

Les éléments à télécharger sont disponibles à l'adresse suivante :
<http://www.editions-eni.fr>
Saisissez la référence de l'ouvrage **EP4DASR** dans la zone de recherche
et validez. Cliquez sur le titre du livre puis sur le bouton de téléchargement.

Chapitre 1 Introduction

1. Introduction	19
2. Contenu de l'ouvrage	21
3. Codes et solutions techniques présentées	22

Chapitre 2 Premiers pas avec R

1. Installation des composants	25
1.1 Installation et lancement de R	25
1.2 Installation et lancement de RStudio	26
1.3 Installation de nouveaux packages	27
1.4 Installation de packages : compléments	29
2. Prise en main de R	31
2.1 R, une calculatrice efficace	31
2.2 R, un langage vectorisé	33
2.3 Fonctions agissant sur des vectors	35
2.3.1 Une première analyse rapide des données	36
2.3.2 Quelques statistiques simples sur les vectors	37
2.3.3 Trier un vector	38
2.3.4 Diverses fonctions avec somme, produit, min et max	39
2.4 Types de données simples	40
2.4.1 Les booléens	40
2.4.2 Les ensembles	43
2.4.3 Les listes	44
2.4.4 Les facteurs	46
2.4.5 Les tableaux	54

2 ————— Data Scientist et langage R

IA, Machine Learning et Statistiques

2.5	Les fonctions	62
2.5.1	Création et utilisation d'une fonction simple	63
2.5.2	Création d'un opérateur à partir d'une fonction de deux variables	64
2.5.3	Usage des fonctions et portée des variables	64
2.5.4	Application des fonctions sur les matrices : apply.	68
2.5.5	Compléments utiles.	69
2.6	Structures de contrôle.	74
2.6.1	Des instructions communes avec d'autres langages	74
2.6.2	Parcourir une matrice via des boucles for.	76
2.7	Les chaînes de caractères.	79
2.8	Le formatage des nombres	82
2.9	Les dates et les temps	82
2.10	Mesure de la durée d'un algorithme.	83
2.11	Les nombres complexes	88
2.11.1	Manipulations de base des nombres complexes.	88
2.11.2	Visualisation de nombres complexes	88
2.12	Programmation orientée objet	90
2.12.1	Classes et objets, en bref	91
2.12.2	Constructeurs.	96
2.12.3	Héritage	98
2.12.4	Objets mutables.	101
2.12.5	Gestion de pile : implémentation orientée objet avec RC.	103
3.	Manipulation des données	108
3.1	Lecture des données : les bases.	108
3.2	Manipulation des colonnes d'un data.frame.	110
3.3	Calculs simples sur un data.frame	111
3.3.1	Calculs sur les colonnes et les lignes.	111
3.3.2	Manipulation des lignes.	112
3.3.3	Application : comparaison effectifs de classes et Khi-2.	116
3.3.4	Création de colonnes calculées	118
3.3.5	Tri d'un data.frame via order()	122

Chapitre 3

Machine Learning

1. Petit bestiaire de la Data Science	125
1.1 Les fondamentaux	125
1.1.1 Apprentissage et classification	125
1.1.2 Petit vocabulaire graphique du Machine Learning	126
1.1.3 Régressions, formulations mathématiques simplifiées	140
1.1.4 Régression linéaire généralisée	142
1.1.5 Arbres de décision, pruning, élagage	142
1.1.6 Clustering, k-means	144
1.1.7 k-NN	144
1.1.8 Modèles paramétriques	144
1.1.9 Lazy algorithm (algorithme paresseux)	145
1.1.10 Overfitting : sur-détermination, sur-apprentissage	145
1.1.11 Validation croisée, régularisation, bagging	146
1.1.12 Optimisation, méthode du gradient	147
1.1.13 Algorithme glouton (greedy algorithm)	149
1.1.14 Programmation linéaire, simplexe, point intérieur	149
1.1.15 Estimation via la méthode de Monte-Carlo	151
1.1.16 Entropie, indépendance et information mutuelle	152
1.1.17 Discrétisation	154
1.2 Méthodes "ensemble"	155
1.2.1 Random forest	155
1.2.2 AdaBoost (adaptative boosting)	156
1.3 Lois de probabilité et distribution	157
1.3.1 Généralités	157
1.3.2 Petit bestiaire des lois de probabilité	159
1.4 Les graphes	161
1.4.1 Introduction à l'univers des graphes	161
1.4.2 Notions importantes à assimiler au sujet des graphes	163
2. Notations mathématiques usuelles en Data Science	165
2.1 Notations des paramètres	165
2.2 Autres notations	166
2.2.1 Fonctions et applications... $f(x)$, $d(x,y)$	166
2.2.2 Quelques confusions possibles	167
2.2.3 Notations ayant trait aux opérations entre vecteurs	167

