Algorithmic Learning in a Random World
# Contents

1 Introduction ................................................................. 1  
  1.1 Machine Learning ..................................................... 1  
    1.1.1 Learning Under Randomness ................................. 2  
    1.1.2 Learning Under Unconstrained Randomness ............... 3  
  1.2 A Shortcoming of Statistical Learning Theory ................ 3  
    1.2.1 The Hold-Out Estimate of Confidence ...................... 4  
    1.2.2 The Contribution of This Book ............................. 4  
  1.3 The Online Framework .............................................. 5  
    1.3.1 Online Learning ............................................. 5  
    1.3.2 Online/Offline Compromises ................................. 6  
    1.3.3 One-Off and Offline Learning ............................... 6  
    1.3.4 Induction, Transduction, and the Online Framework ....... 7  
  1.4 Conformal Prediction .............................................. 9  
    1.4.1 Nested Prediction Sets ..................................... 9  
    1.4.2 Validity .................................................... 10  
    1.4.3 Efficiency .................................................. 11  
    1.4.4 Conditionality ............................................. 12  
    1.4.5 Flexibility of Conformal Predictors ....................... 12  
  1.5 Probabilistic Prediction Under Unconstrained Randomness ..... 13  
    1.5.1 Universally Consistent Probabilistic Predictor .......... 13  
    1.5.2 Probabilistic Prediction Using a Finite Dataset .......... 14  
    1.5.3 Venn Prediction .......................................... 14  
    1.5.4 Conformal Predictive Distributions ........................ 14  
  1.6 Beyond Randomness .................................................. 15  
    1.6.1 Testing Randomness ....................................... 15  
    1.6.2 Online Compression Models ................................ 15  
  1.7 Context .............................................................. 16
Part I  Set Prediction

2 Conformal Prediction: General Case and Regression ........................................ 19
  2.1 Confidence Predictors ....................................................................................... 19
    2.1.1 Assumptions ................................................................................................. 19
    2.1.2 Simple Predictors and Confidence Predictors .............................................. 20
    2.1.3 Validity .......................................................................................................... 22
    2.1.4 Randomized Confidence Predictors ............................................................... 24
    2.1.5 Confidence Predictors Over a Finite Horizon ................................................. 25
    2.1.6 One-Off and Offline Confidence Predictors ................................................ 26
  2.2 Conformal Predictors ......................................................................................... 27
    2.2.1 Bags .............................................................................................................. 27
    2.2.2 Nonconformity and Conformity ...................................................................... 28
    2.2.3 p-Values ........................................................................................................ 29
    2.2.4 Definition of Conformal Predictors ............................................................... 30
    2.2.5 Validity .......................................................................................................... 31
    2.2.6 Smoothed Conformal Predictors .................................................................... 31
    2.2.7 Finite-Horizon Conformal Prediction ............................................................ 32
    2.2.8 One-Off and Offline Conformal Predictors .................................................. 33
    2.2.9 General Schemes for Defining Nonconformity .............................................. 34
    2.2.10 Deleted Conformity Measures ..................................................................... 36
  2.3 Conformalized Ridge Regression ........................................................................ 37
    2.3.1 Least Squares and Ridge Regression ............................................................. 37
    2.3.2 Basic CRR ..................................................................................................... 38
    2.3.3 Two Modifications .......................................................................................... 42
    2.3.4 Dual Form Ridge Regression .......................................................................... 44
  2.4 Conformalized Nearest Neighbours Regression ................................................ 46
  2.5 Efficiency of Conformalized Ridge Regression .................................................. 48
    2.5.1 Hard and Soft Models .................................................................................... 49
    2.5.2 Bayesian Ridge Regression ............................................................................ 49
    2.5.3 Efficiency of CRR ........................................................................................ 50
  2.6 Are There Other Ways to Achieve Validity? ....................................................... 53
  2.7 Conformal Transducers ..................................................................................... 54
    2.7.1 Definitions and Properties of Validity ........................................................... 54
    2.7.2 Normalized Confidence Predictors and Confidence Transducers .................. 56
  2.8 Proofs .................................................................................................................. 58
    2.8.1 Proof of Theorem 2.2 .................................................................................... 58
    2.8.2 Proof of Theorem 2.7 .................................................................................... 58
    2.8.3 Proof of Theorem 2.10 .................................................................................. 63
  2.9 Context ............................................................................................................... 63
    2.9.1 Exchangeability vs Randomness .................................................................... 63
    2.9.2 Conformal Prediction ..................................................................................... 64
    2.9.3 Two Equivalent Definitions of Nonconformity Measures ............................. 66
2.9.4 The Two Meanings of Conformity in Conformal Prediction ........................................ 66
2.9.5 Examples of Nonconformity Measures .................................................. 67
2.9.6 Kernel Methods .................................................................................... 67
2.9.7 Burnaev–Wasserman Programme ...................................................... 68
2.9.8 Completeness Results ........................................................................ 68

3 Conformal Prediction: Classification and General Case ..................... 71
3.1 Criteria of Efficiency for Conformal Prediction ................................... 71
3.1.1 Basic Criteria ....................................................................................... 72
3.1.2 Other Prior Criteria ............................................................................ 72
3.1.3 Observed Criteria ............................................................................... 74
3.1.4 Idealised Setting .................................................................................. 75
3.1.5 Conditionally Proper Criteria of Efficiency ........................................ 77
3.1.6 Criteria of Efficiency that Are not Conditionally Proper ................. 78
3.1.7 Discussion ........................................................................................... 82
3.2 More Ways of Computing Nonconformity Scores ................................. 84
3.2.1 Nonconformity Scores from Nearest Neighbours ......................... 84
3.2.2 Nonconformity Scores from Support Vector Machines .................. 86
3.2.3 Reducing Classification Problems to the Binary Case .................... 86
3.3 Weak Teachers ....................................................................................... 87
3.3.1 Imperfectly Taught Predictors .............................................................. 88
3.3.2 Weak Validity ....................................................................................... 89
3.3.3 Strong Validity .................................................................................... 90
3.3.4 Iterated Logarithm Validity ................................................................. 91
3.3.5 Efficiency ............................................................................................ 91
3.4 Proofs ...................................................................................................... 92
3.4.1 Proofs for Sect. 3.1 ............................................................................ 92
3.4.2 Proofs for Sect. 3.3 ............................................................................ 99
3.5 Context .................................................................................................... 105
3.5.1 Criteria of Efficiency ......................................................................... 105
3.5.2 Examples of Nonconformity Measures ............................................. 105
3.5.3 Universal Predictors .......................................................................... 106
3.5.4 Weak Teachers .................................................................................. 106

