LA DATA SCIENCE POUR MODÉLISER LES SYSTÈMES COMPLEXES

Alain Chautard

LA DATA SCIENCE POUR MODÉLISER LES SYSTÈMES COMPLEXES

Optimiser la prédiction, l'estimation et l'interprétation

DUNOD

À mes collègues de travail, en souvenir de nos travaux fructueux et des moments agréables que nous avons partagés.

Direction artistique : Nicolas Wiel
Conception graphique de la couverture : Elisabeth Riba
Mise en page : Lumina Datamatics, Inc.

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.
Le Code de la propriété intellectuelle du 1er juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée.
Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).

© Dunod, 2022
11 rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-083087-9

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).
Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

Table des matières

© Dunod - Toute reproduction non autorisée est un délit.

2 CAS D'ÉTUDES

© Dunod - Toute reproduction non autorisée est un délit.

© Dunod - Toute reproduction non autorisée est un délit.

Introduction

Les entreprises se doivent désormais de réfléchir à la pertinence de leur organisation (s'adapter à des environnements économiques complexes) et d'améliorer leur efficacité sur les marchés (rester compétitif, établir de nouvelles offres, aller vers de nouveaux clients, fidéliser…). Cela implique une double démarche de développement : l'une concernant les produits, l'autre concernant les systèmes d'information, points d'appui des organisations.

Courant novateur de l'algorithmie, vaste domaine à mi-chemin entre l'informatique et les mathématiques, la **data science** est devenue un outil indispensable aux entreprises. En offrant un grand nombre de techniques, sur fond de concepts informatiques pointus tel que le Big Data, elle permet une expertise des données et de leurs traitements. En effet, les systèmes manipulent de plus en plus de données et cette large combinatoire exige des méthodes d'une performance accrue permettant de reconnaître l'information, donnée utile au fonctionnement du système, de l'extraire et enfin de la traiter.

I.1 Ingénierie des systèmes et gestion de la complexité

L'**ingénierie des systèmes** est au cœur de ses préoccupations. L'ingénieur système est le chef d'orchestre qui gère l'ensemble du cycle de vie du produit (processus qui part de l'offre jusqu'à la mise sur le marché : offre, développement, production, support client, et faisant intervenir les métiers de l'entreprise) et des métiers participant au développement (compétences métier et algorithmie). La conception d'un système est une collaboration entre trois compétences/métiers (savoir-faire métier, data science et ingénierie système) afin de :

- définir cette information et comment l'inscrire dans l'architecture du système ;
- définir les modes d'extraction ;
- définir ses relations avec d'autres données ;
- définir les traitements sur cette donnée.

Un consensus tripartite est ainsi nécessaire sur l'ensemble du développement. Ceci implique que les collaborateurs de ces différents métiers parlent un même langage et partagent un certain nombre de connaissances et compétences. Ainsi, le data scientist doit posséder des compétences en systémique, tout comme l'ingénieur système doit

en posséder en data science. Un développement harmonieux et efficace sera permis par une communication fluide (lien fort, communication et collaboration étroites, co-construction et partage de savoirs) entre trois types de compétences :

- **le savoir-faire métier** : il définit la pertinence des données associées au métier, leur importance, ce qui relève de l'information, ce qui est critique, les simplifications à apporter ;
- **la compétence en systèmes** : l'ingénieur système conçoit une architecture qui associe l'extraction des données utiles, leur traitement et leur exploitation (tableau de bord, prédiction…). Il est le chef d'orchestre technique du projet, le lien entre les différentes compétences métier, informatiques ou algorithmiques ;
- **la compétence en data science** : il est nécessaire de connaître l'actualité des techniques d'architecture informatique et de traitement des données. La data science apporte également ses compétences en data visualisation, de manière à construire une interface utilisateur qui réponde au besoin du client, externe ou interne.

L'ingénierie système technique (systèmes embarqués) et l'ingénierie des systèmes d'information sont deux domaines séparés et spécialisés :

- les ingénieurs système temps réel sont sur la ligne produit, en face des clients et des équipes de développement ;
- le système d'exploitation est en support sur l'organisation interne de l'entreprise (direction des opérations) et des processus.

