TELECOM ParisTech

Une école de l'IMT

CHAIRE MACHINE LEARNING FOR BIG DATA

PROGRAMME

Journée de la Chaire

19 avril 2018

BNP PARIBAS La banque d'un monde qui change

criteo

PSA GROUPE

SAFRAN AEROSPACE · DEFENCE · SECURITY

Valeo SMART TECHNOLOGY FOR SMARTER CARS

DÉROULEMENT DE LA JOURNÉE

MATIN

09h00 - *Accueil petit-déjeuner - Salle E200*

09h30 - Introduction à la séance du matin par Stephan Clémençon, Porteur de la Chaire Machine Learning for Big Data

09h45 - **A Multivariate Extreme Value Theory Approach to Anomaly Clustering and Visualization**
Par Maël Chiapino, LTCI, Télécom ParisTech

10h15 - **JOFER : joint output fisher embedding**
Par Moussab Djerrab, LTCI, Télécom ParisTech

10h45 - **Sélection d'Hyperparamètre par Safe GridSearch : Complexité et Garantie**
Par Eugène Ndiaye, LTCI, Télécom ParisTech

11h15 - *Pause Café*

11h30 - **Ranking Median Regression: Learning to Order through Local Consensus**
Par Anna Korba, LTCI, Télécom ParisTech

12h00 - **Infinite task learning in vector-valued RKHS**
Par Alex Lambert, LTCI, Télécom ParisTech

12h30 - Conclusion par Pascale Massot, Directrice de l'Incubateur ParisTech Entrepreneurs

12h35 - *Déjeuner buffet - Hall Barrault*

Exposition des start up de l'**Incubateur ParisTech Entrepreneur**

Askhub - Beyable - Invenis - Softlaw
Aizimov - Adrock.Tv - .OGO Security

DÉROULEMENT DE LA JOURNÉE

APRÈS-MIDI

14h00 - Introduction à la séance de l'après-midi par François Roueff, Professeur à Télécom-ParisTech

14h15 - **Learning Data Representations through Kernel Autoencoders**
Par Pierre Laforgue, LTCI, Télécom ParisTech

14h45 - **Bandits manchots et risque de défaut**
Par Mastane Achab, LTCI, Télécom ParisTech

15h15 - **Une théorie probabiliste de l'apprentissage supervisé de similarité pour l'optimisation en un point de la courbe ROC**
Par Robin Vogel, LTCI, Télécom ParisTech

15h45 - *Pause Café*

16h15 - **Variable selection in multivariate linear model taking into account the dependance**
Par Marie Perrot-Dockes, MIA, AgroParisTech

16h45 - **Variational inference of Stochastic Block Model from sampled data**
Par Timothée Tabouy, MIA, AgroParisTech

17h15 - Conclusion par Florence d'Alché-Buc, Professeur à Télécom ParisTech

Learning Data Representations through Kernel Autoencoders

Par Pierre Laforgue, LTCI, Télécom ParisTech

Quelle que soit la tâche de machine learning envisagée, le choix des variables utilisées pour décrire les observations joue un rôle crucial dans la performance des algorithmes. Une première réponse à cette problématique est le feature engineering. Néanmoins, ce processus peut s'avérer long, et il requiert de nombreuses interactions avec lesdits experts. À l'inverse, le Representation Learning est le domaine du machine learning qui vise à construire ces variables d'intérêt de manière automatique et non-supervisée. Ce travail propose un nouvel algorithme de representation learning, inspiré de l'Autoencoder, mais dont les fonctions élémentaires appartiennent à des Espaces de Hilbert à Noyau Reproduisant à Valeurs Vectorielles. Ce nouvel algorithme, appelé Kernel Autoencoder (KAE) est résoluble grâce à un Representer Theorem énoncé et démontré spécialement pour ce problème. Enfin, la flexibilité des fonctions à noyaux nous permet d'élargir le champ d'application des autoencoders à tout type de données, là où les réseaux classiques se restreignent à R^d.