4 Modifications of Conformal Predictors ................................................. 107
4.1 The Topics of This Chapter ..................................................................... 107
4.2 Inductive Conformal Predictors ............................................................... 108
4.2.1 Inductive Conformal Predictors in the Online Mode ....................... 108
4.2.2 Inductive Conformal Predictors in the Offline and Semi-Online Modes ................................................................. 109
4.2.3 The General Scheme for Defining Nonconformity ......................... 112
4.2.4 Normalization and Hyper-Parameter Selection ............................ 113
4.3 Further Ways of Computing Nonconformity Scores ............................ 114
4.3.1 Nonconformity Measures Considered Earlier .................................. 115
4.3.2 De-Bayesing ...................................................................................... 116
4.3.3 Neural Networks and Other Multiclass Scoring Classifiers ........................................ 117
4.3.4 Decision Trees and Random Forests ................................................................. 118
4.3.5 Binary Scoring Classifiers .............................................................................. 118
4.3.6 Logistic Regression ......................................................................................... 119
4.3.7 Regression and Bootstrap ............................................................................. 120
4.3.8 Training Inductive Conformal Predictors .................................................. 120

4.4 Cross-Conformal Prediction ........................................................................... 122
4.4.1 Definition of Cross-Conformal Predictors .................................................. 122
4.4.2 Computational Efficiency ........................................................................... 123
4.4.3 Validity and Lack Thereof for Cross-Conformal Predictors ............................. 123

4.5 Transductive Conformal Predictors ............................................................... 125
4.5.1 Definition ...................................................................................................... 125
4.5.2 Validity ........................................................................................................ 126

4.6 Conditional Conformal Predictors ................................................................. 126
4.6.1 One-Off Conditional Conformal Predictors ............................................. 127
4.6.2 Mondrian Conformal Predictors and Transducers .................................. 128
4.6.3 Using Mondrian Conformal Transducers for Prediction ............................ 130
4.6.4 Generality of Mondrian Taxonomies ...................................................... 131
4.6.5 Conformal Prediction .................................................................................. 132
4.6.6 Inductive Conformal Prediction .................................................................. 133
4.6.7 Label-Conditional Conformal Prediction .................................................. 134
4.6.8 Object-Conditional Conformal Prediction ................................................ 134

4.7 Training-Conditional Validity ........................................................................ 135
4.7.1 Conditional Validity .................................................................................... 135
4.7.2 Training-Conditional Validity of Inductive Conformal Predictors ................. 137

4.8 Context ............................................................................................................ 140
4.8.1 Computationally Efficient Hedged Prediction ....................................... 140
4.8.2 Specific Learning Algorithms and Nonconformity Measures ................. 140
4.8.3 Training Conformal Predictors .................................................................. 140
4.8.4 Cross-Conformal Predictors and Alternative Approaches ....................... 140
4.8.5 Transductive Conformal Predictors ............................................................ 141
4.8.6 Conditional Conformal Predictors .............................................................. 141

Part II Probabilistic Prediction

5 Impossibility Results ......................................................................................... 145
5.1 Introduction ...................................................................................................... 145
5.2 Diverse Datasets ............................................................................................. 146
5.3 Impossibility of Estimation of Probabilities .................................................. 147
5.3.1 Binary Case ................................................................................................. 147
5.3.2 Multiclass Case .......................................................................................... 148
5.4 Proof of Theorem 5.2 .................................................. 149
  5.4.1 Probability Estimators and Statistical Tests ............. 149
  5.4.2 Complete Statistical Tests ..................................... 150
  5.4.3 Restatement of the Theorem in Terms of Statistical Tests .................................................. 150
  5.4.4 The Proof of the Theorem .................................... 152

5.5 Context ............................................................................. 152
  5.5.1 More Advanced Results ............................................ 152
  5.5.2 Density Estimation, Regression Estimation, and Regression with Deterministic Objects .................... 153
  5.5.3 Universal Probabilistic Predictors ............................ 154
  5.5.4 Algorithmic Randomness Perspective ....................... 154

6 Probabilistic Classification: Venn Predictors .......................... 157
  6.1 Introduction ............................................................... 157
  6.2 Venn Predictors ......................................................... 158
    6.2.1 Validity of One-Off Venn Predictors ..................... 159
    6.2.2 Are There Other Ways to Achieve Perfect Calibration? 160
    6.2.3 Venn Prediction with Binary Labels and No Objects ..... 162
  6.3 A Universal Venn Predictor .......................................... 163
  6.4 Venn–Abers Predictors ................................................ 163
    6.4.1 Full Venn–Abers Predictors ................................ 163
    6.4.2 Inductive Venn–Abers Predictors ......................... 165
    6.4.3 Probabilistic Predictors Derived from Venn Predictors 166
    6.4.4 Cross Venn–Abers Predictors .............................. 168
    6.4.5 Merging Multiprobability Predictions into a Probabilistic Prediction ........................................ 169
  6.5 Proofs ........................................................................... 170
    6.5.1 Proof of Theorem 6.4 .......................................... 170
    6.5.2 PAVA and the Proof of Lemma 6.6 ......................... 172
    6.5.3 Proof of Proposition 6.7 ....................................... 173
  6.6 Context ........................................................................... 178
    6.6.1 Risk and Uncertainty ............................................ 178
    6.6.2 John Venn, Frequentist Probability, and the Problem of the Reference Class ............................ 178
    6.6.3 Online Venn Predictors Are Calibrated ................... 179
    6.6.4 Isotonic Regression .............................................. 179

7 Probabilistic Regression: Conformal Predictive Systems ............. 181
  7.1 Introduction ............................................................... 181
  7.2 Conformal Predictive Systems ..................................... 182
    7.2.1 Basic Definitions ................................................. 182
    7.2.2 Properties of Validity ......................................... 185
    7.2.3 Simplest Example: Monotonic Conformity Measures 186
    7.2.4 Criterion of Being a CPS .................................... 187
  7.3 Least Squares Prediction Machine .................................. 188
## 7.3.1 Three Kinds of LSPM .................................................. 188
## 7.3.2 The Studentized LSPM in an Explicit Form .................. 190
## 7.3.3 The Offline Version of the Studentized LSPM ............... 191
## 7.3.4 The Ordinary LSPM ................................................ 192
## 7.3.5 Asymptotic Efficiency of the LSPM ............................ 193
## 7.3.6 Illustrations ...................................................... 196
## 7.4 Kernel Ridge Regression Prediction Machine .................... 198
## 7.4.1 Explicit Forms of the KRRPM .................................. 198
## 7.4.2 Limitation of the KRRPM ...................................... 201
## 7.5 Nearest Neighbours Prediction Machine ........................ 202
## 7.6 Universal Conformal Predictive Systems ........................ 204
## 7.6.1 Definitions ...................................................... 205
## 7.6.2 Universal Conformal Predictive Systems .................... 205
## 7.6.3 Universal Deterministic Predictive Systems ................ 206
## 7.7 Applications to Decision Making ................................ 206
## 7.7.1 A Standard Problem of Decision Making ..................... 207
## 7.7.2 Examples ....................................................... 208
## 7.7.3 Asymptotically Efficient Decision Making .................. 209
## 7.7.4 Dangers of Overfitting ........................................ 209
## 7.8 Computationally Efficient Versions .............................. 210
## 7.8.1 Inductive Conformal Predictive Systems .................... 210
## 7.8.2 Cross-Conformal Predictive Distributions .................. 213
## 7.8.3 Practical Aspects .............................................. 214
## 7.8.4 Beyond Randomness ............................................ 215
## 7.9 Proofs and Calculations .......................................... 216
## 7.9.1 Proofs for Sect. 7.2 ............................................ 216
## 7.9.2 Proofs for Sect. 7.3 ............................................ 217
## 7.9.3 Proof of Theorem 7.16 ........................................ 220
## 7.9.4 Proofs for Sect. 7.8 ............................................ 221
## 7.10 Context ................................................................ 222
## 7.10.1 Conformal Predictive Distributions .......................... 222
## 7.10.2 Conformal Predictive Distributions with Kernels .......... 222
## 7.10.3 Venn Prediction for Probabilistic Regression ............. 223
## 7.10.4 Universal Consistency and Universality ................... 223
## 7.10.5 Various Notions of Convergence in Law .................... 223
## 7.10.6 Decision Theory ................................................ 223
## 7.10.7 Postprocessing of Predictive Distributions ............... 224

