

Clustering des nœuds d'un réseau dans le cadre des Stochastic et Latent Blockmodels

Jean-Benoist Léger

Laboratoire MaIAGE.

UR 1404

INRA Jouy-en-Josas, France

`jean-benoist.leger@hds.utc.fr`

Mots clefs : Réseau, Stochastic Block Model.

L'analyse de la topologie d'un réseau, bipartite ou non, peut être réalisée au moyen d'un clustering des nœuds du graphe, ou d'un biclustering dans le cas d'un réseau bipartite. Les modèles statistiques de base pour la réalisation de ce clustering sont le Stochastic Block Model pour le cas général et le Latent Block Model pour le cas des réseaux bipartites.

À l'origine, introduits dans le cas des réseaux binaires, ces modèles ont été étendus au cas des réseaux valués, avec diverses lois de probabilité sur les arrêtes du graphe. De plus, ces modèles ont été étendus de telle manière à prendre en compte des covariables sur les couples de nœuds, permettant de prendre en compte des connaissances extérieures dans le processus de modélisation.

Cette présentation introduira le package `blockmodels` qui réalise l'inférence dans ces modèles au moyen d'un EM Variationnel, et l'exploration du nombre de groupes et la sélection à l'aide d'un critère ICL. Des cas d'utilisation seront présentés.

Les subtilités d'implémentation, les variations entre les différents modèles (Bernouilli, gaussien, Poisson, avec ou sans covariables) seront introduites, ainsi que les différentes techniques utilisées, telles que l'estimation parallèle et les optimisations à la compilation par la spécialisation de template, qui permettent d'analyser des réseaux de plusieurs dizaines de milliers de nœuds dans un temps raisonnable.

English version: Clustering of network nodes with Stochastic and Latent Blockmodels.

Analysis of the topology of a network, bipartite or not, can be realized by a clustering of the nodes, or by a biclustering in the case of a bipartite network. Basic statistical models for the realization of this clustering are the Stochastic Block Model for the general case and the Latent Block Model for the case of bipartite networks.

Originally, introduced in the case of binary networks, these models have been extended to the case of valued networks, with various probability laws on edges of the graph. Moreover, these models have been extended to use covariates on the pairs of nodes, allowing take into account external knowledge in the process of modelization.

This presentation will introduce the `blockmodels` package that performs the inference in these models with a Variational EM, and the exploration of the number of groups and selection using an ICL criterion. Use cases will be presented.

Implementation details, variations between the different models (Bernouilli, Gaussian, Poisson, with or without covariates) will be introduced, as well as the different techniques used, such as parallel estimation and compilation optimizations by the template specialization, which allows to analyze networks of dozens of thousands nodes in a reasonable time.