

---

# Contents

Preface	<i>page xi</i>
Acronyms	xv
<b>1 Introduction</b>	<b>1</b>
1.1 The De Facto Culture	3
1.2 Motivations for This Book	6
1.3 The De Facto Approach	7
1.4 Broader Issues with Evaluation Approaches	12
1.5 What Can We Do?	16
1.6 Is Evaluation an End in Itself?	18
1.7 Purpose of the Book	19
1.8 Other Takes on Evaluation	20
1.9 Moving Beyond Classification	20
1.10 Thematic Organization	21
<b>2 Machine Learning and Statistics Overview</b>	<b>23</b>
2.1 Machine Learning Overview	23
2.2 Statistics Overview	42
2.3 Summary	72
2.4 Bibliographic Remarks	73
<b>3 Performance Measures I</b>	<b>74</b>
3.1 Overview of the Problem	75
3.2 An Ontology of Performance Measures	81
3.3 Illustrative Example	82
3.4 Performance Metrics with a Multiclass Focus	85
3.5 Performance Metrics with a Single-Class Focus	94
3.6 Illustration of the Confusion-Matrix-Only-Based Metrics Using WEKA	107

Cambridge University Press

978-0-521-19600-0 - Evaluating Learning Algorithms: A Classification Perspective

Nathalie Japkowicz and Mohak Shah

Table of Contents

[More information](#)

viii	Contents	
	3.7 Summary	108
	3.8 Bibliographic Remarks	109
<b>4</b>	<b>Performance Measures II</b>	<b>111</b>
	4.1 Graphical Performance Measures	112
	4.2 Receiver Operating Characteristic (ROC) Analysis	112
	4.3 Other Visual Analysis Methods	131
	4.4 Continuous and Probabilistic Classifiers	137
	4.5 Specialized Metrics	143
	4.6 Illustration of the Ranking and Probabilistic Approaches Using R, ROCR, and WEKA	146
	4.7 Summary	159
	4.8 Bibliographic Remarks	159
<b>5</b>	<b>Error Estimation</b>	<b>161</b>
	5.1 Introduction	163
	5.2 Holdout Approach	164
	5.3 What Implicitly Guides Resampling?	167
	5.4 Simple Resampling	171
	5.5 A Note on Model Selection	177
	5.6 Multiple Resampling	178
	5.7 Discussion	185
	5.8 Illustrations Using R	187
	5.9 Summary	202
	5.10 Bibliographic Remarks	202
	Appendix: Proof of Equation (5.5)	204
<b>6</b>	<b>Statistical Significance Testing</b>	<b>206</b>
	6.1 The Purpose of Statistical Significance Testing	207
	6.2 The Limitations of Statistical Significance Testing	210
	6.3 An Overview of Relevant Statistical Tests	213
	6.4 A Note on Terminology	215
	6.5 Comparing <i>Two Classifiers</i> on a <i>Single Domain</i>	217
	6.6 Comparing <i>Two Classifiers</i> on <i>Multiple Domains</i>	231
	6.7 Comparing <i>Multiple Classifiers</i> on <i>Multiple Domains</i>	239
	6.8 Statistical Tests for Two Classifiers on a Single Domain Based on Resampling Techniques	258
	6.9 Illustration of the Statistical Tests Application Using R	263
	6.10 Summary	289
	6.11 Bibliographic Remarks	290
<b>7</b>	<b>Datasets and Experimental Framework</b>	<b>292</b>
	7.1 Repository-Based Approach	294
	7.2 Making Sense of Our Repositories: Metalearning	300

Cambridge University Press

978-0-521-19600-0 - Evaluating Learning Algorithms: A Classification Perspective

Nathalie Japkowicz and Mohak Shah

Table of Contents

[More information](#)

Contents	ix
7.3 Artificial Data Approach	301
7.4 Community Participation: Web-Based Solutions	304
7.5 Summary	306
7.6 Bibliographic Remarks	306
<b>8 Recent Developments</b>	<b>308</b>
8.1 Performance Metrics	309
8.2 Frameworks for Performance Metrics	312
8.3 Combining Metrics	317
8.4 Insights from Statistical Learning Theory	323
8.5 Other Developments	329
8.6 Summary	330
Appendix: Proof of Theorems 8.1 and 8.2	330
<b>9 Conclusion</b>	<b>335</b>
9.1 An Evaluation Framework Template	336
9.2 Concluding Remarks	349
9.3 Bibliographic Remarks	350
<b>Appendix A: Statistical Tables</b>	<b>351</b>
A.1 The $Z$ Table	351
A.2 The $t$ Table	352
A.3 The $\chi^2$ Table	353
A.4 The Table of Critical Values for the Signed Test	355
A.5 The Wilcoxon Table	356
A.6 The $F$ -Ratio Table	357
A.7 The Friedman Table	361
A.8 The Table of Critical Values for the Tukey Test	362
A.9 The Table of Critical Values for the Dunnett Test	363
<b>Appendix B: Additional Information on the Data</b>	<b>364</b>
<b>Appendix C: Two Case Studies</b>	<b>368</b>
C.1 Illustrative Case Study 1	368
C.2 Illustrative Case Study 2	375
Bibliography	393
Index	403