

UNIVERSITE CLAUDE BERNARD – LYON I

DIPLÔME NATIONAL DE DOCTORAT (Arrêté du 25 mai 2016)

Date de la soutenance : **13 Janvier 2017**

Nom de famille et prénom de l'auteur : **Maxime GASSE**

Titre de la thèse : « Apprentissage de Structure de Modèles Graphiques Probabilistes : Application à la Classification Multi-Label. »



RESUME DE LA THESE

Un modèle graphique probabiliste (PGM en Anglais) permet la représentation compacte d'une distribution multivariée $p(\mathbf{v})$ en exploitant la structure d'indépendance entre les variables, encodée sous la forme d'un graphe. Dans cette thèse, nous nous intéressons au problème spécifique de l'apprentissage de structure de modèles graphiques probabilistes, c'est-à-dire trouver la structure la plus efficace pour représenter une distribution, à partir seulement d'un ensemble d'échantillons $\mathcal{D} \sim p(\mathbf{v})$. Dans une première partie, nous passons en revue les principaux modèles graphiques probabilistes de la littérature, des plus classiques (modèles dirigés, non-dirigés) aux plus avancés (modèles mixtes, cycliques etc.). Puis nous étudions particulièrement le problème d'apprentissage de structure de modèles dirigés (réseaux Bayésiens), et proposons une nouvelle méthode hybride pour l'apprentissage de structure, H2PC (Hybrid Hybrid Parents and Children), mêlant une méthode à base de contraintes (tests statistiques d'indépendance) et une méthode à base de score (probabilité postérieure de la structure).

Dans un second temps, nous étudions le problème de la classification multi-label, visant à prédire un ensemble de catégories (vecteur binaire $\mathbf{y} \in \{0,1\}^m$) pour un objet (vecteur $\mathbf{x} \in \mathbb{R}^d$). Dans ce contexte, l'utilisation de modèles graphiques probabilistes pour représenter la distribution conditionnelle des labels prend tout son sens, particulièrement dans le but minimiser une fonction coût complexe. Nous passons en revue les principales approches utilisant un modèle graphique probabiliste pour la classification multi-label (*Probabilistic Classifier Chain, Conditional Dependency Network, Bayesian Network Classifier, Conditional Random Field, Sum-Product Network*), puis nous proposons une approche générique visant à identifier une factorisation de $p(\mathbf{y}|\mathbf{x})$ en distributions marginales, en s'inspirant des méthodes d'apprentissage de structure à base de contraintes. Nous démontrons plusieurs résultats théoriques, notamment l'unicité d'une décomposition minimale, et l'optimalité de plusieurs algorithmes d'identification sous diverses hypothèses : 1) lorsque p supporte la propriété de Composition ; 2) lorsque p supporte la propriété d'Intersection ; enfin 3) sans aucune condition. Nous mettons en pratique ces résultats et montrons empiriquement qu'identifier une telle décomposition permet d'améliorer le coût espéré avec des fonctions coût complexes telles que la fonction *F-loss* ou la fonction *zero-one loss*.