Néanmoins, les deux axes sont liés. La compétitivité des lignes produit s'appuie sur la réactivité et la proactivité des systèmes d'information. Un système d'information met à la disposition de la ligne produit toutes les informations nécessaires à la performance des équipes produits : capitalisation de données stratégiques et des expertises, adaptation des processus, amélioration continue, aide à la décision, prédiction de coûts et délais, gestion des ressources. Nous rappelons les spécificités sur chaque typologie de système dans le Tableau I.1.

Tableau I.1 – Système temps réel et système d'information

	Système temps réel	Système d'information
Définition	Un système temps réel est un système qui réalise des tâches complexes dans un environnement complexe qu'il capte et auquel il tente de s'adapter. Il donne à l'utilisateur des réponses opérationnelles en vue d'actions sur cet environnement.	Un système d'information traite des données de l'entreprise afin de mesurer des indicateurs, de les exploiter et d'établir des projections sur le futur (tableaux de bord).
Exemples d'application	Fusion de données (guidage d'engins, sciences de la complexité comme l'astronomie, la météorologie, les sciences de la Terre…).	Surveillance et pilotage de processus (conduite de projet, marketing, industrie, logistique, veille technologique, devis et réponse à appel d'offres) Sciences économiques et sociales (analyse de données, modélisation de phénomènes, prédictions).

	Système temps réel	Système d'information
Technologie des données	Dépend du contexte. Quantité de données optimisée (système embarqué) ou vastes quantités de données, voire big Data (support aux opérations, applications scientifiques).	Big Data, bases de données.
Algorithmie	Fusion de données capteurs, automatique, traitement de signal ou d'image, statistiques, aide à la décision.	Construction des indicateurs, mise en forme des tableaux de bord, projections et aide à la décision : data science.
Interface utilisateur	Interface homme/machine (IHM), ergonomie, peu d'informations, temps de réaction (signalétique, alarmes)…	Data visualisation (aide à la décision : tableaux, graphiques)
Méthode de développement	Ingénierie système, cycle de développement en W, méthodes agiles.	Méthodologie de consultant, Merise…

Malgré les domaines d'applications et les spécialisations différentes, certaines compétences sont communes. Ces compétences transversales sont au nombre de six, décrites ci-après.

I.1.1 Une culture de la complexité

Les problèmes linéaires ne représentent que 20 % des problèmes rencontrés. Il existe en effet une classe de phénomènes (80 %), mis en évidence par Paul Lévy puis Benoît Mandelbrot, qui ne se comportent pas comme des processus gaussiens ou linéaires.

Ce sont principalement les problèmes économiques, sociologiques ou d'entreprise (économétrie), ainsi que ceux rencontrés en sciences physiques (météorologie, astronomie, sciences de la Terre, environnement de nombre de systèmes temps réel : par exemple les problèmes de rétrodiffusion et de propagation d'ondes radar en environnement marin) qui font appel à des outils d'analyse comme l'analyse fractale et les processus statistiques à queue lourde (Lois de Lévy…), et auxquelles les techniques classiques de traitement linéaire ne s'appliquent qu'avec de forts biais et incertitudes.

Les architectures de traitement de ce type de données et les méthodes non linéaires pouvant contribuer à la résolution de ces problèmes doivent être connus.

I.1.2 Une approche systémique de la complexité

Si la méthode cartésienne (XVII^{e} siècle) de réduction de la complexité à des composants élémentaires est adaptée à l'étude des systèmes stables constitués par un nombre limité d'éléments en interactions linéaires (décrites par des lois mathématiques proportionnelles, additives), elle ne convient plus pour l'étude des systèmes passé un certain niveau de complexité, d'incertitude et de possible logique émergente. On citera par exemple la biologie, l'économie ou les systèmes sociaux, mais aussi les sciences physiques (astronomie, sciences de la Terre, météorologie, écologie).

