Journées MAS 2010, Bordeaux

Apprentissage statistique

Session organisée par **Madalina Olteanu**

La complexification des données que les statisticiens sont amenés à analyser requiert le développement de nouvelles techniques et modèles. Dans la pratique, les données peuvent sortir du cadre euclidien ou la frontière entre le cadre supervisé et non supervisé peut devenir trop rigide quand les données sont labellisées de manière imprécise ou incertaine. Les méthodes "traditionnelles" d'apprentissage statistique sont alors insuffisantes ou incomplètes.

Le but de cette session est de présenter de nouveaux résultats théoriques et des applications en apprentissage statistique et classification en passant en revue, dans l'ensemble des présentations, le cadre supervisé, non supervisé, mais aussi semi ou partiellement supervisé.

Adresse de l'organisateur :

Madalina Olteanu
SAMM, Université Paris 1
90 Rue de Tolbiac
75013 Paris, France
E-mail : madalina.olteanu@univ-paris1.fr
<http://samm.univ-paris1.fr/OLTEANU-Madalina>

Journées MAS 2010, Bordeaux

Session : Apprentissage statistique

Topographic vector quantization in non-Euclidean spaces

par **Thomas Villmann**

Unsupervised and supervised vector quantization is frequently realized as prototype based approaches, i.e. prototypes represent data, data classes or borders between them. The vector quantization task can be considered under restrictions like visualization abilities, hierarchical data information or other requirements depending on task. Powerful methods origin from biologically motivated models for prototype adaptation usually denoted as learning in this context. Famous examples are the self-organizing map (SOM) and the neural gas (NG) as unsupervised models or the family of learning vector quantizers (LVQ) for supervised problems (classification). Usually, the methods base on the Euclidean space for the data and the Euclidean distance as the dissimilarity measure between the objects. However, this assumption is not realistic in many applications. For example, molecule structures in bioinformatics, sentences in text processing maybe inadequate to capture dissimilarity proper ! ties such that more sophisticated dissimilarity measures are required for data processing. Thus, the models developed under the Euclidean assumptions need to be extended for those problems.

The talk will give an overview about recent trends and developments for neural vector quantization approaches for non-Euclidean data. It will be shown, how the existing models can be redefined such that a respective data processing is possible preserving the biologically motivated idea of the algorithms. In particular we will focus on self-organizing maps and variants thereof as a widely ranged model for topographic vector quantization, i.e. a similarity preserving vector quantization approach. The basic properties of the model are discussed in the light of non-Euclidean assumptions. Exemplary applications will illustrate the usability of the approaches.

Adresse :

Thomas Villmann
Dep. of Mathematics, Natural and Computer Sciences
University of Applied Sciences Mittweida
Technikumplatz 17
09648 Mittweida / Germany
E-mail : thomas.villmann@hs-mittweida.de
<http://www.mathematik-mittweida.de>

Journées MAS 2010, Bordeaux

Session : Apprentissage statistique

Classification de données labellisées de manière imprécise et incertaine

par **Etienne Côme**

Cette intervention présente une solution aux problèmes de classification faisant intervenir des données d'apprentissage labellisées de manière imprécise et incertaine. Tous les exemples utilisés pour l'apprentissage sont donc décrits classiquement par un vecteur de descripteurs x_i , mais aussi par une étiquette imprécise et/ou incertaine m_i, spécifiant de manière douce l'appartenance de l'individu aux différentes classes d'intérêts. Nous proposons d'utiliser comme étiquette une fonction de masse de croyance. Le problème d'apprentissage est donc traité dans le cadre de la théorie des fonctions de croyance, ce qui permet une grande souplesse dans la définition des étiquettes. Cette approche généralise ainsi différents problèmes d'apprentissage couramment rencontrés dans la littérature : *supervisé, non-supervisé, semi-supervisé, partiellement supervisé*, que nous présenterons successivement. Pour résoudre le problème de classification, la solution décrite fait l'hypothèse que les données sont générées suivant un modèle de mélange. Nous montrerons que dans ce cadre il est possible de dériver grâce au théorème de Bayes généralisé un critère qui étend les critères de maximum de vraisemblance associés aux différents problèmes d'apprentissages sus-mentionnés. Nous présenterons également un algorithme de type *EM* dédié à l'optimisation de ce critère et permettant d'obtenir une estimation des différents paramètres du modèle. Finalement, nous donnerons des résultats de simulations mettant en évidence l'intérêt d'une telle approche, en particulier pour traiter élégamment les problèmes résultant d'erreurs d'étiquetages.

Adresse :
Etienne Côme
SAMM, Université Paris 1
90 Rue de Tolbiac
75013 Paris, France
E-mail : Etienne.Come@univ-paris1.fr
<http://samm.univ-paris1.fr/COME-Etienne>

Journées MAS 2010, Bordeaux

Session : Apprentissage statistique

Sélection de modèle pour la classification non supervisée

par **Jean-Patrick Baudry**

Nous rappelons les bases de l'approche de la classification non supervisée par les modèles de mélange. La méthode usuelle repose sur le maximum de vraisemblance et le choix du nombre de classes à former se fait par des critères pénalisés. Nous nous intéressons particulièrement au critère ICL (Biernacki, Celeux et Govaert, 2000), mis au point pour tenir compte de l'objectif de classification et pertinent en pratique. L'étude que nous proposons de ce critère et de la notion de classe sous-jacente repose sur l'introduction d'un cadre de minimisation d'un contraste adapté à ce contexte. Ce faisant nous définissons un nouvel estimateur et une nouvelle famille de critères de sélection de modèles dont nous étudions les propriétés – notamment la consistance. La calibration de ces critères peut se faire par l'heuristique de pente (Birgé et Massart, 2006). Divers aspects pratiques de leur mise en œuvre sont discutés et leur compo ! rtement pratique illustré par des simulations.

Adresse :
Jean-Patrick Baudry
Laboratoire de Mathématiques, Université Paris Sud
Bâtiment 430, bureau 25
91405 Orsay Cedex, France
E-mail : jean-patrick.baudry@math.u-psud.fr
<www.math.u-psud.fr/~baudry>

Journées MAS 2010, Bordeaux

Session : Apprentissage statistique

Méthode bayésienne de classification des séquences Barcode basée sur la coalescence

par **Nicolas Vergne**

Le DNA Barcoding a pour but d'assigner une espèce à un individu à partir de sa séquence d'ADN à un certain locus, généralement une partie du gene mitochondrial COI. Nous avons étudié plusieurs méthodes d'assignation, principalement des méthodes supervisées : k plus proches voisins (k-nn), CART, Random Forest. Nous avons ensuite élaboré une nouvelle méthode bayésienne de classification des séquences, basée sur la coalescence. Cette méthode, performante en terme de sensibilité et de spécificité, a le double avantage de donner un pourcentage de fiabilité à une assignation et de permettre de combiner rigoureusement l'information venant de plusieurs gènes. Nous évaluons la performance de cette méthode dans des situations variées (temps de séparation entre deux espèces, taux de mutation, taille des échantillons) et en comparaison avec d'autres méthodes (par exemple 1-nn avec bagging).

Adresse :
Nicolas Vergne
INRA, Unité MIA
Domaine de Vilvert
78352 Jouy-en-Josas Cedex, France
E-mail : nicolas.vergne.pro@gmail.com
<http://w3.jouy.inra.fr/unites/miaj/public/perso/NicolasVergne.html>