Offres de Postdoc/Ingénieur de recherche en Statistique & Sciences des données

Projet AStERiCs: Apprentissage Statistique à l'Echelle pour la Représentation et la Classification non-supervisées

Contexte et présentation générale du projet:

La disponibilité des données en masse révolutionne les questions relatives à leur traitement, analyse, exploitation et valorisation par les acteurs du numérique (académiques, entreprises, acteurs politiques, etc). La problématique principale est celle de l'élaboration de modèles originaux et génériques permettant représentation et classification de données massives, et celle du développement d'algorithmes efficaces optimisés à l'échelle pour les obtenir. Ce contexte de traitement et d'analyse à grande échelle rompt en effet avec la façon selon laquelle se posait classiquement la question de la construction et de l'inférence des modèles à partir de données brutes; la plupart de ceux de l'état de l'art se trouvent en effet inopérants à l'échelle, aussi bien d'un point de vue théorique, que pratique : problèmes d'inférence d'un très grand nombre de paramètres (fléau de la dimension), et/ou incapacité en temps et/ou en mémoire de mettre en œuvre des algorithmes centralisés classiques pour de très gros volumes de données, etc.

AStERiCs est un projet de recherche fondamentale financé dans le cadre du dispositif RIN (Réseaux d'Intérêts Normands)-Recherche dont l'objectif structurel est de fédérer la recherche scientifique en Normandie dans le domaine de la *science statistique des données*, en s'appuyant sur une démarche scientifique pluridisciplinaire impliquant modélisation mathématique, inférence, représentation et classification de données issues d'environnements complexes, hétérogènes, dynamiques et incertains. AStERiCs vise à élaborer un cadre scientifique et technique, complet, pour traiter, analyser, exploiter et valoriser des données massives, complexes, hétérogènes, dynamiques et peu ou non-annotées. Le but est de transformer des données en connaissances sous forme de représentations précises des informations liées aux données, de catégorisations pertinentes de telles informations, jusqu'à la valorisation de celles-ci en révélant/restaurant le modèle générateur des données. Le projet AStERiCs traite ainsi le problème de la grande échelle, sous tous ses aspects de modélisation et d'inférence. Plus précisément, les axes de recherche traitent des grands thèmes suivants : Statistique, Apprentissage, Analyse de données, Classification, Optimisation, Traitement du signal, Grande dimension.

Descriptifs des postes :

Offre 1 : PostDoc 18 mois :

L'objectif de la recherche à laquelle participera ce postdoc est d'élaborer, dans un cadre non-supervisé, des modèles permettant la représentation et la classification de données continues dont la dimension peut être infinie, et de développer des algorithmes d'inférence optimisés.

On s'intéressera principalement aux mélanges parcimonieux infinis. Les mélanges sont une famille de modèles à variables latentes qui permet de modéliser explicitement les données observées comme une observation marginale d'un couple intégrant une variable latente (manquante) représentant généralement une structure cachée, possiblement hiérarchique (qui peut être restaurée étant donné un estimateur du modèle). L'estimation peut s'effectuer de façon paramétrique ou non. On s'intéressera au cas ou l'espace de la variable latente peut être de dimension infinie.

L'inférence de ce modèle dans ce contexte à l'échelle (très grande dimension) nécessite le contrôle du problème d'optimisation lors de l'estimation par maximum de vraisemblance et suggère de nouvelles stratégies de régularisation (cadre non-supervisé, possible problème d'identifiabilité). Pour cela on pourra s'appuyer sur des critères de log-vraisemblance pénalisés où la pénalité devra tenir compte des données manquantes (problème de sélection non-supervisée de variables) et de sa possible organisation en hiérarchie (élaboration d'une régularisation structurée). Ces problèmes de régularisation dans un contexte non-supervisé (classification et sélection simultanée de variables) sont assez récents (Devijver (2015a,b)., G. Celeux, et al. (2011), Ruan et al. (2011), Witten & Tibshirani, (2010)), pour ce qui concerne notamment les données fonctionnelles (Devijver (2015b)). On pourra s'intéresser à l'extention de ses modèles aux cas de modèles de mélanges d'experts régularisés (travaux menés actuellement).

