Statistique en grande dimension pour la génomique
Projets 2016-2017

L. Jacob, F. Picard, N. Pustelnik, V. Viallon

Contents

1 Graph Kernels for Molecular Structure-Activity Relationship Analysis with Support Vector Machines

Description. http://cbio.ensmp.fr/ jvert/svn/bibli/local/Mahe2005Graph.pdf

Les noyaux définis positifs sont des mesures de similarités entre objets. On peut montrer qu'un noyau entre deux objets est équivalent à un produit scalaire pris sur une certaine description (vectorielle ou fonctionnelle) de ces objets. Cette propriété fait des noyaux un outil utile pour appliquer des outils statistiques ne dépendant des données que via leurs produits scalaires à des objets complexes comme des polymères ou des molécules se prêtant difficilement à des descriptions vectorielles explicites. Ce papier présente un noyau pour molécules (vues comme des graphes), et son application à la prédiction de mutagénicité.

Travail demandé. (1/2) Oral 10 minutes + 5 min de questions. Expliquer le principe des noyaux définis positifs. Expliquer le noyau introduit dans l'article: principe et algorithme.

(2/2) Comparer une méthode d'apprentissage statistique de votre choix utilisant le noyau du papier à une méthode utilisant des descripteurs plus naifs (`http://www.predictive-toxicology.org/data/cpdb_mutagens/properties.tab`)

Données décrites ici: `http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=70A0E98B0D568D91665E197C59DA34EE?doi=10.1.1.411.7584&rep=rep1&type=pdf`

Disponibles ici: `http://www.predictive-toxicology.org/data/cpdb_mutagens/`

Code: `http://chemcpp.sourceforge.net/html/index.html`

Fichier. mahe.2005.pdf

Tuteur. L. Jacob

2 A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis

Description. Ce papier présente une famille de méthodes d'analyse exploratoire, qui maximisent certains critères statistiques (variance, covariance) sur des données. Ces maximisations sont faites sous certaines contraintes (sparsité, positivité), censées conduire à des résultats plus interprétables biologiquement.

Travail demandé. (1/2) Oral ∼10 minutes + 5 min de questions. Expliquer le principe de la CCA, et de la sparse CCA (section 4 du papier biostatistics). Principes statistiques, motivation biologique, algorithme. (2/2) Implémentation: sparse CCA sur données synthétiques (code dispo sur la page de D. Witten). Generer des données X, Y dont certaines combinaisons linéaires (éventuellement sparse) de colonnes sont corrélées. Etudier l'effet du nombre d'échantillons, de la dimension de X et Y, des contraintes sur le résultat obtenu.

Fichier. witten.2009.pdf

Tuteur. L. Jacob

3 Efficient RNA Isoform Identification and Quantification from RNA-Seq Data with Network Flows

Description. Ce papier présente un algorithme permettant de calculer efficacement un estimateur de l'expression des isoformes d'un gène dans des données RNA-Seq. Cet estimateur (de type maximum a posteriori) implique la maximisation d'une vraisemblance sous contraintes l1 dans un espace de très grande dimension. Le calcul de l'estimateur est rendu possible par l'équivalence de ce problème de maximisation comme un problème d'optimisation de flot sur un graphe.

Travail demandé. (1/2) Oral ∼10 minutes + 5 minutes de questions. Expliquer le problème biologique et l'estimateur proposé (loss de Poisson et contrainte l1). Montrer l'équivalence entre le problème de maximisation de la vraisemblance a posteriori et le problème de flot. (2/2) Implémentation: consignes précises à venir. Utiliser le package flipflop pour estimer l'ensemble d'isoformes présents dans les données qui seront fournies.

Fichier. besnard.2013.pdf

Tuteur. L. Jacob

4 On stepwise pattern recovery of the fused Lasso

Description. Cet article s'intéresse au modèle tronqué de suite gaussienne, qui peut être vu comme un cas particulier du modèle de régression, dans lequel $p = n$ et la matrice de design est la matrice identité d'ordre n. En supposant un signal constant par morceau, il est établi que le fused lasso ne permet en général pas la détection des sauts dans le signal, sauf

dans des cas très particuliers. Une modification du fused lasso est proposée (méthode de pré-conditionnement par une transformation de Puffer), qui permet la détection des sauts avec grande probabilité, sous des hypothèses raisonnables. Des résultats de simulation viennent illustrer les résultats théoriques.

