Table des matières

Avant-propos **xxi**

Partie I De l'analyse de données à l'apprentissage machine 1

Chapitre 1 : Concepts fondamentaux en analyse de données et en apprentissage machine **3**

1.1 Terminologie relative aux données . . . 7
1.2 Paradigmes d'AM . . . 8
1.2.1 Apprentissage supervisé . . . 9
1.2.2 Apprentissage non supervisé . . . 10
1.2.3 Apprentissage par renforcement . . . 10
1.2.4 Paradigmes complémentaires ou hybrides . . . 12
1.3 Tâches et techniques de l'AM . . . 13
1.3.1 Classification . . . 14
1.3.2 Régression . . . 15
1.3.3 Regroupement . . . 15
1.3.4 Identification d'associations . . . 15
1.3.5 Détection d'anomalies . . . 15
1.3.6 Apprentissage de politiques de décision . . . 16
1.4 Cadre conceptuel de l'AM . . . 16
1.4.1 Modèles, hypothèses et données . . . 16
1.4.2 Généralisation et surapprentissage . . . 18
1.4.3 Compromis biais–variance . . . 19
1.4.4 Validation et sélection de modèles . . . 19
1.5 Développement d'un modèle d'AM . . . 20
1.5.1 Phase d'entraînement . . . 21
1.5.2 Phase d'inférence . . . 24
1.6 Sélection de caractéristiques . . . 25
1.7 Partitionnement des données . . . 30
1.7.1 Données d'entraînement, de test et de validation . . . 31
1.7.2 Proportions typiques de partitionnement . . . 31
1.7.3 Reproductibilité . . . 33
1.7.4 Stratification . . . 33
1.8 Validation croisée . . . 35

1.8.1 Validation croisée k-fold . . . 36
1.8.2 Validation croisée Leave-One-Out . . . 36
1.8.3 Validation croisée de Monte Carlo . . . 37
1.8.4 Validation croisée One-Subject-Out . . . 37
1.9 Évaluation d'un système d'AM . . . 39
1.9.1 Métriques pour les modèles de classification . . . 39
1.9.2 Métriques pour les modèles de regroupement . . . 47
1.9.3 Métriques pour les modèles de régression . . . 50

Chapitre 2 : Préparation et prétraitement des données 53
2.1 Nature des données . . . 55
2.1.1 Variables qualitatives ou catégorielles . . . 56
2.1.2 Variables quantitatives ou numériques . . . 57
2.1.3 Structure des données . . . 57
2.2 Nettoyage et typage des données . . . 60
2.3 Données manquantes . . . 62
2.3.1 Types de données manquantes . . . 62
2.3.2 Suppression des données manquantes . . . 63
2.3.3 Imputation des données manquantes . . . 64
2.4 Valeurs aberrantes . . . 71
2.4.1 Détection basée sur le score-z . . . 72
2.4.2 Détection basée sur l'intervalle interquartile . . . 74
2.4.3 Représentations graphiques pour la détection de valeurs aberrantes . . . 75
2.4.4 Traitement des valeurs aberrantes . . . 77
2.5 Transformation des variables . . . 79
2.5.1 Mise à l'échelle des variables numériques . . . 79
2.5.2 Transformation des distributions asymétriques . . . 82
2.5.3 Suppression des doublons . . . 83
2.5.4 Encodage des variables catégorielles . . . 84
2.5.5 Discrétisation des variables continues . . . 87
2.6 Visualisation des données . . . 88
2.6.1 Visualisation des distributions . . . 88
2.6.2 Visualisation des corrélations . . . 95

Chapitre 3 : Réduction de la dimension 103
3.1 Motivations et approches de la réduction de dimension . . . 105
3.1.1 Objectifs de la réduction de dimension . . . 106
3.1.2 Le fléau de la dimension . . . 107
3.1.3 Panorama des méthodes de réduction de dimension . . . 108
3.2 Analyse en composantes principales (ACP) . . . 110
3.2.1 Approche géométrique . . . 111
3.2.2 Approche algébrique . . . 114
3.2.3 Composantes principales . . . 116
3.2.4 Choix du nombre de composantes principales . . . 120
3.2.5 Projection des données et interprétation des scores . . . 126
3.3 Analyse discriminante linéaire (ADL) . . . 127
3.3.1 Principe de l'ADL . . . 128

