L'apprentissage face à la malédiction de la grande dimension

Collège de France
Cours 1: **Cartographie des Sciences des Données**
de Stéphane Mallat

Notes de John Zarka et Stéphane Mallat

1 Introduction

L'objectif de l'apprentissage est d'*extraire de la connaissance* des données à l'aide d'algorithmes. Cette démarche est au coeur de toutes les problématiques scientifiques mais les données et les questions sont généralement propres au domaine d'étude considéré (sociologie, biologie, physique, etc.). Le but du cours sera de comprendre quels sont les outils mathématiques génériques au coeur de cette démarche d'apprentissage et d'étudier leur implémentation algorithmique.

Les données s'accompagnent le plus souvent d'informations *a priori* sur leur structure (un *modèle*) qui serviront à l'élaboration des algorithmes. Un aspect fondamental du cours sera de comprendre comment incorporer cette information a priori dans les algorithmes d'apprentissage.

La science des données est une discipline émergente qui a évolué d'une approche essentiellement statistique il y a encore 5-10 ans à un domaine mélangeant maintenant différentes branches de l'informatique et des mathématiques :

- En mathématiques : statistiques, probabilités, analyse harmonique (transformée de Fourier, ondelettes, ...), géométrie (notions de distance, de dimension), théorie des groupes (groupes de symétrie)
- En informatique : intelligence artificielle, base de données, calcul distribué, etc.

Le cours sera essentiellement un cours de mathématiques appliquées où l'informatique sera utilisée comme un outil. On adoptera un point de vue générique où les données peuvent être de nature totalement différentes. On utilisera toutefois le plus souvent des images car elles sont l'exemple le plus simple de données en grande dimension sur lesquelles on a une intuition et que l'on peut visualiser facilement.

Les données peuvent en effet être soit structurées (images, sons) ou moins structurées (texte, certains problèmes physiques) mais également un mélange de ces deux catégories (par ex. une image avec du texte, encore plus difficile).

2 Cartographie du domaine

2.1 Notations

A travers le cours, on notera $x(u)$ le signal qui est indexé par une variable u. Par exemple dans le cas d'un son ou d'une série temporelle, u représente le temps et x le signal variant en fonction de u.

u peut être :

- une variable discrétisée $u \in \mathbb{Z}^l$ avec par exemple $l = 1$ dans le cas d'une série temporelle et $l = 2$ dans le cas d'une image avec $u = (u_1, u_2)$. x représente dans ce cas un vecteur $x \in \mathbb{R}^d$ si u appartient à une partie finie de $\mathbb{Z}^l$ de cardinal d.
- une variable réelle $u \in \mathbb{R}^l$ dans le cas d'un signal analogique. x représente dans ce cas une fonction de u.

On se baladera constamment entre le cas discret et le cas continu, le cas continu présentant l'avantage de pouvoir utiliser les notions de régularité ou de dérivée alors que le cas discret est beaucoup moins structuré.

On oscillera également dans le cas fini où $x \in \Omega$ avec $\Omega \subseteq \mathbb{R}^d$ entre modélisation probabiliste et déterministe. Le modèle probabiliste est beaucoup plus riche que le déterministe car il définit non seulement le support Ω du signal x mais également comment ce dernier se repartit sur Ω suivant une certaine loi de probabilité de densité $p(\cdot)$ par rapport à la mesure de Lebesgue. A contrario, le modèle déterministe est un modèle pour lequel on ne connaît que le support Ω du signal x, ce qui équivaut à un modèle de probabilité uniforme sur Ω.

On notera également :

— $\Phi(x)$ une *représentation* d'un signal x

— $x \sim p(\cdot)$ lorsque la variable aléatoire x suit une loi de densité $p(\cdot)$ par rapport à la mesure de Lebesgue

— $\tilde{x}$, $\tilde{p}$ et $\tilde{f}$ sont respectivement des estimations ou approximations d'un signal x, d'une densité de probabilité p ou d'une fonction f.