4 _____ Data Scientist et langage R

IA, Machine Learning et Statistiques

3.	Création de notre premier modèle ML	168
3.1	Analyse statistique et visuelle des données	169
3.1.1	Visualisation globale	169
3.1.2	Visualisation des variables numériques 2 à 2 avec mention des classes	171
3.1.3	Corrélations entre variables numériques	174
3.1.4	Séparation par classe, utilisation de ggplot2 et qplot	176
3.1.5	Visualisation 3D, relation entre trois variables numériques	177
3.1.6	Graphiques par paires	178
3.2	Nettoyage des données	179
3.3	Choix empirique du type de modèle à entraîner	182
3.4	Mise au point d'un modèle	182
3.5	Interprétation et qualification du modèle	184
3.5.1	Lecture de l'arbre	184
3.5.2	Matrice de confusion et accuracy	185
4.	Recettes de représentation graphique des données	185
4.1	De l'usage d'une grammaire graphique	186
4.2	Histogrammes évolués	188
4.2.1	Distribution multiclasse	188
4.2.2	Mixture de plusieurs distributions par classe	190
4.2.3	Visualisation de la densité d'une distribution	191
4.2.4	Une autre mixture par classe	192
4.2.5	Une variable, mais un histogramme pour chaque classe	193
4.2.6	Graphique avec une densité par classe	194
4.3	Diagramme par paires et à facettes	196
4.3.1	Diagramme par paires, version simple	197
4.3.2	Classes en configuration XOR	198
4.3.3	Diagramme par paires avec "factors"	199
4.3.4	Facettes et échelle logarithmique	202
5.	Prise en compte de données et faits entachés d'incertitudes	203
5.1	Quelques notions fondatrices	203
5.1.1	Phénomène aléatoire	203
5.1.2	Probabilité, variable aléatoire et distribution	204
5.1.3	Un peu de mathématiques : notations et définitions utiles en probabilité	208
5.1.4	Moments d'une variable aléatoire discrète X	215
5.1.5	Premières considérations sur les erreurs et estimations	222

5.2	Se familiariser avec ses données	223
6.	Matrices et vecteurs, au cœur du Machine Learning	227
6.1	Conventions, notations, utilisations basiques	227
6.2	Matrices, vecteurs : vers une introduction à la notion d'apprentissage supervisé	237
6.3	Plus loin dans la manipulation des matrices avec R	238
6.3.1	Opérations basiques	238
6.3.2	Quelques savoir-faire utiles sur les matrices de R	239
6.3.3	Normes de vecteurs et normes de matrices	243
6.3.4	Matrices et vecteurs : diverses syntaxes utiles.	247
7.	Estimation et inférence	251
7.1	Positionnement du problème d'estimation	252
7.1.1	Formulation générale du problème.	252
7.1.2	Application et reformulation du problème d'estimation	255
7.2	Les indicateurs d'écart utilisés en Machine Learning	263
7.2.1	MSE, RMSE, SSE, SST.	263
7.2.2	MAE, ME	264
7.2.3	NRMSE/NRMSD, CV_RMSE	266
7.2.4	SDR.	267
7.2.5	Accuracy, R2	267
8.	Mise en pratique : apprentissage supervisé	270
8.1	Préparation	270
8.2	Tester des hypothèses, p_value	272
8.2.1	Analyse graphique interactive avec iplots	272
8.2.2	Test de Breusch-Pagan et zoom sur p_value	273
8.3	Création d'un modèle (régression linéaire multiple)	274
8.4	Établissement d'une prédiction	275
8.5	Étude des résultats et représentations graphiques	276
8.6	Indicateurs courants - calculs.	279
8.7	Étude du modèle linéaire généré	280
8.8	Discussion autour de notre modèle linéaire	286
8.9	Utilisation d'un modèle "Random Forest".	287
9.	Vers l'expertise en Machine Learning conventionnel	290
9.1	Parcours théorique accéléré	290
9.1.1	Linéarité	290
9.1.2	Erreurs in et out, notion de VC dimension	292