3.3.2 Matrices de dispersion . . . 128
3.3.3 Projection des données . . . 129
3.4 Positionnement multidimensionnel (MDS) . . . 132
3.4.1 Principe et formulation . . . 132
3.4.2 Fonction de stress et optimisation . . . 134
3.4.3 Choix du nombre de dimensions . . . 137
3.5 t-Distributed Stochastic Neighbor Embedding (t-SNE) . . . 139
3.5.1 Modélisation probabiliste en grande dimension . . . 140
3.5.2 Perplexité . . . 140
3.5.3 Similarités en basse dimension . . . 141
3.5.4 Optimisation . . . 141
3.5.5 Limites fondamentales de t-SNE . . . 142

Partie II Apprentissage supervisé 147

Chapitre 4 : Modèles de régression 149
4.1 Terminologie . . . 151
4.2 Régression linéaire . . . 151
4.2.1 Fonction de coût . . . 154
4.2.2 Entraînement du modèle de régression linéaire . . . 156
4.2.3 Méthode analytique . . . 157
4.2.4 Méthode itérative . . . 163
4.2.5 Variantes de la descente de gradient . . . 166
4.2.6 Évaluation par généralisation . . . 173
4.3 Régression polynomiale . . . 174
4.3.1 Transformation des variables . . . 175
4.3.2 Entraînement du modèle de régression polynomiale . . . 175
4.4 Régularisation . . . 179
4.4.1 Régression ridge . . . 180
4.4.2 Régression lasso . . . 181
4.4.3 Elastic net . . . 181
4.5 Régression logistique . . . 182
4.5.1 Fonction de coût : perte logarithmique . . . 185
4.5.2 Entraînement du modèle de régression logistique . . . 186
4.5.3 Évaluation du modèle de régression logistique . . . 193
4.6 Régression logistique multinomiale . . . 194
4.6.1 Fonction de coût : entropie croisée . . . 195
4.6.2 Entraînement de la régression logistique multinomiale . . . 196

Chapitre 5 : Méthodes bayésiennes en apprentissage machine 201
5.1 Fondements probabilistes . . . 203
5.2 Théorème de Bayes . . . 207
5.2.1 Règle de décision . . . 208
5.2.2 Fonction de coût et risque conditionnel . . . 212
5.2.3 Classificateur à taux d'erreur minimal . . . 214
5.2.4 Fonctions discriminantes . . . 214
5.3 Inférence bayésienne . . . 217

5.3.1 Estimation par maximum a posteriori (MAP) . . . 218
5.3.2 Estimation à erreur quadratique moyenne minimale (MMSE) . . . 218
5.3.3 Distribution prédictive . . . 219
5.3.4 Lois conjuguées et modèle Beta-Bernoulli . . . 219
5.4 Classificateur naïf bayésien . . . 221
5.4.1 Hypothèse d'indépendance conditionnelle . . . 222
5.4.2 Variante gaussienne . . . 222
5.4.3 Variante multinomiale . . . 222
5.4.4 Variante de Bernoulli . . . 223
5.5 Réseaux bayésiens . . . 226
5.5.1 Structure et factorisation . . . 226
5.5.2 Inférence dans un réseau bayésien . . . 227

