Kernalytics - Méthodes à noyaux et modélisation

Vincent KUBICKI - InriaTech

Inria Lille - Nord Europe

4 août 2019

Kernalytics - Généralités

- Contrat Coreye.
- Preuve de concept pour utiliser les noyaux.
- Pallier aux difficultés liées au C de KernSeg.
- Scala et programmation fonctionelle.
- Gestion des données multivariées.

Segmentation à noyaux

- Adaptation directe de "New Efficient Algorithms for Multiple Change-Point Detection with kernels", de Celisse et al.
- Implémentation utilisant la programmation dynamique.
- Sélection du nombre optimal de segments via une heuristique de pente.

Modularité

Commun

- Matrice de Gram
- KerEval
- ...

Noyaux

- Linéaire
- Polynomial
- Gaussien
- Laplacien

Algorithmes

- Régression
- K-Means
- Test d'égalité de distributions
- Segmentation

Structures algébriques

- Tous les noyaux sont génériques, nécessitant un type de donnée et une structure algébrique.
- Exemple : la segmentation de matrice nécessite seulement un produit interne dans l'espace des matrices réelles (Frobenius par exemple).
- Déduction des structures algébriques (i.e. produit interne $\rightarrow$ norme $\rightarrow$ distance).
- Hiérarchie algébrique validée par typage.

Programmation fonctionnelle

Immutabilité

- Aucune variable.
- Aucune boucle.
- Généralisation des lapply, reduce, etc... de R

Composition de fonctions

- Combinaison linéaire de noyaux.
- Code est concis et de haut niveau.
- Simplicité d'intégration d'un nouveau noyau, sous la forme d'une fonction $(X, X) \to \mathbb{R}$
- Simplicité pour ajouter un nouveau type de données, sous la forme d'une fonction Chaîne de caractères $\to X$.

Génie logiciel

- Utilisation d'une d'algèbre linéaire / statistiques : Breeze.
- Code est compilé : beaucoup de vérifications sont effectuées lors de la compilation, et l'exécution est rapide.
- Tests unitaires très faciles à implémenter, gestion des exceptions limpide.
- Code compilé pour la JVM, facilité de déploiement multiplateforme.
- Maintenance plus simple que du C++ (i.e. système de build trivial).

Utilisation

Format des données

- 2 fichiers d'entrée en csv : données et descripteurs
- 1 colonne par variable en données, 1 colonne par noyau en descripteurs
- Plusieurs noyaux possibles pour une même variable, par exemple

Données

- Nom de variable
- Type de donnée
- Une observation par ligne
- ...

Descripteurs

- Nom de variable
- Poids
- Noyau et paramètres

A faire

- Créer un paquet R en utilisant rscala (wrapper léger, prototype déjà testé).
- Intégrer des noyaux plus avancés.