

TABLE DES MATIERES

CHAPITRE 1. OBJECTIF DE LA DATA SCIENCE	7
CHAPITRE 2. LA REGRESSION PENALISEE	11
2.1. LES LIMITES DE LA REGRESSION CLASSIQUE	11
2.2. ANALYSE BASEE SUR LE MODELE LINEAIRE CLASSIQUE	13
2.3. ANALYSE BASEE SUR LA REGRESSION PENALISEE	15
2.3.1. POURQUOI PENALISER ?	15
2.3.2. DIFFERENTS TYPES DE PENALISATION, LEUR MISE EN ŒUVRE ET INTERPRETATION	15
2.4. ET AVEC PYTHON ?	27
2.5. REFERENCES	29
CHAPITRE 3. REGRESSION LINEAIRE AVEC VARIABLES CORRELEES	31
3.1. CONSEQUENCES D'UTILISER DES VARIABLES CORRELEES DANS UNE REGRESSION	31
3.1.1. LA VARIANCE D'UN ESTIMATEUR DEPEND DES CORRELATIONS ENTRE LES VARIABLES	31
3.1.2. EXEMPLE	32
3.2. REGRESSIONS PCR ET PLSR	35
3.2.1. PRINCIPE	35
3.2.2. MISE EN ŒUVRE	38
3.3. PREDICTION DE LA BIOMASSE DU MAÏS AVEC PCR ET PLS	40
3.3.1. OBJECTIF ET DONNEES DISPONIBLES	40
3.3.2. RESULTATS OBTENUS AVEC PCR ET PLS	42
3.4. ET AVEC PYTHON?	47
3.5. CONCLUSION	49
3.6. REFERENCES	49
CHAPITRE 4. LES SEPARATEURS A VASTES MARGES (SVM)	51
4.1. INTRODUCTION	51
4.2. SEPARATEURS A VASTES MARGES (SVM) POUR DISCRIMINER DEUX CLASSES	52
4.2.1. SVM A MARGE DURE	52
4.2.2. SVM A MARGE SOUPLE	57
4.2.3. SVM DANS LES CAS NON LINEAIREMENT SEPARABLES	63
4.3. SVM MULTICLASSES	75
4.3.1. PREMIERE APPROCHE : ONE VS ONE	75
4.3.2. DEUXIEME APPROCHE : ONE VS ALL	76

4.4. APPLICATION SUR R	77
4.5. ET AVEC PYTHON ?	89
4.6. POUR ALLER PLUS LOIN	93
4.7. REFERENCES	94
CHAPITRE 5. ARBRES ET FORETS ALEATOIRES	97
5.1. ARBRES DE DECISION/REGRESSION	97
5.1.1. PRINCIPE GENERAL	97
5.1.2. DEFINITIONS ET EXEMPLE SIMPLE	97
5.1.3. METHODE CART (CLASSIFICATION AND REGRESSION TREE)	101
5.1.4. GESTION DES DONNEES MANQUANTES	104
5.1.5. AVANTAGES - INCONVENIENTS PAR RAPPORT AUX AUTRES METHODES SUPERVISEES	105
5.1.6. UTILISATION DU LOGICIEL R POUR PREDIRE LA BIOMASSE DU MAÏS AVEC CART	106
5.1.7. UTILISATION DE R POUR CLASSER DES BLASTOCYSTES	112
5.1.8. UTILISATION DE PYTHON POUR PREDIRE LA BIOMASSE DU MAÏS	115
5.2. FORETS ALEATOIRES	117
5.2.1. DEFINITION	117
5.2.2. MESURES D'IMPORTANCE	119
5.2.3. GESTION DES DONNEES MANQUANTES	121
5.2.4. AVANTAGES-INCONVENIENTS DES FORETS ALEATOIRES	122
5.2.5. UTILISATION DE R POUR PREDIRE LA BIOMASSE DU MAÏS	122
5.2.6. UTILISATION DE R POUR CLASSER LES BLASTOCYSTES	129
5.2.7. UTILISATION DE PYTHON POUR CLASSER DES BLASTOYSTES	134
5.3. REFERENCES	137
CHAPITRE 6. RESEAUX DE NEURONES	139
6.1. ORIGINES	139
6.2. PRINCIPES DE BASE D'UN NEURONE ET D'UN RESEAU DE NEURONES	140
6.2.1. UN NEURONE	140
6.2.2. RESEAU DE NEURONES	141
6.2.3. LES FONCTIONS D'ACTIVATION	143
6.3. LA RETRO-PROPAGATION	146
6.4. METHODE DE DECROISSANCE DU GRADIENT	148
6.5. ALGORITHME DE RETRO-PROPAGATION	150
6.5.1. CAS DE LA COUCHE DE SORTIE	150
6.5.2. CAS DE LA COUCHE CACHEE	152
6.5.3. SCHEMA DU FONCTIONNEMENT GENERAL DE L'ALGORITHME	153
6.5.4. ILLUSTRATION DU FONCTIONNEMENT DE L'ALGORITHME	154