JOFER : joint output fisher embedding

Par Moussab Djerrab, LTCI, Télécom ParisTech

In recent years several methods to build data continuous representations have been developed. However these latter are always trained on large data corpus which is a problem when we want to achieve specific tasks. Indeed a representation on a specific corpus will lead to better performances for supervised learning task. Getting the right representation while avoiding the training of deep learning network on huge amounts of data is a key problem. In this work we propose a novel Framework (JOFER) that enables to solve classification problems while using existing pre-trained representations. This algorithm tries to leverage local distribution in order to bend the global representation so as to improve classification performances.

Sélection d'Hyperparamètre par Safe GridSearch : Complexité et Garantie

Par Eugène Ndiaye, LTCI, Télécom ParisTech

De nombreux estimateurs d'apprentissage statistique, tels que le Lasso ou la régression logistique, impliquent des paramètres de régularisation qui peuvent être difficiles à calibrer. Les stratégies habituelles consistent à calculer, par une recherche exhaustive sur une grille, une approximation (discrétisée) de l'ensemble des solutions. Malheureusement, la complexité, c'est-à-dire la taille de la grille discrétisée, peut être exponentielle en la dimension du problème dans le pire des cas. Nous revisitons, dans un cadre unifié, les techniques d'approximation de chemin de régularisation pour une tolérance prescrite, et nous montrons que sa complexité est de l'ordre de $O(1 / \sqrt[d]{\epsilon})$ pour les fonctions de perte Uniformément Convexe d'ordre $d>0$ et $O(1/\sqrt {\epsilon}) $ pour les fonctions Auto-Concordantes Généralisées. Cela inclut des exemples classiques tels que les moindres carrés, mais aussi la régression logistique qui, à notre connaissance, n'a pas été traitée par les travaux antérieurs. Enfin, nous utilisons notre technique pour fournir des bornes fines sur l'erreur de validation et des algorithmes pratiques pour la sélection des hyperparamètres avec une garantie de convergence globale pour toute précision ϵ_v fixée sur l'ensemble de test.

Ranking Median Regression: Learning to Order through Local Consensus

Par Anna Korba, LTCI, Télécom ParisTech

Le but de ces travaux est de prédire les préférences d'un individu sur un catalogue d'objets (indexés par {1, . . . , n}), étant donné des variables explicatives, par exemple des caractéristiques de cet individu (ex:sexe, âge, catégorie socio-professionnelle). Ces préférences peuvent prendre la forme d'une liste ordonnée ou classement («ranking») sur les objets (l'individu classe ces objets par ordre de préférence, du plus apprécié au moins apprécié). Pour ce problème, nous développons des méthodes locales d'apprentissages (k plus proches voisins et arbres de décision) adaptés à ces labels ordonnés (les classements). Ces méthodes construisent une partition de l'espace des variables explicatives («features») sur la base des données d'apprentissage. Dans le cas de la classification ou de la régression, pour une nouvelle données (ex: un nouvel individu), ces méthodes locales lui attribuent soit le label majoritaire (en classification) soit la valeur moyenne (en régression) des données labellisées appartenant à la même cellule de la partition. Dans notre cas, l'agrégation de labels (ici des classements), fait appel à des méthodes plus complexes. Heureusement, dans nos travaux précédents, qui seront rappelés dans cette présentation, nous avons étudié des méthodes efficaces d'agrégation de classements, qui nous permettent de justifier la validité des algorithmes proposés ici..

Infinite task learning in vector-valued RKHS

Par Alex Lambert, LTCI, Télécom ParisTech

Machine learning and statistics have witnessed the construction and the tremendous success of several efficient numerical techniques giving the entire solution path of parameterized problem families, such as LARS and parametric task learning. Unfortunately, these schemes typically have to be designed one-by-one, in a task-specific way. In this talk, we propose a generic approach, called Infinite Task Learning, to learn the solution paths in vector-valued reproducing kernel Hilbert spaces. We provide generalization guarantees to the suggested scheme and demonstrate its efficiency in cost-sensitive classification, conditional quantile regression and density level set estimation.

