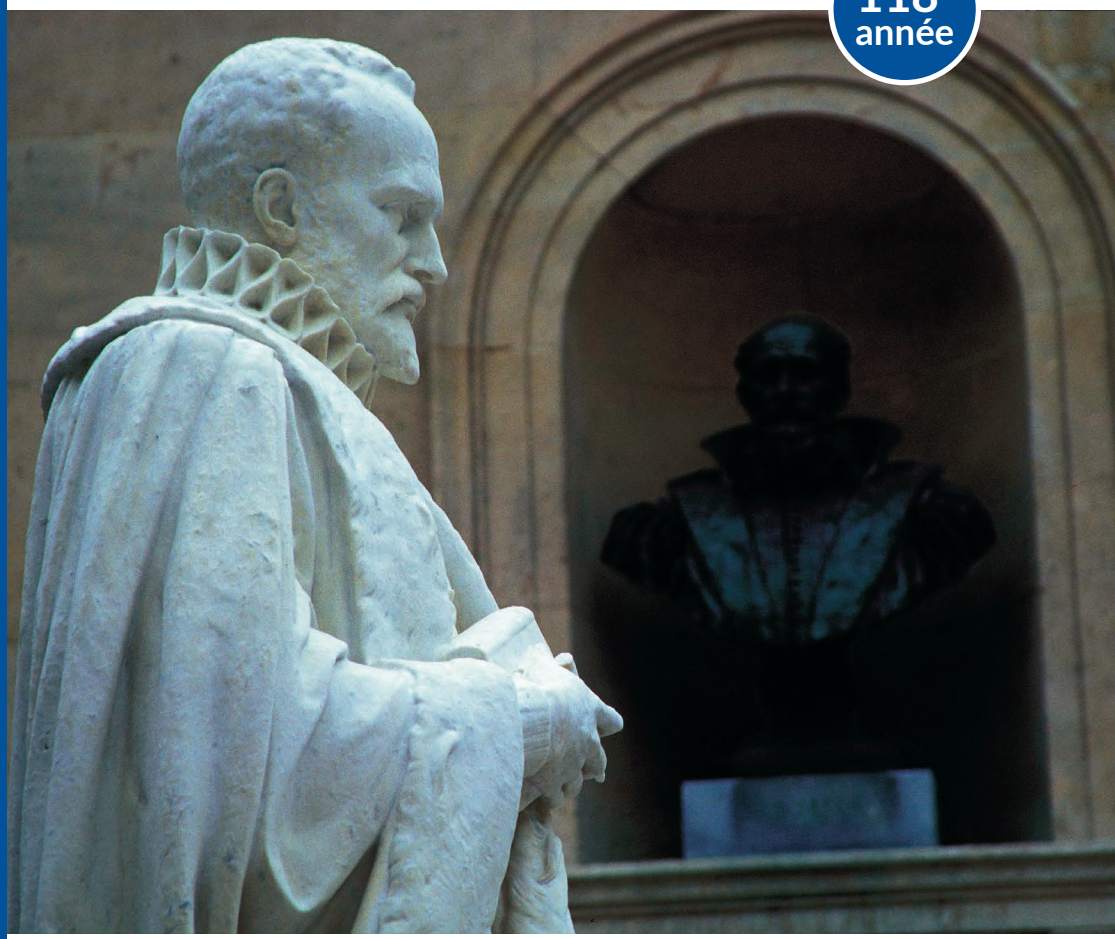
ANNUAIRE du **COLLÈGE DE FRANCE** 2017 - 2018

Résumé des cours et travaux

118e année



COLLÈGE DE FRANCE
1530

Sciences des données

Stéphane Mallat

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

Mots-clés : apprentissage, données, prédictions, réseaux de neurones

La série de cours et séminaires « L'apprentissage face à la malédiction de la grande dimension » est disponible, en audio et/ou en vidéo, sur le site internet du Collège de France (https://www.college-de-france.fr/site/stephane-mallat/course-2017-2018.htm). La leçon inaugurale *Sciences des données et apprentissage en grande dimension* publiée sous forme imprimée (Collège de France/Fayard, 2018) est également disponible en audio et vidéo (https://www.college-de-france.fr/site/stephane-mallat/inaugural-lecture-2017-2018.htm).

