Résumés des communications J-STAR 2015

Agrégation PAC bayésienne

Pierre Alquier (Ensae ParisTech)
Bornes PAC-bayésiennes en agrégation : introduction, et aspects algorithmiques
(travail en collaboration avec James Ridgway et Nicolas Chopin)

Introduites par McAllester (1998) en 'machine learning', les bornes PAC-Bayésiennes sont devenues un outil puissant pour analyser les performances théoriques de procédures d'agrégation d'estimateurs : Catoni (2007), Dalalyan et Tsybakov (2008)... L'implémentation pratique de ces procédures est souvent faite via des méthodes de Monte-Carlo, assez lentes en pratique. Dans cet exposé, après une introduction générale aux bornes PAC-Bayésiennes, je proposerai l'utilisation d'une approche dite 'variationnelle', pour remplacer les algorithmes de Monte Carlo par des algorithmes d'optimisation, pour lesquels on peut obtenir une garantie théorique sur l'erreur de prévision.

Lucie Montuelle (Université Paris Diderot - Paris VII)
Agrégation PAC-bayésienne d'estimateurs affines
(travail en collaboration avec Erwan Le Pennec)

Dans le cadre du modèle de régression à design fixe, une stratégie d'estimation consiste en l'agrégation d'estimateurs. L'estimateur agrégé est la moyenne par rapport à une mesure bien choisie des estimateurs de la collection. Souvent, le poids accordé par la mesure à l'estimateur décroît en fonction de son risque. Les poids exponentiels, proportionnels à $\exp(-r/T)$ où r est un estimateur du risque, sont largement utilisés. Lorsque les estimateurs sont des fonctions affines des données, Barron et Leung, ainsi que Dalalyan et Salmon, ont majoré l'écart moyen entre le risque de l'estimateur agrégé et celui du meilleur estimateur de la collection, via une inégalité oracle exacte en espérance. Ce contrôle ne peut être obtenu en probabilité. Avec une pénalisation du risque, Dai et al. ont fourni une inégalité oracle inexacte en probabilité. Dans cet exposé, une inégalité oracle exacte en probabilité sera produite à condition de prendre en compte le rapport signal sur bruit dans la pénalisation. Un continuum sera établi entre les deux types de résultat. Je me concentrerai sur l'exemple des estimateurs par projection.

Benjamin Guedj (INRIA Lille)
Une approche PAC-bayésienne du ranking binaire en grande dimension
(travail en collaboration avec Sylvain Robbiano)

Le ranking binaire est un problème d'apprentissage supervisé classique, qui consiste à apprendre à ordonner des observations comme leurs labels. Cet objectif passe bien souvent par la construction de fonctions dites de scoring. Je présente dans cet exposé une approche d'agrégation linéaire sur un dictionnaire de fonctions déterministes, à l'aide d'outils PAC-bayésiens. Notre approche est soutenue par des résultats oracles en probabilités faisant apparaître les vitesses minimax pour le ranking introduites par Clémençon et Robbiano (2011), ainsi que des termes permettant une adaptation à la dimension intrinsèque du problème d'apprentissage.

Boosting - Bagging / Forêts aléatoires

Eric Matzner-Løber (Ensae-Ensai Formation continue / Université Rennes 2)
Méthodes itératives et estimateurs de la famille des k plus proches voisins
(travail en collaboration avec Pierre-André Cornillon et Nicolas Hengartner)

Dans cet exposé nous rappellerons les principes de l'estimateur de réduction de biais itéré (ibr) et nous montrerons que cet estimateur itératif ne peut pas utiliser comme lisseurs des lisseurs de la famille des k plus proches voisins (k-ppv, k-ppv mutuels, k-ppv symétrisés).

Christine Tuleau Malot (Université de Nice Sophia-Antipolis)
Forêts aléatoires : des données de grandes dimensions aux données massives
(travail en collaboration avec Robin Genuer, Jean-Michel Poggi et Nathalie Villa)

Les forêts aléatoires, introduites par Breiman et Cutler, sont une méthode d'apprentissage par arbres, dans la lignée de l'algorithme CART, qui peut être mise en oeuvre aussi bien dans un cadre de régression que dans un cadre de classification. De par le fait que les forêts aléatoires mêlent bagging et sélection de variables, cette méthode s'adapte naturellement aux données de grande dimension. Cependant depuis quelques années, en sus des données de grande ou ultra grande dimension, un intérêt tout particulier est porté aux données massives. Une question naturelle est donc d'essayer d'adapter les forêts aléatoires et les notions connexes comme l'importance des variables au cadre des données massives.