6 ————— Data Scientist et langage R

IA, Machine Learning et Statistiques

9.1.3	Hyperplans, séparabilité avec marges	295
9.1.4	Kernel Trick, noyaux, transformations, feature space	297
9.1.5	Problèmes de la régression : introduction à la régularisation	300
9.2	Vers une pratique mature du Machine Learning	305
9.2.1	Cross validation : k-fold CV	305
9.2.2	Naive Bayes	307
9.2.3	C4.5 et C5.0	313
9.2.4	Mise au point d'un modèle de classification multiclassées	314
9.2.5	Support Vector Machines (SVM)	319
9.2.6	Clusterisation, k-means	323
9.3	Tracer une courbe ROC	332
9.4	GAM : généralisation de LM/GLM	334

Chapitre 4

Feature engineering pour le Machine Learning

1.	Travaux préliminaires sur les données	337
2.	Appréhension directe et rapide d'un dataset	338
3.	Analyse de la conformation des distributions par rapport à la loi normale	342
4.	Dépendance linéaire entre variables	344
5.	Mise en évidence de différences sur les distributions	346
6.	Points étrangers	349
7.	Tris et agrégats	350
7.1	Tris automatisés	350
7.2	Calculs sur les agrégats	355
7.3	Extraction, pivot et tri rapide de données d'un dataframe	356
7.4	Utilisation des jointures	357
7.4.1	Enrichissement d'un dataframe par une jointure	357
7.4.2	Jointures pour décodage ou transformation - ex. : factor to integer	359
7.5	Applications de fonctions variables sur les données	360
7.6	Compléments en matière de métaprogrammation	362
7.6.1	Quoting appliqué aux graphiques	362
7.6.2	Définition d'une macro	364

8. Feature Engineering et Data Science	367
8.1 Positionnement du problème.	367
8.2 Ce sur quoi il faut être très attentif.	368
8.2.1 La qualité de la distribution	368
8.2.2 La nature des features	368
8.3 Maîtriser la dimensionnalité	371
8.4 Une solution éprouvée : la PCA.	373
8.5 Un exemple simple d'utilisation de la PCA.	374
9. PCA classique, éléments mathématiques.	381
9.1 Les valeurs inconnues et les features mal conditionnées	383
9.2 Création de nouvelles features.	386
10. Réduction des données (data reduction)	389
11. Réduction de la dimensionnalité et entropie.	390
11.1 Description théorique du problème.	390
11.2 Implémentation en R et discussion.	392

Chapitre 5 R et Python

1. Une polémique stérile, mais une vraie question : R ou Python ?	401
2. Exemples de codes R et Python comparables	404
2.1 Codes triviaux.	405
2.2 Packages.	405
2.3 Fonctions simples	406
2.4 Structures de données	407
2.5 Booléens.	409
2.6 Programmation de base	410
2.7 Création de fonctions.	410
2.8 Blocs et indentation	411
2.9 Datasets.	412
3. Accéder à Python à partir de R	414
3.1 Installation et initialisation du contexte technique.	414
3.2 Interactions R et Python, du point de vue R	415
3.2.1 Partage de variables et exécution de code Python dans le biotope R	415
3.2.2 Bibliothèques Python dans du code R, exemples NLP	416

8 _____ Data Scientist et langage R

IA, Machine Learning et Statistiques

4. Considérations sur les cas difficiles.	422
---	-----

Chapitre 6

Business intelligence avec Shiny

1. Introduction	425
2. Créer sa première application Shiny	425
2.1 Introduction	425
2.2 Installation dans RStudio et préambule	426
2.3 Organisation des fichiers de l'application	428
2.4 Syntaxe des inputs et outputs	431
2.5 Mise en page et organisation visuelle du contenu	438
2.6 Modifier le thème de l'application	450
2.7 Générer l'application	451
2.8 Déployer l'application sur le Web	452
2.9 Shiny Dashboard	453

Chapitre 7

Prédire une série temporelle

1. Introduction	463
1.1 Introduction aux séries temporelles	464
1.2 Modèle stationnaire	465
1.2.1 Processus stationnaire : les bases	465
1.2.2 Processus autorégressif AR : aller plus loin	467
1.2.3 Considérations (très) utiles	468
1.3 Processus non stationnaires	470
1.3.1 Le modèle ARIMA	470
1.3.2 Processus saisonniers : SARIMA	471
1.3.3 Modèles ARCH et GARCH	472
1.3.4 Convolution et filtres linéaires	474
1.4 Mise en pratique	476
1.4.1 Les bases de la manipulation des séries temporelles en R	477
1.4.2 Étude de séries temporelles	481
1.4.3 Prédictions sur ARIMA (AR MA SARIMA)	487
1.5 Mini-bestiaire ARIMA	490

