6ème Journée Probabilités/Statistique Dijon-Besançon

Livret d'informations

Lundi 5 juillet 2021

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

LmB
laboratoiredemathématiquesdebesançon
UNIVERSITÉ DE FRANCHE-COMTÉ • CNRS • UMR 6623

cnrs

IMB
INSTITUT DE MATHÉMATIQUES
DE BOURGOGNE

1 Programme détaillé du lundi 5 juillet

Lundi 5 juillet		
09h15 - 10h00	Accueil - Café	
10h00 - 10h35	Modèles IDLA avec un nombre infini de sources et une forêt IDLA	Arnaud Rousselle
10h40 - 11h15	Robust Shortest Path Problem in presence of uncertainty	Chifaa Dahik
11h20 - 11h30	Pause café	
11h30 - 12h05	Comparaison de trajectoires qualitatives avec des processus semi-Markoviens : une application en analyse sensorielle	Cindy Frascolla
12h10 - 14h00	Pause déjeuner - Café-Photo de groupe	Rendez-de-vous à préciser
14h00 - 14h35	Amélioration de l'estimation d'un total en sondages par des estimateurs assistés de forêts aléatoires	Mehdi Dagdoug
14h40 - 15h15	Résolution et sélection des hyperparamètres pour la Support Vector Régression linéaire sous contraintes	Quentin Klopfenstein
15h20 - 15h35	Pause café	
15h35 - 16h10	Infinitesimal tree-based gradient boosting	Jean-Jil Duchamps

2 Résumés des interventions

2.1 Lundi 5 juillet : matin

Modèles IDLA avec un nombre infini de sources et une forêt IDLA

Auteur : Arnaud Rousselle[1], Nicolas Chenavier[2] et David Coupier[3]
Affiliation : [1]LmB, Université Bourgogne Franche-Comté, France.
[2]Université du Littoral, France.
[3]Institut des Mines Télécom Lille-Douai, France.
Heure : 10h00 - 10h35

Résumé : Trouvant une motivation dans un article de chimie de Meakin et Deutch (1986) portant sur l'amélioration de la régularité de la surface de plaquettes métalliques, le modèle d'agrégation limitée par diffusion interne (IDLA) a été introduit dans la littérature mathématique par Diaconis et Fulton (1993) et Lawler, Bramson et Griffeath (1992). Il s'agit d'un modèle de croissance aléatoire construit, dans le cas classique, à partir de marches aléatoires sur $\mathbf{Z}^d$ émises successivement depuis 0 (la $n+1^{\mathrm{e}}$ particule ne partant qu'après que la n^{e} ait contribué à l'agrégat). Pour se modèle et ses généralisations, l'effort a été concentré sur l'établissement de théorèmes de forme asymptotique. Par ailleurs, on peut associer naturellement un arbre aléatoire enraciné en 0 à ce processus en retraçant la généalogie des ajouts à l'agrégat : lorsqu'une particule contribue à l'agrégat on ajoute dans l'arbre une arête entre le dernier site visité dans l'agrégat courant et le site qu'elle y ajoute. Cet arbre, dont l'étude est délicate du fait de son caractère radial en particulier, n'a pour le moment pas été étudié dans la littérature. Dans cet exposé, on introduira une forêt aléatoire basée sur le processus IDLA dont la loi est stationnaire sous les translation verticale et que l'on conjecture pouvoir fournir une approximation de l'arbre IDLA de dimension 2 loin de l'origine dans la direction $(1,0)$. Dans cet optique, on introduira et étudiera des agrégats IDLA ayant pour sources tous les points de $\{0\} \times \mathbf{Z}$ et non seulement l'origine comme dans le cas classique.

Robust Shortest Path Problem in presence of uncertainty

Auteur : Chifaa Dahik[1], Jean-Marc Nicod[1], Landy Rabehasaina[2] et Zeina Al Masry[1]
Affiliation : [1]FEMTO, Université Bourgogne Franche-Comté, France.
[2]LmB, Université Bourgogne Franche-Comté, France.
Heure : 10h40 - 11h15

Résumé : We address a specific class of combinatorial problems with correlated cost coefficients belonging to an ellipsoidal uncertainty set. An absolute robust problem in these settings is a well-known NP-Hard problem. To tackle this problem, we propose a heuristic approach based on the Frank-Wolfe (FW) algorithm. In our approach, we take a radically different perspective on FW by looking at the exploration power of the integer inner iterates of the method. Experimental tests have been realized for the robust shortest path problem. Comparisons with the optimal solution given by the mixed integer second order cone programming solver of CPLEX have also been provided. Another evaluation of the quality of the heuristic solution is proposed by a Semi-Definite Programming (SDP) relaxation. The corresponding SDP problem is solved by a "decomposition

through formalization in a product space" algorithm with sparse computations. Our findings are illustrated by comprehensive numerical experiments.

