Optimisation multi-noyaux pour la classification supervisée

 Proposition de stage - BAC+5

September 26, 2012

Sujet

Le sujet de ce stage est l'étude de l'intégration de données de natures diverses (graphe, données numériques, données textuelles, données qualitatives) pour réaliser des tâches de classification supervisée. Le stagiaire utilisera, pour cela, des méthodes dites "à noyaux" (type SVM, Boser et al, 1992) et combinera les différentes données par le biais d'une combinaison linéaire de noyaux (voir Lanckriet et al, 2004). La combinaison linéaire sera optimisée par des méthodes numériques, type descente de gradient (comme dans Rakotomamonjy et al., 2007). Les méthodes qui pourront être adaptées à ce cas sont AFD, SVM, nuées dynamiques.

Application : Les approches seront testées sur : 1/ des données simulées : pour valider les diverses méthodes, le stagiaire simulera des jeux de données et comparera plusieurs approches en terme de performance sur la tâche de prédiction. 2/ dans le cadre supervisé, la prédiction de liens dans un réseau. Les données proposées pour réaliser cette tâche sont issues d'un jeu de données public et les liens représentent des régulations gènes/cibles (problème similaire à celui présenté dans Bleakley et al., 2007).

Références

- Bleakley, K. and Biau, G. and Vert, J.P. (2007) Supervised reconstruction of biological networks with local models. Bioinformatics, 23(13), i57-i65.
- Boser, B. and Guyon, I. and Vapnik, V. (1992) A training algorithm for optimal margin classifiers. In Proceedings of 5th annual ACM Workshop on COLT. D. Haussler Editor, ACM Press, 144-152,
- Lanckriet, G.R.G. and Cristianini, N. and Bartlett, P. and El Ghaoui, L. and Jordan, M.I. (2004) Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research, 5, 27-72.
- Rakotomamonjy, A. and Bach, F. and Canu, S. and Grandvalet, Y. (2007) More efficiency in multiple kernel learning. In Proceedings of the 24 th International Conference on Machine Learning.

Pré-requis

D'un point de vue théorique, le stagiaire devra être familié de la problématique de la classification supervisée (l'algorithme SVM pourra être étudié au cours du stage) ainsi que des méthodes d'optimisation courantes (méthodes de descente de gradient, par exemple). Il devra également être familié de la programmation R (des connaissances en programmation C seraient un plus mais ne sont pas indispensables).

Informations pratiques

Niveau : M2 statistique ou modélisation mathématique ; les candidatures d'étudiants de M1 motivés seront également considérées.
Localisation du stage : INRA de Toulouse (Auzeville), Département MIA, Unité BIA.

Contacts : Nathalie Villa-Vialaneix (nathalie.villa@toulouse.inra.fr)
Christine Cierco-Ayrolles (christine.cierco@toulouse.inra.fr).