Chapitre 6 : Arbres de décision **231**
6.1 Structure d'un arbre de décision et définitions . . . 233
6.2 Critères de construction des arbres de décision . . . 236
6.2.1 Critères pour la classification . . . 237
6.2.2 Critères pour les arbres de régression . . . 247
6.2.3 Entraînement des arbres de décision . . . 251
6.3 Algorithmes classiques d'arbres de décision . . . 256
6.3.1 Algorithme ID3 . . . 256
6.3.2 Algorithme C4.5 . . . 257
6.3.3 Algorithme CART . . . 257
6.3.4 Algorithme CHAID . . . 261
6.4 Propriétés statistiques et limites des arbres de décision . . . 262
6.4.1 Complexité algorithmique . . . 262
6.4.2 Biais et variance . . . 263
6.4.3 Consistance et convergence . . . 263
6.4.4 Surapprentissage et sous-apprentissage . . . 264
6.4.5 Sensibilité aux données . . . 266
6.4.6 Limites structurelles . . . 268
6.5 Arbres de décision et émergence des modèles ensemblistes . . . 270

Chapitre 7 : La méthode des k plus proches voisins **273**
7.1 Formalisation . . . 275
7.1.1 Notation et principe . . . 275
7.1.2 Le choix de la distance . . . 276
7.1.3 Le vote pondéré par la distance . . . 276
7.1.4 Le k-PPV comme estimateur de la probabilité a posteriori . . . 277
7.2 La sensibilité aux échelles des variables . . . 278
7.3 Le choix du nombre de voisins . . . 283
7.3.1 Du surapprentissage au sous-apprentissage . . . 283
7.3.2 Sélection de k par validation croisée . . . 284
7.4 Aspects algorithmiques et limites . . . 285
7.4.1 Le coût de la prédiction . . . 285
7.4.2 Le fléau de la dimension . . . 286
7.4.3 Autres limites pratiques . . . 287

7.5 Les k plus proches voisins en régression 287

Chapitre 8 : Machines à vecteurs de support **291**
8.1 Principe des machines à vecteurs de support 293
8.1.1 Hyperplans séparateurs et intuition géométrique 293
8.1.2 Principe de la marge maximale 294
8.1.3 Notion de vecteurs de support 295
8.2 SVM linéaire à marge rigide 295
8.2.1 Formulation primale 295
8.2.2 Formulation duale et conditions d'optimalité 296
8.3 Marge souple et cas non séparables 299
8.3.1 Variables d'écart et problème primal pénalisé 299
8.3.2 Perte charnière 300
8.3.3 Rôle du paramètre de régularisation 301
8.4 Astuce du noyau et noyaux usuels 303
8.4.1 Motivation et principe 303
8.4.2 Condition de Mercer 304
8.4.3 Noyaux usuels 304
8.5 Machines à vecteurs de support pour la régression 308
8.5.1 Tube ε et perte ε-insensible 308
8.5.2 Formulation du problème 308
8.6 Aspects pratiques 310
8.6.1 Classification multiclasse 310
8.6.2 Standardisation des données 311
8.6.3 Choix des hyperparamètres 312
8.7 Propriétés, comparaisons et limites 313
8.7.1 Forces des SVM 313
8.7.2 Limites des SVM 314

Chapitre 9 : Réseaux de neurones artificiels **317**
9.1 Perceptron 319
9.1.1 Structure du perceptron 319
9.1.2 Entraînement du perceptron 322
9.1.3 Limites du perceptron 325
9.2 Perceptron multicouche 326
9.2.1 Structure du MLP 326
9.2.2 Entraînement du perceptron multicouche 327
9.2.3 Fonctions d'activation 332
9.3 Perceptron multicouche « profond » 338

Chapitre 10 : Méthodes ensemblistes **345**
10.1 Méthodes de bagging 347
10.1.1 Principe du bootstrap 348
10.1.2 Effet sur la variance : justification 348
10.1.3 Bagging en régression 349
10.1.4 Bagging en classification 349
10.1.5 Remarque sur l'erreur Out-of-Bag 350
10.1.6 Forêts aléatoires 351