2.2 Vision synthétique

L'apprentissage en général mélange trois grands domaines distincts qui correspondent à des données et des applications de nature différentes :

1. Traitement du signal
2. Modélisation / Apprentissage non supervisé
3. Prédiction / Apprentissage supervisé

On retrouve systématiquement dans ces trois grands domaines les notions d'estimation/approximation, de données et de dimension, bien qu'ils aient différentes applications. Les grandes lignes de ces notions sont résumées dans le tableau 1.

	Traitement du signal	Modélisation / Apprentissage non supervisé	Prédiction / Apprentissage supervisé
Estimation / approximation	Estimation $\tilde{x}(u)$ d'un signal $x(u)$ de d coefficients à partir de données z	Modèle aléatoire X de densité de proba $p(x)$ Estimation $\tilde{p}$ de p	Estimation $\tilde{y}$ de la réponse $y \in \mathcal{A}$ à une question posée sur $x \in \mathbb{R}^d$. classification : $\mathcal{A}$ est discret régression : $\mathcal{A}$ continu
Données	$z = Ax + b$ A opérateur, b bruit	$(x_i)_{i=1,\dots,n}$ $x_i \in \mathbb{R}^d$ i.i.d $x_i \sim p(\cdot)$	Données $(x_i, y_i)_{i=1,\dots,n}$ $x_i \in \mathbb{R}^d$, $y_i \in \mathcal{A}$
Dimension des variables	$u \in \mathbb{Z}^l$ ou $u \in \mathbb{R}^l$ a ℓ dimensions avec $1 \leq l \leq 4$	x a d dimensions avec $d \sim 10^6 \to 10^{23}$	x a d dimensions avec $d \sim 10^6 \to 10^{23}$
Applications	Débruitage, problèmes inverses, compressed sensing, compression	Traitement du signal, synthèse de signaux, physique statistique	Perception (vision, reconnaissance de parole), médecine, sociologie, physique, neurophysiologie, intelligence artificielle en général

TABLE 1 – Cartographie des sciences des données

Traitement du signal En traitement du signal on veut calculer une estimation $\tilde{x}(u)$ d'un signal $x(u)$ ayant d coefficients, à partir de mesures z. La dimension d est typiquement de l'ordre de 10^6. Ainsi un son a de l'ordre de 10^4 échantillons par seconde, une image 10^6 pixels. Les mesures z sont souvent obtenues à partir du signal original x par la transformation d'un opérateur A et l'ajout d'un bruit b. Supprimer le bruit b est une application de débruitage. L'opérateur de mesure A va typiquement perdre de l'information sur x que l'on aimerait récupérer en utilisant certaines informations a priori. Il s'agit de problèmes inverses. Le compressed sensing consister à minimiser le nombre de mesures tout en ajustant l'opérateur A afin que ces mesures soient potentiellement suffisantes pour pouvoir calculer une bonne estimation $\tilde{x}$ de x à partir de z. La compression est une

autre application du traitement du signal, où il s'agit de définir une représentation z qui contient le moins de bits possibles, à partir de laquelle on peut récupérer une bonne approximation $\tilde{x}$ de x.

Le signal $x(u)$ est une fonction de la variable u qui a ℓ dimensions avec $1 \leq \ell \leq 4$. Pour un signal temporel, $\ell = 1$, pour une image $\ell = 2$, pour une vidéo ou un bloc de données tridimensionnelles en séismique ou en imagerie médicale, $\ell = 3$ et si le bloc de données tridimensionnelles évolue dans le temps alors $\ell = 4$. Cette dimension reste donc petite.

