Apprentissage statistique pour des données fonctionnelles

Encadrants : Christophe Denis[(1)]**, Charlotte Dion**[(2)]
[(1)] Université Paris Est Marne-la-Vallée, LAMA, 77454 Marne-la-Vallée Cedex 2
[(2)] Sorbonne Université, LPSM, 75005 Paris
charlotte.dion@upmc.fr, christophe.denis@u-pem.fr

Pré-requis : Estimation non-paramétrique, apprentissage statistique, équations différentielles stochastiques, bonne connaissance du logiciel R ou Python

Mots clés : Processus de diffusion, classification, estimation non-paramétrique.

Description du sujet de stage : La classification de trajectoires est un domaine important de recherche en Statistique, les récentes avancées technologiques (notament l'utilisation de capteurs) rendant très facile la collecte de ce type de données [6]. En particulier, les données fonctionnelles peuvent être modélisées par des processus de diffusion. Par exemple, la dynamique de la vitesse cellulaire en biologie [7] ou le prix d'une action financière [5] sont décrits généralement par des équations différentielles stochastiques.

Dans ce contexte, la mise en place d'une procédure de classification adaptée à ce modèle est un réel enjeu. En effet, des méthodes générales existent [2, 4]. Une procédure de classification basée sur cette modélisation a été proposée dans [3] dans le cas où les classes sont discriminées par le coefficient de dérive (*drift*). Plus précisément, des résultats ont été obtenu dans le cas où la fonction de dérive est supposée paramétrique. La méthode proposée s'appuie sur le principe de minimisation de risque empirique [8, 1].

L'objectif principal du stage sera d'étendre les résultats obtenus dans le cas non-paramétrique. La mise en place d'une telle procédure permettre de traiter des données de nature très différentes. Une partie importante du stage sera dédiée à l'implémentation de la méthode, puis à la comparaison avec des méthodes actuelles d'apprentissage de données fonctionnelles.

Objectifs

1. Écrire une procédure statistique répondant au problème.
2. Implémenter l'algorithme de classification proposé et évaluer sa performance d'un point de vue numérique.
3. Étudier le risque de mauvaise classification de cette nouvelle procédure d'un point de vue théorique.
4. Comparer avec les algorithmes classiques (k plus proches voisins, réseaux de neurones).

Selon les goûts de l'élève, l'accent sera mis sur le point 3 ou sur le point 4.

Références

[1] P. Bartlett, M. Jordan, and J McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473) :138–156, 2006.

[2] G. Biau, F. Bunea, and M. Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 51(6) :2163–2172, 2005.

[3] Christophe Denis, Charlotte Dion, and Miguel Martinez. Consistent procedures for multiclass classification of discrete diffusion paths. *Scandinavian Journal of Statistics, to appear*, 2019.

[4] Sébastien Gadat, Sébastien Gerchinovitz, and Clément Marteau. Optimal functional supervised classification with separation condition. *arXiv preprint arXiv :1801.03345*, 2018.

[5] Damien Lamberton and Bernard Lapeyre. *Introduction to stochastic calculus applied to finance*. Chapman and Hall/CRC, 2011.

[6] James O Ramsay and Bernard W Silverman. *Applied functional data analysis : methods and case studies*. Springer, 2007.

[7] P. Romanczuk, M. Bär, W. Ebeling, B. Lindner, and L. Schimansky-Geier. Active brownian particles. *The European Physical Journal Special Topics*, 202(1) :1–162, 2012.

[8] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct) :1225–1251, 2004.