2. Gestion des dates informatiques	493
3. Co-intégration de données économiques	495

Chapitre 8 Géospatial et Machine Learning

1. Pourquoi étudier les représentations cartographiques ?	503
2. Accéder à l'information géographique	504
3. Création de cartes statiques avec R	508
4. Création de cartes dynamiques avec R.	511
5. Données spatiales, gradient, laplacien et convolution	513
5.1 Préliminaire, créer un échantillon : LHS (hypercube latin)	513
5.2 Travailler sur des données spatiales.	516
5.2.1 Champ et variable régionalisée.	516
5.2.2 Calcul d'un gradient en dimension 3	517
5.2.3 Introduction au calcul numérique du Laplacien et des convolutions	519
5.3 Détermination d'un variogramme.	522
6. Krigeage (kriging)	524
6.1 La théorie, en bref.	524
6.2 Implémentation en R, utilisant un algorithme génétique pour l'optimisation.	526

Chapitre 9 Feature engineering des images

1. Manipulation d'images dans le cadre d'une Machine Learning	529
1.1 Création, visualisation, lecture et écriture d'images.	530
1.2 Transformation des images	535
1.2.1 Exemples de manipulation de la couleur et des intensités	535
1.2.2 Exemples de manipulation de la géométrie de l'image	538
1.2.3 Application de filtres sur les images	541

10 ————— Data Scientist et langage R

IA, Machine Learning et Statistiques

Chapitre 10 Mathématiques pour l'IA

1. Introduction	543
2. Notations et vocabulaire mathématiques importants pour l'IA	545
2.1 Vecteurs	547
2.2 Normes d'un vecteur	548
2.3 Distance issue d'une norme	549
2.4 Indices et convention d'Einstein	550
2.5 Produit scalaire	551
2.5.1 Notations du produit scalaire	551
2.5.2 Norme naturelle issue d'un produit scalaire	552
2.6 Les matrices	553
2.6.1 Rappels sur le produit de matrices	553
2.6.2 Transformation d'un vecteur par une matrice	555
2.7 Forme linéaire	556
2.8 Application multilinéaire	557
2.8.1 Formulation théorique : multilinéarité vs tenseurs	558
2.8.2 Exemple numérique d'une contraction débouchant sur un scalaire	560
2.9 Changement de base	562
2.10 Tenseurs covariants et contravariants	562
2.11 Vocabulaire de la différentiation	565
2.12 Méthode d'optimisation utilisant le gradient	571
3. Taxonomie des tâches de Machine Learning courantes, loss et métriques associées	573
3.1 Tâches générales de Machine Learning	575
3.1.1 Régression	575
3.1.2 Classification	578
3.2 Tâches liées aux données multimédias (images, sons, vidéos)	581
3.2.1 Classification d'images	581
3.2.2 Détection d'objets	581
3.2.3 Segmentation sémantique	582
3.2.4 Détection de visages	582
3.2.5 Reconnaissance faciale	583
3.2.6 Génération d'images	583

3.3	Tâches liées aux données textuelles	585
3.3.1	Tâches NLP (Natural Language Processing) "classiques".	585
3.3.2	Tâches NLP lors de la mise au point de vos propres modèles .	586
3.4	Tâches liées aux données temporelles	588
4.	Calculs numériques dans l'esprit Matlab	588
4.1	Considérations pratiques	588
4.2	Un aperçu de la précision de notre machine	590
4.3	Un package possédant les mêmes instructions que Matlab	592
4.4	Quelques compléments sur les complexes	592
4.5	Quelques compléments sur les matrices	593
4.5.1	Création de matrices basiques	594
4.5.2	Création de matrices de nombres aléatoires	596
4.5.3	Calculs sur les matrices d'entiers	596
5.	Un peu d'algèbre linéaire opérationnelle	597
5.1	Exemple de création d'une base orthonormale	597
5.2	Projection sur un sous-espace vectoriel	598
6.	Fonctions et systèmes d'équations, 1 à n variables	600
6.1	Fonctions d'une variable	600
6.1.1	Exploration de divers tracés	600
6.1.2	Autour des racines d'une fonction	603
6.2	Fonctions de plusieurs variables	605
6.2.1	Représentation d'une fonction réelle de deux variables réelles .	605
6.2.2	Recherche de minima d'une fonction en deux dimensions . .	606
6.3	Systèmes d'équations non différentielles, linéaires ou non	608
6.3.1	Système d'équations linéaires	608
6.3.2	Système d'équations non linéaires non différentielles, $n = m$.	611
7.	Dérivation de fonctions	612
7.1	Dérivée symbolique et numérique avec R-base	612
7.2	Dérivées en utilisant des packages spécifiques	614
7.3	Dérivée de fonctions spéciales	615
8.	Autour de l'intégration	616
8.1	Calcul d'une intégrale multiple	616
8.2	Autres intégrales, dont intégrale impropre	617

