Statistiques et Sciences des Données

Journée Big Data, SPC
30 novembre 2015

Tableau général

- Apport théorique et algorithmique des statistiques pour l'étude des grands jeux de données
  - Nature interdisciplinaire par essence, besoin de travailler avec un spécialiste des données
  - Thèmes actuels: biologie/médecine, astronomie, ville connectée, histoire, ...
  - Incorporation naturelle dans des demandes de financement
  - Besoins indirects en infrastructure
- Prise en compte des méthodes pour grands jeux de données dans les formations
  - DU Big Data à l'IUT Paris Descartes
  - Masters de Modélisation et Statistique de l'UFR Math-Info de Paris Descartes

Apprentissage statistique sur données de grande taille

Problématique : les données de grande taille obligent à adapter les méthodes pour tenir en compte la complexité algorithmique.

Apprentissage statistique:
- Classification supervisée
- Classification non supervisée
- Estimation de paramètres (modèle linéaire, points de rupture, ...)
- Prédiction

Big Data:
- Algorithmes linéaires ou quasi-linéaires
- Méthodes on-line

Exemple 1: Comparaison de graphes issus de métagénomique

Collaboration: Paris Diderot (informatique) et UPMC (biologie)

Type de données: Graphes de comparaison de courtes séquences, prélevés dans des estomacs de lézards de types différents.

Taille des données: 10 GB, 1 million de graphes, 100-1000 sommets par graphe

Problème appliqué: Caractériser les graphes qui se ressemblent (ou pas) pour comparer les bactéries présentes

Problème statistique: Classification d'un très grand nombre de données complexes

- Choix de résumés **pertinents et rapides à déterminer** pour les objets à classer
- Classification **efficace**

Principal défi: Les algorithmes standards sont trop lents pour les **données de grande taille**.

Apprentissage statistique en grande dimension (n«p)

Problématique : les données de grande dimension sont aujourd'hui omni-présentes dans les domaines applicatifs mais leur traitement reste complexe.

Classification supervisée et non supervisée:
- modèles Gaussiens HD pour le clustering et la classification,
- clustering discriminatif permettant la visualisation des clusters,

Estimation de la dimension intrinsèque :
- dans le cas du modèle probabilistic PCA (PPCA),

Sélection de variables en clustering et régression:
- clustering parcimonieux par pénalisation ℓ_1,
- régression grande dimension sparse.

Exemple 2: Recherche de marqueurs génétiques sanguins pré-diagnostic dans le cancer du sein

Collaboration: Université de Tromso (épidémiologie et informatique)

Type de données: Données RNA, miRNA, méthylation dans des cellules sanguines avant le diagnostic du cancer du sain

Taille des données: Plusieurs dizaines de milliers de relevés dans quelques centaines d'échantillons

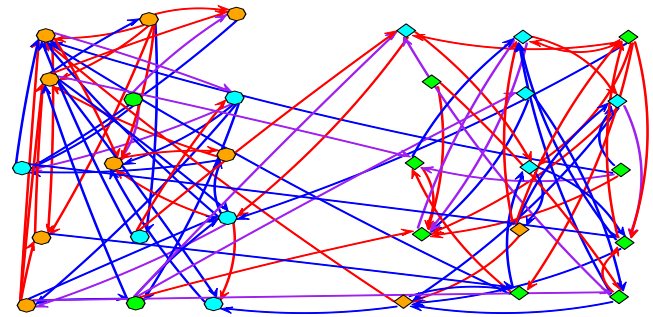
Problème appliqué: Déterminer les gènes/miRNA/zones méthylées dont les changements sont annonciateurs d'un diagnostic précis: screening/clinique et local/métastasé

Problème statistique:
- Sélection de variables
- Prédiction

Principal défi: Travailler en **grande dimension**: le nombre d'observations est bien inférieur au nombre de paramètres à inférer.

Apprentissage statistique sur données atypiques

Problématique: les données modernes sont de types variés (réseaux, fonctions, mixtes) et il est nécessaire de proposer des méthodes adaptées à cette variété de types.



