Journée d'automne de l'AFZ
Elevage de précision
Compte-rendu de discussions
Intervention de Céline Lévy-Leduc
Analyse de données en grande dimension

Question du public :

Dans votre premier exemple, ne peut-on pas poser une contrainte sur le nombre de ruptures à trouver ?

Céline Lévy-Leduc, AgroParisTech MMIP :

Dans le premier exemple développé, le nombre de sauts n'est pas connu à priori, on suspecte seulement leur présence. On utilise une méthode de sélection des modèles statistiques pour le définir. Le nombre et la position des ruptures sont donc fixés statistiquement, et sont optimaux grâce au modèle. Dans le second exemple, en revanche, le nombre de ruptures est connu d'avance grâce à la modélisation antérieure. Par exemple, on sait que la croissance se fait en trois phases. Le but est ici de travailler sur la qualité d'ajustement du modèle à la réalité, sur le moment où ont lieu ces ruptures.

Question du public :

Quel support est utilisé pour les bases de données et le stockage des données ? On atteint ici les limites d'Excel en la matière. Quels outils sont disponibles pour travailler sur les modèles statistiques ?

Céline Lévy-Leduc :

Le stockage se fait sur des fichiers .txt ou .zip, beaucoup moins lourds à manipuler. Pour le traitement statistique, le logiciel R est encore utilisé, notamment certains packages particuliers. A noter que ces packages sont écrits à la fois en R et en C++ voire en C, ce qui permet de gagner en rapidité de travail par un codage en langage peu élaboré. Par exemple, le package glmnet a été codé par des chercheurs en statistiques pour traiter beaucoup de données, il a donc été codé en C++ pour aller plus vite sur de grandes quantités d'informations.

Question du public :

Partir à l'aveugle dans du data mining assure-t-il un résultat ? Ne vaudrait-il pas mieux définir un objectif avant ? D'expérience, il faut une question précise et un modèle précis, sinon il est impossible de tomber sur des résultats valables.

Vous avez parlé de modèles avec beaucoup de p (paramètres) et peu de n (individus/ données). Ces modèles sont-ils utilisables avec beaucoup de n ?

Céline Lévy-Leduc :

Les modèles explicités servent si, en proportion, il y a beaucoup plus de variables que de données, par exemple des millions de variables pour quelques milliers d'observations. Il y a donc une notion de proportion à prendre en compte. Il faut par ailleurs bien noter que ces méthodes ont été développées car le modèle linéaire ne marche pas lorsque p est supérieur à n. Dans le cas ou n est très supérieur à p, les méthodes expliquées sont toujours applicables, mais dans une optique différente, en sélection de variables par exemple. On ne trouve cependant pas un optimum comme avec la méthode des moindres carrés, il faut prendre une décision un peu arbitraire.

Question du public :

Ces méthodes sont-elles sorties des laboratoires ? Sont-elles déjà utilisées en entreprise ?

Céline Lévy-Leduc :

Même s'il y a déjà eu des thèses CIFRE les utilisant, on ne sait pas si ces méthodes sont largement répandues en entreprise. Elles sont pour la plupart dans le domaine public, comme les packages R par exemple, et sont donc utilisables par tous et entièrement libres. Cependant, une discussion avec un spécialiste reste nécessaire pour voir si la méthode est pertinente. Les instituts de recherche les utilisent beaucoup, notamment l'Institut Curie et la recherche en santé de manière générale, en ce qui concerne le traitement du cancer.

Propos recueillis par Maximin Bonnet, François Gaudin et Luc Métayer,
étudiants à AgroParisTech (EDEN - Elevages et filières Durables Et iNnovants)