© Dunod - Toute reproduction non autorisée est un délit.

Une autre approche est requise, fondée sur de nouvelles représentations de la réalité, prenant en compte l'instabilité, la fluctuation, le chaos, le désordre, le flou, l'ouverture, la créativité, la contradiction, l'ambiguïté, le paradoxe… Pour rendre compte de la complexité, la systémique impose l'appréhension concrète de concepts qui lui sont propres : vision globale, système, niveau d'organisation, interaction, rétroaction, régulation, finalité, évolution. La démarche systémique moderne est largement influencée par la mondialisation, qui a stimulé la prise de conscience de la complexité (du cosmos, des organismes vivants, des sociétés humaines, et des systèmes artificiels conçus par les hommes). Elle a évolué vers l'étude de la complexité, avec une attention particulière accordée aux systèmes dynamiques (évolutifs) et a donné lieu à de nombreuses applications, en biologie, en écologie, en économie, dans le management des organisations…

Contrairement à l'analyse de la méthode de Descartes, la méthode systémique explore les concepts :

- par une approche globale (schémas, réseaux, modèles) ;
- par une démarche descendante (facteur d'échelle) du plus grand au plus petit des niveaux d'organisation (à l'image d'un hologramme ou d'une fractale) ;
- par une démarche transverse (pensée en arborescence et raisonnement par analogies) ;
- par une étude des relations (entrées, sorties, fonctions de transfert). plus que des objets.

Les grands précurseurs de ce domaine sont Norbert Wiener[1] et Ludwig Von Bertalanffy[2]. La systémique est un « Macroscope » (Joël de Rosnay) permettant de résoudre des problèmes. Il n'est pas étonnant qu'on retrouve cette méthode au cœur des compétences de l'ingénierie système.

Tableau I.2 – Approche analytique *versus* approche systémique

Approche analytique (faible complexité)	Approche systémique (forte complexité)
Analyse : décompose en éléments simples (atomes). Isole : se concentre sur les éléments.	Synthèse, relie : se concentre sur les interactions entre les éléments. Démarche descendante par niveaux d'organisation ou de structure de moins en moins complexes (hologramme, fractale).
Indépendante de la durée : les phénomènes sont supposés réversibles. Observateur indépendant et sans interaction (neutralité, pas d'objectif préconçu).	Intègre la durée et l'irréversibilité : l'expérimentateur est en interaction avec le système (action qui modifie le système, objectif qui oriente la vue de l'observateur)
La validation des faits se réalise par la preuve expérimentale dans le cadre d'une théorie.	La validation des faits se fait en comparant le modèle avec la réalité.

1. *La cybernétique. Information et régulation dans le vivant et la machine*, Paris, Le Seuil, 2014 pour la présente édition (première édition en 1948).
2. *Théorie générale des systèmes,* Malakoff, Dunod, 2012 pour la présente édition (première édition en 1968).

Approche analytique (faible complexité)	Approche systémique (forte complexité)
Théorie détaillée répondant à la construction d'une connaissance.	Modèle simplifié répondant à un objectif donné.
Déduction : la théorie déduit des principes qui se valident par les expérimentations et enrichissent la théorie.	Induction : les expérimentations conduisent au modèle et à une théorie qui amène à de nouvelles expérimentations, puis un nouveau modèle…
Modèle précis et détaillé, mais difficilement utilisable en pratique.	Modèles globaux, approchés avec un objectif défini (question). Ne peuvent pas servir de base de connaissance, mais sont très utiles en pratique.
Théorie détaillée répondant à la construction d'une connaissance.	Modèle répondant à un objectif donné.
S'appuie sur la précision des détails.	S'appuie sur la perception globale.
Modifie une variable à la fois : démarche séquentielle.	Modifie des groupes de variables simultanément : parallélisme.
Approche efficace lorsque les interactions sont linéaires et faibles.	Approche efficace lorsque les interactions sont non linéaires et fortes.
Conduit à un enseignement par discipline (juxta disciplinaire).	Conduit à un enseignement pluridisciplinaire.
Conduit à une action programmée dans les détails (planification détaillée).	Conduit à une planification par objectifs.

