Mixture Models for Graph Clustering

Sofiane Ezzehi [1] Bastien Le Chenadec [1] Theïlo Terrisse [1]

École des Ponts ParisTech

Context

The Stochastic Block Model

The **Stochastic Block Model** is a *mixture* model for graphs. Each node is assigned to a *class*, and conditionally on these classes the probability of an edge between two nodes i and j only depends on their class memberships. This model draws inspiration from mixture models for distribution of degrees, and the Erdös-Rényi model which deals with the probability for two given nodes of being connected.

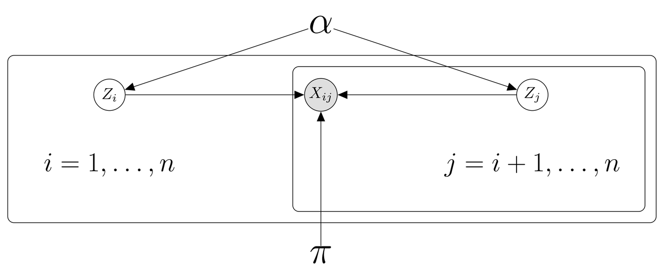

Figure 1. Graphical model of the SBM

Formally, we consider an undirected graph with n nodes and no self-loops. We denote X the adjacency matrix of this graph, thus $X_{ij} \in \{0, 1\}$ denotes the existence of an edge between nodes i and j.

The model is parametrized by Q the number of classes, $\alpha \in [0,1]^Q$ the prior distribution on the classes, and $\pi \in [0,1]^{Q\times Q}$ the probability of an edge between two nodes of different classes. We also introduce the random variables $Z_i \in \{0,1\}^Q$ for $i \in [\![1, n]\!]$, that represent the membership of node i to each class. The prior distributions on Z and X are given by :

$$\begin{cases} \forall i \in [\![1,n]\!], \quad \sum_{q=1}^{Q} Z_{iq} = 1 \quad \text{(unique class)} \\ \forall q \in [\![1,Q]\!], \quad \mathbb{P}(Z_{iq} = 1) = \alpha_q \quad \text{(class distribution)} \\ \forall q, l \in [\![1,Q]\!], \quad \forall i \neq j \in [\![1,n]\!], \quad \mathbb{P}(X_{ij} = 1 \mid Z_{iq} = 1, Z_{jl} = 1) = \pi_{ql} \quad \text{(edge probability)} \end{cases} \tag{1}$$

The variational Expectation-Maximization algorithm

In [1], the authors propose a variational Expectation-Maximization (EM) algorithm to fit the SBM. The log-likelihood of the model writes :

$$\log \mathcal{L}(X, Z) = \sum_i \sum_q Z_{iq} \log \alpha_q + \frac{1}{2} \sum_{i \neq j} \sum_{q,l} Z_{iq} Z_{jl} \, \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{1 - X_{ij}} \tag{2}$$

This likelihood does not lead to a closed form solution for the parameters of the model. Using an EM algorithm, there is no closed form solution for the E-step either. Instead the authors search for an approximated distribution R_X of Z given X among the family of product of multinomial distributions, thus introducing the random variables $\tau_i \in [0,1]^Q$ for $i \in [\![1, n]\!]$. This approximation leads to maximizing the following lower bound of $\log \mathcal{L}(X)$:

$$\mathcal{J}(R_X) = \log \mathcal{L}(X) - \mathrm{KL}[R_X(\cdot), P(\cdot|X)] \tag{3}$$

The authors are able to derive a closed form solution for the M-step of the algorithm, and a fixed point relation for the E-step :

$$\begin{cases} \hat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_l \left[\pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{1 - X_{ij}} \right]^{\hat{\tau}_{jl}} & \text{(E-step)} \\ \hat{\alpha}_q = \frac{1}{n} \sum_i \tau_{iq} & \text{(M-step)} \\ \hat{\pi}_{ql} = \frac{\sum_{i \neq j} \tau_{iq} \tau_{jl} X_{ij}}{\sum_{i \neq j} \tau_{iq} \tau_{jl}} & \text{(M-step)} \end{cases} \tag{4}$$

Algorithm 1 Variational Expectation-Maximization Algorithm

```
Input: Adjacency matrix X, number of communities Q
Initialize: Initialize τ_iq for i ∈ [[1, n]], q ∈ [[1, Q]]
while not converged do
    E-step:
    for i ∈ [[1, n]] do
        for q ∈ [[1, Q]] do
            Fixed-point algorithm:
            τ̂_iq ∝ α_q ∏_{j≠i} ∏_l [π_ql^{X_ij} (1 − π_ql)^{1−X_ij}]^{τ̂_jl}    ▷ Update class memberships

    M-step:
    for q ∈ [[1, Q]] do
        α̂_q = (1/n) ∑_i τ_iq                                            ▷ Update class proportions
        π̂_ql = ∑_{i≠j} τ_iq τ_jl X_ij / ∑_{i≠j} τ_iq τ_jl               ▷ Update edge probabilities
Return: α̂, π̂, τ̂
```

Variants

Experiments on an SBM dataset

Second set of experiments

References

[1] J.-J. Daudin, F. Picard, and S. Robin.
A mixture model for random graphs.
Statistics and Computing, 18(2):173–183, June 2008.