

Apprentissage Causal pour les Humanités Numériques

RESPONSABLES: Ph. Caillou, C. Furtlehner, I. Guyon, P. Tubaro

LABORATOIRE: Laboratoire de Recherche en Informatique CNRS UMR 8623 & INRIA Saclay

ADRESSE: Université Paris-Sud, 91405 Orsay Cedex

MAIL: Prenom.Nom@lri.fr

URL: <http://www.lri.fr/~sebag/Stages/Causal.pdf>

RÉSUMÉ :

La science des données peut contribuer à la recherche en sciences humaines et sociales (SHS, incluant notamment sociologie, économie, histoire) en proposant des modèles reflétant les données disponibles. Cependant, beaucoup de modèles peuvent être construits à partir de grandes données ; du point de vue de l'apprentissage, la question est de savoir si ces modèles sont de bonne qualité statistique et assez généraux, i.e. s'ils rendent compte d'autres données que celles qui ont été utilisées pour construire ces modèles.

Du point de vue SHS, la question est de savoir si les modèles appris par la science des données *ont du sens*. Cette notion, difficile à caractériser, sera considérée dans ce stage sous l'angle de la causalité (pour fixer les idées : l'altitude d'une ville *cause* sa température moyenne; sa température moyenne *ne cause pas* son altitude).

Objectif du stage

Considérons un ensemble de données. L'application motivant cette étude est la modélisation de la qualité de la vie au travail (données collectées par le Ministère du Travail, décrites par 350 variables – responsabilités, formation, stress, absentéisme, etc – recueillies pour 33 000 personnes – questionnaire anonymisé). Une partie des relations causales connues sera fournie au data scientifique par le sociologue. Se fondant sur ces relations, un modèle sera construit.

Nous cherchons un critère de falsifiabilité, permettant d'établir si la démarche suivie en sciences des données est scientifique du point de vue SHS, ou non. Le critère étudié ici sera donnée par la fraction des relations causales (connues du sociologue, mais non fournies au data scientifique) présentes dans le modèle construit. Plus cette fraction sera élevée, plus on en déduira que la démarche du data scientifique est scientifique au sens des SHS.

Approche

Considérons la construction d'un réseau Bayésien : un graphe orienté, dont les noeuds sont les variables, et les arcs indiquent une relation statistique (causalité ou corrélation entre les deux variables). L'objectif est de déterminer un graphe sans cycles, compatible avec les relations causales connues (e.g. Var_1 est indépendante de Var_2 conditionnellement à Var_3), qui approche au mieux la distribution empirique des données.

On se ramènera à un problème d'optimisation booléen sous contraintes.

Ce stage peut être traité sous l'angle théorique ou sous l'angle empirique. L'étude théorique demande de très bonnes capacités d'analyse. L'étude empirique demande d'excellentes capacités informatiques (programmation C++ ou Python, en partant de bibliothèques existantes).

Références

Sofia Triantafillou, Ioannis Tsamardinos: Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research* 16: 2147-2205 (2015)

James Cussens: Bayesian network learning by compiling to weighted MAX-SAT. *UAI* 2008: 105-112