12 ————— Data Scientist et langage R

IA, Machine Learning et Statistiques

9. Fonctions spéciales et équations différentielles	618
9.1 Autres fonctions spéciales	618
9.1.1 Fonctions d’Airy et de Bessel.	618
9.1.2 La fonction d’erreur erf	620
9.2 Équation différentielle ordinaire normalisée d’ordre 1	620
9.3 Équation différentielle du second degré (problème aux limites)	621
9.3.1 EQD linéaire (2nd degré).	621
9.3.2 EQD non linéaire (2nd degré)	622
10. Éléments pratiques de calcul différentiel	626
10.1 Fonctions réelles d’un réel	626
10.2 Fonctions réelles d’un vecteur	628
10.2.1 Calcul des dérivées partielles.	628
10.2.2 Calculs d’opérateurs utilisant les dérivées partielles	630
10.3 Fonctions vectorielles d’un vecteur	631

Chapitre 11

Réseaux neuronaux et Gradient Boosting

1. Réseau neuronal - premiers pas vers le Deep Learning	635
1.1 Introduction aux réseaux neuronaux	635
1.1.1 Composition de fonction.	635
1.1.2 Description succincte d’un modèle NN de classification	636
1.1.3 Backpropagation et règle de composition des différentielles . . .	639
1.1.4 Réduction de l’overfitting du NN.	641
1.1.5 Définir la sortie du modèle - softmax.	642
1.1.6 Évaluation du modèle	643
1.1.7 Utilisation du modèle instancié	643
1.2 Création d’un réseau neuronal de classification binaire.	644
2. Gradient Boosting et Generalized Boosted Regression	648
2.1 Les grands principes	648
2.2 Les paramètres et les usages du package GBM	649
2.2.1 Covariance	649
2.2.2 Loss	649
2.2.3 Optimisation de l’algorithme	650
2.3 Mise en pratique	650

Chapitre 12

Deep Learning avec PyTorch ou Keras/TensorFlow

1. Considérations préliminaires sur l'utilité d'un framework puissant comme PyTorch ou TensorFlow	653
1.1 Croissance rapide du coût de l'entraînement des modèles du Deep Learning	653
1.2 Formats de nombres optimisés pour l'IA	654
1.3 Deux frameworks puissants à notre disposition	655
2. Présentation de PyTorch dans l'environnement R	657
3. PyTorch en Python versus sa version R (torch/R)	661
4. Opérations tensorielles fondamentales, illustrées avec torch.	661
4.1 Création d'un tenseur.	661
4.2 Indexation.	663
4.3 Matrice identité	664
4.4 Rename (renommer les dimensions)	664
4.5 Matrice nulle.	665
4.6 Matrice diagonale	666
4.7 Slicing (tranchage)	666
4.8 Reshape (changement de forme)	667
4.9 Transpose (transposition)	667
4.10 Permute (permutation des dimensions)	668
4.11 Expand (diffusion/broadcast)	669
4.12 Concat (concaténation)	670
4.13 Stack (empilement)	670
4.14 Flip (inversion)	671
4.15 Roll (décalage circulaire)	671
4.16 Split (découpage)	672
4.17 Squeeze/unsqueeze.	673
4.18 Flatten (aplatissement)	674
4.19 Sum (somme)	675
4.20 Mean (moyenne)	675
4.21 Std pour standard deviation (écart-type)	676
4.22 Maximum/minimum	676
4.23 Cumsum (somme cumulée)	677
4.24 Argmax/argmin.	677
4.25 Norm (norme)	678

