists

Alice Zheng and Amanda Casari

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Feature Engineering for Machine Learning

by Alice Zheng and Amanda Casari

Copyright © 2018 Alice Zheng, Amanda Casari. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editors: Rachel Roumeliotis and Jeff Bleiel
Production Editor: Kristen Brown
Copyeditor: Rachel Head
Proofreader: Sonia Saruba

Indexer: Ellen Troutman
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

April 2018: First Edition

Revision History for the First Edition

2018-03-23: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781491953242* for release details.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Feature Engineering for Machine Learning*, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc.

While the publisher and the authors have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the authors disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-95324-2

[LSI]

Table of Contents

Preface. vii

1. The Machine Learning Pipeline. 1

Data 1

Tasks 1

Models 2

Features 3

Model Evaluation 3

2. Fancy Tricks with Simple Numbers. 5

Scalars, Vectors, and Spaces 6

Dealing with Counts 8

Binarization 9

Quantization or Binning 10

Log Transformation 15

Log Transform in Action 19

Power Transforms: Generalization of the Log Transform 23

Feature Scaling or Normalization 29

Min-Max Scaling 30

Standardization (Variance Scaling) 31

ℓ^2 Normalization 32

Interaction Features 35

Feature Selection 38

Summary 39

Bibliography 39

3. Text Data: Flattening, Filtering, and Chunking. 41

Bag-of-X: Turning Natural Text into Flat Vectors 42

Bag-of-Words 42
Bag-of-n-Grams 45
Filtering for Cleaner Features 47
Stopwords 48
Frequency-Based Filtering 48
Stemming 51
Atoms of Meaning: From Words to n-Grams to Phrases 52
Parsing and Tokenization 52
Collocation Extraction for Phrase Detection 52
Summary 59
Bibliography 60

4. The Effects of Feature Scaling: From Bag-of-Words to Tf-Idf. 61
Tf-Idf : A Simple Twist on Bag-of-Words 61
Putting It to the Test 63
Creating a Classification Dataset 64
Scaling Bag-of-Words with Tf-Idf Transformation 65
Classification with Logistic Regression 66
Tuning Logistic Regression with Regularization 68
Deep Dive: What Is Happening? 72
Summary 75
Bibliography 76

5. Categorical Variables: Counting Eggs in the Age of Robotic Chickens. 77
Encoding Categorical Variables 78
One-Hot Encoding 78
Dummy Coding 79
Effect Coding 82
Pros and Cons of Categorical Variable Encodings 83
Dealing with Large Categorical Variables 83
Feature Hashing 84
Bin Counting 87
Summary 94
Bibliography 96

6. Dimensionality Reduction: Squashing the Data Pancake with PCA. 99
Intuition 99
Derivation 101
Linear Projection 102
Variance and Empirical Variance 103
Principal Components: First Formulation 104
Principal Components: Matrix-Vector Formulation 104

General Solution of the Principal Components 105
Transforming Features 105
Implementing PCA 106
PCA in Action 106
Whitening and ZCA 108
Considerations and Limitations of PCA 109
Use Cases 111
Summary 112
Bibliography 113

7. Nonlinear Featurization via K-Means Model Stacking. 115
k-Means Clustering 117
Clustering as Surface Tiling 119
k-Means Featurization for Classification 122
Alternative Dense Featurization 127
Pros, Cons, and Gotchas 128
Summary 130
Bibliography 131

8. Automating the Featurizer: Image Feature Extraction and Deep Learning. 133
The Simplest Image Features (and Why They Don't Work) 134
Manual Feature Extraction: SIFT and HOG 135
Image Gradients 135
Gradient Orientation Histograms 139
SIFT Architecture 143
Learning Image Features with Deep Neural Networks 144
Fully Connected Layers 144
Convolutional Layers 146
Rectified Linear Unit (ReLU) Transformation 150
Response Normalization Layers 151
Pooling Layers 153
Structure of AlexNet 153
Summary 157
Bibliography 157

9. Back to the Feature: Building an Academic Paper Recommender. 159
Item-Based Collaborative Filtering 159
First Pass: Data Import, Cleaning, and Feature Parsing 161
Academic Paper Recommender: Naive Approach 161
Second Pass: More Engineering and a Smarter Model 167
Academic Paper Recommender: Take 2 167
Third Pass: More Features = More Information 173

Academic Paper Recommender: Take 3 174
Summary 176
Bibliography 177

A. Linear Modeling and Linear Algebra Basics. 179

Index. 193