Annuaire du Collège de France

122e année

2021
2022

Résumé des cours et travaux

COLLÈGE
DE FRANCE
1530

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 « Information et complexité » est disponible en audio et en vidéo sur le site internet du Collège de France (https://www.college-de-france.fr/fr/agenda/cours/information-et-complexite), de même que la série de séminaires qui lui est associée (https://www.college-de-france.fr/fr/agenda/seminaire/information-et-complexite). Sont également disponibles en ligne les liens vers des articles étudiés lors du cours (https://www.di.ens.fr/~mallat/CoursCollege.html) ainsi que des notes de cours (https://www.di.ens.fr/~mallat/College/Cours-2022-Mallat-Jean-Eric-Campagne.pdf).

ENSEIGNEMENT

COURS – INFORMATION ET COMPLEXITÉ

Le cours introduit une approche mathématique de l'apprentissage statistique à travers l'estimation par maximum de vraisemblance, la théorie de l'information et la construction de modèles d'approximation. Les apprentissages non supervisé et supervisé passent par l'estimation de distributions de probabilités en grande dimension, à partir des données d'apprentissage. Cela nécessite de construire des modèles paramétrés, définis par une information *a priori*. Il peut s'agir de réseaux de neurones profonds dont l'architecture est spécifiée.

S. Mallat, « Sciences des données », *Annuaire du Collège de France. Résumé des cours et travaux*, 122e année : *2021-2022*, 2025, p. 41-50, https://journals.openedition.org/annuaire-cdf/20417.

Le cours soulève les questions fondamentales de modélisation en grande dimension et leur formalisation mathématique à travers des mesures d'information. Il introduit les notions d'information de Fisher pour l'inférence de modèle par maximum de vraisemblance et d'information de Shannon pour la prédiction et le codage. L'information de Shannon est fondée sur une notion de concentration et de mesure d'incertitude par l'entropie.

La construction de classes de modèles se base sur des hypothèses concernant la structure des distributions et leurs invariants. On s'intéresse particulièrement aux données « complexes » qui mettent en jeu de nombreuses échelles de variabilité, que ce soit des images, des sons, des séries temporelles ou des données qui proviennent de la physique. Le cours étudie des applications à la compression de signaux et d'images ainsi qu'à l'apprentissage non supervisé.

Cours 1 – Information et complexité

Le 19 janvier 2022

Les problèmes de classification peuvent se modéliser de façon déterministe ou probabiliste. Ces deux approches se distinguent par la façon de représenter l'information *a priori*. En grande dimension, les modèles stochastiques offrent des représentations souvent plus fines des données en mettant en évidence les phénomènes de concentration qui sont au cœur des probabilités et des statistiques.

La première partie du cours concerne les bases des statistiques mathématiques, qui ont été établies il y a un siècle par Ronald Fisher à travers les notions d'estimation consistante, de modèle paramétrique, de maximum de vraisemblance et d'information de Fisher. L'apprentissage des réseaux de neurones se fait le plus souvent par maximisation de la vraisemblance des paramètres.

La seconde partie du cours concerne la théorie de l'information de Shannon qui permet de quantifier l'information intrinsèque apportée par des données, à travers la notion d'entropie. L'une des applications est la compression de données. La théorie de la complexité est ici considérée à travers la recherche de structures parcimonieuses ou de régularités permettant de construire des paramétrisations de distributions de probabilités en grande dimension.

Cours 2 – Estimation par maximum de vraisemblance

Le 26 janvier 2022

Les statistiques sont essentiellement fondées sur des phénomènes de concentration qui sont des conséquences de la loi des grands nombres. On revoit ici la loi faible des grands nombres et la consistance de l'estimateur d'un paramètre. Le cours introduit l'estimation des paramètres d'une distribution de probabilités par maximum de vraisemblance, par annulation du score. Cet estimateur est appliqué aux distributions de Laplace et gaussiennes ainsi qu'à la régression logistique.

Cours 3 – Optimisation et modèles exponentiels

Le 2 février 2022

Maximiser la vraisemblance revient à minimiser une fonction de coût qui est l'opposée de la log-vraisemblance. Cette minimisation peut se calculer par descente de gradient. Sa convergence dépend du hessien de la fonction de coût. La convergence est garantie si le hessien est strictement positif et la vitesse de convergence exponentielle dépend de son conditionnement.

On considère le cas particulier des distributions exponentielles définies par une énergie de Gibbs qui dépend linéairement des paramètres. On démontre que le hessien est alors toujours positif mais peut être mal conditionné. L'optimisation des paramètres peut s'interpréter comme un déplacement sur une variété riemannienne, ce qui est le point de vue de la géométrie de l'information. On considère le cas particulier des distributions gaussiennes multivariées.

