Algorithms of the Intelligent Web

HARALAMBOS MARMANIS
DMITRY BABENKO


1. What is the intelligent web?  1
  1.1 Examples of intelligent web applications  3
  1.2 Basic elements of intelligent applications  4
  1.3 What applications can benefit from intelligence?  6
    Social networking sites 6 • Mashups 7 • Portals 8 • Wikis 9
    Media-sharing sites 9 • Online gaming 10
  1.4 How can I build intelligence in my own application?  11
    Examine your functionality and your data 11 • Get more data from
    the web 12
  1.5 Machine learning, data mining, and all that  15
  1.6 Eight fallacies of intelligent applications  16
    Fallacy #1: Your data is reliable 17 • Fallacy #2: Inference happens
    instantaneously 18 • Fallacy #3: The size of data doesn’t matter 18
    Fallacy #4: Scalability of the solution isn’t an issue 18 • Fallacy #5:
    Apply the same good library everywhere 18 • Fallacy #6: The
    computation time is known 19 • Fallacy #7: Complicated models are
    better 19 • Fallacy #8: There are models without bias 19
4 Clustering: grouping things together 121

4.1 The need for clustering 122

User groups on a website: a case study 123 • Finding groups with a SQL order by clause 124 • Finding groups with array sorting 125

4.2 An overview of clustering algorithms 128

Clustering algorithms based on cluster structure 129 • Clustering algorithms based on data type and structure 130 • Clustering algorithms based on data size 131

4.3 Link-based algorithms 132

The dendrogram: a basic clustering data structure 132 • A first look at link-based algorithms 134 • The single-link algorithm 135 • The average-link algorithm 137 • The minimum-spanning-tree algorithm 139

4.4 The k-means algorithm 142

A first look at the k-means algorithm 142 • The inner workings of k-means 143

4.5 Robust Clustering Using Links (ROCK) 146

Introducing ROCK 146 • Why does ROCK rock? 147

4.6 DBSCAN 151

A first look at density-based algorithms 151 • The inner workings of DBSCAN 153

4.7 Clustering issues in very large datasets 157

Computational complexity 157 • High dimensionality 158

4.8 Summary 160

4.9 To Do 161

4.10 References 162

5 Classification: placing things where they belong 164

5.1 The need for classification 165

5.2 An overview of classifiers 169

Structural classification algorithms 170 • Statistical classification algorithms 172 • The lifecycle of a classifier 173

5.3 Automatic categorization of emails and spam filtering 174

NaiveBayes classification 175 • Rule-based classification 188
5.4 Fraud detection with neural networks 199
   A use case of fraud detection in transactional data 199 • Neural networks overview 201 • A neural network fraud detector at work 203
   The anatomy of the fraud detector neural network 208 • A base class for building general neural networks 214

5.5 Are your results credible? 219

5.6 Classification with very large datasets 223

5.7 Summary 225

5.8 To do 226

5.9 References 230
   Classification schemes 230 • Books and articles 230

6 Combining classifiers 232

6.1 Credit worthiness: a case study for combining classifiers 234
   A brief description of the data 235 • Generating artificial data for real problems 239

6.2 Credit evaluation with a single classifier 243
   The naïve Bayes baseline 243 • The decision tree baseline 245 • The neural network baseline 247

6.3 Comparing multiple classifiers on the same data 250
   McNemar's test 251 • The difference of proportions test 253
   Cochran's Q test and the F test 255

6.4 Bagging: bootstrap aggregating 257
   The bagging classifier at work 258 • A look under the hood of the bagging classifier 260 • Classifier ensembles 263

6.5 Boosting: an iterative improvement approach 265
   The boosting classifier at work 266 • A look under the hood of the boosting classifier 268

6.6 Summary 272

6.7 To Do 273

6.8 References 277

7 Putting it all together: an intelligent news portal 278

7.1 An overview of the functionality 280

7.2 Getting and cleansing content 281
   Get set. Get ready. Crawl the Web! 281 • Review of the search prerequisites 282 • A default set of retrieved and processed news stories 284
7.3 Searching for news stories 286
7.4 Assigning news categories 288
   Order matters! 289 • Classifying with the NewsProcessor class 294
   Meet the classifier 295 • Classification strategy: going beyond low-
   level assignments 297
7.5 Building news groups with the NewsProcessor class 300
   Clustering general news stories 301 • Clustering news stories within
   a news category 305
7.6 Dynamic content based on the user's ratings 308
7.7 Summary 311
7.8 To do 312
7.9 References 316

appendix A Introduction to BeanShell 317
appendix B Web crawling 319
appendix C Mathematical refresher 323
appendix D Natural language processing 327
appendix E Neural networks 330
index 333