Comparaison de trajectoires qualitatives avec des processus semi-Markoviens : une application en analyse sensorielle

Auteur : Cindy Frascolla[1], Hervé Cardot [1], Guillaume Lecuelle [2] et Pascal Schlich[2]
Affiliation : [1]IMB, Université Bourgogne Franche-Comté, France.
[2]Centre des Sciences du Goût et de l'Alimentation, INRAE.
Heure : 11h30 - 12h05

Résumé : Nous nous intéressons à des tests d'hypothèse pour des panels de processus semi-Markoviens, motivés par une application en analyse sensorielle. Pour modéliser les différentes sensations perçues au cours de la dégustation d'un produit, Lecuelle et al. (2018) ont proposé d'utiliser les processus semi-Markoviens. Pour déterminer si deux produits testés sont perçus différemment, on introduit dans ce travail un test statistique basé sur le test du rapport de vraisemblance entre deux modèles semi-Markoviens. Pour construire la zone de rejet, trois approches sont évaluées : deux approches basées sur des techniques ré-échantillonnage (bootstrap paramétrique et permutations) et une approche asymptotique basée sur la loi du rapport de vraisemblance. Ce travail est découpé en deux parties : une partie méthodologique où nous comparons les trois approches à partir de simulations et de tests réalisés sur des jeux de données réelles de dégustations de chocolats et de fromages et une partie théorique. Pour cette dernière, nous étudions la convergence asymptotique lorsque le nombre L de trajectoires tend vers l'infini, des estimateurs du maximum de vraisemblance des paramètres des processus semi-Markoviens et la distribution asymptotique du rapport de vraisemblance. Nous considérons deux modèles d'observation pour chaque trajectoire : celui d'un processus semi-Markovien absorbant et celui où chaque trajectoire est composée d'un nombre aléatoire de transitions.

2.2 Lundi 5 juillet : après midi

Amélioration de l'estimation d'un total en sondages par des estimateurs assistés de forêts aléatoires

Auteurs : Mehdi Dagdoug[1], Camélia Goga[1] et David Haziza[2]
Affiliation : [1]LmB, Université Bourgogne Franche-Comté, France.
[2]Université de Montréal, Canada.
Heure : 14h00 - 14h35

Résumé : De nos jours, les enquêtes par sondage font face à l'émergence de jeux de données complexes et de très grandes tailles. Ce type de nouvelles bases de données soulève de nouveaux défis et l'estimation de paramètres d'intérêt tels que le total, le ratio ou les quantiles basés sur des modèles paramétriques traditionnels peuvent s'avérer inefficaces. Dans ce travail, nous proposons une nouvelle classe d'estimateurs assistés par un modèle et basés sur des forêts aléatoires pour l'estimation du total d'une variable d'intérêt. Sous certaines conditions de régularité sur la variable d'intérêt, la structure des forêts et le plan de sondage, l'estimateur

proposé est asymptotiquement sans biais et consistant pour l'estimation d'un total. Un estimateur consistant de la variance est suggéré et la distribution asymptotique de l'estimateur assisté par des forêts aléatoires est obtenue également permettant ainsi la construction des intervalles de confiance asymptotiques. Les simulations effectuées suggèrent que l'estimateur proposé est généralement efficace et meilleur que les estimateurs basés sur des modèles paramétriques dans le cas de relations complexes et en présence d'un très grand nombre de variables auxiliaires.

Résolution et sélection des hyperparamètres pour la Support Vector Régression linéaire sous contraintes

Auteurs : Quentin Klopfenstein[1]
Affiliations :
[1]IMB, Université Bourgogne Franche-Comté.
Heure : 14h40 - 15h15

Résumé : La Support Vector Regression (SVR) est un estimateur pour des modèles de régression qui permet d'estimer des fonctions linéaires ou non-linéaires. Le problème d'optimisation dual sous-jacent implique la minimisation d'une fonction quadratique sous des contraintes d'inégalités, dites de boîtes. Cet estimateur dépend de deux hyperparamètres $C > 0$ et $\varepsilon > 0$. L'hyperparamètre C contrôle le poids entre la régularisation $\ell - 2$ et la fonction d'attache aux données ($\varepsilon-$insensitive). Dans cet exposé, nous nous proposons d'étudier l'ajout de contraintes linéaires sur l'estimateur SVR classique ce qui permet notamment de faire de la régression non-négative. Nous étudierons le problème d'optimisation sous-jacent et proposerons un algorithme efficace pour sa résolution. Dans un deuxième temps, un algorithme de différentiation implicite sera utilisé pour choisir les hyperparamètres optimaux selon un critère de performance. Enfin, nous terminerons en appliquant les méthodes décrites ci-dessus sur des jeux de données réels, notamment pour une application dans la recherche biomédicale sur l'estimation de proportions de cellules immunitaires au sein d'une tumeur.

Infinitesimal tree-based gradient boosting

Auteurs : Jean-Jil Duchamps[1] et Clément Dombry[1]
Affiliations :
[1]LmB, Université Bourgogne Franche-Comté.
Heure : 15h35 - 16h10

Résumé : We introduce infinitesimal gradient boosting as a limit in the vanishing learning rate asymptotic of the popular tree-based gradient boosting from machine learning. For this purpose, we use a new class of randomized regression tree with randomized binary splitting according to a softmax selection. Our main result is the convergence of the associated stochastic algorithm and the characterization of the limiting algorithm by a nonlinear ordinary differential equation in a infinite dimensional function space. The solution, coined infinitesimal gradient boosting, provides a smooth path in a Sobolev-like function space along which the training error decreases, the residuals are centered and the total variation is well-controlled.