- *Profil recherché* :
  - Être titulaire d'un doctorat en mathématiques appliquées avec une spécialisation confirmée en statistique/apprentissage
  - Avoir une expérience en apprentissage de représentations à partir de données réelles massives
  - Avoir un goût particulier pour le développement d'algorithmes et les applications
  - Maîtriser la programmation Matlab/R/Python

- *Laboratoire d'accueil* : UMR CNRS LMNO

Offre 2 : PostDoc 18 mois :

L'objectif de la recherche à laquelle participera ce postdoc est d'élaborer, dans un cadre non-supervisé, des modèles de représentation et de classification de données principalement discrètes de grande dimension et de développer des algorithmes d'inférence optimisés.
Apportant des réponses précises et flexibles aux problèmes multiformes de classification, les modèles de mélange fini de distributions de probabilité sont devenus aujourd'hui un outil extrêmement étudié et utilisé avec succès dans des disciplines variées (génétique, traitement d'images, astronomie...). On s'appuiera sur des modèles de mélange discret et on pourra commencer par l'extension de travaux de Karlis et Meligkotsidou (2007) et Shi et Valdez (2014) à des données multivariées et surdispersées. On considérera également des modèles de mélanges non- ou semi-paramétriques en se basant sur des méthodes à noyau comme dans Benaglia et al. (2009). On cherchera à calibrer le paramètre de lissage du noyau (la fenêtre) par des méthodes de sélection récentes comme dans Goldenshluger et Lepski (2011) ou Lacour et al (2017). Il a été prouvé que ces méthodes surpassent les critères classiques de validation croisée en termes de temps de calcul (e.g., Chagny et Roche 2015), ce qui est crucial dans ce contexte de données à grande échelle. Les résultats pourront également être étendus à la modélisation séquentielle des données par des méthodes markoviennes. L'objectif des algorithmes développés est le traitement de données génomiques (de type RNA-Seq) et en particulier l'analyse différentielle de gènes.

- *Profil recherché* :
  - Être titulaire d'un doctorat en mathématiques appliquées avec une spécialisation confirmée en statistique/apprentissage
  - Avoir une expérience en apprentissage de représentations à partir de données réelles massives
  - Avoir un goût particulier pour le développement d'algorithmes et les applications
  - Maîtriser la programmation R/Matlab/Python

- *Laboratoire d'accueil* : UMR CNRS LMRS

Offre 3 : Ingénieur de Recherche (IGR) 18 mois :

L'un des objectifs majeurs du projet est la création d'une plateforme scientifique et technique ouverte et à visibilité internationale autour de l'apprentissage non-supervisé sur masses de données (BigData). Celle-ci proposera une architecture complète (pré-traitement, représentation, classification/catégorisation, visualisation) et considérera au moins les aspects suivants des données en masse : données de grande dimension, de gros volume, hétérogènes, non-annotées, dynamiques. Elle proposera des algorithmes originaux de traitement, d'analyse et d'exploitation de données hétérogènes de différents types (continues, longitudinales/fonctionnelles, discrètes, signaux/images) avec un calcul déporté et de haute performance (le CRIANN sera sollicité dans le cadre du projet pour avoir des ressources de calcul distribué haute performance). Elle proposera trois niveaux d'accessibilité (i) entreprise (ii) intermédiaire (e.g. étudiant) et (iii) chercheur, selon les besoins respectifs.

La personne recrutée participera en collaboration avec le reste du personnel du projet à la création de cette plateforme selon les étapes suivantes. La première concerne le prototypage d'algorithmes déjà développés par le LMNO, leur intégration dans cette plateforme sur diverses applications réelles (veille environnementale, séries temporelles, séquences génomiques, etc.), ainsi que leur promotion auprès d'entreprises régionales. Il s'agit d'algorithmes de classification non-supervisée à base de modèles à variables latentes pour différents types de données de grande dimension. La deuxième étape consiste à contribuer à l'intégration à la plateforme des algorithmes développés dans le cadre du projet. Cette seconde tâche, plus fondamentale, portera en particulier sur le passage à grande- échelle de modèles de clustering distribués. Pour cela on s'appuyera sur la théorie du ré-echantillonnange (bootstrap) pour inférer un modèle à variables latentes (e.g. mélange de lois) à partir d'un gros volume de données, pour lequel paralléliser le calcul est une façon naturelle de s'y prendre

surtout pour des données traitées en mode batch. Les questions à traiter dans ce contexte sont principalement *i)* l'obtention de garanties et de nouvelles stratégies d'agrégation d'estimateurs, i.e comment obtenir à moindre perte un estimateur comme une agregation de plusieurs estimateurs issus d'échantillons bootstrap, et *ii)* traiter le problème de sélection de modèles qui dans ce cas distribué consiste à agréger des critères de sélection, construits à partir de petits sous-échantillons pour avoir des pseudo-critères de gros échantillons.
Les principales missions techniques sont : (i) Prototypage d'algorithmes d'apprentissage non-supervisé (ii) Cloud computing distribué haute performance (iii) Intégration et interfaçage web.