### Part III Testing Randomness

8 Testing Exchangeability .................................................. 227

8.1 Testing Exchangeability .............................................. 227

8.1.1 Exchangeability Supermartingales .......................... 228

8.1.2 Conformal Test Martingales .................................. 229

8.2 Testing for Concept and Label Shift in Anticausal Classification . 233
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.3</td>
<td>Multistage Nonrandomness Detection</td>
<td>239</td>
</tr>
<tr>
<td>8.3.1</td>
<td>CUSUM-Type Change Detection</td>
<td>241</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Shiryaev–Roberts Change Detection</td>
<td>241</td>
</tr>
<tr>
<td>8.4</td>
<td>When Do We Retrain?</td>
<td>243</td>
</tr>
<tr>
<td>8.4.1</td>
<td>The Ville Procedure for the Wine Quality Dataset</td>
<td>244</td>
</tr>
<tr>
<td>8.4.2</td>
<td>CUSUM and Shiryaev–Roberts for Wine Quality</td>
<td>248</td>
</tr>
<tr>
<td>8.4.3</td>
<td>Fixed and Variable Schedules</td>
<td>250</td>
</tr>
<tr>
<td>8.5</td>
<td>Proofs</td>
<td>255</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Proof of Corollary 8.9</td>
<td>255</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Proofs for Sect. 8.3</td>
<td>256</td>
</tr>
<tr>
<td>8.5.3</td>
<td>A Proof for Sect. 8.4</td>
<td>258</td>
</tr>
<tr>
<td>8.6</td>
<td>Context</td>
<td>259</td>
</tr>
<tr>
<td>8.6.1</td>
<td>Conformal Test Martingales</td>
<td>259</td>
</tr>
<tr>
<td>8.6.2</td>
<td>The Other Conformal Martingales</td>
<td>260</td>
</tr>
<tr>
<td>8.6.3</td>
<td>Optimality of Sequential Testing Procedures</td>
<td>261</td>
</tr>
<tr>
<td>8.6.4</td>
<td>Standard Procedures for Change Detection and Their Optimality</td>
<td>261</td>
</tr>
<tr>
<td>8.6.5</td>
<td>Deciding When to Retrain</td>
<td>262</td>
</tr>
<tr>
<td>8.6.6</td>
<td>Avoiding Numerical Problems</td>
<td>262</td>
</tr>
<tr>
<td>8.6.7</td>
<td>Low-Dimensional Dynamic Models</td>
<td>263</td>
</tr>
<tr>
<td>8.6.8</td>
<td>Islands of Randomness</td>
<td>263</td>
</tr>
<tr>
<td>9</td>
<td>Efficiency of Conformal Testing</td>
<td>265</td>
</tr>
<tr>
<td>9.1</td>
<td>Conformal Testing with a Known Critical Region</td>
<td>265</td>
</tr>
<tr>
<td>9.1.1</td>
<td>Randomness Probability vs Exchangeability Probability</td>
<td>266</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Conformal Probability</td>
<td>267</td>
</tr>
<tr>
<td>9.2</td>
<td>Conformal Testing with a Known Alternative</td>
<td>269</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Bayes–Kelly Algorithm</td>
<td>270</td>
</tr>
<tr>
<td>9.2.2</td>
<td>Bayes–Kelly Algorithm with a Composite Alternative</td>
<td>274</td>
</tr>
<tr>
<td>9.2.3</td>
<td>Changepoint Alternatives</td>
<td>275</td>
</tr>
<tr>
<td>9.2.4</td>
<td>More Natural Conformal Test Martingales for Changepoint Alternatives</td>
<td>282</td>
</tr>
<tr>
<td>9.2.5</td>
<td>Symmetric Markov Alternatives</td>
<td>284</td>
</tr>
<tr>
<td>9.2.6</td>
<td>Asymmetric Markov Alternatives</td>
<td>289</td>
</tr>
<tr>
<td>9.2.7</td>
<td>Mixing Over All Markov Alternatives</td>
<td>291</td>
</tr>
<tr>
<td>9.3</td>
<td>Testing the Validity of Putative Test Martingales</td>
<td>294</td>
</tr>
<tr>
<td>9.4</td>
<td>Proofs and Discussions</td>
<td>298</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Proof of Proposition 9.2</td>
<td>298</td>
</tr>
<tr>
<td>9.4.2</td>
<td>Proof of Proposition 9.5</td>
<td>299</td>
</tr>
<tr>
<td>9.4.3</td>
<td>An Alternative Interpretation of the Bayes–Kelly Algorithm and the Proof of Proposition 9.9</td>
<td>300</td>
</tr>
<tr>
<td>9.5</td>
<td>Context</td>
<td>301</td>
</tr>
<tr>
<td>9.5.1</td>
<td>Randomness and Exchangeability: Algorithmic Perspective</td>
<td>301</td>
</tr>
<tr>
<td>9.5.2</td>
<td>Testing Exchangeability and Randomness</td>
<td>303</td>
</tr>
</tbody>
</table>
## 11.5 Proofs

- **11.5.1 Proof of Theorem 11.2** ........................................... 353
- **11.5.2 Proof of Proposition 4.4** ..................................... 355
- **11.5.3 Proof of Theorem 8.8** ........................................ 355
- **11.5.4 Proof of Proposition 11.6** .............................. 357

## 11.6 Context

- **11.6.1 Kolmogorov’s Programme** .................................. 359
- **11.6.2 Repetitive Structures** ....................................... 360
- **11.6.3 Exchangeability Model** ..................................... 361
- **11.6.4 Gaussian Models** ............................................ 362
- **11.6.5 Markov Model** .................................................. 362