A Multivariate Extreme Value Theory Approach to Anomaly Clustering and Visualization

Par Maël Chiapino, LTCI, Télécom ParisTech

In a wide variety of situations, anomalies in the behaviour of a complex system, whose health is monitored through the observation of a random vector X = (X1, ..., Xd) valued in Rd, correspond to the simultaneous occurrence of extreme values for certain subgroups $\alpha \subset \{1, ..., d\}$ of variables Xj. Under the heavy-tail assumption, which is precisely appropriate for modelling these phenomena, a statistical method for identifying such events/subgroups has been recently developed in (Goix et al. 2016), relying on the concept of angular measure in multivariate extreme value theory, which characterizes the dependence structure of the Xj's in the extremes. It is the purpose of this paper to exploit this approach further, by means of a mixture model that permits to describe the distribution of extremal observations and where the anomaly type α is viewed as a latent variable. In particular, the model enables to assign to any such point X a posterior probability for each anomaly type α, defining implicitely a similarity measure between anomalies. A procedure based on the EM algorithm is also proposed here to infer the parameters of the mixture model from a (truncated) training dataset and it is explained at length how the corresponding posterior similarity measure estimates permit to obtain an informative planar representation of anomalies using standard graph-mining tools. The relevance and usefulness of the 2-d visual display thus designed is illustrated on real datasets, in the aeronautics application domain.

Bandits manchots et risque de défaut

Par Mastane Achab, LTCI, Télécom ParisTech

La performance de beaucoup de méthodes d'apprentissage statistique dépend du choix d'une métrique adéquate sur l'espace d'entrée. L'apprentissage de similarité (ou apprentissage de métrique) vise à construire une telle fonction à partir de données d'entraînement de manière à ce que les observations associées à la même (resp. à différentes) classe(s) soient aussi proches (resp. éloignées) que possible. Dans cet exposé, l'apprentissage de similarité est étudié en tant qu'ordonnancement biparti de paires d'observations, dont l'objectif est de ranger les éléments d'une base de données dans l'ordre décroissant de leur probabilité d'être dans la même classe qu'une donnée requête, en utilisant les scores de similarité. Un critère de performance naturel est alors l'optimisation en un point de la courbe ROC, qui consiste à maximiser le taux de vrais positifs sous un taux de faux positifs fixé. Nous étudions cette nouvelle perspective sur l'apprentissage de similarité avec une formulation probabiliste rigoureuse. La version empirique de ce problème induit une optimisation sous contrainte mettant en jeu des U-statistiques, pour lequel nous dérivons des vitesses d'apprentissage universelles ainsi que des vitesses rapides sous une hypothèse de bruit sur la distribution des données. Nous adressons aussi le problème de mise à l'échelle en analysant l'effet d'approximations basées sur des méthodes d'échantillonnage. Nos résultats théoriques sont illustrés par des expériences numériques.

Une théorie probabiliste de l'apprentissage supervisé de similarité pour l'optimisation en un point de la courbe ROC.

Par Robin Vogel, LTCI, Télécom ParisTech

La performance de beaucoup de méthodes d'apprentissage statistique dépend du choix d'une métrique adéquate sur l'espace d'entrée. L'apprentissage de similarité (ou apprentissage de métrique) vise à construire une telle fonction à partir de données d'entraînement de manière à ce que les observations associées à la même (resp. à différentes) classe(s) soient aussi proches (resp. éloignées) que possible. Dans cet exposé, l'apprentissage de similarité est étudié en tant qu'ordonnancement biparti de paires d'observations, dont l'objectif est de ranger les éléments d'une base de données dans l'ordre décroissant de leur probabilité d'être dans la même classe qu'une donnée requête, en utilisant les scores de similarité. Un critère de performance naturel est alors l'optimisation en un point de la courbe ROC, qui consiste à maximiser le taux de vrais positifs sous un taux de faux positifs fixé. Nous étudions cette nouvelle perspective sur l'apprentissage de similarité avec une formulation probabiliste rigoureuse. La version empirique de ce problème induit une optimisation sous contrainte mettant en jeu des U-statistiques, pour lequel nous dérivons des vitesses d'apprentissage universelles ainsi que des vitesses rapides sous une hypothèse de bruit sur la distribution des données. Nous adressons aussi le problème de mise à l'échelle en analysant l'effet d'approximations basées sur des méthodes d'échantillonnage. Nos résultats théoriques sont illustrés par des expériences numériques.