Enseignement

Cours et séminaires – L'apprentissage face à la malédiction de la grande dimension

Les sciences des données ont pour objectif « d'extraire de la connaissance » de données numériques, avec des algorithmes. Les applications sont considérables, pour stocker, analyser et valoriser les masses de données : images, sons, textes, mesures physiques ou données d'Internet. On distingue deux types de problèmes : la prédiction et la modélisation. Les prédictions sont faites par des algorithmes d'apprentissage statistique, qui sont à l'origine du renouveau de l'intelligence artificielle. Un modèle décrit la variabilité des données et permet d'en générer des nouvelles. Les mathématiques ont ici pour but de comprendre sous quelles conditions il est possible d'apprendre et donc de généraliser, ou de construire des modèles, tandis que l'informatique a pour objectif de développer des algorithmes qui résolvent ces problèmes.

Le premier cours de la chaire pose le cadre mathématique et algorithmique de ce domaine, en dégageant les questions et techniques importantes pour l'apprentissage. La

difficulté principale de la prédiction ou de la modélisation vient du grand nombre de variables des données, souvent plus d'un million, à l'instar du nombre de pixels d'une image. Cette grande dimension génère une explosion combinatoire des possibilités de prédiction ou de modélisation. On fait face à cette malédiction de la grande dimension avec des algorithmes qui utilisent de l'information *a priori* sur certaines régularités du problème. Le cours introduit des outils mathématiques et algorithmiques permettant de spécifier et d'exploiter cette régularité, pour prédire ou modéliser.

Leçon inaugurale – Sciences des données

11 janvier 2018

La leçon inaugurale introduit les principes des algorithmes d'apprentissage et certains grands problèmes posés par cette discipline. Une question centrale est de comprendre les conditions dans lesquelles il est possible d'apprendre et donc de généraliser à partir d'exemples, suivant la complexité des données et de la fonction qu'il faut prédire. Cela met en jeu l'existence de régularités connues *a priori*, utilisées par des algorithmes comme l'apprentissage à noyaux, ou qui peuvent être découvertes par des algorithmes comme les réseaux de neurones. Ces notions de régularité font appel à diverses notions mathématiques comme la parcimonie, la séparation d'échelles par ondelettes, ou des notions d'invariance relativement à l'action de groupes de symétries. Au cours des dernières années, les réseaux de neurones profonds ont obtenu des résultats spectaculaires dans de nombreux domaines d'applications, que les mathématiques sont loin de pouvoir expliquer. Ces succès posent aussi des questions sociétales et éthiques importantes, qui sont évoquées en lien avec les problèmes scientifiques sous-jacents.

Cours 1 – Cartographie des sciences des données

17 janvier 2018

Le premier cours effectue une cartographie des sciences des données, qui regroupent trois grands domaines : le traitement du signal, la modélisation de données, ainsi que la prédiction. Le cours introduit les grands enjeux de chacun de ces domaines, ainsi que les notions mathématiques et informatiques auxquels ils font appel.

En traitement du signal, on veut calculer une estimation d'un signal x ayant d coefficients, à partir de mesures. La dimension d est typiquement supérieure à un million, que ce soit un son, une image ou toute autre observation. Les problèmes inverses ont pour but d'améliorer la qualité des signaux. Un instrument de mesure effectue une transformation du signal d'entrée et ajoute des erreurs, autrement dit du bruit. Inverser la transformation tout en réduisant le bruit nécessite d'utiliser des informations *a priori* sur les propriétés du signal. La compression des signaux est une autre application, dont le but est de réduire le nombre de bits pour coder les signaux, afin de limiter l'espace de stockage ou le temps de transmission. Là encore, il s'agit d'exploiter une information *a priori* sur la structure des signaux.

La modélisation consiste à capturer la nature et la variabilité des données. Cela se fait en estimant la distribution des données observées. Cette distribution est caractérisée par un modèle aléatoire dont on suppose qu'il a une densité de probabilité. C'est une fonction du grand nombre d de variables de chaque donnée. La difficulté principale vient de cette grande dimension. La construction de tels

modèles est nécessaire pour optimiser les algorithmes de traitement du signal, pour la physique statistique, ou pour la synthèse de nouvelles données. Cette modélisation est également utile pour faire de la prédiction.

Une prédiction calcule une estimation de la réponse y à une question, à partir d'une donnée x qui peut inclure de nombreuses variables. Par exemple, y peut être le nom d'un animal qui apparaît dans une image x, ou un diagnostic estimé à partir de données médicales x. L'apprentissage supervisé optimise le paramétrage d'algorithmes de prédictions, en utilisant de nombreux exemples composés de données x pour lesquels on connaît la réponse y.

