# Sparsification de modèles de classification pour des problèmes à large-échelle

Encadrants :
- Massih-Reza Amini (UGA, LIG, Massih-Reza.Amini@imag.fr)
- Ioannis Partalas (Viseo, ioannis.partalas@viseo.com)

Résumé : In large-scale scenarios like for example, ad-click prediction, text and gene classification, much attention has been given to the deployment of linear models, mostly due to their simplicity and efficiency. In such scenarios, the vector representation of data is often sparse and the size of the feature space exceeds the size of the available training examples. Also, in several applications constraints in both space and test-time prediction may apply making cumbersome the maintenance of large models. For instance, for the DMOZ dataset of the LSHTC challenge which contains over 27,000 classes and over half a million of features a linear model would require approximately 124 Gb of memory [2]. Besides this, many features in such datasets are correlated or uninformative and can harm the performance of a predictive model. In this project, we aim to study a simple yet effective method for sparsifying a posteriori linear models for large-scale text classification. The objective is to maintain high performance while reducing the prediction time by producing very sparse models. This is especially important in real-case scenarios where one deploys predictive models in several machines across the network and constraints apply on the prediction time.

Bibliographie :

[1] Sparsification of Linear Models for Large-Scale Text Classification Simon Moura, Ioannis Partalas, Massih-Reza Amini, In Conférence sur l'Apprentissage Automatique, 2015.

[2] LSHTC: A Benchmark for Large-Scale Text Classification Ioannis Partalas, Aris Kosmopoulos, Nicolas Baskiotis, Thierry Artieres, George Paliouras, Eric Gaussier, Ion Androutsopoulos, Massih-Reza Amini, Patrick Galinari, In CoRR, volume abs/1503.08581, 2015.

[3] Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". Journal of the Royal Statistical Society, Series B 58 (1): 267?288.

Pré-requis : Cours de Statistique de 1A et 2A.

Nombre de groupes : 1 Nombre d'étudiants par groupe : 2 ou 3