Classification dans les réseaux:
- modèles supervisés et semi-supervisés d'espaces latents,
- détection de communautés non observées dans des réseaux,
- comparaison de sous-réseaux grâce à leur structure en clusters,

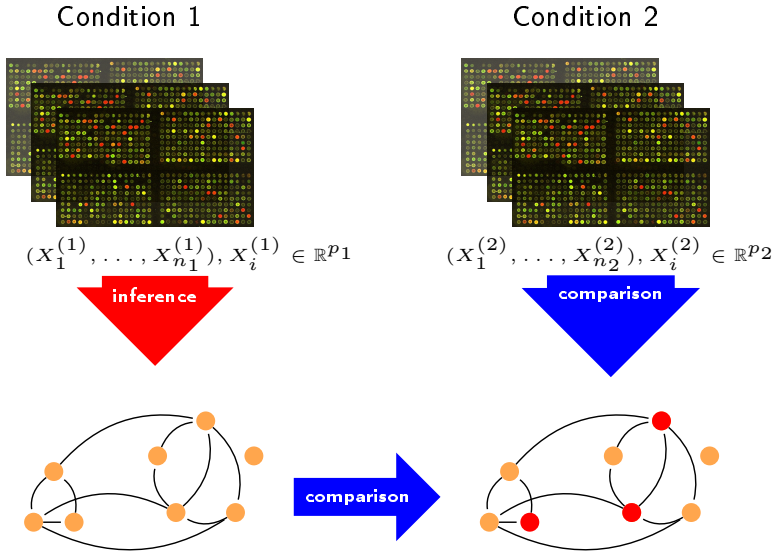
Classification de données fonctionnelles:
- modélisation et classification dans des sous-espaces fonctionnels,

Classification de données hétérogènes:
- modélisation par processus gaussiens parcimonieux et estimation au travers de fonctions noyaux.

Exemple 3: Inférence et perturbation de réseaux de régulation

Collaboration: Institut Curie (oncologie), Génopole et Université de Montpellier (bioinformatique).

Type de données: Expression des gènes/mutations/nombre de copies de l'ADN dans des cellules normales et tumorales de la vessie



Taille des données: Plusieurs dizaines de milliers de mesures dans quelques centaines d'échantillons

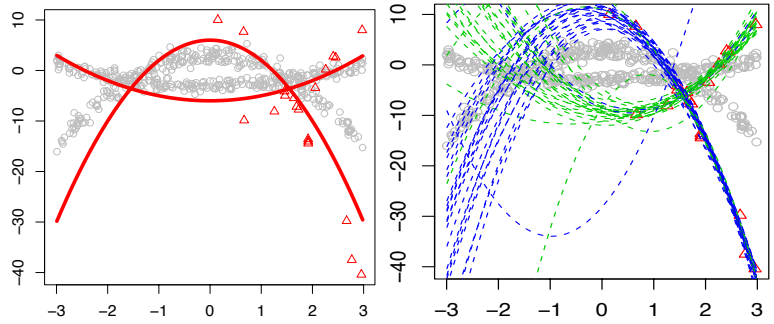
Problème appliqué: Déterminer les gènes dont la régulation est problématique dans les tumeurs et les régulateurs responsables

Problème statistique:
- Inférer un réseau de référence dans les cellules normales
- Définir un modèle de perturbation dans les cellules tumorales

Principal défi: Travailler en **grande dimension**, intégrer des **données hétérogènes**, objet organisés en **réseaux** .

Apprentissage statistique adaptatif

Problématique : dans de nombreuses situations pratiques, le phénomène à modéliser est évolutif (dans le temps) ou son observation est biaisée (labels incertains).



Modèles adaptatifs en régression :
- modèles paramétriques pour la régression adaptative,
- modèle bayésien pour les mélanges de régressions adaptatives,

Classification supervisée avec labels incertains :
- classification robuste par mélange de mélanges,
- classification robuste par modèle d'espace latent,

Classification supervisée avec classes non observées :
- méthodes de découverte de classes de nouveautés.