Robin Genuer (Université de Bordeaux)
Analyse du biais de forêts purement aléatoires
(travail en collaboration avec Sylvain Arlot)

Les forêts aléatoires sont une méthode d'apprentissage statistique très performante et très utilisée dans des domaines d'application divers. Cependant, leur analyse théorique est encore un problème ouvert aujourd'hui. Une première approche est d'étudier des méthodes simplifiées, comme les forêts purement aléatoires, pour tenter d'apporter des explications des performances de la méthode initiale. Dans ce travail, on étudie l'erreur d'approximation (le biais) de modèles de forêts purement aléatoires dans un cadre de régression, en se concentrant sur l'influence du nombre d'arbres dans la forêt. Sous des hypothèses de régularité sur la fonction de régression, nous montrons que le biais de la forêt infinie décroît plus rapidement (en fonction la taille de chaque arbre) que celui d'un arbre. Ainsi, les forêts infinies atteignent une vitesse de convergence (en fonction du nombre d'observations) strictement plus rapide. De plus, nos résultats permettent de donner un nombre d'arbres minimum suffisant pour atteindre la même vitesse qu'une forêt infinie. Enfin, notre analyse montre un résultat annexe explicitant un lien entre le biais d'une forêt purement aléatoire et celui d'un estimateur à noyau.

Agrégation linéaire

Frédéric Lavancier (Université de Nantes)
Une procédure générale pour combiner des estimateurs
(travail en collaboration avec Paul Rochet)

Nous proposons une procédure, utilisable dans un cadre général, visant à combiner plusieurs estimateurs d'une même quantité. Dans l'esprit du model averaging ou du forecast averaging, l'estimateur final est une moyenne pondérée des estimateurs initiaux, où les poids somment à un. Les poids optimaux minisant l'erreur quadratique moyenne (EQM) s'expriment en fonction de la matrice des EQM (croisés) des estimateurs initiaux. L'estimateur final est obtenu en estimant cette matrice, en général à partir des mêmes données que celles utilisées pour construire les estimateurs initiaux. Nous obtenons un contrôle à horizon fini de la différence entre l'oracle et l'estimateur final, d'où s'en suit l'optimalité asymptotique de notre procédure sous des conditions raisonnables sur la matrice EQM empirique. Cette méthode est illustrée sur des modèles paramétriques et semi-paramétriques, ajustés à des données i.i.d ou des processus stochastiques, d'où il résulte que l'estimateur final est dans la plupart des cas bien meilleur que chacun des estimateurs initiaux.

Guillaume Lecué (Ensae ParisTech)
Quelques aspects géométriques et stochastiques du problème d'agrégation
(travail en collaboration avec S. Mendelson et P. Rigollet)

Dans le modèle de régression à design aléatoire, nous considérerons la procédure de minimisation du risque empirique (ERM) comme procédure d'agrégation pour les trois types de problème d'agrégation (MS, convexe et linéaire). Nous présenterons sa propriété d'optimalité pour le problème d'agrégation convexe. Nous étudierons sa sous-optimalité pour les problèmes d'agrégation MS et linéaire. Nous présenterons ensuite deux procédures d'agrégation optimales pour le problème d'agrégation de type MS en présentant le compromis géométrie/complexité. Finalement, nous introduirons la propriété de petite boule pour le problème d'agrégation linéaire.

Arnak Dalalyan (Ensae ParisTech)
Garanties théoriques pour le calcul approché de l'agrégat par poids exponentiels

Il a été démontré dans plusieurs contextes que l'agrégat par poids exponentiels est statistiquement optimal et préférable à bien d'autres estimateurs. Cependant, le calcul effectif de l'agrégat par poids exponentiels représente un problème difficile qui constitue un obstacle à son utilisation. Le but de cet exposé est de présenter des résultats récents sur la complexité computationnelle du calcul approché de l'agrégat par poids exponentiel dans le cas où le risque (pénalisé) utilisé dans les poids est fortement convexe.

Agrégation de tests - Tests multiples

Etienne Roquain (Université Pierre et Marie Curie - Paris VI)
Estimation d'un facteur de translation dans un modèle de test multiple
(travail en collaboration avec Sylvain Delattre)

Le problème de test multiple devient particulièrement épineux lorsque des variables latentes viennent perturber les mesures observées. Il convient ainsi d'essayer d'estimer ces "facteurs" pour les faire disparaître dans les statistiques de tests. Afin d'appréhender le problème, nous étudierons le cas simple d'un facteur de translation avec un bruit gaussien.

Pierre Neuvial (Université d'Évry Val d'Essonne, Genopole)
Inférence post hoc en test multiple
(travail en collaboration avec Gilles Blanchard et Etienne Roquain)

Lorsque l'on teste simultanément un grand nombre d'hypothèses nulles, une pratique courante dans les applications (notamment en génomique ou en neuro-imagerie) consiste à (i) sélectionner un sous-ensemble d'hypothèses candidates, puis (ii) raffiner cette sélection à l'aide de connaissances a priori. Le contrôle de mesures de risque classiques en test multiple comme le False Discovery Rate ne fournit aucune garantie statistique sur les ensembles d'hypothèses obtenus par ce processus. Ce fossé entre les besoins des applications et les garanties fournies par les méthodes actuelles motive le développement de procédures dites post hoc, c'est-à-dire pour lesquelles les ensembles d'hypothèses sélectionnés peuvent être définis par l'utilisateur de la procédure, après avoir "vu les données". Goeman et Solari (Stat. Science, 2011) ont proposé des procédures post hoc reposant sur la notion de "closed testing". Nous illustrerons, dans des modèles simples, des limitations de ces procédures et proposerons une construction alternative reposant sur une nouvelle mesure de risque appelée joint Family-Wise Error Rate. Il s'agit d'un travail en cours.