Séminaire 1 – Challenges de données (I)

17 janvier 2018

Le site web *challengedata.ens.fr* met à disposition des challenges de traitement de données par apprentissage supervisé. Ce séminaire introduit une première partie des challenges qui sont utilisés dans le cadre du cours. 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. Les données mises à disposition sont non confidentielles et les rapports algorithmiques des participants peuvent être mis à la disposition de tous, s'ils le souhaitent, après la clôture de la saison.

Les challenges couvrent un large spectre d'applications, sur des images, sons, textes, données médicales, mesures physiques, données d'Internet. 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 2018. Une clôture intermédiaire a lieu en juin par une évaluation des prédictions sur de nouvelles données de test. La clôture finale est en décembre, avec une remise des prix en janvier 2019.

Cette année, les challenges ont été organisés et supervisés à l'ENS par Mathieu Andreux, Tomas Anglès, Georgios Exarcharkis, Louis Thiry, John Zarka et Sixin Zhang. 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. Lors de cette première session, les six challenges suivant ont été présentés :

- « Prédiction de la volatilité sur des marchés financiers », présenté par Éric Lebigot de la société Capital Fund Management. L'objectif du challenge est de prédire la volatilité de fin de journée d'actions américaines à partir de leur historique de rendements de début de journée ;
- « Identification de célébrités », présenté par Antoine Chassang de la société Reminiz. L'objectif est d'identifier les visages apparaissant dans des vidéos à partir d'un dictionnaire de référence de visages de célébrités ;
- « Prédiction de la production électrique horaire par unité de production en France », présenté par Alexi Bergès de la société Wattstrat. L'objectif est de prédire la production électrique horaire de chaque unité de production en France, à partir des courbes de demandes et de productions renouvelables régionales ainsi que de la disponibilité horaire des unités de production ;
- « Prédiction de réclamations lors de transactions e-commerce », présenté par Vincent Michel de la société PriceMinister – Rakuten France. L'objectif est de

prédire si une transaction *e*-commerce donnera lieu à une réclamation ou non, et si oui, de quel type, à partir des caractéristiques de la transaction ;
• « Prédiction de la réponse attendue à des questions de pharmaceutique », présenté par Emmanuel Bilbault de la société Posos. L'objectif est de catégoriser des questions de pharmaceutique selon le type de réponse attendu ;
• « Prédiction des performances énergétiques de bâtiments », présenté par Sylvain Le Corff de la société Oze-Energies. L'objectif est de prédire les consommations énergétiques ainsi que les températures intérieures de bâtiments à partir des températures extérieures et d'un nombre réduit de paramètres décrivant la structure et les réglages des bâtiments.

Cours 2 – Compromis biais-complexité

24 janvier 2018

Ce cours introduit le fonctionnement d'un algorithme d'apprentissage et le compromis entre le biais et la variance des estimateurs de prédiction. Un algorithme d'apprentissage prend en entrée une donnée x à partir de laquelle il prédit une approximation de la réponse y. Un tel algorithme inclut des paramètres internes qui sont optimisés grâce aux exemples d'entraînement, de façon à ce que la prédiction soit précise sur ces exemples. Le but est que la précision de la prédiction se généralise à d'autres données de « même type » que les exemples d'entraînement. On parle alors d'erreur de *généralisation*. L'algorithme d'apprentissage supervisé calcule la prédiction avec une fonction sélectionnée parmi une classe de fonctions possibles, par exemple en minimisant l'erreur moyenne obtenue sur les exemples d'entraînement. Cette erreur est le risque statistique. L'information *a priori* sur le problème permet de spécifier la classe d'approximation et le calcul du risque.

Le cours commence par une présentation des algorithmes de classification linéaire qui divisent l'espace des donnés en deux parties séparées par un hyperplan. Celui-ci est optimisé afin de minimiser l'erreur de classification. Ces algorithmes s'étendent par un changement de variable qui remplace x par une nouvelle représentation. Un enjeu central est de choisir ce changement de variable afin de minimiser le risque de généralisation.

Le risque de généralisation comporte deux termes. Une erreur d'approximation due au fait que l'on approxime la réponse sur une classe limitée de fonctions. C'est le terme de biais. Le biais augmente lorsque l'on réduit la classe d'approximation. Le second terme d'erreur est dû au fait que l'algorithme minimise l'erreur sur des exemples d'entraînement et non pas sur toutes les données possibles. Cela induit une fluctuation aléatoire qui dépend du choix des exemples d'entraînement, et dont on veut contrôler l'amplitude maximum. Ce terme de variabilité dépend de la complexité de la classe d'approximation. Il diminue lorsque l'on réduit la taille de cette classe d'approximation. L'optimisation de la classe d'approximation résulte donc d'un compromis entre le biais et la variabilité du risque.