## 12 Generalized Venn Prediction and Hypergraphical Models

- **12.1 Introduction** ..................................................... 363
- **12.2 Venn Prediction and Venn-Conditional Conformal Prediction in One-Off Structures** ........................................... 364
  - **12.2.1 Venn Prediction** ............................................. 364
  - **12.2.2 Venn-Conditional Conformal Prediction** .............. 365
- **12.3 Hypergraphical Models** ....................................... 366
  - **12.3.1 Hypergraphical Repetitive Structures** .................. 367
  - **12.3.2 Generality of Finitary Repetitive Structures** .......... 368
  - **12.3.3 Generality of Hypergraphical Models** .................. 368
  - **12.3.4 Junction-Tree Models** ................................... 369
  - **12.3.5 Combinatorics of Junction-Tree Models** .............. 370
  - **12.3.6 Shuffling Datasets** ...................................... 371
  - **12.3.7 Decomposability in Junction-Tree Models** ............ 372
- **12.4 Venn Prediction in Hypergraphical Models** ............ 372
  - **12.4.1 Prediction in Junction-Tree Models** ................... 373
  - **12.4.2 Universality of the Fully Object-Conditional Venn Predictor** .................................. 374
  - **12.4.3 An Experiment with a Simple Causal Network** ....... 375
- **12.5 Conformal Prediction in Hypergraphical Models** ........ 378
  - **12.5.1 Conditional Probability Conformity Measure** ....... 378
  - **12.5.2 LED Datasets** ............................................ 379
  - **12.5.3 Hypergraphical Assumptions for LED Datasets** ....... 380
  - **12.5.4 Results for the LED Datasets** ........................... 381
  - **12.5.5 Label-Conditional Conformal Prediction for Hypergraphical OCMs** .............................. 384
  - **12.5.6 Further Results for the LED Datasets** ................. 385
  - **12.5.7 Summary of the Experiments** .......................... 388
- **12.6 Context** ....................................................... 388
  - **12.6.1 Conformal Prediction in Hypergraphical Models** .... 388
  - **12.6.2 Additive Models** ........................................ 389

## 13 Contrasts and Perspectives

- **13.1 Introduction** .................................................... 391
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.2 Inductive Learning</td>
<td>392</td>
</tr>
<tr>
<td>13.2.1 Jacob Bernoulli’s Learning Problem</td>
<td>393</td>
</tr>
<tr>
<td>13.2.2 Statistical Learning Theory</td>
<td>396</td>
</tr>
<tr>
<td>13.2.3 The Quest for Data-Dependent Bounds</td>
<td>399</td>
</tr>
<tr>
<td>13.2.4 The Hold-Out Estimate</td>
<td>401</td>
</tr>
<tr>
<td>13.3 Transductive Learning</td>
<td>402</td>
</tr>
<tr>
<td>13.3.1 Student and Fisher</td>
<td>402</td>
</tr>
<tr>
<td>13.3.2 Conformal Predictive Distributions in the Gaussian Model</td>
<td>405</td>
</tr>
<tr>
<td>13.3.3 Tolerance Regions</td>
<td>405</td>
</tr>
<tr>
<td>13.3.4 Dempster–Hill Procedure</td>
<td>407</td>
</tr>
<tr>
<td>13.3.5 Transduction in Statistical Learning Theory</td>
<td>409</td>
</tr>
<tr>
<td>13.3.6 Validity and Efficiency in Probabilistic Prediction</td>
<td>412</td>
</tr>
<tr>
<td>13.4 Bayesian Learning</td>
<td>413</td>
</tr>
<tr>
<td>13.4.1 Bayesian Ridge Regression</td>
<td>413</td>
</tr>
<tr>
<td>13.4.2 An Illustration</td>
<td>415</td>
</tr>
<tr>
<td>13.5 Further Development of Conformal Prediction</td>
<td>417</td>
</tr>
<tr>
<td>13.5.1 Moving Beyond Exchangeability</td>
<td>418</td>
</tr>
<tr>
<td>13.5.2 Causality</td>
<td>418</td>
</tr>
<tr>
<td>13.5.3 Anomaly Detection</td>
<td>419</td>
</tr>
<tr>
<td>13.6 Proof of Proposition 13.3</td>
<td>420</td>
</tr>
<tr>
<td>13.7 Context</td>
<td>421</td>
</tr>
<tr>
<td>13.7.1 Inductive Prediction</td>
<td>421</td>
</tr>
<tr>
<td>13.7.2 Transductive Prediction</td>
<td>421</td>
</tr>
<tr>
<td>13.7.3 Online Transduction</td>
<td>422</td>
</tr>
<tr>
<td>13.7.4 Bayesian Prediction</td>
<td>422</td>
</tr>
<tr>
<td>13.7.5 Conformal Prediction: Further Developments</td>
<td>422</td>
</tr>
<tr>
<td>A Probability Theory</td>
<td>423</td>
</tr>
<tr>
<td>A.1 Basics</td>
<td>423</td>
</tr>
<tr>
<td>A.1.1 Kolmogorov’s Axioms</td>
<td>423</td>
</tr>
<tr>
<td>A.1.2 Convergence</td>
<td>425</td>
</tr>
<tr>
<td>A.2 Independence and Products</td>
<td>425</td>
</tr>
<tr>
<td>A.2.1 Products of Probability Spaces</td>
<td>426</td>
</tr>
<tr>
<td>A.2.2 Randomness Model</td>
<td>427</td>
</tr>
<tr>
<td>A.3 Expectations and Conditional Expectations</td>
<td>427</td>
</tr>
<tr>
<td>A.4 Markov Kernels and Regular Conditional Distributions</td>
<td>429</td>
</tr>
<tr>
<td>A.5 Exchangeability</td>
<td>431</td>
</tr>
<tr>
<td>A.5.1 De Finetti’s Representation Theorem</td>
<td>431</td>
</tr>
<tr>
<td>A.5.2 Conditional Probabilities Given a Bag</td>
<td>432</td>
</tr>
<tr>
<td>A.6 Theory of Martingales</td>
<td>433</td>
</tr>
<tr>
<td>A.6.1 Basic Results</td>
<td>434</td>
</tr>
<tr>
<td>A.6.2 Limit Theorems</td>
<td>435</td>
</tr>
<tr>
<td>A.6.3 Hoeffding’s Inequality</td>
<td>436</td>
</tr>
<tr>
<td>A.7 Bibliographical Remarks</td>
<td>437</td>
</tr>
</tbody>
</table>
A.7.1 Conditional Probabilities ........................................... 438  
A.7.2 Martingales ............................................................. 438  
A.7.3 Hoeffding’s Inequality .............................................. 438  

B  Datasets and Computations ............................................. 439  
B.1 USPS Dataset .............................................................. 439  
B.2 Wine Quality Dataset .................................................... 440  
B.3 Bank Marketing Dataset ................................................ 440  
B.4 Normalization ............................................................... 440  
B.5 Randomization and Reshuffling ...................................... 441  
B.6 Bibliographical Remarks ................................................ 442  

