Table des matières

Remerciements – Acknowledgements 3

Au lecteur non-mathématicien 5

Avertissement 7

Foreword 9

Chapitre 1. Introduction 15
1.1. Rééchantillonnage 16
1.1.1. L'heuristique d'Efron 17
1.1.2. Un peu d'histoire 19
1.1.3. Champs d'application 20
1.1.4. Avantages et limites 20
1.1.5. Résultats théoriques 21
1.2. Sélection de modèles 23
1.2.1. Cadre de la prédiction 23
1.2.2. Sélection de modèles 28
1.2.3. Pénalisation 33
1.2.4. Contributions de la thèse 38
1.3. Sélection de modèles par rééchantillonnage 39
1.3.1. Validation croisée 39
1.3.2. Pénalités par rééchantillonnage 43
1.3.3. Contributions de la thèse 46
1.4. Régions de confiance et tests par rééchantillonnage 51
1.4.1. Position du problème 51
1.4.2. Motivations pratiques 56
1.4.3. Méthodes par rééchantillonnage 59
1.4.4. Contributions de la thèse 62

Notations 65
Conventions 65
Abbreviations 65
Mathematical notations (Chap. 2 to 9) 65
Histogram framework (Sect. 6.3, p. 152 and Sect. 6.8.1, p. 171) 67
Confidence regions and tests (Chap. 10) 67

Chapter 2. Optimal model selection 69
2.1. Model selection for prediction 70
2.1.1. The prediction framework 70
2.1.2. Sharp oracle inequalities 71

2.2. A gap between theory and practice 74
2.2.1. Theory: hold-out, aggregation and local Rademacher complexities 74
2.2.2. Practice: V-fold cross-validation 77
2.3. Accurate calibration of penalties 78
2.3.1. A practical algorithm for calibration of penalties 79
2.3.2. The slope heuristics 79
2.3.3. Our contributions 80
2.4. Contributions on V-fold and other resampling procedures 81
2.4.1. Performances of V-fold cross-validation 81
2.4.2. V-fold and resampling penalties 82

Chapter 3. Slope heuristics 87
3.1. Introduction 87
3.2. Framework 88
3.2.1. Regression 88
3.2.2. Model selection 88
3.2.3. Histograms 89
3.3. Theoretical results 90
3.3.1. Optimal penalties 90
3.3.2. Minimal penalties 92
3.3.3. Comments 92
3.4. Practical use of slope heuristics: data-driven penalties 93
3.4.1. Computation of $\widehat{K}_{\min}$ 93
3.4.2. Definition of $\widehat{K}_{\min}$ 93
3.4.3. Shape of the penalty 94
3.4.4. Large number of models 95
3.5. Proofs 95
3.5.1. Proof of Thm. 3.1 95
3.5.2. Proof of Thm. 3.2 98
3.5.3. Proof of Prop. 3.1 100
3.5.4. Concentration inequalities used in the main proofs 101
3.5.5. Additional results needed 102
3.5.6. Proof of Prop. 3.3 102
3.5.7. Proof of Prop. 3.4 103

Chapter 4. Limitations of linear penalties 107
4.1. Introduction 107
4.2. Non-linearity of the ideal penalty in the histogram framework 108
4.3. Suboptimality of linear penalization 109
4.4. Simulation study 109
4.4.1. Framework of Prop. 4.1 110
4.4.2. Twelve more experiments 110
4.5. Proofs 117

Chapter 5. V-fold cross-validation 119
5.1. Introduction 119
5.2. Performance of V-fold cross-validation 121
5.2.1. General framework 121

5.2.2. The histogram regression case 121
5.3. An alternative V-fold algorithm: V-fold penalties 123
5.3.1. Definition of V-fold penalties 123
5.3.2. Expectations 126
5.3.3. Theoretical results for histograms 127
5.4. Simulations 128
5.4.1. Experimental setup 128
5.4.2. Results and comments 129
5.4.3. Additional experiments 131
5.5. Discussion 131
5.6. Probabilistic tools 133
5.6.1. Expectations of inverses of binomials 133
5.6.2. Concentration of inverses of multinomials 133
5.6.3. Moment inequalities for some U-statistics 134
5.7. Proofs 135
5.7.1. Notations 135
5.7.2. Expectations 135
5.7.3. Proof of Thm. 5.1 140
5.7.4. Concentration results 143
5.7.5. Technical lemmas 146
5.7.6. Expectation of inverses of binomials 147

Chapter 6. Resampling penalties 149
6.1. Introduction 149
6.2. A general model selection algorithm 151
6.3. The histogram regression case 152
6.3.1. A modified algorithm for histograms 152
6.3.2. Explicit formulas 153
6.3.3. Examples of resampling weights 154
6.4. Main results 156
6.4.1. Oracle inequalities 156
6.4.2. Adaptivity to the hölderian regularity 158
6.5. Simulations 160
6.5.1. Experimental setup 160
6.5.2. Results and comments 160
6.6. Discussion 162
6.6.1. Practical implementation 162
6.6.2. Comparison with other procedures 166
6.6.3. General frameworks, including classification 169
6.6.4. Conclusion 169
6.7. Probabilistic tools: expectations of inverses 170
6.7.1. Binomial case 170
6.7.2. Hypergeometric case 170
6.7.3. Poisson case 171
6.7.4. Numerical illustration 171
6.8. Proofs 171
6.8.1. General framework 171