Le cadre statistique d'approximation PAC (probablement approximativement correcte) consiste à établir des conditions pour avoir un risque de généralisation qui devienne arbitrairement petit. Lorsque la classe est de taille finie, on peut calculer une borne supérieure de la fluctuation maximum du risque, en fonction du cardinal de la classe d'approximation, avec une probabilité arbitrairement proche de 1. Cela permet d'établir des conditions pour que l'algorithme d'apprentissage soit probablement approximativement correct.

Séminaire 2 – Challenges de données (II)

24 janvier 2018

Lors de cette deuxième session, sept challenges du site web *challengedata.ens.fr* ont été présentés :

• « Prédiction de maladie à partir du génome », présenté par Gilles Wainrib de la société Owkin en collaboration avec l'Inserm. L'objectif est de prédire si un individu développera une certaine maladie à partir de données spécifiques à son génome ;

• « Prédiction de la production d'énergie éolienne », présenté par Paul Poncet de la société Engie. L'objectif est de prédire la production d'énergie éolienne de quatre éoliennes à partir de leurs différents paramètres de fonctionnement ;

• « Prédiction de la dynamique de liquides surfondus », présenté par François Landes de l'équipe de Giulio Biroli, IPhT CEA et ENS Paris. L'objectif est de prédire la mobilité des différentes particules au sein d'un système à partir de leurs positions et vitesses initiales ;

• « Classification en stade de sommeil », présenté par Valentin Thorey de la société Rythm. L'objectif est de prédire le stade de sommeil d'un individu à partir de 30 secondes de signaux EEG, accéléromètre et oxymètre de pouls ;

• « Prédiction de la saturation d'huile résiduelle », présenté par Jean-François Lecomte de l'IFPEN. À partir de la structure poreuse 3D d'une roche, l'objectif est de déterminer son type et de prédire la quantité d'huile (pétrole brut) résiduelle piégée ainsi que la distribution en taille des clusters d'huile résiduelle ;

• « Prédiction de l'équipe vainqueur d'un match de NBA », présenté par Sébastien Loustau de la société LumenAI. L'objectif est de prédire l'équipe remportant un match de NBA à partir d'indicateurs clés (score, fautes, etc.) du déroulement de la première mi-temps relevés à chaque seconde ;

• « Prédiction d'approbation de publications », présenté par Nies Lubbers de la société Dassault Systèmes. L'objectif est de prédire, pour chaque utilisateur d'une application sociale et collaborative, les cinq publications qu'il est le plus susceptible d'aimer, à partir d'un historique d'approbation des publications par un ensemble d'utilisateurs.

Cours 3 – Malédiction de la grande dimension

31 janvier 2018

Ce cours montre que l'approximation de fonctions localement régulières nécessite d'avoir un nombre d'exemples qui croît de façon exponentielle avec la dimension des données, ce que l'on appelle « la malédiction de la grande dimension ». Si la réponse y associée à une donnée x est unique, alors c'est une fonction $y = f(x)$. La prédiction de y revient donc à approximer la fonction $f(x)$ en choisissant une approximation dans une classe prédéfinie. Le choix de cette classe ainsi que l'amplitude de l'erreur d'approximation dépend de la régularité de f.

L'algorithme du plus proche voisin approxime $f(x)$ par la valeur $f(x')$ où x' est l'exemple le plus proche de x. L'approximation est donc une fonction constante par morceaux. On établit le lien entre l'erreur d'approximation, la régularité de f, le nombre d'exemples d'entraînement et la dimension d des données. Le cours se concentre sur l'approximation de fonctions Lipchitz. Sous cette hypothèse de régularité locale, pour que l'erreur soit inférieure à c avec n exemples, il faut que l'espace des

données puisse être couvert par n boules de rayon c. On démontre que le nombre n de boules croît comme c^{-d}. Cette croissance exponentielle en d est la malédiction de la grande dimension. Dès que d est plus grand que 10, on n'a typiquement jamais assez d'exemples pour atteindre une précision c petite. Pour pouvoir approximer précisément des fonctions f en grande dimension, il faut que celles-ci aient une régularité globale beaucoup plus forte qu'une régularité lipchitzienne.

Cours 4 – Réduction de la dimension et débruitage

31 janvier 2018

Pour éviter la malédiction de la grande dimension, on peut essayer de réduire la dimension des données, si elles appartiennent à un sous-ensemble de dimension plus petite que la dimension d de l'espace d'origine. La réduction de la dimension des données est au cœur du traitement du signal avec de nombreuses applications pour la compression, le débruitage et les problèmes inverses. Il s'agit d'approximer le signal avec un nombre minimum de variables.