C  FAQ ............................................................................. 443  
C.1 Unusual Features of Conformal Prediction ....................... 443  
C.2 Independence of Errors .................................................. 444  
C.3 Ways of Applying Conformal-Type Predictors ................... 445  
C.4 Conformal Prediction vs Standard Methods ...................... 446  

References ........................................................................ 449  
Index .................................................................................. 467
Preface to the Second Edition

The second edition contains three new chapters, and the old chapters have been revised, updated, and, in many cases, expanded. The approximate correspondence between the chapters in the first and second editions is summarized in Table 1. These are the main changes in the second edition:

- The structure of the book has become clearer. Now it is divided into four parts, the first two dealing with prediction under the assumption of randomness, the third part devoted to testing the assumption of randomness, and the fourth devoted to prediction under generalized randomness.
- A new chapter, Chap. 7, has been added in Part II. It treats a new subject, conformal predictive distributions and their application to decision-making.
- The part of the book devoted to testing (Part III) has been greatly expanded. It has two new chapters, Chaps. 9 and 10.
- Sections 3.2–3.5 of the first edition have been removed, since they are based on criteria of efficiency for conformal prediction different from the ones that we recommend in the second edition. This material has been replaced by a thorough analysis of various criteria of efficiency (Sect. 3.1 in the second edition).

<table>
<thead>
<tr>
<th>Table 1</th>
<th>The chapters in the second edition (from 1 to 13) and their counterparts in the first edition (given in parentheses, from 1 to 10); “corr. to” stands for “roughly corresponds to”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 (corr. to 1)</td>
</tr>
<tr>
<td></td>
<td>2 (corr. to 2)  3 (corr. to 3)  4 (corr. to 4)</td>
</tr>
<tr>
<td>Part I</td>
<td>5 (corr. to 5)  6 (corr. to 6)  7 (new)</td>
</tr>
<tr>
<td>Part II</td>
<td>8 (corr. to 7)  9 (new)  10 (new)</td>
</tr>
<tr>
<td>Part III</td>
<td></td>
</tr>
<tr>
<td>Part IV</td>
<td>11 (corr. to 8)  12 (corr. to 9)  13 (corr. to 10)</td>
</tr>
</tbody>
</table>

xvii
• We have simplified the definition of validity of Venn prediction (Chap. 6 of the second edition, corresponding to Chap. 6 of the first edition). The price to pay is that our new notion of validity is much more basic; however, it is also more accessible to readers who are not familiar with game-theoretic probability [317, 320].

The development of the theory of conformal prediction has been active, especially over the last several years. The goal of covering all new developments would have been too ambitious, and in the new edition, we concentrated on the work (often our own) that is most closely connected with the material of the first edition. Some of the new developments are briefly described in a new section, Sect. 13.5.

In the new edition, all misprints and errors in the first edition, known to us today, have been corrected. We are grateful to Stijn Vanderlooy, Fabio Stella, David Lindsay, Misha Dashevsky, and Dima Devetyarov for pointing out some of them. Discussions and correspondence with numerous colleagues have led to several clarifications and are gratefully appreciated; those colleagues include Dmitry Adamskiy, Ernst Ahlberg, Anastasios Angelopoulos, Alexander Balinsky, Rina Foygel Barber, Stephen Bates, David Bell, Tony Bellotti, Claus Bendtsen, Henrik Boström, Brian Burford, Evgeny Burnaev, Emanuël Candès, Lars Carlsson, Giovanni Cherubin, Alexey Chervonenkis, Dave Cohen, Nicolo Colombo, David Cooper, A. Philip Dawid, Wang Di, Thomas Dietterich, David L. Dow, Martin Eklund, Paul Embrechts, Ola Enkqvist, Valentina Fedorova, Matteo Fontana, Patrizio Giovannotti, Leo Gordon, Danqiao Guo, Yuri Gurevich, Jan Hannig, Glenn Hawe, Emily Hector, Alan Hutchinson, Drago Indjic, Ian Jacobs, Ulf Johansson, Wouter Koolen, Antonis Lambrou, Martin Larsson, Steffen Lauritzen, Rikard Laxhammar, Jing Lei, Lihua Lei, Pitt Lim, Xiaohui Liu, Philip M. Long, Zhiyuan Luo, Valery Manokhin, Ryan Martin, Lutz Mattner, Sally McLean, Robert C. Merton, Ilya Muchnik, Khuong Nguyen, Ulf Norinder, Ilia Nourdetdinov, Nell Painter, Nicola Paolelli, Harris Papadopoulos, Dusko Pavlovic, Ivan Petej, Aaditya Ramdas, Daljit Rehal, Yaniv Romano, Johannes Ruf, Alessio Sancetta, Teddy Seidenfeld, Jieli Shen, Evgeni Smirnov, James Smith, Ryan Tibshirani, Paolo Toccaeceli, Alexandre Tsybakov, Vladimir Vapnik, Jesus Vega, Denis Volkonskiy, George Vostrov, Vladimir V’yugin, Ruodu Wang, Larry Wasserman, Chris Watkins, Peter Westfall, Bob Williamson, Minge Xie, Meng Yang, Fedor Zhdanov, Chenzhe Zhou, and Andrej Zukov Gregoric. Thanks to the members of the \TeX-\LaTeX Stack Exchange community, first of all David Carlisle, Phelype Oleinik, and Werner for help with \LaTeX. In our computational experiments, we have used \texttt{scikit-learn} [274].

We thank the following companies and funding bodies for generous financial support: AFOSR, Amazon Web Services, AstraZeneca, BBSRC, BMBF, Centrica, Cyprus Research Promotion Foundation, EPSRC, EU Horizon 2020 Research and Innovation programme, Hewlett-Packard, Leverhulme Magna Carta Doctoral

London, UK
London, UK
Newark, NJ, USA
July 2022

Vladimir Vovk
Alexander Gammerman
Glenn Shafer
Preface to the First Edition

This book is about prediction algorithms that learn. The predictions these algorithms make are often imperfect, but they improve over time, and they are hedged: they incorporate a valid indication of their own accuracy and reliability. In most of the book we make the standard assumption of randomness: the examples the algorithm sees are drawn from some probability distribution, independently of one another. The main novelty of the book is that our algorithms learn and predict simultaneously, continually improving their performance as they make each new prediction and find out how accurate it is. It might seem surprising that this should be novel, but most existing algorithms for hedged prediction first learn from a training dataset and then predict without ever learning again. The few algorithms that do learn and predict simultaneously do not produce hedged predictions. In later chapters we relax the assumption of randomness to the assumption that the data come from an online compression model. We have written the book for researchers in and users of the theory of prediction under randomness, but it may also be useful to those in other disciplines who are interested in prediction under uncertainty.