14 _____ Data Scientist et langage R

IA, Machine Learning et Statistiques

4.26	Hadamard (multiplication par élément)	679
4.27	Addition (avec broadcasting)	680
4.28	Comparaisons (>, >=, <, <=, ==)	681
4.29	Topk (les k plus grandes valeurs)	681
4.30	Produit scalaire — dot product (1D)	682
4.31	Produit vectoriel (cross product - 3D)	682
4.32	Multiplication matricielle (mm)	683
4.33	Outer product (produit externe)	683
4.34	Produit de Kronecker	684
4.35	Inversion de matrice (inv)	686
4.36	Pinverse (pseudo-inverse de Moore-Penrose)	686
4.37	Déterminant (det)	687
4.38	Diagonale et trace	688
4.39	Sort (tri)	689
4.40	Clamp (limitation dans des intervalles donnés)	690
4.41	Masquage et remplissage (masked fill)	690
4.42	Meshgrid	691
4.43	Clause where (sélection conditionnelle entre deux tenseurs)	693
4.44	BMM (Batch Matrix Multiplication)	694
4.45	Logsumexp (log de la somme exponentielle)	696
5.	Notation d'Einstein	697
5.1	Produit dyadique de deux vecteurs	698
5.2	Produit d'Hadamard de deux matrices	699
5.3	Produit matriciel	699
5.4	Application d'une matrice sur un vecteur	700
5.5	Somme des éléments d'un tenseur	701
5.6	Trace d'une matrice	701
5.7	Transposée d'une matrice	702
5.8	Produit de Kronecker de deux matrices	702
6.	Création d'un réseau de neurones simple (NN "vanilla") avec PyTorch . .	704
7.	Un réseau de neurones effectuant une régression, avec Neuralnet puis avec Torch	705

Chapitre 13

Heuristiques clés : logique floue - swarm - CNN

1. Systèmes flous	717
1.1 Introduction	717
1.2 La logique floue en pratique (systèmes experts)	726
2. Essaim (swarm)	732
2.1 Swarm et optimisation : l'algorithme PSO	732
2.1.1 Présentation de PSO	733
2.1.2 Description de PSO	733
2.2 Mise en pratique de PSO	735
3. Historique et agencement des concepts liés aux réseaux de neurones qui déboucheront sur les LLM	738
4. Le "Légo" des réseaux neuronaux - CNN et compagnons	740
4.1 Classification d'images en noir et blanc, CNN, convolution et pooling	746
4.2 Réseaux neuronaux récurrents (RNN)	750
4.3 Variational Autoencoders (VAE)	752
4.3.1 Espace latent	752
4.3.2 Modèle VAE	754
4.4 Generative Adversarial Networks (GAN)	757
4.5 Les principales fonctions d'activation à notre disposition	760

Chapitre 14

Traitement du langage naturel - NLP

1. Introduction	761
2. Analyse sémantique latente et SVD : "embedding" naturel	761
2.1 Aspects théoriques	762
2.1.1 SVD : généralités	763
2.1.2 Une justification de la décomposition SVD	763
2.1.3 SVD dans le contexte LSA	764
2.1.4 Interprétation	764
2.1.5 Alternative non linéaire, Isomap (MDS, géodésique, variété, manifold)	765

16 _____ Data Scientist et langage R

IA, Machine Learning et Statistiques

2.2	Mise en pratique	766
2.2.1	Initialisation	766
2.2.2	Au cœur de LSA	768
2.2.3	Résultats	769
2.2.4	Manipulations, interprétations récréatives et non fondées. . .	772
3.	Briques et outils à notre disposition pour l'embedding et l'utilisation de textes	774

Chapitre 15 Knowledge Graphs

1.	Introduction	777
2.	Premiers pas	777
2.1	Quelques notions et notations complémentaires basiques	778
2.2	Manipulations simples de graphes avec R	779
2.3	Structure des graphes	790
3.	Graphes et réseaux (sociaux).	792
3.1	Analyse des réseaux sociaux : concepts de base	793
3.2	Mise en pratique	796
3.3	Détection de communautés	798
4.	Embedding de graphes	805
5.	GNN (Graph Neural Networks)	806

Chapitre 16 LLM, Transformers, GPT, RAG et Agents

1.	Introduction aux LLM	809
2.	Entraînement des modèles LLM	811
3.	LLM basés sur l'architecture Transformer	813
3.1	Synthèse de l'architecture Transformers	813
3.1.1	Encodeur	814
3.1.2	Décodeur	814
3.1.3	Générateur	815
3.1.4	La technique de connexion résiduelle, skip connection.	815
3.2	Un peu plus loin avec l'architecture Transformer	816