Cours 4 – Information de Fisher

Le 9 février 2022

Sous hypothèses de régularité, on démontre que l'estimateur par maximum de vraisemblance est consistant. L'information de Fisher est définie comme la variance du score, qui est le gradient de la log-vraisemblance. On montre que c'est aussi le hessien de la négative log-vraisemblance. Celle-ci est additive pour des variables aléatoires indépendantes.

Le résultat principal est la borne de Cramér-Rao. Celle-ci exprime une borne inférieure de la variance d'un estimateur des paramètres d'une loi de probabilité, à partir de l'inverse de l'information de Fisher. Trouver une bonne paramétrisation d'une loi de probabilité revient donc à minimiser cette inverse. On démontre plus précisément que, sous hypothèses de régularité, l'estimation asymptotique par maximum de vraisemblance a une distribution asymptotiquement gaussienne dont la covariance est l'inverse de l'information de Fisher.

Cours 5 – Théorie de l'information de Shannon

Le 16 février 2022

Dans son article sur la théorie mathématique des communications, Claude Shannon définit l'entropie comme une mesure d'incertitude qui est la complexité descriptive d'une variable aléatoire. Cette notion correspond à l'entropie en physique statistique, laquelle est notamment au cœur de la thermodynamique. Elle dépend du nombre de configurations possibles du système physique. Par opposition à l'information de Fisher, l'approche de Shannon est non paramétrique. Mathématiquement, l'entropie de Shannon est l'espérance de la négative log-probabilité.

On établit les propriétés de l'entropie jointe et conditionnelle de paires de variables aléatoires ainsi que leur information mutuelle et leur entropie relative. On démontre qu'un grand nombre de variables aléatoires indépendantes de même probabilité se concentrent avec forte probabilité dans un ensemble typique dont le cardinal est proportionnel à l'entropie de cette distribution de probabilité. Ce résultat permet de démontrer que la taille d'un code optimal est bornée par l'entropie, et que l'on peut s'approcher arbitrairement de cette borne par un codage typique.

Cours 6 – Codage entropique et modèles d'entropie maximum

Le 23 février 2022

On introduit les codes instantanés, définis sur des alphabets finis de symboles, et les codes de préfixe, qui peuvent être représentés par des arbres binaires. On démontre le théorème de Shannon à travers le lemme de Kraft, qui montre que la longueur minimum de ces codes est bornée inférieurement par l'entropie. Le code de Shannon est un code instantané qui s'approche arbitrairement de l'entropie lorsqu'il est calculé sur des blocs dont la longueur augmente. Le code de préfixe optimal s'obtient par l'algorithme de Huffman.

La notion d'entropie s'étend à des variables à valeurs réelles par la notion d'entropie différentielle, qui n'est pas toujours positive. On démontre un résultat d'équipartition asymptotique qui vérifie que la densité jointe d'un grand nombre de variables aléatoires indépendantes est quasiment constante sur des ensembles typiques, dont le volume dépend de l'entropie. L'entropie définit donc le volume du domaine dans lequel une variable aléatoire est concentrée.

L'entropie de Shannon se relie à l'information de Fisher à travers les modèles de probabilités d'entropie maximum. On définit un modèle de probabilités à partir d'observables qui sont des moments correspondant à l'espérance de fonctions des données. Le théorème de Boltzmann démontre que la distribution d'entropie maximum est une distribution de probabilités exponentielle, dont les paramètres peuvent aussi être calculés par maximum de vraisemblance.

Cours 7 – Compression par transformées orthogonales

Le 2 mars 2022

Les algorithmes de compression de signaux audio et d'images sont souvent fondés sur une transformation orthogonale qui produit de nombreux petits coefficients pouvant être approximés par des zéros. Les coefficients obtenus sont quantifiés puis représentés sous forme binaire par un codage entropique. On étudie l'erreur introduite par ces quantificateurs sous hypothèse de quantification haute résolution.

On démontre que le quantificateur qui minimise l'erreur sous contrainte d'entropie bornée est un quantificateur uniforme. Pour effectuer un codage optimal dans une base orthogonale, il est aussi nécessaire de faire une allocation de bits pour les

différents coefficients qui doivent être codés. L'allocation optimale pour une métrique euclidienne est obtenue avec le quantificateur le plus simple : celui qui quantifie tous les coefficients avec le même pas de quantification uniforme. Pour s'adapter à la sensibilité auditive ou visuelle, les pas de quantification sont proportionnels à des poids qui approximent la métrique perceptuelle à partir d'une distance euclidienne pondérée.