10.2 Méthodes de boosting . . . 354
10.2.1 Principe général du boosting . . . 354
10.2.2 AdaBoost . . . 355
10.2.3 Gradient boosting . . . 356
10.2.4 XGBoost (eXtreme Gradient Boosting) . . . 357
10.2.5 Limites du boosting . . . 357
10.3 Méthodes de stacking . . . 359
10.3.1 Principe du stacking . . . 359
10.3.2 Construction des jeux d'apprentissage . . . 359
10.3.3 Choix du méta-apprenant . . . 360
10.3.4 Blending . . . 360
10.3.5 Vote d'apprenants hétérogènes . . . 360
10.4 Comparaison des méthodes ensemblistes . . . 362
10.4.1 Choix du nombre d'apprenants . . . 363
10.4.2 Principaux hyperparamètres . . . 365
10.4.3 Quand éviter les méthodes ensemblistes . . . 365

Chapitre 11 : Interprétabilité et explicabilité des modèles d'apprentissage machine 367
11.1 Enjeux de l'explicabilité . . . 369
11.1.1 Contrôle de la cohérence décisionnelle . . . 369
11.1.2 Acceptabilité sociale et opérationnelle . . . 369
11.1.3 Détection des biais et équité . . . 369
11.1.4 Responsabilité éthique et juridique . . . 370
11.2 Approches de l'explicabilité . . . 370
11.2.1 Approches intrinsèques et post-hoc . . . 370
11.2.2 Explications globales et locales . . . 370
11.2.3 Approches agnostiques et spécifiques au modèle . . . 371
11.3 Importance par permutation et dépendance partielle . . . 374
11.3.1 Importance par permutation . . . 374
11.3.2 Dépendance partielle . . . 375
11.4 LIME (Local Interpretable Model-agnostic Explanations) . . . 379
11.5 SHAP (SHapley Additive exPlanations) . . . 381
11.5.1 Importance globale par agrégation des valeurs SHAP . . . 384
11.5.2 Explication locale d'une prédiction avec SHAP . . . 385
11.6 LRP (Layer-wise Relevance Propagation) . . . 387
11.7 Explications par l'exemple . . . 392
11.7.1 Explications contrefactuelles . . . 392
11.7.2 Prototypes et observations influentes . . . 394
11.8 Limites et bonnes pratiques . . . 395
11.9 L'explicabilité au-delà de l'apprentissage supervisé . . . 396

Partie III Apprentissage non supervisé 399

Chapitre 12 : Techniques de regroupement 401
12.1 Principes fondamentaux du regroupement . . . 403
12.1.1 Compacité intra-cluster . . . 403

12.1.2 Séparation inter-clusters . . . 404
12.1.3 Rôle de la distance et de la représentation des données . . . 405
12.2 Méthodes de partitionnement . . . 406
12.2.1 Principe du regroupement par partitionnement . . . 407
12.2.2 Algorithme des k-moyennes . . . 407
12.2.3 Autres algorithmes de partitionnement . . . 414
12.3 Méthodes hiérarchiques . . . 418
12.3.1 Principes généraux . . . 418
12.3.2 Regroupement hiérarchique agglomératif . . . 418
12.3.3 Regroupement hiérarchique divisif . . . 419
12.3.4 Avantages et limites . . . 420
12.4 Méthodes basées sur la densité . . . 423
12.4.1 Principe général . . . 424
12.4.2 Algorithme DBSCAN . . . 424
12.4.3 Extensions et variantes . . . 426
12.4.4 Avantages et limites . . . 426
12.5 Évaluation et validation des regroupements . . . 427
12.5.1 Mesures internes . . . 427
12.5.2 Mesures externes . . . 428
12.5.3 Validation relative et sélection du nombre de clusters . . . 429
12.5.4 Limites et enjeux . . . 431

Chapitre 13 : Règles d'association 433
13.1 Cadre formel et notations . . . 435
13.2 Définition d'une règle d'association . . . 436
13.3 Évaluation des règles d'association . . . 437
13.3.1 Support . . . 437
13.3.2 Confiance . . . 438
13.3.3 Lift . . . 439
13.3.4 Autres mesures d'intérêt . . . 439
13.4 Extraction des ensembles d'items fréquents . . . 440
13.4.1 Principe général . . . 440
13.4.2 Algorithmes classiques d'extraction . . . 444
13.4.3 Représentations condensées . . . 450
13.5 Génération des règles d'association . . . 450
13.5.1 Processus de génération des règles d'association . . . 450
13.5.2 Interprétation et sélection des règles . . . 451
13.6 Propriétés, avantages et limites . . . 456
13.6.1 Avantages . . . 456
13.6.2 Limites . . . 456
13.6.3 Positionnement parmi les méthodes d'AM . . . 457
13.6.4 Des règles d'association aux systèmes de recommandation . . . 457