This book has its roots in a series of discussions at Royal Holloway, University of London, in the summer of 1996, involving AG, Vladimir Vapnik and VV. Vapnik, who was then based at AT&T Laboratories in New Jersey, was visiting the Department of Computer Science at Royal Holloway for a couple of months as a part-time professor. VV had just joined the department, after a year at the Center for Advanced Study in Behavioral Sciences at Stanford. AG had become the head of department in 1995 and invited both Vapnik and VV to join the department as part of his programme (enthusiastically supported by Norman Gowar, the college principal) of creating a machine learning centre at Royal Holloway. The discussions were mainly concerned with Vapnik’s work on support vector machines, and it was then that it was realized that the number of support vectors used by such a machine could serve as a basis for hedged prediction.

Our subsequent work on this idea involved several doctoral students at Royal Holloway. Ilia Nouretdinov has made several valuable theoretical contributions. Our other students working on this topic included Craig Saunders, Tom Melluish, Kostas Proedrou, Harris Papadopoulos, David Surkov, Leo Gordon, Tony Bellotti,
Daniil Ryabko, and David Lindsay. The contribution of Yura Kalnishkan and Misha Vyugin to this book was less direct, mainly through their work on predictive complexity, but still important. Thank you all.

GS joined the project only in the autumn of 2003, although he had earlier helped develop some of its key ideas through his work with VV on game-theoretic probability; see their joint book—*Probability and Finance: It's Only a Game!*—published in 2001.

Steffen Lauritzen introduced both GS and VV to repetitive structures. In VV’s case, the occasion was a pleasant symposium organized and hosted by Lauritzen in Aalborg in June 1994. We have also had helpful conversations with Masafumi Akahira, Satoshi Aoki, Peter Bramley, John Campbell, Alexey Chervonenkis, Philip Dawid, José Gonzáles, Thore Graepel, Gregory Gutin, David Hand, Fumiyasu Komaki, Leonid Levin, Xiao Hui Liu, George Loizou, Zhiyuan Luo, Per Martin-Löf, Sally McClean, Boris Mirkin, Fionn Murtagh, John Shawe-Taylor, Sasha Shen’, Akimichi Takemura, Kei Takeuchi, Roger Thatcher, Vladimir V’yugin, David Waltz, and Chris Watkins.

Many ideas in this book have their origin in our attempts to understand the mathematical and philosophical legacy of Andrei Nikolaevich Kolmogorov. Kolmogorov’s algorithmic notions of complexity and randomness (especially as developed by Martin-Löf and Levin) have been for us the main source of intuition, although they almost disappeared in the final version. VV is grateful to Andrei Nikolaevich for steering him in the direction of compression modelling and for his insistence on its independent value.

We thank the following bodies for generous financial support: EPSRC through grants GR/L35812, GR/M14937, GR/M16856, and GR/R46670; BBSRC through grant 111/BIO14428; MRC through grant S505/65; the Royal Society; the European Commission through grant IST-1999-10226; NSF through grant 5-26830.

University of London (VV, AG, and GS)  
Rutgers University (GS)  
July 2004

Vladimir Vovk  
Alexander Gammerman  
Glenn Shafer
Notation and Abbreviations

Sets, Bags, and Sequences

∅                   The empty set
N                   The positive integer numbers, {1, 2, ...}
N₀                  The nonnegative integer numbers, {0, 1, ...}
Z                   The integer numbers
Q                   The rational numbers
R                   The real numbers
Rₚ                  The extended real numbers, R ∪ {−∞, ∞}
{z₁, ..., zₙ}       Set (each element enters only once)
{z₁, ..., zₙ}       Bag (can contain more than one copy of the same element); Sect. 2.2.1
(z₁, ..., zₙ)       Sequence (the parentheses and commas may be omitted)
□                    The empty sequence
[a₁, ..., aₙ]       The set of all infinite continuations of a finite sequence a₁, ..., aₙ
|A|                  The size of a set or bag A
Zⁿ                   The set of all sequences of elements of Z of length n
Zⁿ(ₙ)               The set of all bags of elements of Z of size n
Z*                  The set of all finite sequences of elements of Z
Z*(*)               The set of all bags (always finite) of elements of Z
Z∞                   The set of all infinite sequences of elements of Z
Z²                   The set of all subsets of a set Z
Y²                   The set of all functions of the type X → Y

Stochastics

P                    Probability
E                    Expectation
\[ \mathbf{P}(Z) \text{ The set of all probability distributions on } Z \text{ (measurable space)} \]
\[ \mathbf{B}_\delta \text{ The Bernoulli distribution on } \{0, 1\} \text{ with parameter } \delta: \mathbf{B}_\delta\{1\} = \delta \text{ and } \mathbf{B}_\delta\{0\} = 1 - \delta \]
\[ \text{Ber}(\pi) \text{ The power distribution } \mathbf{B}_\infty^{\pi} \text{ on } \{0, 1\}^\infty \]
\[ \mathbf{U} \text{ The uniform distribution on } [0, 1] \]
\[ \mathbf{N}_{\mu, \sigma^2} \text{ The normal distribution on } \mathbb{R} \text{ with mean } \mu \text{ and variance } \sigma^2 \text{ (often written as } \mathbf{N}(\mu, \sigma^2) \text{)} \]
\[ \Phi \text{ The distribution function of the standard normal distribution } \mathbf{N}_{0,1} \]
\[ \phi \text{ The density function of the standard normal distribution } \mathbf{N}_{0,1} \]
\[ t_{\delta,n} \text{ The } (1 - \delta)\text{-quantile of the } t\text{-distribution: } \mathbb{P}\{\xi \geq t_{\delta,n}\} = \delta, \text{ where } \xi \text{ has Student’s } t\text{-distribution with } n \text{ degrees of freedom} \]
\[ z_\delta \text{ The } (1 - \delta)\text{-quantile of the standard normal distribution: } \mathbb{P}\{\xi \geq z_\delta\} = \delta, \text{ where } \xi \text{ has the normal distribution } \mathbf{N}_{0,1} \]
\[ Q_X \text{ Marginal distribution of } Q \text{ on } X \text{ (Sect. 3.1.4)} \]
\[ Q_{Y|X} \text{ Conditional } Q\text{-distribution on } Y \text{ given } x \in X \text{ (Sect. 3.1.5)} \]
\[ \Omega \leftrightarrow Z \text{ Markov kernel from } \Omega \text{ to } Z \text{ (Sect. A.4)} \]
\[ P f^{-1} \text{ The image of } P \text{ under a mapping } f \text{ (Sect. A.1.1)} \]
\[ \rho \text{ Variation distance between probability distributions (Sect. 7.6)} \]
\[ \xrightarrow{\text{law}} \text{ Convergence in law (Sect. A.1.2)} \]

**Machine Learning**

- **X** Object space (Sect. 2.1.1)
- **Y** Label space, \(|Y| > 1\) (Sect. 2.1.1)
- **Z** The example space (\(Z = X \times Y\), Sect. 2.1.1)
- **H** Feature space (with the feature mapping \(F: X \rightarrow H\); Sect. 2.3.4)

**Programming**

- \(+\) is used in the sense of Python: \(a += b\) is equivalent to \(a := a + b\)
- \(-\) \(a -= b\) is equivalent to \(a := a - b\)