Cours 8 – Parcimonie pour la compression audio et JPEG pour l'image

Le 9 mars 2022

Les bases de cosinus sont des transformées de Fourier qui symétrisent les conditions aux bords afin de définir des signaux périodiques qui ne sont pas discontinus aux bords. Pour coder un signal audio, celui-ci est découpé en tranches temporelles que l'on représente dans une base orthogonale de cosinus. Les coefficients de cosinus sont comprimés par quantification et codage entropique. L'erreur perceptuelle est minimisée par une technique de masquage. Pour la compression d'images, le standard JPEG les découpe en fenêtres de 8 par 8 pixels qui sont représentées dans une base de cosinus séparable. Ce standard spécifie la quantification et le codage entropique des coefficients.

Ces algorithmes opèrent le plus souvent dans un régime de haute compression où l'hypothèse de quantification haute résolution n'est plus valable. Ceci introduit un phénomène d'approximation non linéaire qui gouverne l'efficacité de ces algorithmes de compression. On peut calculer ce terme d'approximation non linéaire en fonction de la parcimonie des coefficients dans la base orthogonale. On obtient ainsi le lien entre l'erreur produite par la quantification et le nombre de bits du codage. Le comportement pour un fort taux de compression est totalement différent de celui obtenu pour un faible taux de compression. Ces résultats montrent que l'on améliore la compression en optimisant la base afin d'augmenter la parcimonie des coefficients. Le standard JPEG-2000 améliore ainsi les performances du standard JPEG en remplaçant la base de cosinus par une base orthogonale d'ondelettes.

SÉMINAIRE (EN RELATION AVEC LE SUJET DU COURS)

Séminaires 1 et 2 – Challenges de données 2022

Les 19 et 26 janvier 2022

Pour les étudiants et participants au cours, le site web challengedata.ens.fr met à disposition de nouveaux challenges de traitement de données par apprentissage supervisé pour la saison 2022. Ces challenges sont proposés par des entreprises ou des scientifiques et sont issus de problématiques concrètes rencontrées 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 2022. Une clôture intermédiaire a lieu en mars 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'École normale supérieure et à l'Institut Louis Bachelier par Simon Coste et Marine Neyret, avec la participation de Florentin Guth, Rudy Morel, 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 douze challenges suivants ont été organisés et présentés lors des deux premières séances des séminaires :

- « Can you predict the tide? » présenté par Simon Coste pour Louis Thiry (Inria) ;
- « Prediction of missing Bid-Ask spread values » présenté par Romain Picon, de la société CFM ;
- « Data centric movie reviews sentiment classification » présenté par Maxime Duval, de la société Kili Technology ;
- « Semantic segmentation of industrial facility point cloud » présenté par Guillaume Terrasse, de la société EDF R&D ;
- « Learning factors for stock market returns prediction » présenté par Adrien Hardy, de la société QRT ;
- « Return forecasting of cryptocurrency clusters » présenté par Hugo Schnoering, de la société Napoleon X ;
- « Bankers and markets » présenté par Olivier Croissant, de la société Natixis ;
- « Predicting odor compound concentrations » présenté par Yannick Deleuze, de la société Veolia ;
- « Real estate price prediction » présenté par Louis Boulanger, de l'institut Louis Bachelier ;
- « What do you see in the stock market data? » présenté par Iris Lucas, de la société AMF ;
- « Learning biological properties of molecules from their structure » présenté par Aymeric Basset pour Robert Fraczkiewicz, de la société Simulations Plus ;
- « Solar forecasting using Copernicus radiation images » présenté par Philippe Blanc (MINES ParisTech, PSL).

Séminaire 3 – Prix des challenges de la saison 2021

Le 2 février 2022

Au cours de la première partie, certains lauréats des Challenges 2019 ont présenté leurs algorithmes ainsi que les résultats obtenus. Une remise des prix a été effectuée pour les gagnants des challenges de la saison 2021. La liste des gagnants est indiquée ci-dessous :

- « Stock trading: Prediction of auction volumes » (CFM) : Franck Zibi, Thibaud Blondel et Alexandre Jardillier ;
- « Detecting sleep apnea from raw physiological signals » (Dreem) : Léo Heidelberger, Louis Bouvier, Clement Grisi et Thibault Blanc ;
- « Land cover predictive modeling from satellite images » (Preligens) : Gauthier Hamon, Maxime Colignon, Yann Feunteun et Raphaël Basler ;
- « Assessing uncertainty in air quality predictions » (Oze Energies) : Raphaël Cousin, Hugues Van Assel ;
- « Reconstruction of liquid asset performance » (QRT) : Thibaud Blondel et Pierre-Alain Reigneron ;
- « EV charging stations usage analysis » (Planète Oui) : François Ledée, Eliott Girard, Alan Gany et Nicolas Dieu ;
- « Who are the high-frequency traders? » (AMF) : Nicolas Huynh et Dan Berrebbi ;
- « Sinusoid segmentation in subsurface images » (Schlumberger) : Thibaud Blondel et Amadou Dioulde Barry.

