Méthodes à noyau, Applications en Bio-informatique

Master Recherche M2

"Mathématique, Vision, Apprentissage"
ENS Cachan, 2005/2006

Jean-Philippe Vert

Jean-Philippe.Vert@mines.org

Double objectif

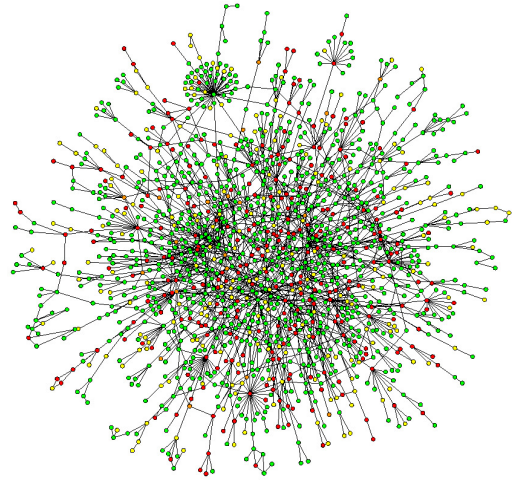
- Coeur mathématique: théorie des *noyaux positifs* et des *méthodes à noyau* en apprentissage statistique
- Exemples d'application traités: analyse de *données biologiques* (bio-informatique)

Noyaux et méthodes à noyau

- *Noyaux positif*: un moyen astucieux de *représenter des données* par une fonction de similarité $K(x, y)$
- *Méthodes à noyau*: des algorithmes puissants qui utilisent la représentation par noyau pour la classification, la régression (*Support Vector Machines*), la réduction de dimensionalité (*kernel-PCA*), la comparaison de données hétérogènes (*kernel-CCA*), la classification non supervisée (*kernel-clustering*), la séparation de signaux (*kernel-ICA*), etc...
- Applicables dans de nombreux domaines (images, textes, finance, ..., *bio-informatique*)

Bio-informatique

- *Des données*: génomes séquencés, structures 3D de protéines, réseaux d'interaction, expression des gènes...
- *Des problèmes*: fonctions, structure des protéines, prédiction de voies métaboliques, modélisation de systèmes biologiques, compréhension des pathologies (cancer, infection virale...)



Bio-informatique et noyaux

Les données de bio-informatique sont souvent:

- en *grande dimension* (ex: expression de genes)
- *structurées* (ex: séquences d'ADN, structures de molécules, graphes d'interaction...)
- *hétérogènes* (vecteurs, séquences, graphes...)

L'utilisation de noyaux est bien adaptée à ces cas, et fournit d'excellents résultats. Nécessite le développement d'outils spécialisés

Contenu du cours

- *Introduction* à la biologie moléculaire et à la génomique
- *Noyaux positifs*: définition, propriétés, espaces de Hilbert à noyau reproduisant, kernel trick, representer theorem
- *Méthodes à noyau*: kernel PCA, SVM, kernel logistic regression, kernel least square, kernel CCA
- *Noyaux*: pour séquences, pour graphes, noyau de diffusion, noyau de convolution, noyau de semi-groupe, noyaux marginalisés...
- *Applications*: classification de séquences, inférence sur des graphes, sélection de genes...

Organisation du cours

- $8 \times 2,5h$ le mardi de 10h à 12h30
- `http://cg.ensmp.fr/~vert/teaching/2005master`
  - slides
  - emploi du temps
  - références à télécharger
  - informations projet
- Un examen écrit (50%) et un projet (50%)
- Me contacter: `Jean-Philippe.Vert@mines.org`