Annuaire du Collège de France

121e année

2020
2021

Résumé des cours et travaux

COLLÈGE
DE FRANCE
1530

Annuaire du Collège de France
Cours et travaux du Collège de France

121 | 2024
2020-2021

Sciences des données

Stéphane Mallat

Édition électronique
URL : https://journals.openedition.org/annuaire-cdf/19182
DOI : 10.4000/12kti
ISBN : 978-2-7226-0778-1
ISSN : 2109-9227

Éditeur
Collège de France

Édition imprimée
Date de publication : 18 novembre 2024
Pagination : 37-45
ISBN : 978-2-7226-0777-4
ISSN : 0069-5580

Ce document vous est fourni par Collège de France

Référence électronique
Stéphane Mallat, « Sciences des données », *L'annuaire du Collège de France* [En ligne], 121 | 2024, mis en ligne le 01 octobre 2024, consulté le 28 novembre 2024. URL : http://journals.openedition.org/annuaire-cdf/19182 ; DOI : https://doi.org/10.4000/12kti

SCIENCES DES DONNÉES

Stéphane Mallat

Membre de l'Institut (Académie des sciences)
et de l'Académie des technologies,
professeur au Collège de France

La série de cours « Représentations parcimonieuses » est disponible en audio et vidéo, sur le site internet du Collège de France (https://www.college-de-france.fr/fr/agenda/cours/representations-parcimonieuses), ainsi que le colloque « The representation of language in brains and machines » (https://www.college-de-france.fr/agenda/colloque/the-representation-of-language-in-brains-and-machines).

ENSEIGNEMENT

Cette année, il n'y a pas eu de séminaire en parallèle du cours, car l'enseignement de janvier à mars 2021 ne s'est pas fait en présence du public à cause des restrictions sanitaires. Cependant, des challenges de données ont été organisés ainsi qu'un colloque en collaboration avec Stanislas Dehaene et Luigi Rizzi sur « La représentation du langage dans le cerveau et les machines ».

COURS – REPRÉSENTATIONS PARCIMONIEUSES

Introduction

L'utilisation de représentations parcimonieuses est au cœur de la démarche scientifique de modélisation, à travers le concept philosophique du rasoir d'Occam, mais la parcimonie est aussi fondamentale pour la construction de modèles de basse

S. Mallat, « Sciences des données », *Annuaire du Collège de France 2020-2021. Résumé des cours et travaux*, 121e année, Paris, Collège de France, 2024, p. 37-45, https://doi.org/10.4000/annuaire-cdf.19182.

dimension pour le traitement de données. Le cours montre l'équivalence mathématique des notions de « parcimonie », d'« approximation », et de « régularité », définies par des opérateurs linéaires ou non linéaires.

Les représentations parcimonieuses ont de nombreuses applications pour la compression de données, le débruitage, la résolution de problèmes inverses, ainsi que pour l'apprentissage statistique. Le cours est consacré aux propriétés mathématiques des approximations parcimonieuses qui font appel à l'analyse harmonique, la transformée de Fourier et les bases d'ondelettes. On étudie aussi les approximations en grande dimension avec des réseaux de neurones à deux couches.

Cours 1 – Le triangle « Régularité, Approximation, Parcimonie »

Le 13 janvier 2021

La réduction de la dimensionnalité est au cœur de de la modélisation et de l'analyse de données, que ce soit pour représenter les données x ou des fonctions de ces données $f(x)$. On veut construire des modèles ayant le moins de variables possible de x ou de f, ce que l'on appelle « une représentation parcimonieuse ». Cela correspond au principe du rasoir d'Occam en sciences et en philosophie, « les explications les plus simples sont les meilleurs ». Ce cours d'introduction rappelle ces notions et explique pourquoi la réduction de la dimensionnalité est une étape importante de la compression de signaux x avec le moins de bits possibles, mais aussi pour la suppression de bruit additionné à un signal x inconnu ou pour des problèmes inverses, ou bien il faut estimer x à partir de mesures incomplètes et bruitées. C'est aussi le cas pour les problèmes de classification ou de régression, où il s'agit d'approximer la fonction $f(x)$ avec un nombre limité de données d'entraînement. Dans tous les cas, la réduction de dimensionnalité permet de réduire le nombre de paramètres à estimer, par exemple ceux d'un réseau de neurones pour l'approximation de f.