**Confidence Prediction**

- \(\epsilon\) Significance level
- \(\Gamma_n^\epsilon\) The prediction set at trial \(n\) (Sect. 2.1.3)
- \(\text{err}_n^\epsilon\) The indicator of error at trial \(n\) (Eq. (2.8))
- \(\text{Err}_n^\epsilon\) The cumulative number of errors up to trial \(n\) (Eq. (2.9))
OE$^\varepsilon_n$  The cumulative observed excess up to trial $n$ (Protocol 3.1)
OE$^\varepsilon_n$  The cumulative observed fuzziness up to trial $n$ (Protocol 3.1)
$\tau_n$  The $n$th random number used by a randomized confidence predictor (Sect. 2.1.4)

Other Notations

$1_A$  The indicator function of a set or property $A$, i.e., $1_A = 1$ if $A$ holds and $1_A = 0$ if not
$f|_A$  The restriction of a function or Markov kernel $f$ to a subset $A$ of its domain
diam $A$  The diameter (largest distance between points) of $A$
$\text{co } A$  The convex hull of a set $A$ in a linear space
$u \cdot v$  The scalar product of vectors $u$ and $v$
$I_n$  The identity $n \times n$ matrix ($n$ is omitted if clear from the context)
$X'$  Matrix $X$ transposed
$X^{-1}$  The inverse of matrix $X$
$\text{rank } X$  The rank of matrix $X$
$u \vee v$  The maximum of $u$ and $v$, also denoted $\max(u, v)$
$u \wedge v$  The minimum of $u$ and $v$, also denoted $\min(u, v)$
$u^+$  $u \vee 0$
$u^-$  $(-u) \vee 0$
$F(t^-)$  The limit of $F(u)$ as $u$ approaches $t$ from below
$F(t^+)$  The limit of $F(u)$ as $u$ approaches $t$ from above
$f_n = O(g_n)$  $\lim \sup_{n \to \infty} (f_n / g_n) < \infty$ (used for $f_n, g_n > 0$)
$f_n = \Theta(g_n)$  $f_n = O(g_n)$ and $g_n = O(f_n)$

Abbreviations

AA  Aggregating Algorithm
ACP  Adjusted conditional probability (idealized conformity measure)
ASP  Adjusted signed predictability (idealized conformity measure)
a.s.  Almost surely (i.e., with probability one)
BJK  Bayes–Kelly–Jeffreys (martingale)
BRR  Bayesian ridge regression
CCP  Cross-conformal predictor
CCPS  Cross-conformal predictive system
CLS  Conformalized least squares
CP  Conditional probability (idealized conformity measure)
CPD  Conformal predictive distribution
CPS  Conformal predictive system
CRR  Conformalized ridge regression
CSD  Cumulative sum diagram
CTM  Conformal test martingale
DIR  Direct isotonic regression
DPS  Deterministic predictive system
FOCVP Fully object-vonditional Venn predictor
GCM  Greatest convex minorant
ICP  Inductive conformal predictor
ICPS Inductive conformal predictive system
IID  Independent and identically distributed
KRRPM Kernel ridge regression prediction machine
LCCT Label-conditional conformal transducer
LSPM Least squares prediction machine
MCP  Mondrian conformal predictor
MCT  Mondrian conformal transducer
NNR  Nearest neighbours regression
OCM  Online compression model
OOS  One-off structure
PAVA Pair-adjacent violators algorithm
PDMS Predictive decision-making system
RPD  Randomized predictive distribution
RPS  Randomized predictive system
SP   Signed predictability (idealized conformity measure)
SVM  Support vector machine
USPS For the USPS dataset, see Sect. B.1
w.r. with respect
Chapter 1
Introduction

Abstract In this introductory chapter, we sketch the existing work in machine learning on which we build and then outline the contents of the book.

Keywords Machine learning · Conformal prediction · Venn prediction · Assumption of randomness

1.1 Machine Learning

The rapid development of computer technology during the last several decades has made it possible to solve ever more difficult problems in a wide variety of fields. The development of software has been essential to this progress. The painstaking programming in machine code or assembly language that was once required to solve even simple problems has been replaced by programming in high-level object-oriented languages. We are concerned with the next natural step in this progression—the development of programs that can learn, i.e., automatically improve their performance with experience.

The need for programs that can learn was already recognized by Alan Turing [358] in 1950, who argued that it may be too ambitious to write from scratch programs for tasks that even humans must learn to perform. Consider, for example, the problem of recognizing hand-written digits. We are not born able to perform this task, but we learn to do it quite robustly. Even when the hand-written digit is represented as a grey-scale matrix, as in Fig. 1.1, we can recognize it easily, and our ability to do so scarcely diminishes when it is slightly rotated or otherwise perturbed. We do not know how to write instructions for a computer that will produce equally robust performance.

The essential difference between a program that implements instructions for a particular task and a program that learns is adaptability. A single learning program may be able to learn a wide variety of tasks: recognizing hand-written digits and faces, diagnosing patients in a hospital, estimating house prices, etc.

Recognition, diagnosis, and estimation can all be thought of as special cases of prediction. A person or a computer is given certain information and asked to
predict the answer to a question. A broad discussion of learning would go beyond prediction to consider the problems faced by a robot, who needs to act as well as predict. The literature on machine learning, has emphasized prediction, however, and the research reported in this book is in that tradition. We are mostly interested in algorithms that learn to predict well.

1.1.1 Learning Under Randomness

One learns from experience. This is as true for a computer as it is for a human being. In order for there to be something to learn there must be some stability in the environment; it must be governed by constant, or evolving only slowly, laws. And when we learn to predict well, we may claim to have learned something about that environment.

The traditional way of making the idea of a stable environment precise is to assume that it generates a sequence of examples randomly from some fixed probability distribution, say $Q$, on a fixed space of possible examples, say $Z$. These mathematical objects, $Z$ and $Q$, describe the environment.

The environment can be very complex; $Z$ can be large and structured in a complex way. This is illustrated by the USPS dataset from which Fig. 1.1 is drawn (see Sect. B.1 in Appendix B). Here an example is any $16 \times 16$ image with 31 shades of grey, together with the digit the image represents (an integer between 0 to 9). So there are $31^{16 \times 16} \times 10$ (this is approximately $10^{382}$) possible examples in the space $Z$.

In most of this book, we assume that each example consists of an object and its label. In the USPS dataset, for example, an object is a grey-scale matrix like the one in Fig. 1.1, and its label is the integer between 0 and 9 represented by the grey-scale matrix.

In the problem of recognizing hand-written digits and other typical machine-learning problems, it is the space of objects, the space of possible grey-scale images, that is large. The space of labels is either a small finite set (in what is called classification problems) or the set of real numbers (regression problems).

When we say that the examples are chosen randomly from $Q$, we mean that they are independent in the sense of probability theory and all have the distribution $Q$. They are independent and identically distributed (IID). We call this the randomness assumption (or, sometimes, statistical randomness assumption, when we want to emphasize that it is not related to Kolmogorov’s algorithmic randomness).