Chapitre 14 : Détection d'anomalies 459
14.1 Cadre général . . . 462
14.2 Méthodes statistiques . . . 464
14.2.1 Principe . . . 464

14.2.2 Distance de Mahalanobis et enveloppe elliptique . . . 464
14.2.3 Avantages et limites . . . 465
14.3 Méthodes fondées sur la distance . . . 466
14.3.1 Principe et formalisation . . . 466
14.3.2 Avantages et limites . . . 466
14.4 Méthodes fondées sur la densité . . . 467
14.4.1 Principe . . . 467
14.4.2 Construction du facteur local d'aberrance . . . 467
14.4.3 Avantages et limites . . . 468
14.5 Méthodes fondées sur l'isolation . . . 468
14.5.1 Principe . . . 468
14.5.2 Construction de la forêt et score d'isolation . . . 469
14.5.3 Avantages et limites . . . 470
14.6 Méthodes fondées sur une frontière . . . 471
14.6.1 Principe . . . 471
14.6.2 Comparaison visuelle des cinq familles . . . 471
14.6.3 Avantages et limites . . . 472
14.7 Méthodes fondées sur la reconstruction . . . 472
14.7.1 Principe . . . 472
14.7.2 Erreur de reconstruction d'une analyse en composantes principales . . . 473
14.7.3 Avantages et limites . . . 475
14.8 Choisir une famille . . . 475
14.9 Évaluation des détecteurs d'anomalies . . . 477
14.9.1 Courbe ROC ou courbe précision-rappel . . . 479
14.9.2 Précision et rappel au rang k . . . 480
14.9.3 Régler un détecteur sans étiquettes . . . 481
14.10 Anomalies contextuelles et collectives . . . 482
14.11 Interprétation des anomalies . . . 484

Partie IV Apprentissage par renforcement 487

Chapitre 15 : Programmation dynamique et méthodes de Monte Carlo 489
15.1 Fondements de l'apprentissage par renforcement . . . 491
15.1.1 L'interaction agent-environnement . . . 491
15.1.2 Le compromis exploration-exploitation . . . 491
15.1.3 Retour, actualisation et tâches épisodiques . . . 494
15.1.4 Le processus de décision markovien . . . 494
15.1.5 Fonctions de valeur et équations de Bellman . . . 496
15.1.6 Optimalité . . . 497
15.2 Programmation dynamique . . . 497
15.2.1 Évaluation itérative d'une politique . . . 497
15.2.2 Amélioration de politique et itération sur les politiques . . . 499
15.2.3 Itération sur les valeurs . . . 500
15.2.4 Portée et limites . . . 501
15.3 Méthodes de Monte Carlo . . . 502
15.3.1 Évaluation par Monte Carlo . . . 503

15.3.2 Contrôle et maintien de l'exploration 505
15.3.3 Apprentissage hors politique 508
15.4 Programmation dynamique ou Monte Carlo 509

Chapitre 16 : Apprentissage par différence temporelle **511**
16.1 L'idée de la différence temporelle 512
16.1.1 De Monte Carlo à TD(0) 512
16.1.2 Monte Carlo ou différence temporelle 514
16.2 Contrôle par différence temporelle 516
16.2.1 SARSA, sur politique 516
16.2.2 Q-learning, hors politique 517
16.3 Sur politique ou hors politique, la falaise 518
16.4 Deux raffinements 522
16.4.1 SARSA espéré 522
16.4.2 Biais de maximisation et double Q-learning 524
16.5 Traces d'éligibilité et TD(λ) 525
16.6 Régler les hyperparamètres 527

Mot de fin **531**

Bibliographie **533**

Index **541**