Séminaire 4 – Cosmologie et complexité

Brice Ménard (université John Hopkins), le 9 février 2022

Le séminaire commence par présenter une synthèse de la démarche à suivre pour tester des modèles et estimer leurs paramètres. Il montre ensuite comment ces concepts sont utilisés pour analyser des données à différents niveaux de complexité. Il présente au final des applications en astrophysique et cosmologie.

Séminaire 5 – Estimation du fond diffus cosmologique : analyse en composantes indépendantes et géométrie de l'information

Jean-François Cardoso (CNRS), le 16 février 2022

Si l'on dispose de cartes du ciel observées à différentes longueurs d'onde, il est possible de les combiner pour révéler le « fond diffus cosmologique » : un instantané de l'Univers primordial. Cette opération est un exemple de « séparation de composantes » : l'art et la manière d'extraire de plusieurs observations des signaux élémentaires sous-jacents.

Le séminaire porte sur l'analyse en composantes indépendantes (ACI) : une méthode de séparation ne s'appuyant que sur l'hypothèse de composantes sous-jacentes indépendantes et non-gaussiennes. L'analyse de la vraisemblance de l'ACI fait apparaître quelques notions clés de la théorie de l'information : entropie, divergence de Kullback-Leibler, information mutuelle... Nous verrons comment ces quantités sont reliées de façon naturelle dans une interprétation géométrique : celle de la géométrie de l'information, construite sur la divergence de Kullback-Leibler et dont la métrique est l'information de Fisher.

Séminaire 6 – Une introduction à la géométrie de l'information

Frank Nielsen (Sony Computer Science Laboratories), le 23 février 2022

La géométrie de l'information étudie les structures géométriques, les distances et les notions d'invariance d'une famille de distributions de probabilités appelée le « modèle statistique ». Un modèle statistique paramétrique peut se traiter comme une variété riemannienne, en l'équipant du tenseur métrique de Fisher qui induit la distance de Rao.

Cette structure riemannienne sur la variété de Fisher-Rao fut par la suite généralisée par une structure duale reposant sur des paires de connexions affines couplées à la métrique de Fisher. Cette structure duale permet d'expliquer l'interaction étroite entre les estimateurs pour l'inférence en statistique (maximum de vraisemblance) et la génération de modèles statistiques paramétriques (familles exponentielles obtenues par le principe de l'entropie maximale), et met en jeu un théorème de Pythagore généralisé. On illustrera des applications de la géométrie de l'information en statistique et en apprentissage de réseaux de neurones.

Séminaire 7 – Fuites d'information et attaques par canaux cachés

Olivier Rioul (Telecom Paris), le 2 mars 2022

Les définitions d'entropie de Shannon, de divergence de Kullback-Leibler et d'information mutuelle de Fano ont été utilisées par Shannon sous « forme opérationnelle » pour résoudre les problèmes de codage (compression et transmission). D'autres types de problèmes font appel à d'autres notions, comme l'information de Fisher pour l'estimation paramétrique. Choisir *a priori* un critère comme l'entropie n'est pas forcément optimal pour résoudre un problème donné.

Le séminaire s'intéresse à mesurer des fuites d'information dans les systèmes cryptographiques. Ici, la notion de maximum *a posteriori* (MAP) est importante, mais l'entropie classique n'a pas réellement de définition opérationnelle. On présente alors une théorie de l'« alpha-information » fondée sur l'entropie de Rényi, l'entropie conditionnelle d'Arimoto et l'information de Sibson. Elle englobe la théorie de l'information classique pour alpha = 1, l'information de Hartley pour alpha = 0 et le MAP pour alpha infini, tout en préservant des inégalités essentielles de traitement de

données et de Fano. Ces inégalités permettent d'évaluer les limites de toute attaque à partir de mesures divulguées par un canal caché en relation à une croyance *a priori* sur le secret (sans accès aux mesures).

Séminaire 8 – Stockage des images numériques : l'ADN est-il l'avenir de l'archivage des mégadonnées ?