4. Création d'un modèle GPT from scratch	821
5. Utilisation d'un LLM en local	824
5.1 Ollama	824
5.2 Ollama et Shiny	826
5.3 Ollama et Lucie : personnalisation de l'interaction d'un LLM	826
5.4 Paramètres détaillés de configuration des LLM	830
5.5 Perplexité	835
6. Améliorer les réponses du LLM via le RAG	835
6.1 Le RAG (Retrieval Augmented Generation)	835
6.2 RAG naïf avec Ollama	838
7. Utilisation d'agents pour améliorer les IA basées sur les LLM	842
7.1 Les agents sont vos amis	842
7.2 Création d'un agent	845
8. En guise de conclusion	851
 Index	 853

Les éléments à télécharger sont disponibles à l'adresse suivante :

<http://www.editions-eni.fr>

Saisissez la référence ENI de l'ouvrage **RIRANADO** dans la zone de recherche et validez. Cliquez sur le titre du livre puis sur le bouton de téléchargement.

Avant-propos

1. Introduction 9
2. À qui s'adresse cet ouvrage ? 10
3. Comment utiliser cet ouvrage ? 11

Chapitre 1

Le langage R

1. Histoire du langage R 15
2. Installation de R et de RStudio 17
 - 2.1 Comprehensive R Archive Network (CRAN) 17
 - 2.2 Installation de R via le CRAN 17
 - 2.3 Installation de RStudio 27
 - 2.4 Paramétrage de RStudio 32
 - 2.5 Trucs et astuces avec RStudio 34
3. Packages. 36
 - 3.1 Qu'est-ce qu'un package ? 36
 - 3.2 Installation d'un package 37
 - 3.3 Utilisation d'un package, d'une fonction 39
 - 3.4 Trouver de l'aide à propos d'un package ou d'une fonction ... 40
4. Travail sous forme de projet 42
5. Réalisation de rapports avec R Markdown 46

2 ————— Langage R et statistiques

Initiation à l'analyse de données

6. Initiation au codage en R	50
6.1 Nomenclature du langage R	50
6.2 Importer des données	55
6.3 Manipuler les données	58
6.4 Concaténation de tables	61

Chapitre 2 Les statistiques

1. Notions clés sur les statistiques	67
1.1 Histoire rapide de la discipline des statistiques	67
1.2 Définitions des mots-clés importants	68
2. Initiation à la réalisation de graphiques avec {ggplot2} dans R	72
3. Lois de probabilités statistiques	78
3.1 Introduction aux lois de probabilités	78
3.2 Loi uniforme	82
3.3 Loi binomiale	85
3.4 Loi de Poisson	87
3.5 Loi normale	89
3.6 Loi exponentielle	91
3.7 Résumé des paramètres liés aux lois de probabilités	94
4. Réalisation de statistiques descriptives dans R	95
4.1 Principe des statistiques descriptives	95
4.2 Statistiques descriptives univariées dans R	95
4.2.1 Type de données	95
4.2.2 Variable qualitative	97
4.2.3 Variable quantitative	104
4.3 Statistiques descriptives bivariées dans R	116
4.3.1 Deux variables qualitatives	117
4.3.2 Une variable qualitative et une variable quantitative	118
4.3.3 Deux variables qualitatives	121
5. Détection et visualisation des données manquantes dans R	123

Chapitre 3 La régression

1. Définitions et limites de la régression	127
2. Régression linéaire simple	128
2.1 Objectif de la régression linéaire simple	128
2.2 Réalisation d'une régression linéaire simple dans R	130
2.3 Limites d'application de la régression linéaire simple	132
2.3.1 Visualisation de la corrélation entre les données	132
2.3.2 Indépendances des données	133
2.3.3 Analyse des résidus	133
2.4 Intervalle de confiance de la régression dans R	136
3. Généralisation de la régression paramétrique dans R	137
3.1 Régression polynomiale	137
3.2 Régressions exponentielle et logarithmique	143
3.3 Introduction à la régression logistique binomiale	148
3.4 Régression multiple	153
4. Régression non paramétrique dans R	163
4.1 Définitions et limites de la régression non paramétrique	163
4.2 Régression non paramétrique univariée	166
4.3 Modèles additifs généralisés	168