Modélisation La modélisation consiste à capturer la nature et la variabilité des données. Cela se fait en estimant la distribution des données x dans l'espace. Cette distribution est caractérisée par un modèle aléatoire X, dont on suppose qu'il a une densité de probabilité $p(x)$ relativement à la mesure de Lebesgue. Le problème est de calculer un estimateur $\tilde{p}(x)$ de $p(x)$. La construction de tels modèles est nécessaire en traitement du signal, est au coeur de la physique statistique, et sert aussi à la synthèse de signaux. Cette modélisation est également utile pour faire de la prédiction. Une telle modélisation se fait à partir de n observations de signaux x_i que l'on suppose être des réalisations indépendantes de X et qui suivent donc la distribution de probabilité p. La densité de probabilité $p(x)$ est une fonction de x qui a d variables. Comme d est souvent de l'ordre du million, c'est une fonction d'un très grand nombre de variables. La difficulté principale pour calculer une bonne estimation $\tilde{p}$ de p vient de cette grande dimension.

Prédiction La prédiction consiste à calculer une estimation $\tilde{y}$ de la réponse y à une question à partir de la donnée x. Cette réponse appartient à un alphabet $\mathcal{A}$. Cela peut être un réel qui appartient à un intervalle $\mathcal{A} = [a, b] \subset \mathbb{R}$. Cette prédiction est alors un problème de *régression*. Dans un problème de *classification*, $\mathcal{A}$ est l'ensemble de toutes les classes possibles, par exemple des noms d'animaux pour la reconnaissance d'un animal y dans une image x. L'estimation se fait à l'aide de n exemples de données $x_i \in \mathbb{R}^d$ pour lesquelles on connaît la réponse y_i. On appelle cela de l'apprentissage supervisé car on fournit la réponse y_i avec la donnée x_i, ce qui est une forme de supervision, alors que dans les problèmes de modélisation on fournit juste la donnée x_i. C'est pour cela que les problèmes de modélisation sont considérés comme de l'apprentissage non-supervisé.

Les applications de l'apprentissage supervisé sont considérables. Cela inclut notamment tous les problèmes de perception (vision, audition), le diagnostic médical, mais aussi l'analyse de données dans toutes les sciences dures ou sociales ou humaines. L'apprentissage supervisé est à l'origine du renouveau de l'intelligence artificielle.

Si la réponse y est unique pour une donnée x alors on peut l'écrire comme une fonction de x : $y = f(x)$. Estimer $\tilde{y}$ revient à estimer la fonction f, qui dépend de d variables. Tout comme pour la modélisation il s'agit donc d'estimer une fonction de très grande dimension, d'où la difficulté de ce problème. Nous allons voir que comme x a un grand nombre de variables, il appartient à un espace énorme et donc que tous les exemples x_i sont très loin les uns des autres. Il est donc difficile d'interpoler les valeurs connues de f, à moins de connaître a priori des propriétés de régularité très forte sur f.

2.3 Mathématiques et informatique

L'étude de ces trois domaines fait appel à des notions mathématiques très diverses en statistiques, probabilités, en analyse harmonique, mais aussi en géométrie notamment à travers la théorie des groupes. L'informatique est surtout un outil de calcul en traitement du signal, alors que la modélisation et la prédiction font appel à des questions de recherche plus fondamentales pour l'informatique.

	Traitement du signal	Modélisation / Apprentissage non supervisé	Prédiction / Apprentissage supervisé
Mathématiques	Analyse harmonique (Fourier, ondelettes), statistiques, optimisation	avec en plus des probabilités (grande déviations), géométrie sur des groupes simples (translation)	avec en plus de la géométrie sur des groupes plus complexes
Informatique	Utilisation de librairies standards et CPU	Mémoire distribuée, réalité virtuelle	Intelligence artificielle. Calcul parallèle sur GPU et principalement langage Python

TABLE 2 – Outils mathématiques et informatiques utilisés.

En se déplaçant du traitement du signal vers la prédiction, les problèmes mathématiques deviennent de plus en plus difficiles. Historiquement, les outils de traitement du signal ont commencé à être étudiés dès les années 30 mais la discipline s'est surtout développée à partir des années 1950 notamment suite aux travaux fondateurs de Shannon. Il s'agit le plus souvent d'approximer un signal de la meilleure qualité possible à partir de données incomplètes, dégradées ou comprimées. Le cadre mathématique est aujourd'hui relativement bien compris mais beaucoup de questions restent ouvertes. En particulier, les problèmes inverses sont un domaine de recherche très actif.