On étudie l'approximation de signaux x décomposés dans une base orthonormale. Une approximation linéaire approxime x à partir des M premiers vecteurs de la base. L'erreur d'approximation est reliée à la vitesse de décroissance des coefficients de x dans la base. Une approximation non linéaire réduit l'erreur d'approximation en sélectionnant les M coefficients de x ayant les plus grandes amplitudes. La sélection de ces coefficients est une fonction non linéaire de x. L'erreur d'approximation dépend alors de la décroissance de l'amplitude des coefficients de x ordonnés.

Le cours introduit une application au débruitage, où le signal est contaminé par un bruit additif supposé être gaussien et blanc. Un algorithme linéaire de réduction de bruit effectue une projection du signal bruité dans un espace de dimension M, qui est choisi afin d'approximer le signal au mieux, tout en éliminant une large proportion de bruit. L'optimisation de la dimension M résulte d'une optimisation des erreurs de biais et de variance. Le biais est l'erreur d'approximation de x tandis que la variance mesure l'énergie du bruit qui n'est pas éliminée par la projection. Une approche non linéaire est aussi étudiée, par seuillage des coefficients du signal bruité dans une base orthonormale.

Cours 5 – Analyse de Fourier, filtrage et échantillonnage

14 février 2018

La base orthogonale de Fourier joue un rôle particulier pour représenter des signaux car elle diagonalise les opérateurs linéaires qui sont covariants par translation. Ces opérateurs sont des convolutions que l'on appelle « filtrage » en traitement du signal. Ce cours revoit la définition et les propriétés des séries de Fourier ainsi que de l'intégrale de Fourier, en montrant le lien entre la décroissance des coefficients de Fourier et la régularité d'une fonction.

Le théorème d'échantillonnage de Nyquist-Shannon est présenté comme un théorème d'approximation linéaire dans la base de Fourier où la fonction est approximée par une projection orthogonale sur les basses fréquences. Cette approximation peut se réécrire dans une base de sinus cardinaux. On montre l'effet des erreurs aux hautes fréquences, que l'on appelle *aliasing*.

Le théorème d'échantillonnage est ensuite généralisé en remplaçant l'espace d'approximation et la base de sinus cardinaux par d'autres espaces générés par

d'autres bases orthogonales, comme les espaces de fonctions constantes par morceaux ou de splines linéaires.

Séminaire 3 – Le débruitage d'images en quelques formules

Jean-Michel Morel (ENS Paris-Saclay), le 14 février 2018

Les images sont des matrices de pixels, dont les valeurs sont proportionnelles à un compte de photons. Ce compte est un processus stochastique dû à la nature quantique de la lumière. Donc toutes les images sont bruitées. Des algorithmes numériques ont été proposés pour améliorer le rapport du signal à bruit. Ces algorithmes de débruitage nécessitent d'établir un modèle du bruit et des images. Il est relativement facile d'établir un modèle du bruit. Obtenir un bon modèle statistique de l'image est beaucoup plus difficile. Les images sont un reflet du monde et ont donc le même niveau de complexité.

Le séminaire montre que l'on est probablement proche de comprendre les statistiques des images à l'échelle des patches de pixels. Des algorithmes récents, basés sur des modèles non paramétriques de patches de 8×8 pixels, obtiennent de très bons résultats. Des arguments mathématiques et expérimentaux indiquent que l'on est probablement proche des meilleurs résultats atteignables. Cette hypothèse est validée par la convergence des résultats obtenus par toutes les techniques récentes. Les trois approches principales sont présentées : l'approche bayésienne, le seuillage d'opérateurs linéaires et les modèles d'autosimilarités. La plupart des algorithmes de débruitage peuvent être testés sur n'importe quelle image sur le site *Image Processing On Line* (www.ipol.im/).

Cours 6 – Transformées et bases d'ondelettes

21 février 2018

Une transformée en ondelettes permet de réduire la dimensionnalité d'un signal en le décomposant sur une famille de petites ondes, qui sont dilatées et translatées. La transformée en ondelette calcule la corrélation entre un signal et ces ondelettes, à toutes les échelles et positions. Ce cours présente la transformée en ondelette continue et la transformée dyadique où les échelles sont limitées à des puissances de 2, ainsi que les bases d'ondelettes orthonormales. On résume les propriétés principales de ces trois types de représentations multi-échelles pour des signaux et des images. La condition de Littlewood-Paley garantit que la transformée en ondelette dyadique est un opérateur unitaire et donc inversible. Les bases d'ondelettes orthogonales se construisent avec une structure d'espaces vectoriels emboîtés que l'on appelle « multirésolutions ». Ces multirésolutions sont reliées à l'existence de filtres qui gouvernent les passages d'une échelle à l'autre, et qui permettent d'implémenter l'algorithme de transformée rapide en ondelettes.