La démarche descendante de l'approche systémique peut se voir comme une démarche arborescente ou fractale qui conduit à des niveaux d'organisation ou de structure suffisamment simples pour être approchés par la démarche analytique. Les deux démarches sont donc complémentaires et peuvent être utilisées de manière hiérarchisée dans l'arbre des niveaux de complexité : les branches terminales peuvent être abordées par la démarche analytique.

I.1.3 La modélisation

La modélisation permet une représentation de la complexité et offre un support compréhensible par les différents métiers de la conception. La méthode systémique prend forme dans le processus de modélisation (Jean Louis Le Moigne – 1990[1]), qui utilise le langage graphique et permet l'élaboration de modèles (représentations simplifiées adaptées à un but particulier) qualitatifs (en forme de « cartes ») et la construction de modèles dynamiques, quantifiés, opérables sur ordinateur et débouchant sur la simulation.

I.1.4 L'algorithmie

L'algorithmie ((data science, traitement de l'information : signal, image, statistique), intelligence artificielle) est utilisée comme technologie de traitement des données. Il existe deux méthodes pour construire des algorithmes :

1. *La modélisation des systèmes complexes,* Malakoff, Dunod, 1990.

© Dunod - Toute reproduction non autorisée est un délit.

- la première consiste à s'appuyer sur une idée préconçue du comportement du système ou de son modèle. C'est la technique « *model driven* » qui s'appuie sur les lois physiques du système et de son évolution telles qu'on les perçoit (construction théorique basée sur une hypothèse initiale) ;
- l'autre méthode s'appuie sur les données telles qu'elles sont, sans idée préconçue sur le système. Peu d'hypothèses sont faites sur le système, seul son comportement dans les conditions d'utilisation est retenu. Cette méthode est expérimentale et itérative (construction empirique). L'algorithmie classique jusqu'aux années 1990 est plutôt « *model driven* ». Depuis une vingtaine d'années, la data science est plutôt « *data driven* ».

I.1.5 La data visualisation et l'ergonomie

L'ergonomie permet la mise en forme des données dans un format facilitant l'utilisation d'un outil pour l'utilisateur final. Ensemble de fonctionnalités, l'ergonomie est un outil servant la communication. Par exemple, les données vont être présentées et communiquées de manière à faciliter la lecture, le sens et l'interprétation, l'exploitation (compréhension, aide à la décision). Certains outils sont très connus : un diagramme, une jauge, une cartographie, un nuage de point… La lisibilité est améliorée par des gammes de couleurs harmonieuses et des indications suffisamment claires pour permettre de comprendre l'information en quelques secondes.

I.1.6 L'intelligence artificielle

L'intelligence artificielle (IA) est utilisée comme support à la décision (apprentissages, moteurs d'inférence, bases de règles…). L'apprentissage est un système dont la réponse évolue en fonction des données qui y sont injectées : c'est le *Machine Learning*. L'IA se présente comme la mise en œuvre d'agents intelligents, d'entités autonomes capables de percevoir leur environnement et d'interagir avec lui. Ils sont capables d'apprendre, d'analyser, d'utiliser des connaissances et de prendre des décisions.

Historiquement, les premières IA n'étaient pas réellement « apprenantes ». Elles utilisaient au mieux des fonctions de parcours de graphe (heuristiques) combinées avec des moteurs de règles. Aujourd'hui, le *Machine Learning* s'appuie sur des algorithmes (principalement statistiques) pour permettre à une machine « d'apprendre » à partir d'un certain nombre de réponses correctes connues au départ (échantillon ou base d'apprentissage). Sans cette base de données disponible, souvent très volumineuse, l'apprentissage est impossible.

Le *Machine Learning* fait appel à différentes techniques : les réseaux de neurones, des outils récents qui s'appuient sur la technologie Big Data, disponibles sur le marché du logiciel.

