
Programming Collective Intelligence

Building Smart Web 2.0 Applications

Toby Segaran

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Paris • Sebastopol • Taipei • Tokyo

Table of Contents

Preface	xiii
1. Introduction to Collective Intelligence	1
What Is Collective Intelligence?	2
What Is Machine Learning?	3
Limits of Machine Learning	4
Real-Life Examples	5
Other Uses for Learning Algorithms	5
2. Making Recommendations	7
Collaborative Filtering	7
Collecting Preferences	8
Finding Similar Users	9
Recommending Items	15
Matching Products	17
Building a del.icio.us Link Recommender	19
Item-Based Filtering	22
Using the MovieLens Dataset	25
User-Based or Item-Based Filtering?	27
Exercises	28
3. Discovering Groups	29
Supervised versus Unsupervised Learning	29
Word Vectors	30
Hierarchical Clustering	33
Drawing the Dendrogram	38
Column Clustering	40

K-Means Clustering	42
Clusters of Preferences	44
Viewing Data in Two Dimensions	49
Other Things to Cluster	53
Exercises	53
4. Searching and Ranking	54
What's in a Search Engine?	54
A Simple Crawler	56
Building the Index	58
Querying	63
Content-Based Ranking	64
Using Inbound Links	69
Learning from Clicks	74
Exercises	84
5. Optimization	86
Group Travel	87
Representing Solutions	88
The Cost Function	89
Random Searching	91
Hill Climbing	92
Simulated Annealing	95
Genetic Algorithms	97
Real Flight Searches	101
Optimizing for Preferences	106
Network Visualization	110
Other Possibilities	115
Exercises	116
6. Document Filtering	117
Filtering Spam	117
Documents and Words	118
Training the Classifier	119
Calculating Probabilities	121
A Naïve Classifier	123
The Fisher Method	127
Persisting the Trained Classifiers	132
Filtering Blog Feeds	134

Improving Feature Detection	136
Using Akismet	138
Alternative Methods	139
Exercises	140
7. Modeling with Decision Trees	142
Predicting Signups	142
Introducing Decision Trees	144
Training the Tree	145
Choosing the Best Split	147
Recursive Tree Building	149
Displaying the Tree	151
Classifying New Observations	153
Pruning the Tree	154
Dealing with Missing Data	156
Dealing with Numerical Outcomes	158
Modeling Home Prices	158
Modeling “Hotness”	161
When to Use Decision Trees	164
Exercises	165
8. Building Price Models	167
Building a Sample Dataset	167
k-Nearest Neighbors	169
Weighted Neighbors	172
Cross-Validation	176
Heterogeneous Variables	178
Optimizing the Scale	181
Uneven Distributions	183
Using Real Data—the eBay API	189
When to Use k-Nearest Neighbors	195
Exercises	196
9. Advanced Classification: Kernel Methods and SVMs	197
Matchmaker Dataset	197
Difficulties with the Data	199
Basic Linear Classification	202
Categorical Features	205
Scaling the Data	209

Understanding Kernel Methods	211
Support-Vector Machines	215
Using LIBSVM	217
Matching on Facebook	219
Exercises	225
10. Finding Independent Features	226
A Corpus of News	227
Previous Approaches	231
Non-Negative Matrix Factorization	232
Displaying the Results	240
Using Stock Market Data	243
Exercises	248
11. Evolving Intelligence	250
What Is Genetic Programming?	250
Programs As Trees	253
Creating the Initial Population	257
Testing a Solution	259
Mutating Programs	260
Crossover	263
Building the Environment	265
A Simple Game	268
Further Possibilities	273
Exercises	276
12. Algorithm Summary	277
Bayesian Classifier	277
Decision Tree Classifier	281
Neural Networks	285
Support-Vector Machines	289
k-Nearest Neighbors	293
Clustering	296
Multidimensional Scaling	300
Non-Negative Matrix Factorization	302
Optimization	304

A. Third-Party Libraries	309
B. Mathematical Formulas	316
Index	323