L'approximation non linéaire d'un signal dans une base orthogonale d'ondelettes s'obtient en ne gardant que les grands coefficients d'ondelettes. Cela revient à construire une approximation qui s'adapte à la régularité locale du signal. La décroissance de l'erreur d'approximation d'un signal est reliée à sa régularité lipchitzienne. Des applications sont montrées pour le débruitage de signaux et d'images, par seuillage des coefficients d'ondelettes. Cela permet de restaurer les zones régulières de l'image ainsi que ses contours.

Séminaire 4 – S'attaquer à une compétition d'apprentissage : méthodologie et exemples pratiques

Pierre Courtiol (Owkin), le 21 février 2018

Participer à une compétition de *machine learning* demande à la fois des compétences poussées en informatique et une connaissance des modèles de *machine learning* du point de vue mathématique et algorithmique. Cet exposé explique le processus itératif permettant d'obtenir des bons résultats lors d'une compétition de *machine learning*.

La méthodologie proposée se décompose en cinq phases, répétées jusqu'à la fin de la compétition. Elle commence par une revue de l'état de l'art sur le sujet, en matière de publications scientifiques et de compétitions similaires. Suit une exploration des données, pour comprendre leurs structures et avoir une première idée des *features* ayant un pouvoir prédictif. La troisième phase construit une représentation des données qui optimise ces *features* : c'est ce qu'on appelle le *feature engineering*. Après avoir construit une procédure d'évaluation des modèles, impliquant par exemple une validation croisée (k-fold), il reste à créer une batterie de modèles, les comparer et les combiner pour obtenir le meilleur modèle prédictif possible. Un *data scientist* émet ensuite des hypothèses sur les nouvelles *features* qui pourraient apporter une représentation plus pertinente des données, et les intègre en répétant cette méthodologie pour améliorer les résultats jusqu'à la fin de la compétition.

Atteindre d'excellents classements lors de compétitions de *machine learning* nécessite donc une connaissance précise des modèles pour les paramétrer au mieux et pour connaître leurs limites, mais également de la créativité pour construire une représentation des données susceptible de contenir un maximum d'informations pertinentes.

Cours 7 – Apprentissage bayésien et linéaire à noyaux

7 mars 2018

Dans un cadre stochastique bayésien, l'estimation optimale d'une réponse y à partir de données x s'obtient en maximisant la probabilité conditionnelle de y sachant x. Cependant, l'estimation de cette probabilité conditionnelle souffre à nouveau de la malédiction de la dimensionnalité si on suppose seulement qu'elle est localement régulière. Il faut donc introduire des conditions de régularité beaucoup plus fortes.

Beaucoup d'algorithmes d'apprentissage linéarisent l'estimation de y en effectuant un changement de variable qui transforme le vecteur x de dimension d en un vecteur $\Phi(x)$ de dimension d'. L'estimation de y se fait à partir du produit scalaire $<w, \Phi(x)> + b$ où le vecteur w et le biais b sont optimisés afin de minimiser le risque empirique calculé sur les données d'apprentissage. Le calcul de w en fonction des données d'apprentissage s'obtient en inversant une matrice d'affinité qui explicite la corrélation entre les données d'apprentissage. Pour un risque quadratique, le théorème de représentation démontre que le w optimal s'obtient par combinaison linéaire des $\Phi(x')$, où les x' sont les exemples d'apprentissage.

Afin de contrôler l'erreur de généralisation, le risque empirique peut être régularisé en introduisant une pénalité de Tikhnonov, proportionnelle à la norme de w au carré. Cette régularisation garantit que l'inversion de la matrice d'affinité est stable. De façon générale, on montre qu'une estimation stable de y en fonction de x a nécessairement de bonnes propriétés de généralisation.

Cours 8 – Régression à noyaux et optimisation convexe

7 mars 2018

Les algorithmes de régression linéaires peuvent devenir très flexibles en introduisant au préalable un changement de variable $\Phi(x)$. Après ce changement de variable, une régression linéaire peut se réécrire à partir des valeurs prises par le noyau $k(x,x') = < (x), \Phi(x') >$ sur les données d'apprentissage. La régression qui minimise l'erreur empirique se calcule par la résolution d'un problème d'optimisation convexe.