I.2 Présentation de l'ouvrage

Dans cet ouvrage, nous allons présenter la data science, son histoire, ses algorithmes et ses méthodes. Par ailleurs, nous allons décrire trois compétences communes de l'ingénierie des systèmes (système temps réel et système d'information) : complexité,

systémique, modélisation. Nous allons les définir et voir leur application sur les cycles de développement.

Afin d'illustrer ces compétences et les appliquer, nous avons choisi de développer trois exemples au fil de l'ouvrage :

- un **système temps réel** (fusion de données, système radar) ;
- un **système d'information** (pilotage adaptatif de l'entreprise) ;
- un **système financier** (gestion de placements financiers par adaptation à l'évolution des indicateurs boursiers).

Ces trois problèmes complexes font appel à la data science comme moyen de développement. Les exemples choisis couvrent un large spectre de problématiques pluri disciplinaires :

- la **data fusion**, une technologie qui s'applique à toutes les sciences et aux systèmes temps réel ;
- un **pilotage adaptatif des processus de l'entreprise** (contrôle statistique de processus), qui répond aux problématiques actuelles de compétitivité ;
- la **réalisation de produits financiers optimaux**, qui passe par une adaptation aux mouvements des marchés. Les autres compétences (algorithmie, visualisation et intelligence artificielle) seront abordées en toile de fond dans chacune des applications.

Pour chacun de ces trois exemples, nous développerons les étapes de la méthode systémique nécessaire, afin de montrer comment elle peut s'adapter à différentes applications.

© Dunod - Toute reproduction non autorisée est un délit.

1

DATA SCIENCE : HISTOIRE ET MÉTHODES

1 La data science

Dans ce chapitre, nous allons rappeler les principes, méthodes et outils de la data science. Ensuite, nous verrons comment appliquer le Machine Learning et quelques algorithmes statistiques (estimateurs de lois à queues lourdes) utiles aux applications spécifiques que nous souhaitons traiter.

Depuis 2010, le volume de données créées et exploitées dans le monde augmente de 30 % par an. Le volume de données numériques créées ou répliquées à l'échelle mondiale a été multiplié par plus de trente au cours de la dernière décennie, passant de 2 zettaoctets en 2010 à 64 zettaoctets en 2020. Si l'on souhaitait sauvegarder les 64 zettaoctets de données générées en 2020, 640 millions des plus gros disques SDD actuellement commercialisés (100 To de stockage) seraient nécessaires !

Afin de traiter cette masse de données toujours plus imposante, les méthodes informatiques se sont développées à grande vitesse (data science, intelligence artificielle…), un accent particulier ayant été mis en 2019 sur le Machine Learning. Les budgets américains et les publications dans le domaine augmentent de 25 % par an [19]. Les leaders mondiaux sont l'Inde, la Chine et les États-Unis.

1.1 Définition

Science des données, la data science s'intéresse à l'art de les extraire, de les traiter et de les visualiser sous différentes formes : nombres, logiques, textes… Elle s'appuie sur des outils mathématiques, de statistiques, d'informatique (c'est donc principalement une « science des données numériques ») et de visualisation des données.

Sur la Figure 1.1, sont indiquées les interactions de la data science avec son environnement. Cette interaction est la base de toute modélisation prudente, et d'une méthodologie rigoureuse.