Chapitre 4 Les tests d'hypothèses

1. Vocabulaire lié aux tests d'hypothèses	173
2. Démarche du test d'hypothèses	177
3. Test de Shapiro-Wilk	182

4 ————— Langage R et statistiques

Initiation à l'analyse de données

4. Test du Khi-2	186
4.1 Paramétrique ou non paramétrique, le test du Khi-2 ?	186
4.2 Loi du Khi-2 à k degrés de liberté.	187
4.3 Test du Khi-2 d'adéquation	189
4.4 Test du Khi-2 d'indépendance	193
4.5 Test du Khi-2 d'homogénéité.	199

Chapitre 5

La comparaison à une valeur théorique

1. Quand comparer à une valeur théorique ?	205
2. Théorème central limite.	206
2.1 Histoire et utilisation du théorème central limite	206
2.2 Énoncé du théorème central limite	206
3. Comparaison d'une proportion à une valeur théorique	209
3.1 Exemples de comparaison d'une proportion à une valeur théorique	209
3.2 Test de comparaison d'une proportion à une valeur théorique avec R	212
4. Comparaison d'une moyenne à une valeur théorique	218
4.1 Test de Student	218
4.2 Exemple d'utilisation du test de Student dans R	222
4.3 Test de Student apparié	224
4.4 Test de Wilcoxon apparié.	228
5. Comparaison d'une variance à une valeur théorique	231
5.1 Variance d'une variable	231
5.2 Test de comparaison de la variance à une valeur théorique	232
5.3 Exemple de comparaison d'une variance à une valeur dans R	233
6. Test de Kolmogorov-Smirnov	235

Chapitre 6

La comparaison de deux groupes

1. Généralités sur la comparaison de deux groupes	245
2. Comparaison de proportions de groupes dans R	246
2.1 Utilisation de la comparaison de proportions.	246
2.2 Test z à deux proportions et test du Khi-2	247
2.3 Test exact de Fisher	251
2.4 Test de McNemar	254
3. Comparaison de variances de deux groupes dans R	257
3.1 Mise en situation de la comparaison de variances	257
3.2 Test F de Fisher	259
3.3 Test de Bartlett	261
3.4 Test de Levene	264
3.5 Test de Fligner-Killeen	267
4. Comparaison de moyennes de deux groupes dans R	270
4.1 Choix du test de comparaison de moyennes à utiliser.	270
4.2 Test t de Student	271
4.3 Test t de Welch	275
4.4 Test de Wilcoxon-Mann-Whitney	277
5. Comparaison de moyennes/médianes de plus de deux groupes . . .	279
5.1 Pourquoi ne pas comparer deux à deux ?	279
5.2 ANOVA : analyse de variance dans R	281
5.3 Test de Kruskal-Wallis dans R	287

6 ————— Langage R et statistiques

Initiation à l'analyse de données

Chapitre 7

Les modèles linéaires généralisés

1. Modèles linéaires classiques	291
1.1 Rappel	291
1.2 Les tests post-hoc	294
2. Généralisation des modèles linéaires en R	299
2.1 Utilisation des modèles linéaires généralisés.	299
2.2 Régression logistique binomiale	302
2.3 Régression de Poisson.	308
3. Modèles linéaires mixtes	315
3.1 Généralités et utilisation des modèles linéaires mixtes	315
3.2 Réalisation de modèles linéaires mixtes dans R	317
4. Traitement des données manquantes dans les GLM	329
5. Modèles non linéaires à effets fixes ou mixtes	330

Chapitre 8

L'analyse en composantes principales

1. Utilisation de l'ACP	331
2. Bases mathématiques de l'ACP	336
2.1 Approche par les individus.	336
2.2 Approche par les variables	343
3. Mise en place de l'ACP	346
3.1 Déroulement de l'ACP	346
3.2 Sélection des variables actives	347
3.3 Réduction des variables	348
3.4 Réalisation de l'ACP dans R.	349
3.4.1 Lignes de code et sorties automatiques.	349
3.4.2 Représentativité des axes.	352
3.4.3 Graphique des individus.	355
3.4.4 Graphique des variables	356

3.5	Sélection du nombre de dimensions à interpréter dans R	358
3.5.1	Significativité des dimensions obtenues par l'ACP dans R	358
3.5.2	Nombre de dimensions à interpréter.	359
3.6	Interprétation des sorties de l'ACP et indicateurs supplémentaires dans R	361
3.7	Ajout de variables supplémentaires dans R	364
4.	Limites et ouverture des analyses factorielles.	374
4.1	Données manquantes en ACP	374
4.2	Analyse factorielle des correspondances	383
4.3	Pour conclure sur les analyses factorielles.	388
	Index	389