Variable selection in multivariate linear model taking into account the dependance

Par Marie Perrot-Dockes, MIA, AgroParisTechh

Dans cette présentation, nous proposons une nouvelle méthode de sélection de variables dans le cadre du modèle linéaire multivarié qui prend en compte la dépendance potentielle entre les réponses. Nous proposons d'estimer dans un premier temps la matrice de covariance des réponses puis d'utiliser cet estimateur dans un critère Lasso afin d'avoir un estimateur parcimonieux de la matrice des coefficients. Nous étudions les propriétés théoriques et numériques de notre estimateur. Plus précisément, nous donnons des conditions que l'estimateur de la matrice de covariance et son inverse doivent satisfaire afin que les positions non nulles de la matrice des coefficients soient retrouvées quand la taille de Σ n'est pas fixe et peut tendre vers l'infini. Notre approche est implémentée dans le package R MultiVarSel disponible sur le CRAN (*The Comprehensive R Archive Network*). Nous étudions ensuite les performances de notre approche en la comparant à d'autre méthodes sur des jeux des données simulés. Ces simulations montrent que la prise compte de la dépendance entre réponses dans le critère Lasso permet d'améliorer drastiquement les performances dans la plupart des cas.

Variational inference of Stochastic Block Model from sampled data

Par Timothée Tabouy, MIA, AgroParisTech

Here we deal with non-observed dyads during the sampling of a network and consecutive issues in the Stochastic Block Model (SBM) inference. We review sampling designs and recover Missing At Random (MAR) and Not Missing At Random (NMAR) conditions for SBM. We introduce several variants of the variational EM (VEM) algorithm for inferring the SBM under various sampling designs (MAR and NMAR) all available as an \texttt{R} package on github at \url{this https URL}. Model selection criteria based on Integrated Classification Likelihood (ICL) are derived for selecting both the number of blocks and the sampling design. We investigate the accuracy and the range of applicability of these algorithms with simulations. We finally explore a real-world networks from biology (protein-protein interaction network), where the interpretations considerably depends on the sampling designs considered.

START-UP EXPOSANTES

AskHub est une plateforme SaaS de Bot in the Bot fournissant aux éditeurs de chatbots des plugins conversationnels intelligents pré-entrainés sur étagère.

Beyable est une plateforme SaaS de génération de prospects sur un site web via une meilleure connaissance des visiteurs, en particulier les visiteurs anonymes.

Invenis est un logiciel d'analyse Big Data permettant le traitement de données massives sans aucune connaissance technique, à travers une interface graphique très conviviale et simple d'emploi.

Softlaw développe un logiciel d'aide à la revue, à l'analyse et à l'exploitation de documents juridiques et administratifs à destination des professionnels du Droit et des services juridiques et administratifs d'entreprises de toutes tailles.

Aizimov est un Assistant IA qui rédige les meilleurs emails prospectifs qui soient pour que nos clients puissent se concentrer sur l'essentiel: gagner de nouveaux projets.

AdRockTv est une startup technologique spécialisée dans la publicité embarquée dans l'image du contenu éditorial.

Ogo Security propose des solutions de Cybersécurité des sites internet et applications web a base d'intelligence Artificielle, pour 9,90 € par mois.

TELECOM ParisTech

Une école de l'IMT

CHAIRE MACHINE LEARNING FOR BIG DATA

machinelearningforbigdata.telecom-paristech.fr/fr/

Réseau EDUROAM
login : telecom804190-4903
mot de passe : LnLHVhc8

Avril 2018 - Journée de la Chaire Machine Learning for Big Data - Conception graphique : Pôle Communication Scientifique - Direction de la recherche - Contact : Samuel Pujol