6.8.2. Proof of Thm. 6.1 174
6.8.3. Proof of Thm. 6.1: alternative assumptions 174
6.8.4. Proof of Thm. 6.2 175
6.8.5. Proof of Lemma 6.4 177
6.8.6. Resampling constants 181
6.8.7. Concentration inequalities 183
6.8.8. Technical lemmas 186
6.8.9. Expectations of inverses 188

Chapter 7. The classification case 193
7.1. Introduction 193
7.2. Framework 195
7.2.1. General framework 195
7.2.2. Main assumptions 196
7.2.3. Binary classification 197
7.2.4. Bounded regression 197
7.2.5. Resampling schemes 198
7.3. Main results 198
7.3.1. A recipe of oracle inequalities 198
7.3.2. Concentration inequalities 199
7.3.3. Program for further research 200
7.4. Practical application 201
7.4.1. Tuning parameters 202
7.4.2. Comparison with other penalties 203
7.5. Proofs 204
7.5.1. Proof of Prop. 7.2 204
7.5.2. Proof of Thm. 7.1 204
7.5.3. $\varepsilon_{\star,m}$ as a function of n and m 205

Chapter 8. Appendix on resampling penalties 207
8.1. Uniqueness and existence of $\widehat{s}_m^W$ 207
8.1.1. Existence 207
8.1.2. Uniqueness: histogram case 207
8.1.3. Uniqueness: general case 208
8.2. Resampling and structural constraints on the penalties 209
8.3. Other assumption sets for oracle inequalities for RP 211
8.3.1. Bounded case 211
8.3.2. Unbounded case 211
8.3.3. Sufficient condition for $(\mathbf{A}\delta)$ 212
8.4. Resampling Penalties with general weights 213
8.4.1. Expectations 213
8.4.2. Concentration inequality 214
8.5. Useful concentration inequalities 215
8.6. Moments, Exponential moments and Concentration 216
8.6.1. Exponential moments $\Rightarrow$ concentration 216
8.6.2. Moments $\Rightarrow$ concentration 217
8.6.3. Moments *vs.* Exponential moments 217
8.6.4. Concentration $\Rightarrow$ moments 217

8.7. Expectation of inverses: symmetric case 218
Lower bound 219
Small values of n 219
Upper bound 219
8.8. Concentration of inverses of multinomials: proofs 219
Lower deviations 220
Upper deviations 220
8.9. Moment inequalities for some U-statistics 223
Moments of Z_1 223
Moments of Z_2 223
8.10. Approximation properties of histograms 225
8.10.1. Results 225
8.10.2. Proofs 227

Chapter 9. On the constant in front of global penalties 233
9.1. Introduction 233
9.2. Lower bounds on $R_Z(\mathcal{F})$ 235
9.2.1. A factor 2 between theory and practice 235
9.2.2. Tight theoretical bounds 235
9.3. Upper bounds on $R_Z(\mathcal{F})$ 237
9.3.1. Two points classes 237
9.3.2. Independent classes 238
9.3.3. Rademacher complexities 239
9.4. Discussion 239
9.5. Proofs 240

Chapter 10. Resampling-based confidence regions and multiple tests 245
10.1. Introduction 246
10.1.1. Goals and motivations 246
10.1.2. Two approaches to our goal 247
10.1.3. Relation to previous work 247
10.1.4. Notations 248
10.2. Confidence region using concentration 249
10.2.1. Main result 249
10.2.2. Comparison in expectation 251
10.2.3. Concentration around the expectation 251
10.2.4. Resampling weight vectors 252
10.2.5. Practical computation of the thresholds 253
10.3. Confidence region using resampled quantiles 255
10.3.1. Main result 255
10.3.2. Practical computation of the resampled quantile 256
10.4. Application to multiple testing 257
10.4.1. Multiple testing and connection with confidence regions 258
10.4.2. Background on step-down procedures 259
10.4.3. Using our confidence regions to build step-down procedures 260
10.4.4. Uncentered quantile approach for two-sided testing 260
10.5. Simulations 262
10.5.1. Confidence balls 263

10.5.2. Multiple testing 264
10.6. Discussion and concluding remarks 266
10.6.1. Confidence regions and tests 266
10.6.2. FWER versus FDR in multiple testing 267
10.6.3. About the variances of the coordinates 267
10.6.4. Conclusion 267
10.7. Proofs 268
10.7.1. Confidence regions using concentration 268
10.7.2. Quantiles 271
10.7.3. Multiple testing 274
10.7.4. Exchangeable resampling computations 274
10.7.5. Non-exchangeable weights 276

Chapter 11. Conclusions, open problems and prospects 281
11.1. Why should resampling be used? 281
11.1.1. A general purpose device 281
11.1.2. A naturally adaptive device 282
11.1.3. Choice of the resampling scheme 283
11.2. Advances in the non-asymptotic theory of resampling 284
11.2.1. Concentration results 284
11.2.2. Expectations 285
11.3. Optimal calibration of penalties 285
11.3.1. Flexibility of penalization 285
11.3.2. Calibration of resampling penalties with the slope heuristics 286
11.3.3. Need for overpenalization 287
11.3.4. Larger families of models 288
11.3.5. Global penalties 289
11.4. Confidence regions and multiple testing 289
11.4.1. Quantiles without additional term 289
11.4.2. Unknown noise-level 290
11.4.3. Non-gaussian or asymmetric data 290

Bibliographie 291