Le cours va aborder ce sujet de trois points de vue différents mais équivalents : l'existence de *régularités*, l'*approximation* en basse dimension, et la construction de représentation *parcimonieuses*. Ces trois notions sont les sommets du triangle RAP (Régularité, Approximation, Parcimonie) dont le cours va étudier les propriétés mathématiques, avec des approches linéaires et non linéaires.

Cours 2 – Approximations linéaires et analyse de Fourier

Le 20 janvier 2021

Le cours commence par montrer la différence entre des approximations linéaires et non linéaires dans des bases orthonormées. L'approximation linéaire d'un signal x se fait en sélectionnant un nombre limité M de coefficients de décomposition dans une base, alors qu'une approximation non linéaire peut adapter le choix de ces M coefficients en fonction de x, notamment en choisissant les plus grands. Un exemple d'approximations linéaires s'obtient avec un échantillonnage uniforme d'une

fonction, par opposition à un échantillonnage adaptif et non linéaire, qui s'adapte à la régularité locale de la fonction.

Dans un cadre linéaire, le problème d'approximation en basse dimension est abordé en étudiant la régularité sous-jacente. La régularité d'une fonction *x(t)* peut se définir par l'existence de *k* dérivées d'énergies finies, ce qui correspond à la régularité de Sobolev. Les propriétés des dérivées sont caractérisées dans une base qui diagonalise l'opérateur de dérivation. Cet opérateur étant covariant par translation, on démontre qu'il s'agit de la base de Fourier. Le cours fait un bref rappel des propriétés des bases et de l'intégrale de Fourier. La régularité de Sobolev est définie sur l'intégrale de Fourier en une dimension.

Cours 3 – Grande dimension et composantes principales

Le 27 janvier 2021

On commence par construire l'extension séparable d'une base pour des fonctions de plusieurs variables, et définir la régularité de Sobolev en dimension *p* quelconque à partir de la décroissance des coefficients de Fourier capturée par une série convergente. On démontre l'équivalence entre cette régularité et la décroissance de l'erreur d'approximation à partir de *M* coefficients de Fourier en dimension *p*. Ce théorème met en évidence la malédiction de la dimensionnalité, qui nécessite un nombre de coefficients *M* qui grandit exponentiellement avec la dimension, pour atteindre une erreur d'approximation fixée. Dans le cas de la base de Fourier, ces approximations peuvent aussi s'interpréter comme des filtrages qui ne gardent que les basses fréquences.

L'optimalité des approximations linéaires est étudiée dans un cadre probabiliste à partir des composantes principales. Lorsque *x* est un vecteur aléatoire, on démontre que l'erreur d'approximation linéaire à partir de *M* coefficients dans une base orthonormée ne dépend que de sa matrice de covariance. On démontre le théorème d'approximation de Karhunen-Loève, à savoir que les bases qui minimisent l'approximation linéaire sont des bases qui diagonalisent la matrice de covariance. Les vecteurs d'une telle base sont appelés des « composantes principales ». Si *x(t)* est stationnaire, alors sa matrice de covariance est une matrice de convolution diagonalisée dans la base de Fourier. Les composantes principales sont donc des sinusoïdes. Les approximations linéaires de Fourier sont optimales dans ce cas. Ceci termine le parcours du triangle entre Régularité, Approximation et Parcimonie, dans le cas linéaire.

Cours 4 – Approximations non linéaires et réseaux de neurones

Le 3 février 2021

Ce cours aborde le triangle Régularité, Approximation et Parcimonie dans un cadre non linéaire. L'approximation non linéaire optimale de *x* dans une base

orthonormée revient à sélectionner les coefficients de x dans la base, de plus grandes amplitudes. On démontre que la vitesse de décroissance de l'erreur d'approximation dépend de la vitesse de décroissance des coefficients ordonnés, ce qui peut être spécifié avec des normes l^p.

Le cours applique ces résultats d'approximations non linéaires aux réseaux de neurones à une couche cachée. Il démontre que l'apprentissage d'un tel réseau revient à calculer une approximation non linéaire, qui dépend de la non-linéarité ponctuelle utilisée dans le réseau. Dans le cas où cette non-linéarité est une sinusoïde, l'apprentissage calcule une approximation non linéaire dans une base de Fourier. De telles approximations sont optimales dans des espaces de Barron, que l'on caractérise en fonction de la vitesse de décroissance de l'erreur d'approximation. Cependant, l'utilisation de ces espaces donne des bornes pessimistes, car ils ne tiennent pas en compte le fait que les données x se concentrent dans des ensembles typiques qui sont beaucoup plus petits que l'espace global. En grande dimension, pour capturer cette concentration il faut définir des modèles probabilistes et approximer les distributions de probabilités sous-jacentes. Ceci sera le sujet du cours de l'année prochaine.