Marco Antonini (CNRS), le 9 mars 2022

Le stockage des données numériques devient un défi pour l'humanité en raison de la durée de vie relativement courte des dispositifs de stockage. De plus, l'augmentation exponentielle de la génération de données numériques crée le besoin de construire constamment de nouvelles ressources pour gérer leur archivage. Des études récentes suggèrent l'utilisation de la molécule d'ADN comme un nouveau candidat prometteur qui pourrait contenir théoriquement 215 pétaoctets dans un seul gramme. Toute information numérique peut être synthétisée en ADN *in vitro* et stockée dans de minuscules capsules spéciales qui proposent une fiabilité de stockage de plusieurs centaines d'années. La séquence d'ADN stockée peut être récupérée à tout moment à l'aide de machines spéciales appelées « séquenceurs ». L'ensemble de ce processus est très difficile, car la synthèse de l'ADN est coûteuse et le séquençage est sujet à des erreurs. Cependant, des études ont montré qu'en respectant plusieurs règles dans le codage, la probabilité d'erreur de séquençage est réduite. Par conséquent, le codage de l'information numérique n'est pas trivial, et les données d'entrée doivent être efficacement compressées avant leur codage afin de réduire le coût élevé de la synthèse.

Dans cette présentation, nous parlerons de l'état de l'art en matière de stockage de données ADN pour l'encodage efficace de données numériques dans un code quaternaire constitué des quatre bases ADN A (Adénine), T (Thymine), C (Cytosine) et G (Guanine). Nous présenterons également une nouvelle solution prometteuse de codage d'images numériques dans de l'ADN synthétique que nous avons développée au laboratoire I3S au cours des cinq dernières années et qui prend en compte les contraintes liées au stockage des données sur ADN tout en optimisant le compromis entre qualité de compression et coût de synthèse.

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 pour la classification aussi bien que pour la génération de données. En 2021-2022, trois types de résultats ont été obtenus. Le premier, avec F. Guth et J. Zarka, concerne la construction d'architectures de réseaux de neurones profonds qui atteignent les meilleures performances de l'état de l'art pour la classification d'images, au moyen de filtres spatiaux calculés avec des ondelettes, et qui ne sont donc pas appris. L'apprentissage est réduit à des filtres le long des canaux du réseau.

Le second type de problème concerne la modélisation de processus stationnaires, non gaussiens. Nous avons montré qu'une représentation par *scattering* en ondelettes permettait de construire des modèles stochastiques interprétables pour des textures visuelles, des champs physiques ou des séries temporelles. Nous avons ainsi fait de la génération de textures d'images complexes avec A. Brochard et S. Zhang, mais aussi construit des modèles pour l'analyse de paramètres cosmologiques avec une équipe du Flatiron Institute, et enfin modélisé différents types de séries temporelles avec R. Morel, G. Rochette, R. Leonarduzzi et J.-P. Bouchaud.

Dans le dernier projet, nous avons montré avec T. Marchand, M. Ozawa et G. Biroli que l'on peut modéliser des champs physiques multi-échelles avec des modèles de basse dimension en établissant un lien avec le groupe de renormalisation, développé en physique pour expliquer les transitions de phases. Le résultat principal est la factorisation de distributions de probabilités en produit de probabilités conditionnelles, sur des coefficients d'ondelettes orthogonaux. En collaboration avec F. Guth, S. Coste et V. de Bortoli, nous avons appliqué ce résultat pour la génération d'images, avec des techniques de diffusion par équations différentielles stochastiques.

PUBLICATIONS

Brochard A., Zhang S. et Mallat S., « Generalized rectifier wavelet covariance models for texture synthesis », International Conference on Learning Representations (ICLR), 2022, https://arxiv.org/abs/2203.07902.

Guth F., Zarka J. et Mallat S., « Phase collapse in neural networks », International Conference on Learning Representations (ICLR), 2022, https://arxiv.org/abs/2110.05283.

Eickenberg M., Allys E., Moradinezhad Dizgah A., Lemos P., Massara E., Abidi M., Hahn C.H., Hassan S., Regaldo-Saint Blancard B., Ho S., Mallat S., Andén J. et Villaescusa-Navarro F., « Wavelet moments for cosmological parameter estimation », 2022, https://arxiv.org/abs/2204.07646.

Morel R., Rochette G., Leonarduzzi R., Bouchaud J.-P. et Mallat S., « Scale dependencies and self-similar models with wavelet scattering spectra », 2022, https://arxiv.org/abs/2204.10177.

Marchand T., Ozawa M., Biroli G. et Mallat S., « Wavelet conditional renormalization group », 2022, https://arxiv.org/abs/2207.04941.

Guth F., Coste S., de Bortoli V. et Mallat S., « Wavelet score-based generative modeling », 2022, https://arxiv.org/abs/2208.05003.