- *Profil recherché* :
  - Être titulaire d'un doctorat en mathématiques appliquées ou en informatique avec une spécialisation confirmée en apprentissage statistique non-supervisé et analyse de données complexes
  - Avoir une expérience en apprentissage de modèles à variables latentes sur des données réelles massives
  - Avoir un goût pour les applications et une expérience dans le prototype de code et l'intégration logicielle
  - Maîtriser la programmation Matlab/R/Python et les environnements big data (Hadoop/Spark, MapReduce)
  - Compétences souhaitées: cloud computing, systèmes OLAP, technos web
- *Laboratoire d'accueil* : UMR CNRS LMNO

Offre 4 : Ingénieur de Recherche (IGR) 14 mois :

L'un des objectifs majeurs du projet est la création d'une plateforme scientifique et technique ouverte et à visibilité internationale autour de l'apprentissage non-supervisé sur masses de données (BigData). Celle-ci proposera une architecture complète (pré-traitement, représentation, classification/catégorisation, visualisation) et considérera au moins les aspects suivants des données en masse : données de grande dimension, de gros volume, hétérogènes, non-annotées, dynamiques. Elle proposera des algorithmes originaux de traitement, d'analyse et d'exploitation de données hétérogènes de différents types (continues, longitudinales/fonctionnelles, discrètes, signaux/images) avec un calcul déporté et de haute performance (le CRIANN sera sollicité dans le cadre du projet pour avoir des ressources de calcul distribué haute performance). Elle proposera trois niveaux d'accessibilité (i) entreprise (ii) intermédiaire (e.g. étudiant) et (iii) chercheur, selon les besoins respectifs.

La personne recrutée participera en collaboration avec le reste du personnel du projet à la création de cette plateforme selon les étapes suivantes. La première concerne le prototypage d'algorithmes déjà développés par le LMRS, leur intégration dans cette plateforme sur diverses applications réelles, ainsi que leur promotion auprès d'entreprises régionales. Il s'agit d'algorithmes de classification non-supervisée à base de modèles à variables latentes pour des données discrètes (e.g. génomiques) et de traitement de données fonctionnelles par des méthodes non-paramétriques. La deuxième étape consiste à participer à intégrer les algorithmes qui seront développés dans le cadre du projet à la plateforme. Cette seconde tâche sera menée en collaboration principalement avec le LMNO sur les modèles de mélanges distribués régularisés avec des applications en veille environnementale/séquences génomiques.

Les principales missions techniques sont : (i) Prototypage d'algorithmes d'apprentissage non-supervisé (ii) Cloud computing distribué haute performance (iii) Intégration et interfaçage web.

- *Profil recherché* :
  - Être titulaire d'un doctorat en mathématiques appliquées ou en informatique avec une spécialisation confirmée en apprentissage statistique non-supervisé et analyse de données complexes
  - Avoir une expérience en statistique non-paramétrique et modèles à variables latentes sur données massives
  - Avoir un goût pour les applications et une expérience dans le prototype de code et l'intégration logicielle
  - Maîtriser la programmation Matlab/R/Python et les environnements big data (Hadoop/Spark, MapReduce)
  - Compétences souhaitées: cloud computing, systèmes OLAP, technos web
- *Laboratoire d'accueil* : UMR CNRS LMRS

En plus des missions décrites plus haut, les personnes recrutées participeront aux tâches suivantes :

- Travail collaboratif avec les différents membres du projet
- Rédaction de rapports et articles de recherche
- Manifestations scientifiques organisées autour du thème du projet

Informations complémentaires :

- *Date de commencement prévue* : début 2018
- *Rémunération* : environ 2100 euros net/mois
- *Contact* : faicel.chamroukhi@unicaen.fr ; Envoyez votre CV complet en précisant l'offre qui vous intéresse.

Quelques références en lien avec le projet :

F. Chamroukhi, (2017) “Skew *t* mixture of experts”, *Neurocomputing*, V266, pp. 390-408.

F. Chamroukhi, (2016) “Unsupervised learning of regression mixture models with unknown number of components”, *Journal of Statistical Computation and Simulation*, V86.12, pp. 2308-2334.

