

Massih-Reza Amini • Renaud Blanch • Marianne Clausel
Jean-Bapstiste Durand • Eric Gaussier • Jérôme Malick
Christophe Picard • Vivien Quéma • Georges Quénot

DATA SCIENCE

Cours et exercices

EYROLLES



Table des matières

Table des figures	IX
Liste des algorithmes	XIII
CHAPITRE 1	
Introduction	1
CHAPITRE 2	
Prétraitement des données	5
2.1 Prétraitement des données	
textuelles	5
2.1.1 Segmentation	6
2.1.2 Normalisation et filtrage	7
2.1.3 Filtrage	10
2.2 Prétraitement des données image	
ou vidéo	11
2.2.1 Représentations globales	12
2.2.2 Représentations locales	16
2.2.3 Agrégation de représentations locales	17
2.2.4 Représentations apprises	19
2.3 Exercices	19
CHAPITRE 3	
Gestion de données large-échelle et systèmes distribués	21
3.1 Les limites des systèmes	
traditionnels de gestion de données	21

3.1.1	Les besoins liés au traitement de grandes masses de données	22
3.1.2	Limites des architectures incrémentales	23
3.2	L'architecture Lambda pour le traitement de grandes masses de données	24
3.3	La couche batch : traitement de données par lots	27
3.3.1	Caractéristiques du jeu de données principal	28
3.3.2	Stockage du jeu de données principal	29
3.3.3	Traitement de données par lots	30
3.4	La couche service : stockage et requêtes sur les vues batch	33
3.4.1	Remarque préliminaire sur le stockage des vues batch	33
3.4.2	Stockage des vues batch	35
3.5	La couche vitesse : traitement de flux de données, stockage et requêtes sur les vues temps réel	36
3.5.1	Traitement de flux de données	37
3.5.2	Stockage des vues temps réel	38
3.6	Conclusion	38
CHAPITRE 4		
	Calcul haute performance	41
4.1	Introduction	41
4.1.1	Motivations	42
4.1.2	Hierarchies du parallélisme	44
4.1.3	Classification des plateformes	47
4.1.4	Coûts de communication	48
4.2	Principes de conception des algorithmes	50
4.2.1	Techniques de décomposition	51
4.2.2	Caractéristiques des tâches et des interactions	53
4.2.3	Équilibrage des ressources	54
4.2.4	Modèles d'algorithmes parallèles	56

4.3 Modèles analytiques	57
4.3.1 Métriques de performances	58
4.3.2 Passage à l'échelle des systèmes parallèles	59
4.3.3 Effet de la granularité	61
4.3.4 Notion d'iso-efficacité	62
4.4 Conclusion	62

CHAPITRE 5

Optimisation pour l'analyse de données	65
5.1 Apprentissage et optimisation	66
5.2 Introduction à l'optimisation	72
5.2.1 Problèmes d'optimisation	72
5.2.2 Analyse convexe pour impatientes	74
5.2.3 Algorithmes d'optimisation	77
5.3 Algorithmes en science des données	82
5.3.1 Algorithmes incrémentaux	83
5.3.2 Algorithmes distribués	87
5.3.3 Au-delà de ce chapitre	91
5.4 Exercices	91

CHAPITRE 6

Décomposition matricielle/tensorielle	95
6.1 Motivations	95
6.2 La SVD	96
6.2.1 Quelques rappels d'algèbre linéaire	96
6.2.2 Approximation de rang faible	97
6.2.3 SVD et analyse en composantes principales	99
6.2.4 Algorithme pour déterminer la SVD d'une matrice	100
6.3 Décomposition en matrices non négatives	104
6.3.1 Algorithme de Seung et Lee	105
6.3.2 Algorithme des moindres carrés alternés	107
6.3.3 Comparaison de la SVD et de la NMF	107
6.4 Décomposition tensorielle	108
6.4.1 Décomposition canonique polyadique	109

6.5	Conclusion	111
6.6	Exercices	111
CHAPITRE 7		
	Modèles génératifs	115
7.1	Motivations	115
7.1.1	Modèles graphiques	117
7.1.2	Modèles à variables latentes	117
7.2	Introduction à la statistique bayésienne	120
7.2.1	Généralités	120
7.2.2	Algorithmes génériques pour l'inférence bayésienne	123
7.3	Inférence dans les modèles à variables latentes	127
7.3.1	Modèles probabilistes graphiques	127
7.3.2	Mélanges	129
7.3.3	Analyse en composantes principales probabiliste	132
7.3.4	Chaînes de Markov cachées	134
7.3.5	Modèles hiérarchiques à variables latentes	136
7.4	Références	138
7.5	Exercices	140
CHAPITRE 8		
	Modèles discriminants	145
8.1	Approches supervisées	146
8.1.1	Modèles binaires	147
8.1.2	Modèles multi-classes	161
8.2	Approches semi-supervisées	164
8.2.1	Modèles graphiques	165
8.2.2	Modèles de mélange	171
8.2.3	Modèles discriminants	171
8.3	Exercices	172

CHAPITRE 9

Deep learning	177
9.1 Neurone formel	178
9.2 Réseaux simples	179
9.2.1 Perceptron	180
9.2.2 ADALINE	183
9.2.3 Perceptrons multicouches (PMC)	185
9.3 Réseaux à propagation avant	186
9.3.1 Composition de fonctions	186
9.3.2 Fonction-objectif et descente de gradient stochastique par mini-lots	187
9.3.3 Calcul des gradients par rétro-propagation de l'erreur	188
9.3.4 Architecture modulaire	190
9.3.5 Réseaux quelconques	195
9.3.6 Différentiation automatique	196
9.4 Réseaux convolutifs	197
9.4.1 Couche de convolution	197
9.4.2 Changements de résolution	199
9.4.3 Passage à des couches complètement connectées	200
9.4.4 Un exemple : AlexNet	200
9.5 Optimisations supplémentaires	202
9.5.1 Traitement par mini-lots	202
9.5.2 Moment	202
9.5.3 Fonctions d'activation	203
9.5.4 Dropout	204
9.5.5 Normalisation de lots	204
9.5.6 Augmentation de données	205
9.6 Réseaux pour la catégorisation d'images	206
9.7 Exercices	207

CHAPITRE 10

Visualisation interactive d'information	211
10.1 Introduction	211
10.2 Des données au graphique	214
10.2.1 Les données	214

10.2.2 L'image	216
10.2.3 Encodage visuel	224
10.3 Encodages avancés	227
10.3.1 Utilisation multiple des variables graphiques	227
10.3.2 Encodage des liens entre individus	230
10.4 Pour aller plus loin	234
10.5 Exercices	235
Bibliographie	239
Index	251