La minimisation de fonctions convexes sous contraintes est au cœur de beaucoup de problèmes d'apprentissage. La convexité garantit que l'ensemble des minima est convexe. De plus, il existe un unique minimum si la convexité est stricte. La minimisation d'une fonction sous contraintes d'inégalités peut s'exprimer en définissant un Lagrangien qui associe des variables duales à chacune des contraintes. Ces variables sont les multiplicateurs de Lagrange. On montre qu'un point-selle du lagrangien correspond à un minimum qui satisfait les contraintes. Si la fonction que l'on minimise ainsi que les contraintes sont convexes, alors la condition de point-selle du lagrangien est nécessaire et suffisante pour obtenir une solution du problème d'optimisation. Les conditions de Kuhn et Tucker s'obtiennent en annulant les dérivées partielles du Lagrangien relativement aux données, et en assurant que les dérivées relativement aux multiplicateurs sont négatives.

Cours 9 – Classification à noyaux et SVM

14 mars 2018

Les algorithmes de classification à noyaux donnent un cadre mathématique et algorithmique relativement simple pour développer des algorithmes d'apprentissage. Ils séparent deux classes en ajustant un hyperplan séparateur, après avoir effectué un changement de variable qui associe a une donnée x, un vecteur $\Phi(x)$. Les *support vector machines* optimisent la position de l'hyperplan en minimisant le risque empirique régularisé par un critère de marge. La marge mesure la distance minimum entre les points de chacune des classes et l'hyperplan. Cette minimisation peut se réécrire comme un problème d'optimisation convexe sous contraintes linéaires, qui dépend des produits scalaires $k(x,x') = <x,x'>$ entre les données d'apprentissage.

Le même type de résultat s'obtient avec un changement de variable $\Phi(x)$ en remplaçant le noyau par $k(x,x') = < \Phi(x), \Phi(x') >$. L'optimisation peut s'effectuer directement à partir des valeurs du noyau, en calculant les variables duales du Lagrangien associé à la minimisation du risque régularisé. Le théorème de Mercer prouve que n'importe quel noyau défini positif s'obtient avec un changement de variable $\Phi(x)$. La difficulté principale est de trouver un changement de variables qui permet de réduire le risque de généralisation. On étudie les propriétés des noyaux polynomiaux et des noyaux gaussiens.

Séminaire 5 – L'apprentissage fédéré pour les données médicales

Gilles Wainrib (Owkin), le 14 mars 2018

Les technologies d'apprentissage fédéré ont trouvé leurs premières applications avec la publication par Google d'un article concernant les algorithmes intégrés au clavier des téléphones Android. Plutôt que de faire remonter aux serveurs de Google

des données concernant les conversations des utilisateurs, les claviers Android effectuent l'entraînement des modèles prédictifs qui permettent de faire de la correction d'orthographe ou de la suggestion sur le téléphone lui-même, en ne faisant remonter que les modifications des modèles prédictifs inférées à partir des données des conversations de l'utilisateur.

En plus des considérations liées à la confidentialité des données sur lesquelles est réalisé l'entraînement des algorithmes, ces techniques peuvent être utiles pour des cas d'usage où les données ne peuvent pas franchir les frontières nationales, ou dans des situations où un consortium d'entreprises souhaiterait coconstruire un modèle prédictif sans qu'aucune n'ait à dévoiler d'informations importantes à ses concurrents.

Les technologies d'apprentissage fédéré présentent également un potentiel formidable pour générer de nouvelles découvertes médicales. Aujourd'hui, il n'existe pas de méthode répondant à toutes les contraintes qui s'imposent : un tel système doit avoir une faible bande passante, garantir la confidentialité des données transmises, gérer l'asynchronie entre les nœuds et la non-représentativité des *batches*. Des technologies d'apprentissage fédéré en pair-à-pair intégrant des communications compressées sont actuellement développées pour libérer le potentiel de l'apprentissage fédéré sur des données médicales.

Cours 10 – Descente de gradient et réseaux de neurones

21 mars 2018

La minimisation d'une fonction convexe peut se faire avec un algorithme de descente de gradient qui additionne itérativement un vecteur colinéaire au gradient de la fonction. Si la fonction que l'on minimise est Lipschitz et strictement convexe, alors on montre que la descente de gradient converge vers l'unique minimum. On étudie la convergence de cet algorithme lorsque la fonction convexe est quadratique. Dans ce cas, la vitesse de convergence dépend du rapport entre la plus petite et la plus grande valeur propre du hessien.