Table des matières	5
6.6. LES RESEAUX DE NEURONES A CONVOLUTION (CNN)	158
6.6.1. LE KERNEL	158
6.6.2. LE POOLING	160
6.6.3. EVOLUTION DES FEATURES MAPS DANS LE CNN	160
6.6.4. DROPOUT	161
6.7. LIBRAIRIE KERAS POUR PYTHON ET R	162
6.7.1. APPLICATION 1	162
6.7.2. APPLICATION 2	166
6.8. ET AVEC PYTHON ?	172
6.9. REFERENCES	178
CHAPITRE 7. APPRENTISSAGE NON SUPERVISE	181
7.1. CLUSTERING	181
7.1.1. MESURES DE DISTANCE	181
7.1.2. TYPES DE CLUSTERING	184
7.1.3. EVALUATION ET VALIDATION DU CLUSTERING	192
7.1.4. DEVELOPPEMENTS RECENTS POUR LE CLUSTERING	201
7.2. PATTERN MINING	201
7.2.1. EXTRACTION D'ITEMSET FREQUENTS ET DE REGLES D'ASSOCIATION	202
7.2.2. EXTRACTION DE SEQUENCES FREQUENTES	209
7.3. ET AVEC PYTHON ?	215
7.4. REFERENCES	218
CHAPITRE 8. EVALUATION	221
8.1. OBJECTIF	221
8.2. DIFFERENTS INDICATEURS POUR MESURER LA QUALITE DE PREDICTION	223
8.2.1. CAS DE LA REGRESSION	224
8.2.2. MISE EN ŒUVRE SOUS R	226
8.3. CAS DE LA CLASSIFICATION SUPERVISEE	228
8.3.1. MATRICE DE CONFUSION	228
8.3.2. INDICATEURS	229
8.3.3. METHODE ROC	235
8.3.4. GENERALISATION A PLUS DE DEUX CLASSES	237
8.3.5. MISE EN ŒUVRE SOUS R	238
8.4. METHODES POUR ESTIMER LA QUALITE DE PREDICTION	239
8.4.1. DEMARCHE GENERALE ET PARTITION DU JEU DE DONNEES	239
8.4.2. LA VALIDATION CROISEE	242
8.4.3. PRISE EN COMPTE DE LA STRUCTURE DE DEPENDANCE ENTRE LES INDIVIDUS	243

8.4.4. MISE EN ŒUVRE SOUS R AVEC LE PACKAGE CARET	246
8.5. EXPLORER LE COMPORTEMENT DU MODELE EN FONCTION DES VARIABLES D'ENTREE	249
8.5.1. GRAPHIQUE DE DEPENDANCE PARTIELLE	249
8.5.2. MISE EN ŒUVRE SOUS R	250
8.6. ET AVEC PYTHON ?	252
8.7. REFERENCES	255