Cours 5 – Ondelettes et échantillonnage

Le 10 février 2021

En basse dimension, l'approximation non linéaire est généralement fondée sur l'existence de régularités locales. Le cours montre que les base d'ondelettes jouent un rôle particulier car elles permettent d'obtenir des approximations quasi-optimales de fonctions qui sont localement régulières. La régularité locale peut être spécifiée par un exposant de Lipchitz.

La transformée en ondelettes est définie par projection de $x(t)$ sur des ondelettes qui sont des fonctions localisées. Celles-ci sont déduites d'une ondelette mère, qui est translatée et dilatée, ce qui permet de définir une base d'ondelettes. On démontre que l'exposant de régularité de Lipchitz local s'obtient à partir de la décroissance locale des coefficients d'ondelettes, lorsque l'échelle tend vers 0.

La décomposition d'un signal $x(t)$ dans une base d'ondelettes est aussi reliée à une approximation par échantillonnage de t. Dans le cas où l'ondelette mère est une ondelette de Shannon, dont la transformée de Fourier est l'indicateur d'intervalles, la décomposition dans la base d'ondelettes s'obtient à partir du théorème d'échantillonnage de Shannon.

Cours 6 – Multirésolutions

Le 17 février 2021

La théorie des multirésolutions donne un cadre mathématique pour construire les bases orthonormées d'ondelettes et obtenir un calcul rapide des coefficients de décomposition dans une base d'ondelettes. Le point de départ est l'approximation de

signaux x à différentes échelles 2^j par projections linéaires dans des espaces V_j qui sont emboîtés, et qui définissent une multirésolution. Ces projections sont calculées en construisant des bases orthonormées de chaque espace V_j, en dilatant et translatant une fonction, que l'on appelle la « fonction d'échelle ». Ces fonctions d'échelles sont caractérisées par des filtres discrets $h(n)$ dont on spécifie la transformée de Fourier afin d'obtenir des bases orthonormées de chaque V_j. Le calcul des projections dans ces espaces s'obtient par un algorithme de filtrage par h et de sous-échantillonnage à chaque échelle. La base d'ondelettes de Haar correspond à des approximations constantes par morceaux, calculées par des moyennes successives.

Cours 7 – Bases orthonormées d'ondelettes

Le 3 mars 2021

Le cours introduit la construction des bases d'ondelettes orthogonales ainsi que l'algorithme rapide pour calculer les coefficients de décomposition en ondelettes. On obtient une base d'ondelettes à partir d'une multirésolution en calculant le complément orthogonal W_j de V_j dans V_{j-1}. Chaque espace W_j admet une base orthonormée en dilatant une ondelette mère 2^j et en la translatant. L'union de ces bases à toutes les échelles 2^j définie une base orthonormée d'ondelettes de l'espace L^2. Ces ondelettes se construisent à partir du filtre passe-bas h qui spécifie chaque multirésolution, en introduisant un nouveau filtre g passe-bande qui dépend de h. Ces filtres satisfont une condition de quadrature qui est suffisante pour générer des bases orthonormées d'ondelettes. L'algorithme rapide calcule les coefficients d'ondelettes à travers les échelles en itérant des convolutions avec h et g suivies de sous-échantillonnages.

En dimensions multiples, les bases d'ondelettes s'obtiennent avec un produit tensoriel d'ondelettes d'une seule variable. En deux dimensions pour des images, la base s'obtient avec trois ondelettes qui capturent les variations de l'image dans des directions différentes, à chaque échelle 2^j.

Cours 8 – Parcimonie et compression d'images

Le 10 mars 2021

On étudie l'approximation de signaux et d'images avec des représentations parcimonieuses dans une base d'ondelettes, ainsi que l'application à la compression d'images. La vitesse de décroissance des coefficients d'ondelettes dépend de la régularité locale du signal. Les coefficients d'ondelettes d'une fonction régulière par morceaux ont une grande amplitude au voisinage des singularités. On relie la vitesse de décroissance de l'erreur d'approximation non linéaire dans une base d'ondelettes à la régularité locale du signal, et à la parcimonie de ses coefficients. La régularité locale s'exprime par l'appartenance à des espaces de Besov qui sont caractérisés par les normes l^p des coefficients d'ondelettes. Pour des images, l'erreur d'approximation

dans une base d'ondelettes dépend de la longueur des contours le long desquels l'image est discontinue.