Les mathématiques de la modélisation aléatoires nécessitent de définir des modèles en grande dimension. Les fondations mathématiques ont été établies par le mathématicien russe Kolmogorov. Au coeur de ces questions nous verrons apparaître la théorie des grandes déviations, initiée par les travaux de Shannon et Kolmogorov et qui s'est surtout développée depuis les années 1970. Nous verrons que ce domaine est dominé par deux hypothèses simplificatrices, de Gaussianité et de Markov, qui permettent d'estimer des densités de probabilité en grande dimension. Cependant, le plus souvent ces hypothèses ne sont pas satisfaites. L'enjeu mathématique et algorithmique est d'aller au-delà, en remplaçant ces hypothèses par des cadres plus généraux qui permettent cependant de faire de l'estimation en grande dimension.

Les problèmes de prédiction sont aussi difficiles car on comprend très mal la notion de régularité en grande dimension. Toutefois, si le cadre mathématique de la prédiction est relativement difficile et mal compris, on dispose a contrario de beaucoup d'outils algorithmiques en informatique qui permettent de faire des expériences numériques. Cela peut se faire par exemple en utilisant des fonctions de la librairie Python `scikit-learn`. A l'heure actuelle, l'informatique a beaucoup d'avance sur les mathématiques au sens où l'on dispose d'algorithmes ayant de bonnes performances, comme les réseaux de neurones, mais dont on comprend mal les propriétés mathématiques.

2.4 Problématiques Communes

Au-delà des différences entre les problèmes et les données, on retrouve des problématiques communes entre le traitement du signal, la modélisation et la prédiction. Il s'agit notamment de spécifier la notion de "régularité", de définir des classes de modèles *a priori*, de mesurer leur complexité et de définir des représentations qui permettent de simplifier la résolution du problème. Ces différents aspects sont résumés dans le tableau 3.

Régularité La régularité d'un signal $x(u)$ permet de réduire le nombre de mesures pour restituer ce signal, par exemple le nombre d'échantillons de x. Une courbe régulière nécessite peu de points pour être interpolée. C'est l'idée derrière le théorème d'échantillonnage de Shannon qui caractérise la régularité par la transformée de Fourier d'un signal et calcul une approximation $\tilde{x}(u)$ avec un opérateur linéaire. Cependant, des signaux comme les images présentent généralement des singularités. Une bonne estimation $\tilde{x}$ de x nécessite plus d'échantillons sur les zones non régulières, ce qui nécessite de définir des approximations adaptatives, non linéaires. On verra que cela demande de définir des notions de régularités plus complexes. Les différentes notions de régularité en basse dimension sont relativement bien comprises et correspondent à différents espaces de fonctions comme les espaces de Sobolev, de Hölder ou de Besov.

La régularité de $p(x)$ doit être beaucoup plus forte pour pouvoir l'estimer car x a un grand nombre d de variables. Cette régularité peut être définie par l'invariance de $p(x)$ lorsque x est

	Traitement du signal	Modélisation / Apprentissage non supervisé	Prédiction / Apprentissage supervisé
A priori	Régularité du signal $x(u)$ en fonction de u.	Régularité de $p(x)$ en fonction de x	Régularité de $y = f(x)$
Complexité des modèles	Capacité de compression du signal. Entropie de Kolmogorov et de Shannon.	Entropie de classes de modèles	Entropie, complexité de Rademacher, dimension de Vapnik
Représentation $\Phi(x)$	Coefficients de x dans une base ou dans un dictionnaire redondant. On recherche des représentations *parcimonieuses* (peu de coefficients non nuls).	Estimateurs de moments généralisés, polynomiaux ou autres	Attributs discriminants de x pour estimer y

TABLE 3 – Cartographie des sciences des données

transformé par des opérateurs particuliers. Ainsi on dit que le modèle est stationnaire si $p(x)$ ne change pas lorsque le signal x est translaté.