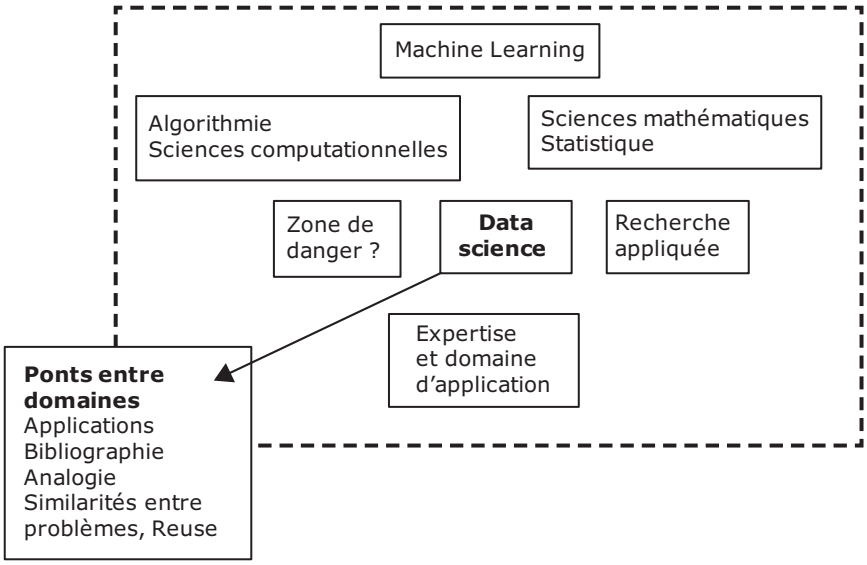

Figure 1.1 – La data science, à la croisée des disciplines

La data science est au cœur de plusieurs centres de connaissances et de compétences, décrite ci-dessous :

- **Les connaissances mathématiques et statistiques**. Nous allons ainsi développer des notions de convergence uniforme, de calcul de lois de probabilité et d'optimisation (Moindres carré, régression, Newton-Raphton…). Les lois de probabilité usuelles sont à décroissance rapide, mais elles feront intervenir dans notre travail des lois dites à queue lourde. Construire un histogramme de ces lois demande plus d'échantillons que dans le cas des courbes à décroissance rapide.
- **L'algorithmie et les sciences de l'informatique**. Il existe toute une famille d'algorithmes depuis le tri et la segmentation jusqu'à des analyses mathématiques et statistiques poussées (classification de données).
- **Les domaines d'application (expertise métiers).** Le data scientist travaille en équipe avec un architecte système et un spécialiste métier. Il est donc amené à connaître le langage et les techniques de ces disciplines. Ces résultats seront validés à la lumière de l'expertise métier. Par exemple, une régression ou une corrélation ne prendront de sens que si elles sont validées par un sens opérationnel. Les opérationnels métiers veilleront à la cohérence des aspects suivants :
  - paramètres ayant une signification physique ou opérationnelle directement analysable en termes d'ordre de grandeur et de signification opérationnelle. Certains algorithmes extraient des paramètres de calcul (par exemple le modèle ARMA). Un travail pour ramener ces paramètres à des aspects directement opérationnels sera effectué ;

- relations de cause à effet : la corrélation entre variables induit des inductions qui ne peuvent trouver leur sens que par une analyse mathématique rigoureuse et l'expertise domaine.

Tableau 1.1 – Le cycle de développement de la data science

Étape	Bonne pratique
Analyse du problème	L'expertise dans le domaine d'application concerné permet de rendre les modèles réalisés interprétables (confiance permise par une validation du modèle).
Spécification des algorithmes	Expertise statistique jointe à un sens opérationnel aigu. Les modèles doivent être rigoureusement justifiés (physique fondamentale et mathématiques avancées).
Data visualisation	Apporter au décideur une information pertinente et utile corroborée par l'application opérationnelle des décisions.
Codage	Simulateur avec données de scénarios, lisibilité, décomposition du problème, évolutivité du code (style de programmation)…
Validation et visualisation des résultats	Data visualisation en cascade afin de maîtriser les différents niveaux de modèle, du résultat global aux étapes. Travail en équipe avec les métiers et les opérationnels.
Documentation	Démarche industrielle rigoureuse : documentation et jeux d'essais commentés.

1.2 Principaux algorithmes utilisés

Depuis les années 2000, période à laquelle la Data Science a été institutionnalisée, de nombreux algorithmes d'analyse de données ont été mis en place. Néanmoins, ces travaux s'appuient sur des recherches antérieures (théorie du signal, traitement de l'image, économétrie…). Il existe de très nombreux algorithmes utiles, aussi nous ne citerons que les principaux (Tableau 1.2).

© Dunod - Toute reproduction non autorisée est un délit.