Un algorithme de compression dans une base d'ondelettes commence par quantifier les coefficients d'ondelettes du signal, puis code leurs valeurs sous forme binaire avec un codage entropique. Pour des représentations parcimonieuses, on démontre que le nombre de bits nécessaire est proportionnel au nombre M de coefficients non nuls dans la base. L'erreur introduite par la quantification des coefficients peut aussi être reliée à M, et donc au nombre de bits utilisés par le code. Le standard de compression d'images JPEG-2000 est implémenté par un tel algorithme dans une base d'ondelettes.

CHALLENGES DE DONNÉES

Pour les étudiants et participants au cours, le site web challengedata.ens.fr met à disposition des nouveaux challenges de traitement de données par apprentissage supervisé pour la saison 2021. Ces challenges sont proposés par des entreprises ou des scientifiques, et sont issus de problématiques concrètes qu'ils rencontrent dans leur activité. Ils s'inscrivent dans un esprit d'échange scientifique, avec un partage de données et d'algorithmes.

Chaque challenge fournit des données labélisées, ainsi que des données de test. Les participants soumettent sur le site web leurs prédictions calculées sur les données de test. Le site calcule un score avec une métrique d'erreur qui est spécifiée. Il fournit un classement aux participants, ce qui permet d'évaluer leurs résultats dans une large communauté. Les challenges commencent le 1er janvier 2021. Une clôture intermédiaire a lieu en juin par une évaluation des prédictions sur de nouvelles données de test. La clôture finale est en décembre.

Cette année, les challenges ont été organisés et supervisés à l'ENS par Rudy Morel et Tanguy Marchand, avec la participation de Florentin Guth, Louis Thiry, Gaspard Rochette, et John Zarka. L'organisation de ces challenges de données est soutenue par la chaire CFM de l'École normale supérieure, et par la Fondation des sciences mathématiques de Paris. Les 11 challenges suivant ont été organisés :

– « Transactions sur actions : prédiction du volume des enchères », présenté par Éric Lebigot de la société CFM. Dans beaucoup de marchés actions, des enchères sont effectuées sur chaque stock en fin de journée. Étant donné les volumes et prix des transactions effectuées au sein d'une journée, le but est de prédire le volume d'action qui va être mis à disposition pour ces enchères en fin de journée ;

– « Détection automatique de l'apnée du sommeil », présenté par Valentin Thorey et Antoine Guillot de la société Dreem. Sur des signaux physiologiques enregistrés durant la nuit (EEG, ECG, signaux respiratoires) le but est de délimiter les périodes d'apnée du sommeil ;

– « Extraction multi-modale de la couleur d'un produit », présenté par Anuvabh Dutt, Aloïs de La Comble, Parantapa Goswami et Laurent Ach de la société Rakuten. Étant donné les textes et images d'annonces internet de produits, le but est de déterminer la/les couleur(s) du produit vendu ;

– « Prédiction du type de sol sur des images satellites », présenté par Jules Bourcier de la société Preligens. Étant donné les images satellites, et 10 types de sols parmi lesquels nuages, zones cultivées, végétations, etc., le but est de prédire la proportion de chaque type de sol sur l'image ;

– « Prédiction de l'incertitude dans la qualité de l'air », présenté par Maurice Charbit, Max Cohen et Sylvain Le Corff de la société Oze-Energies. Le but est de prédire avec précision la qualité de l'air dans un bâtiment, plus particulièrement de prédire à chaque instant une borne inférieure et une borne supérieure sur la température et le taux d'humidité ;

– « Classifications du genre, de l'instrument et de l'ambiance d'un morceau de musique », présenté par François Malabre de la société Mewo. À partir de variables de sortie d'un réseau de neurone développé par l'entreprise, le but est de prédire le genre, l'instrument et l'ambiance d'un morceau de musique ;

– « Reconstruction des performances d'actifs liquides », présenté par Nazih Benoumechiara de la société QRT. À partir des performances d'actifs non liquides et de leur catégorisation, le but est de prédire le signe des performances d'actifs liquides ;