La prédiction de $y = f(x)$ à partir des exemples connus $y_i = f(x_i)$ peut être interprétée comme un problème d'interpolation, en grande dimension. Il faut pour cela avoir des informations a priori très fortes sur la régularité de $f(x)$. Cette régularité peut aussi être spécifiée par des invariants relativement à des groupes de transformations connues, qui sont des symétries de f. Actuellement, on ne sait pas bien définir la régularité en grande dimension. L'extension des notions de régularité définies en basse dimension ne sont pas suffisamment fortes pour capturer les propriétés dont on a besoin.

Complexité On calcule une approximation en la choisissant dans une classe particulière, qui dépend de l'information a priori que l'on a. La complexité d'un modèle est une mesure de la taille de cette classe d'approximation. Plus la taille est grande, plus l'approximation peut être précise et plus la complexité du modèle est grande. Ces notions de complexité sont mesurés par des entropies qui peuvent être définies différemment suivant que la classe soit finie ou infinie, suivant que le modèle soit déterministe ou stochastique. L'entropie de Shannon est définie dans un cadre stochastique où l'on a spécifié une mesure de probabilité. Dans un cadre déterministe, elle est remplacé par l'entropie de Kolmogorov. Pour des problèmes de classification pour lesquels les valeurs de la réponses sont discrètes et finies, la notion de complexité peut se mesurer différemment notamment par la complexité de Rademacher ou la dimension de Vapnik-Chervonenkis. On voit ici qu'il y a beaucoup de façon de définir des modèles et leur complexité, suivant les situations.

Représentations Φ L'utilisation de représentation permet d'expliciter l'information importante pour la résolution d'un problème. Ainsi, la régularité peut se définir par la parcimonie d'une représentation $\Phi(x) = (v'_1, ..., v'_{d'})$ de $x = (v_1, ..., v_d)$ dans une base ou dans un dictionnaire redondant. Parcimonie veut dire que la plupart des nouvelles variables v' sont quasiment nulles. En effet, si x peut être caractérisé par un petit nombre de coefficients dans un système de coordonnées prédéfini (base ou dictionnaire) alors cela veut dire qu'il a relativement peu de variabilité et donc qu'il a une forme de régularité. Les notions les plus classiques de régularité sont définies dans la base de Fourier. Dans une base d'ondelettes, on peut définir des notions plus complexes de régularité adaptatives. Un enjeu fondamental du traitement du signal est d'optimiser la représentation d'un signal afin qu'elle soit la plus parcimonieuse possible. C'est une façon de révéler la régularité de ce signal.

Pour l'estimation d'une densité de probabilité, la représentation $\Phi(x)$ correspond souvent à des estimations de moments. Ces moments se calculent en projetant $p(x)$ sur des familles de fonctions $\{\phi_k(x)\}_k$ prédéfinies :

$$\mathbb{E}[\phi_k(x)] = \int_{\mathbb{R}^d} \phi_k(x)\, p(x)\, dx$$

Dans ce cas, la représentation $\Phi(x)$ est choisie afin d'obtenir des estimations de chacun de ces moments. Les moments les plus utilisés sont la moyenne pour laquelle $\phi_u(x) = x(u)$ ainsi que les moments d'ordre 2 pour lesquels $\phi_{u,u'}(x) = x(u)\, x(u')$. On peut ainsi définir des moments polynomiaux d'ordres supérieurs et bien d'autres moments en choisissant différemment les fonctions

$\phi_k(x)$. Si l'on connaît un certain nombre de moments, on peut alors calculer un estimateur $\tilde{p}(x)$ de la densité $p(x)$ en supposant que l'on ne dispose pas d'information supplémentaire, ce qui revient à maximiser l'entropie de $\tilde{p}(x)$. Ce principe est à la base de la physique statistique et définit $\tilde{p}(x)$ sous la forme d'une énergie de Gibbs. On peut montrer que $\log \tilde{p}(x)$ est alors une combinaison linéaire des $\phi_k(x)$. Une question centrale pour l'estimation de densité de probabilité va être d'optimiser le choix des $\phi_k(x)$ afin d'optimiser l'approximation $\log \tilde{p}(x)$ de $\log p(x)$.