Tableau 1.2 – Principaux algorithmes de data science

Algorithmes	Utilisation	Exemples
Analyse des séries temporelles	Réaliser une prédiction à partir d'un nombre donné d'échantillons.	– ARMA : modèle à variance constante $x(t) = A(t) + \sigma N(t)$ – ARCH, GARCH : modèles à variance évolutive (volatilité) $x(t) = A(t)\left(1 + \sigma N(t)\right)$ – N bruit gaussien décorrélé σ constant.
Régression multiple	Analyser les liens entre une variable à expliquer et plusieurs variables quantitatives explicatives indépendantes comme on l'admet.	– Déterminer les équations d'un ajustement polynomial non-linéaire pour l'analyse des liens entre deux variables quantitatives. – Déterminer les équations de surfaces de tendances. – Analyser la rugosité du relief. – Déterminer les équations polynomiales d'un modèle de correction géométrique applicable à des vecteurs.
Association de données	Distinguer des groupes de données sur un critère ou plusieurs : « Comment ces données forment-elles des groupes et à quels clusters appartiennent-elles ? »	Clustering (regroupement de données) ou reconnaissance de forme (Pattern recognition)
Classification	« À quelle catégorie appartiennent ces données ? »	– Arbres de décision générés par un ordinateur pour répertorier les données dans des catégories définies. – Méthodes probabilistes. – Réseaux de neurones, qui ont démontré de profondes capacités de classification quand ils sont appliqués à de grands volumes de données.
Apprentissage	Il s'agit pour un type de problème donné (estimation, prédiction) de construire un schéma d'approximations successives pour régler les paramètres de l'algorithme. Ce réglage se fait sur un nombre assez grand de configurations. Les paramètres ainsi réglés permettent de résoudre de nombreux cas qui bien sûr font évoluer les réglages.	– Réseaux de neurones. – Machine Learning.

1.3 Application des méthodes d'apprentissage aux cas retenus

Prenons pour base un système stationnaire et ergodique, c'est-à-dire que ses propriétés statistiques ne dépendent pas du temps. Simplement, la corrélation entre deux instants ne dépend que de l'intervalle entre ces instants. Nous pouvons donc définir une loi de probabilité des valeurs numériques observées $pth(x,q)$ où q est le vecteur des paramètres d'entrée.

Le seul accès que nous avons de pth est un histogramme $s(x)$ sur une taille limitée d'échantillon (taille N). Un histogramme définit des classes de valeur, caractérisées par une valeur constante Dx.

Nous définissons donc un problème similaire à celui de l'apprentissage :

$$x(i) = X0 + iDx$$

Et

$$y(i) = \frac{n(x(i))}{N}$$

où $n(x(i)$ représente le nombre de valeurs entre $x(i)$ et $x(i) + Dx$. $Y(i)$ est donc la fréquence mesurée associée à $x(i)$.

Les fonctions f qui sont « optimales » au sens de l'approximation sont la densité de probabilité théorique de x, $pth(x,q)$. La loi est paramétrée par ses « moments » généralisés, eux-mêmes fonctions des paramètres d'entrée q.

$$h(x,q) = pth(x,q)$$

La loi de coûts est celle de la convergence uniforme aux moindres carrés :

$$R_{emp}(q) = \int \left(h(x,q) - s(x)\right)^2 dx$$

Pour déterminer $s(x)$, et en particulier l'intervalle Dx optimal, on se référera à la référence [20].

La convergence uniforme des histogrammes $s(x)$ vers $pth(x,q)$ est assurée par la référence [21].

La convergence uniforme assure la faisabilité de l'apprentissage.

L'apprentissage s'appuie donc sur la fonction $pth(x,q)$. Un raisonnement rigoureux est nécessaire pour construire cette fonction. Nous allons nous appuyer sur une bibliographie abondante avec, en particulier, de fortes analogies avec la physique des phénomènes complexes et systèmes auto organisés. Nous nous référerons donc à une expertise double : mathématique et statistique d'une part, domaine et complexité d'autre part.

© Dunod - Toute reproduction non autorisée est un délit.