F. Chamroukhi, (2016) “Robust Mixture of Experts modeling using the *t* distribution”, *Neural Networks*, V79, pp 20-36.

F. Chamroukhi, (2016) “Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation”, *Journal of Classification*, 33(3):374–411.

F. Chamroukhi, H. Glotin & A. Samé (2013) “Model-based functional mixture discriminant analysis with hidden process regression for curve classification”, *Neurocomputing*, 112:153-163.

F. Chamroukhi, S. Mohammed, D. Trabelsi, L. Oukhellou, Y. Amirat, (2013) “Joint segmentation of multivariate time series with hidden process regression for human activity recognition”, *Neurocomputing*, 120: 633-644.

J.-L Starck, F. Murtagh, M.J. Fadili, (2016) “*Sparse Image and Signal Processing: Wavelets and Related Geometric Multiscale Analysis*”, (2nd Edition, ed.), Cambridge University Press, Cambridge, ISBN 9781107088061.

C. Chesneau, M.J. Fadili, B. Maillot, (2015) “Adaptive estimation of an additive regression function from weakly dependent data”, *J. of Multivariate Analysis*, V.133.1, pp. 77-94.

C. Bérard, M. Seifert, T. Mary-Huard, M-L. Martin-Magniette, (2013) “MultiChIPmixHMM : an R package for ChIP-chip data analysis modeling spatial dependencies and multiple replicates”. *BMC Bioinformatics*, 14 :271.

S. Volant, C. Bérard, M-L. Martin-Magniette, S. Robin, (2014) “Hidden Markov Models with mixture as emission distribution”. *Statistics and Computing*, 24(4):493–504.

G. Chagny, Roche, A. (2015) “Adaptive estimation in the functional nonparametric regression model”, *J. of Multiv. analysis*. V146, pp. 105-118.

G. Chagny, C. Lacour, (2015) “Optimal adaptive estimation of the relative density”, *TEST* 24(3) : 605-631.

G. Chagny (2015) “Adaptive warped kernel estimators”, *Scandinavian Journal of statistics*, 42(2) : 336-360.

A. Channarond, J.-J. Daudin, S. Robin, (2012) “Classification and estimation in the Stochastic Blockmodel based on the empirical degrees”, *Electronic Journal of Statistics*, 6 : 2574-2601

G. Chagny, A.Roche, (2014) “Adaptive and minimax estimation of the cumulative distribution function given a functional covariate”. *Electronic Journal of Statistics*, 8 : 2352-2404.

A. Patel, T. Nguyen, R. Baraniuk (2016) “A Probabilistic Framework for Deep Learning”. In NIPS, Barcelona.

A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan (2014) “A scalable bootstrap for massive data”. *JRSS B*, 76(4):795-816.

D. Witten, and R. Tibshirani, (2010) “A framework for feature selection in clustering”. *Journal of the American Statistical Association*, 105(490):713-726.

L. Ruan, M. Yuan., H. Zou (2011) “Regularized parameter estimation in high-dimensional Gaussian mixture models”. *Neural Computation*, 23:1605-1622.

G. Celeux, M.-L. Martin-Magniette, C. Maugis, A.E. Raftery, (2011) Letter to the editor: “A framework for feature selection in clustering”. *Journal of the American Statistical Association*, 106:383.

E. Devijver (2015a) “An l1-oracle inequality for the Lasso in finite mixture of multivariate Gaussian regression models”. *ESAIM:PS*19. 649-670.

E. Devijver (2015b) “Finite mixture regression: a sparse variable selection by model selection for clustering”, *Electronic Journal of Statistics* 9(2), pp. 2642-2674.

D. Karlis, L. Meligkotsidou, (2007) "Finite mixtures of multivariate Poisson distributions with application". *Journal of Statistical Planning and Inference*, 137(6), pp. 1942-1960.

P. Shi, E.A. Valdez, E. A. (2014) "Multivariate negative binomial models for insurance claim counts". *Insurance: Maths. & Economics*, 55, pp. 18-29.

T. Benaglia, Chauveau, D. Hunter, D. R. (2009) "An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures", *Journal of Computational and Graphical Statistics*, 18(2), pp. 505-526.

A. Goldenshluger, 0. Lepski, (2011) "Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality". *The Annals of Statistics*, 39(3), pp. 1608-1632.

C. Lacour, P. Massart, and V. Rivoirard, (2016). Estimator selection: a new method with applications to kernel density estimation.