Of course, not all work in machine learning is concerned with learning under randomness. In learning with expert advice, for example, randomness is replaced...
1.2 A Shortcoming of Statistical Learning Theory

by a game-theoretic set-up [45, 381]; here a typical result is that the learner can predict almost as well as the best strategy in a pool of possible strategies. In reinforcement learning, which is concerned with rational decision-making in a dynamic environment [351], the standard assumption is Markovian. In this book, we will consider extensions of learning under randomness in Parts III and IV.

1.1.2 Learning Under Unconstrained Randomness

Sometimes we make the randomness assumption without assuming anything more about the environment: we know the space of examples $Z$, we know that examples are drawn independently from the same distribution, and this is all we know. We know nothing at the outset about the probability distribution $Q$ from which each example is drawn. In this case, we say we are learning under unconstrained randomness. Most of the work in this book, like much other work in machine learning, is concerned with learning under unconstrained randomness.

The strength of modern machine-learning methods often lies in their ability to make hedged predictions under unconstrained randomness in a high-dimensional environment, where examples have a very large (or infinite) number of components. We already mentioned the USPS dataset, where each example consists of 257 components ($16 \times 16$ pixels and the label). In machine learning, this number is now considered small, and the problem of learning from the USPS dataset is sometimes regarded as a toy problem.

1.2 A Shortcoming of Statistical Learning Theory

Machine learning has made significant strides in its study of learning under unconstrained randomness. We now have a wide range of algorithms that often work very well in practice: neural networks, decision trees, nearest neighbours algorithms, and naive Bayes methods have been used for decades; newer algorithms include support vector machines, random forests, and boosting, and in recent years neural networks have again become extremely popular in the form of deep learning.

From a theoretical point of view, traditional machine learning’s most significant contributions to learning under unconstrained randomness fall under the umbrella of statistical learning theory. This theory, which began with the discovery of VC dimension by Vapnik and Chervonenkis in the late 1960s and was partially rediscovered independently by Valiant [359] in 1984, has produced both deep mathematical results and learning algorithms that work well in practice (see Vapnik [366] for a review).

Given a “training” set of examples, a prediction algorithm of machine learning produces what we call a prediction rule—a function mapping the objects into the labels. This is illustrated in Fig. 1.4 (where “prediction rule” is referred to as
“general rule”). Usually, the value taken by a prediction rule on a new object is a
*simple prediction*—a guess that is not accompanied by any statement concerning
how accurate it is likely to be. Statistical learning theory does guarantee, however,
that, under some conditions on the prediction algorithm, as the training set becomes
bigger and bigger these predictions will become more and more accurate with
greater and greater probability: they are *probably approximately correct*.

How probably and how approximately? This question has not been answered as
well as we might like by statistical learning theory. This is because the theoretical
results that might be thought to answer it, the bounds that demonstrate arbitrarily
good accuracy with sufficiently large sizes of the training set, are usually too loose
to tell us anything interesting for training sets that we actually have. This happens
in spite of the empirical fact that the predictions often perform very well in practice.
Consider, for example, the problem of recognizing hand-written digits, which we
have already discussed. Here we are interested in giving an upper bound on the
probability that our learning algorithm fails to choose the right digit; we might like
this probability to be less than 0.05, for example, so that we can be 95% confident
that the prediction is correct. Unfortunately, typical upper bounds on the probability
of error provided by the theory, even for relatively clean datasets such as the USPS
dataset we have discussed, are greater than 1; bounds less than 1 can usually be
achieved only for very straightforward problems or with very large datasets. This
is true even for newer results in which the bound on the accuracy depends on the
training set (cf. Sect. 13.2.3).

### 1.2.1 The Hold-Out Estimate of Confidence

Fortunately, there are less theoretical and more effective ways of estimating the
confidence we should have in predictions output by machine-learning algorithms.
One of the most effective is the oldest and most naive: the “hold-out” estimate.
In order to compute this estimate, we split the available examples into two parts,
a training set and a hold-out sample, playing the role of a test set. We apply the
algorithm to the training set in order to find a prediction rule, and then we apply this
prediction rule to the hold-out sample. The observed rate of errors on the hold-out
sample tells us how confident we should be in the prediction rule when we apply it
to new examples (for details, see Sect. 1.2.1).

### 1.2.2 The Contribution of This Book

When we use a hold-out sample to obtain a bound on the probability of error, or
when we use an error bound from statistical learning theory, we are *hedging* the
prediction—we are adding to it a statement about how strongly we believe it. In this
book, we develop a different way of producing hedged predictions. Aside from the
1.3 The Online Framework

The new methods presented in this book are quite general; they can be tried out, at least, in almost any problem of learning under randomness. The framework in which we introduce, and then study, these methods is somewhat unusual, however. Most previous theoretical work in machine learning has been in an offline framework: one uses a batch of old examples to produce a prediction rule, which is then applied to new examples. We begin instead with a framework that is online: one makes predictions sequentially, basing each new prediction on all the previous examples instead of repeatedly using a rule constructed from a fixed batch of examples.

1.3.1 Online Learning

Our basic framework is online because we assume that the examples are presented one by one. Each time, we observe the object and predict the label. Then we observe the label and go on to the next example. We start by observing the first object $x_1$ and predicting its label $y_1$. Then we observe $y_1$ and the second object $x_2$, and predict its label $y_2$. And so on. At the $n$th step, we have observed the previous examples
\[(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\]  

(1.1)

and the new object \(x_n\), and our task is to predict \(y_n\). The quality of our predictions should improve as we accumulate more and more old examples. This is the sense in which we are learning.

### 1.3.2 Online/Offline Compromises

The methods we present in this book are most naturally described and are most amenable to mathematical analysis in the online framework. However, they also extend to relaxations of the online protocol that make it close to the offline setting, and this is important, because most practical problems have at least some offline aspects. If we are concerned with recognizing hand-written zip codes, for example, we cannot always rely on a human teacher to tell us the correct interpretation of each hand-written zip code; why not use such an ideal teacher directly for prediction? The relaxation of the online protocol considered in Sect. 3.3 includes “slow teachers”, who provide the feedback with a delay, and “lazy teachers”, who provide feedback only occasionally. In the example of zip codes recognition, this relaxation allows us to replace constant supervision by using returned letters for teaching or by occasional lessons.

### 1.3.3 One-Off and Offline Learning

We will sometimes discuss prediction algorithms in the simplest possible setting, which we will refer to as one-off learning. A one-off predictor is given a training set (1.1) and a test object \(x_n\); its task is to predict the label \(y_n\) of \(x_n\). It is applied only once.

A one-off predictor can be embedded into both online and offline frameworks (see Fig. 1.3). In the online framework, new examples arrive sequentially, their labels are predicted given their objects, and they are added to the training set, as described earlier.

![Fig. 1.3 One-off, online, and offline learning: two ways of embedding one-off predictors into a bigger learning protocol](image-url)