Lorsque la fonction que l'on minimise est un risque qui s'évalue comme une moyenne de risques calculés sur chaque donnée d'entraînement, alors on peut remplacer cet algorithme de gradient par un algorithme de gradient stochastique, où l'on additionne l'un après l'autre le gradient du risque pour chaque donnée d'entraînement. Cet algorithme s'utilise notamment pour l'apprentissage des réseaux de neurones.

Le cours se finit par une présentation rapide des réseaux de neurones, en introduction du cours de l'année prochaine, et mentionne le théorème d'approximation universel qui démontre que toute fonction peut être approximée par un réseau ayant une couche cachée suffisamment grande. Les réseaux de neurones convolutionnels profonds incluent un grand nombre de couches avec des neurones dont les poids sont invariants par translations. L'analyse de ces réseaux sera au centre du cours de l'année prochaine.

Séminaire 6 – Gradients stochastiques et conditionnels pour les réseaux de neurones

Francis Bach (Inria, ENS), le 21 mars 2018

La plupart des méthodes d'apprentissage supervisé, dont font partie les réseaux de neurones, se formalisent comme un problème d'optimisation dans lequel la moyenne

des erreurs sur les données observées est minimisée par rapport aux paramètres du modèle de prédiction. Cependant, l'apprentissage statistique donne lieu à des problèmes d'optimisation spécifiques, car on minimise une moyenne, ou plus généralement une espérance. Cette spécificité rend naturelle et efficace l'utilisation de méthodes dites « de gradient stochastique », où le modèle est mis à jour très fréquemment, après seulement quelques observations.

L'exposé présente quelques avancées récentes en optimisation par gradient stochastique, qui utilisent la « réduction de variance ». Pour les problèmes « convexes » (correspondant à un réseau de neurones sans couche cachée), ces avancées permettent d'atteindre en théorie et en pratique un taux de convergence exponentiel (dans le nombre d'itérations) vers l'optimum global. L'exposé présente aussi les méthodes dites « de gradient conditionnel », qui permettent un apprentissage incrémental où les neurones sont ajoutés aux modèles les uns après les autres.

Recherche

Je 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. Le but est de dégager les principes mathématiques sous-jacents, liés à la grande dimension des données, en explorant des applications pour des données très diverses comme des images, des sons, des données physiques provenant de cosmologie, de dynamique de fluides ou de chimie quantique.

Les réseaux de neurones profonds ont récemment obtenu des résultats remarquables que l'on comprend mal mathématiquement. L'équipe travaille sur des modèles mathématiques simplifiés permettant d'expliquer ces performances. Ils se fondent notamment sur la transformée de *scattering* qui est un modèle non linéaire de réseaux de neurones basé sur la transformée en ondelettes. Cela permet de calculer des invariants multi-échelles, relativement à l'action de groupes de symétrie de nature différentes suivant les applications. En 2017-2018, une partie importante de la recherche était dédiée à des applications en chimie quantique, pour régresser l'énergie quantiques de molécules, ainsi qu'à la construction de modèle stochastiques pour la synthèse d'images, de sons et de turbulences.

Publications

EICKENBERG M., EXARCHAKIS G., HIRN M. et MALLAT S., « Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3D electronic densities », *in* I. GUYON, U.V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN et R. GARNETT (dir.), *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., 2017, p. 6540-6549, http://papers.nips.cc/paper/7232-solid-harmonic-wavelet-scattering-predicting-quantum-molecular-energy-from-invariant-descriptors-of-3d-electronic-densities.pdf.

BRUNA J. et MALLAT S., « Multiscale sparse microcanonical models », arXiv: 1801.02013 2018.

EICKENBERG M., EXARCHAKIS G., HIRN M., MALLAT S. et THIRY L., « Solid harmonic wavelet scattering for predictions of molecule properties », *The Journal of Chemical Physics*, vol. 148, n° 24, 2018, p. 241-732, DOI : 10.1063/1.5023798 [arXiv: 1805.00571].

ANGLES T. et MALLAT S., « Generative networks as inverse problems with scattering transforms », *International Conference on Learning Representations*, 2018, [arXiv: 1805.06621].

ANDREUX M. et MALLAT S., « Music generation and transformation with moment matching scattering inverse networks », *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018, p. 327-333.

HIRN M., MALLAT S. et POILVERT N., « Wavelet scattering regression of quantum chemical energies », *Multiscale Modeling & Simulation*, vol. 15, nº 2, 2017, p. 827-863, DOI : 10.1137/16M1075454.

MALLAT S., *Sciences des données et apprentissage en grande dimension*, Paris, Collège de France/Fayard, coll. « Leçons inaugurales du Collège de France », nº 276, 2018.