– « Analyse de l'utilisation de bornes de chargement de véhicules électriques à Paris », présenté par Aude Laurent de la société Planète OUI. À partir de l'historique d'utilisation des bornes de chargement à Paris et d'informations contextuelles, le but est de prédire l'état des bornes sur deux semaines (« en chargement », « disponible », « déconnectée » ou « hors service ») ;

– « Qui sont les *traders* haute-fréquence ? », présenté par Iris Lucas de l'institution AMF. À partir de données de transaction sur un groupe de *traders*, le but est de classifier les *traders* en trois catégories : *trader* haute fréquence, *trader* non haute fréquence et *trader* mixte ;

– « Inspection visuelle par ordinateur dans une chaîne de production », présenté par Christophe Sauvanet, Éric Manouvrier, Alexandre Briot, Souhaiel Khalfaoui de la société Valéo. À partir d'images de pièces d'usine fournies par une caméra d'inspection optique, le but est de prédire le statut de la pièce entre valide et hors service ;

– « Segmentation sinusoïdale d'images électromagnétiques », présenté par Ana Escobar et Salma Benslimane de la société Schlumberger. Sur des images électromagnétiques de puits de pétrole sur lesquelles apparaissent des sinusoïdes représentant des formations géologiques, le but est de segmenter avec précision l'amplitude et la position de ces formes.

COLLOQUE – LA REPRÉSENTATION DU LANGAGE DANS LE CERVEAU ET LES MACHINES

Les 24 et 25 juin 2021

Ce colloque a été conjointement organisé avec Stanislas Dehaene et Luigi Rizzi. Le but de cette rencontre était de discuter de la convergence et des divergences entre les modèles statistiques, les approches neuroscientifiques et formelles du langage. La dernière décennie a été marquée par des avancées majeures dans les approches statistiques de l'apprentissage automatique du langage, conduisant à de nouvelles représentations et applications. Dans le même temps, les neurosciences cognitives ont fait des progrès significatifs sur la représentation du langage dans le cerveau, et la linguistique formelle a continué à faire des progrès constants sur la description structurelle du langage. Les trois approches ont suivi leurs propres chemins indépendants, ce qui est naturel, étant donné les différences substantielles dans les méthodologies et les objectifs. Néanmoins, on aurait pu s'attendre à un niveau d'intégration plus élevé. Ce colloque avait pour but de favoriser l'interaction entre ces domaines, à travers des présentations d'un groupe sélectionné de linguistes, psychologues, scientifiques du cerveau et chercheurs en apprentissage automatique. Il y a eu 15 exposés d'experts internationaux en sciences cognitives, linguistique et sciences des données et deux tables rondes, organisés sur deux jours.

RECHERCHE

Stéphane Mallat dirige l'équipe de recherche « Data » à l'École normale supérieure, qui étudie des problèmes de mathématiques appliquées aux sciences des données. Cela couvre l'apprentissage supervisé, l'apprentissage non supervisé ainsi que des problèmes inverses de traitement du signal.

L'équipe travaille sur des modèles mathématiques permettant d'expliquer la performance des réseaux de neurones profonds. Ils se basent notamment sur la transformée en ondelettes. En 2019-2020, une partie importante de la recherche était dédiée à la construction de modèles stochastiques pour la cosmologie, ainsi qu'à l'étude de modèles mathématiques permettant d'expliquer les phénomènes de concentration observés dans les classificateurs par réseaux de neurones.

PUBLICATIONS

Mallat S. et Zhang S., « Maximum entropy models from phase harmonic covariances », *Applied and Computational Harmonic Analysis*, vol. 53, 2021, p. 199-230, https://doi.org/10.1016/j.acha.2021.01.003 [arXiv : 1911.10017].

Allys E., Marchand T., Cardoso J.-F., Villaescusa-Navarro F., Ho S. et Mallat S., « New interpretable statistics for large scale structure analysis and generation », *Physical Review D*, vol. 102, nº 10, 2020, art. 103506, https://doi.org/10.1103/PhysRevD.102.103506.

Brochard A., Błaszczyszyn B., Mallat S. et Zhang S., « Particle gradient descent model for point process generation », 2020, https://doi.org/10.48550/arXiv.2010.14928 [arXiv : 2010.14928].

Zarka J., Guth F. et Mallat S., « Separation and concentration in deep networks », soumis pour les actes de la *9th International Conference on Learning Representations (ICLR 2021)*, 2021.