Dans un cadre de prédiction, on peut interpréter $\Phi(x)$ comme un ensemble d'attributs qui permettent de discriminer les différentes réponses $y = f(x)$. Le but va être de pouvoir calculer une approximation de $f(x)$ aussi simple que possible à partir de $\Phi(x)$, idéalement par combinaison linéaires des variables de $\Phi(x)$. Dans le cas d'un problème de classification, nous verrons qu'il s'agit d'aplatir les frontières de classification. Le calcul de ces représentations est au coeur de l'apprentissage statistique.

3 Challenge Data

Il est difficile de comprendre les enjeux en sciences des données sans développer des algorithmes et faire des expériences numériques sur des données réelles. Pour permettre cela, avec une équipe de doctorants et de post-doc à l'ENS, nous avons mis en place un site web de challenge de données.

3.1 Présentation

Le site Web *challengedata.ens.fr* met à disposition des challenges de traitement de données par apprentissage supervisé. 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 des 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 sont des problèmes de prédiction -régression ou de classification- avec des données réelles, mises à disposition par des entreprises ou des laboratoires de recherche. Ils couvrent un large spectre d'applications, sur des images, sons, textes, données médicales, mesures physiques, données d'Internet, et sont présentés dans des vidéos sur le site du Collège de France. 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. 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 après chaque clôture.

Le site Web *challengedata.ens.fr* offre aussi un support aux Professeurs voulant utiliser ces challenges comme projets pour les élèves de leur cours. L'enseignant peut inscrire son cours sur le site Web, et spécifier une liste de projets pouvant être traités par les élèves dans le cadre du cours. Il a accès aux scores et aux rapports postés par ses élèves. Chaque année les propositions de nouveaux challenges doivent être soumis en Septembre en envoyant un mail à *challenge.data@ens.fr*. Ils sont validés par une équipe de l'École Normale Supérieure.

Cette année les challenges ont été supervisés à l'ENS par *Mathieu Andreux, Tomas Anglès, Georgios Exarcharkis, Louis Thiry, John Zarka, Sixin Zhang*. L'organisation de ces challenges de données est soutenue par la chaire CFM de l'Ecole Normale Supérieure, et par la Fondation des Sciences Mathématiques de Paris.

3.2 Spécification des problèmes

Les challenges sont des problèmes d'apprentissage supervisés. Pour chaque challenge on dispose d'un ensemble de n exemples de données labélisées $\{x_i\}_{i \leq n}$ et $\{y_i\}_{i \leq n}$. Le site fournit par ailleurs un ensemble de n_t données de tests $\{x_{t,i}\}_{i \leq n_t}$ pour lesquelles les réponses $\{y_{t,i}\}_{i \leq n_t}$ ne sont pas fournies. Les candidats doivent fournir une estimation $\{\tilde{y}_{t,i}\}_{i \leq n_t}$ des réponses sur ces données de test. Le site spécifie la fonction de score (risque) qui quantifie l'erreur entre les $y_{t,i}$ et $\tilde{y}_{t,i}$.

Dans un problème de régression, cette erreur peut être l'erreur quadratique moyenne

$$\tilde{R}\{y_{t,i}, \tilde{y}_{t,i}\}_{i\leq n_t} = \frac{1}{n_t}\sum_{i=1}^{n_t} |y_{t,i} - \tilde{y}_{t,i}|^2.$$

L'erreur est cependant souvent mesurée différemment suivant l'application.

Pour chaque soumission d'une estimation de $\{\tilde{y}_{t,i}\}_{i\leq n_t}$, le site web fournit au participant la valeur du score $\tilde{R}\{y_{t,i}, \tilde{y}_{t,i}\}_{i\leq n_t}$ ainsi que son classement parmi les autres participants. Afin d'éviter tout phénomène de surapprentissage sur les données de test, les participants sont limités à deux soumissions de leurs estimations par jour.