Travail demandé. (1/2) Oral $\sim$ 10 minutes + 5 minutes de question. Présenter le modèle tronqué de suite gaussienne et le fused lasso. Réécrire le fused lasso comme un lasso sur une transformation des données. A partir de la condition d'irreprésentabilité pour le lasso vue en cours, en déduire des conditions nécessaires et suffisantes pour la détection des sauts dans le signal par le fused lasso. Présenter le principe de la transformation de Puffer et le résultat du Théorème 3. (2/2) Implémentation : implémenter le fused lasso, et la version préconditionnée, et illustrer leurs performances relatives sur des données simulées.

Fichier. QiangJiaFusedLassoCSDA2016.pdf

5 High-dimensional graphs and variable selection with the lasso

Description. Cet article s'intéresse au cadre de l'estimation de la structure dans les modèles graphiques gaussiens, qui permettent l'étude des relations d'indépendance conditionnelle entre les composantes d'un vecteur gaussien. Les auteurs proposent une méthode approchée pour estimer cette structure, qui repose sur l'utilisation de régressions linéaires pénalisées (Lasso). Des résultats théoriques sont obtenus, et la méthode est illustrée sur des données simulées.

Travail demandé. (1/2) Oral $\sim$10 minutes + 5 min de questions. Après avoir fait quelques rappels sur les vecteurs gaussiens (notamment le théorème de corrélation normale), présenter le principe général de la méthode (en particulier, expliquer pourquoi les indépendances conditionnelles correspondent aux zéros de la matrice de concentration et aux zéros dans les vecteurs de coefficients des régressions linéaires considérées.) Résumer les résultats théoriques du papier et les hypothèses sous lesquelles ils ont été obtenus (en particulier, les rapprocher des hypothèses vues en cours pour établir les propriétés du Lasso). (2/2) Implémentation: Implémenter la méthode et la comparer aux résultats du package glasso sur quelques exemples.

Fichier. meinshausenbuhlmann-2006.pdf

Tuteur. V. Viallon

6 Block coordinate descent algorithms for large-scale sparse multiclass classification

Description. Cet article s'intéresse au problème de sélection de "features" en classification. Les auteurs proposent une extension multiclasse de la fonction de coût "hinge" quadratique et la sélection de "features" s'effectue grâce à une pénalisation de type $\ell_{1,2}$. L'algorithme proposé est une variante de l'algorithme de descente par bloc de coordonnées qui conduit à de meilleures performances que celles aobtenues avec des algorithmes de l'état de l'art tels que FISTA.

Travail demandé. (1/2) Oral $\sim$ 10 minutes + 5 min de questions. Détailler le critère proposé par les auteurs et justifier le choix de chacun des termes. Présenter les deux variantes de l'algorithme de descente par bloc. Quelles sont les garanties de convergence de ces algorithmes ?
(2/2) Reprendre l'exemple étudié en TP (minimisation logistique + l1) et résoudre le problème associé avec les algorithmes proposés dans cet article. Commentez les résultats obtenus.

Fichier. blondel.2013.pdf

Tutrice. N. Pustelnik

---

7 Supervised Feature Selection in Graphs with Path Coding Penalties and Network Flows

Description. Description. Cet article s'intéresse au problème d'apprentissage supervisé en combinant l'hypothèse usuelle de parcimonie avec une hypothèse de structure de graphe. Plus précisément pour un graphe donné a priori sur les variables, les pénalités proposées conduisent à des estimateurs linéaires parcimonieux à l'échelle des groupes de variables, oùles groupes correspondent à des chemins sur le graphe. Autrement dit, le support des estimateurs obtenus en minimisant le risque empirique pénalisé de cette manière correspondra typiquement à un petit nombre de chemins sur le graphe des variables. Une application est proposée en pronostic de cancer du sein, oùl'on souhaite construire un estimateur n'impliquant qu'un petit nombre de composantes connexes sur le graphe décrivant les interactions connues entre les gènes.

Travail demandé. (1/2) Oral $\sim$ 10 minutes + 5 min de questions. Décrire l'intérêt de cette régularisation, sa formulation, sa relaxation convexe, l'opérateur

proximal associé, et l'algorithme utilisé dans le cas convexe.(2/2) Générer des données synthétiques à partir d'un modèle linéaire, où le support de la fonction linéaire correspond à différents nombres de composantes connexes sur un graphe donné. Utiliser le code disponible sur `http://spams-devel.gforge.inria.fr/` pour évaluer sur ces données l'impact de ce type de régularisation en fonction du nombre de composantes connexes. Évaluer également l'impact du paramètre λ sur les propriétés de l'estimateur.

Fichier. mairal.2013.pdf

Tutrice. N. Pustelnik

8 On the convergence of the iterates of "FISTA"

Description. Cet article s'intéresse à la converge des itérées de l'algorithme FISTA, proposé par Beck et Teboulle en 2009. FISTA permet de minimiser une somme de deux fonctions convexes, propres, semi-continues inférieurement dont l'une est de gradient Lipschitz. L'article de Beck et Teboulle s'intéressait principalement au taux de convergence FISTA et à sa convergence en terme de fonctionnelle. Dans l'article que nous vous proposons d'étudier, les auteurs montrent que la convergence des itérées peut être garantie si certaines hypothèses de FISTA sont modifiées. Les performances de l'algorithme sont étudiées sur une exemple de restauration d'images.

Travail demandé. (1/2) Oral $\sim$ 10 minutes + 5 min de questions.
Présenter les caractéristiques du critère pouvant être résolu par FISTA. Détailler les itérations de FISTA ainsi que les hypothèses originalement introduite par Beck et Teboulle pour assurer la convergence de la fonctionnelle. Détailler les modifications faites par Chambolle et Dossal dans les hypothèses de FISTA permettant d'assurer la convergence des itérées. Donner les grandes lignes de la preuve permettant de montrer la convergence des itérées. Pourquoi la convergence des itérées est-elle importante ? Détailler les avantages/inconvénients de FISTA/ISTA/Forward-Backward. (2/2) En utilisant l'exemple étudié en TP (minimisation logistique + l1), comparer l'algorithme forward-backward, FISTA et la version modifiée par les auteurs.

Fichier. chambolle.2014.pdf

Tutrice. N. Pustelnik

9 Controlling the Rate of GWAS False Discoveries

Description. Les auteurs de cet article s'intéressent à la détection de SNPs associés avec une réponse continue. La problématique abordée comporte deux aspects. Dans un premier temps sont exposées les difficultés liées au déséquilibre de liaison, et au fait que les SNPs causaux ne sont pas nécessairement ceux qui font partie du jeu de données. Dans un deuxième temps est abordée la question de la préselection de certains SNPs avant de procéder à l'analyse des liens entre SNPs et le trait d'intérêt. La thématique abordée ici est celle de l'inférence post-sélection (selective inference) qui consiste à effectuer les test sur un sous-ensemble d'hypothèses pré-sélectionnées. Cette thématique est en pleine expansion ces dernières années, et pose des problèmes de définition du risque à contrôler pour prendre en compte la première étape de sélection.

Travail demandé. (1/2) Oral $\sim$ 10 minutes + 5 min de questions. Vous présenterez la problématique de la dépendance entre SNPs voisins en introduisant la notion de déséquilibre de laison et ses conséquences sur les analyses GWAS. Vous étudierez la procédure de contrôle du FDR proposée par les auteurs, et vous intéresserez à sa démonstration. (2/2) Vous mettrez en place un schéma de simulation et vous étudierez les propriétés empiriques de cette méthode.

Fichier. `http://www.genetics.org/content/genetics/205/1/61.full.pdf`

Tuteur. F. Picard