PHYS-467 Assignment 2

November, 14 2023

Instructions You are asked to submit a file `code_assignment_2_your_name.py` (for example `code_assignment_2_lucas_clarte.py`), where the different functions are duly implemented. For questions 2, 5, and 6, you are asked to submit a pdf file `answers_assignment_2_your_name.pdf` containing your answers and the different plots.

Stochastic Block Model The Stochastic Block Model (SBM) is based on the following quantities:

1. The number of possible groups q;
2. The expected fraction of members of each group $\{n_a\}_{a=1}^q$, such that $\sum_{a=1}^q n_a = 1$;
3. The symmetric matrix $p_{ab} \in [0,1]^{q\times q}$ probability of an edge between group a and b.

Given these elements, we can generate a directed graph G with N nodes and adjacency matrix A as follows

1 Assign node i to group a with probability $P(g_i = a) = n_a$, where g_i indicates the group assignment of node i. Repeat $\forall i \in 1, ..., N$.

2 Include an edge between nodes i ad j with probability p_{g_i,g_j} setting $A_{ij} = 1$, and set $A_{ij} = 0$ with probability $1 - p_{g_i,g_j}$. Self-loops are forbidden, i.e. $A_{ii} = 0$.

Question 1 (2 Pt.) Implement the function `generate_data` that takes as argument the parameters N (number of nodes), q (number of groups), $\{n_a\}_{a=1}^q$ (expected fraction of nodes per group, given as a list), and p_{ab} (probability between two groups, given as a `numpy` matrix) and returns a N-dimensional vector g (assignment of each node) and the adjacency matrix A as defined above.

Question 2 (2 Pt.) Write explicitly the posterior distribution

$$P(G, \{g_i\}|\theta) \propto P(G|\{g_i\}, \theta)P(\{g_i\}|\theta) \tag{1}$$

where $\theta = \{q, \{n_a\}, \{p_{ab}\}\}$ and show that the probability distribution over the group assignments can be written as:

$$\mu(\{g_i\}|G, \theta) = P(\{g_i\}|G, \theta) = \frac{e^{-H(\{g_i\}|G,\theta)}}{\sum_{\{g_i\}} e^{-H(\{g_i\}|G,\theta)}} \tag{2}$$

where

$$H(\{g_i\}|G, \theta) = -\sum_i \log(n_{g_i}) - \sum_{i \neq j} \left[A_{ij} \log(p_{g_i,g_j}) + (1 - A_{ij}) \log(1 - p_{g_i,g_j})\right] \tag{3}$$

Question 3 (1 Pt.)

- Write a function `energy` taking as input $\{g_i\}$, $\{n_a\}_{a=1}^q$, p_{ab} and A and returning the value of the energy as in Eq. 3
- Implement a second function `energy_difference`, taking as input two node configurations $\{g_i'\}$ and $\{g_i\}$ and the parameters $\{n_a\}_{a=1}^q$, p_{ab} and A, and returning $H(\{g_i'\}|G, \theta) - H(\{g_i\}|G, \theta)$

Question 4 (3 Pt.) Consider the following setting:

- $N = 100$;
- $q = 2$;
- $n_0 = 0.7$ and $n_1 = 0.3$;
- $p_{01} = p_{10} = 0.3$, $p_{00} = 0.4, p_{11} = 0.5$.

Sample a group assignment $\{g_i^*\}$ and an adjacency matrix A from these parameters. Given A and the parameters θ, we would like to recover g^* from a random initial configuration $\{g^0\}$. To do so, we resort to the Metropolis-Hastings scheme running for T iterations:

- Initially, at $t = 0$, sample g^0 where for each g_i^0, $P(g_i^0 = 1) = n_1$
- At each iteration t: pick an index i uniformly at random in $[0, N]$
- Define $\{g_i'\} = \{g_0^t, ..., 1 - g_i^t, ..., g_N^t\}$. Calculate the energy difference between $\{g_i'\}$ and $\{g_i^t\}$, i.e., $\Delta H =$`energy_difference`$(\{g_i'\}, \{g_i^t\}, \{n_a\}_{a=1}^q, p_{ab}, A)$. With probability $\min(1, \exp(-\Delta H))$, set $\{g_i^{t+1}\} = \{g_i'\}$, otherwise define $\{g_i^{t+1}\} = \{g_i^t\}$

Implement a function named `run_mcmc` returning a sequence of T states $\{g_i^t\}$ for $t = 1, ..., T$.

Question 5 (2 Pt.) For each of the T states output by the `run_mcmc` function, compute

1. The state energy $H(g^t)$
2. The overlap between g^t the ground truth state g^*. The overlap can be defined as $Q(\{g_i\},\{g_i^*\}) = \max_\pi \frac{\frac{1}{N}\sum_i \delta_{g_i^*,\pi(g_i)} - \max_a n_a}{1-\max_a n_a}$, where π ranges over the permutations on q elements.
3. The fraction of non-zero entries in g^t

Plot these quantities as a function of time.

Question 6 (3 Pt.) We now consider the case where p_{ab}, q and A are available but the $\{n_a\}_{a=1}^q$ are not known and must be learnt. It can be shown that the maximization of the posterior distribution over the parameters $P(\theta|G)$ w.r.t. n_a leads to the following update rule:

$$\frac{1}{N}\sum_i \langle \delta_{g_i,a} \rangle = \frac{\langle N_a \rangle}{N} = n_a \qquad \forall a = 1,..,q \tag{4}$$

where by $\langle f(\{g_i\}) \rangle = \sum_{\{g_i\}} f(\{g_i\})\mu(\{g_i\}|G,\theta)$ Given this update rule, use the Expectation-Maximization (EM) algorithm to infer $\{n_a\}_{a=1}^q$. Specifically, assume the ground-truth data are obtained under the setting specified in Question 4. Perform $M = 10$ steps of EM, by using the `run_mcmc` function implemented before. Assume $n_0^0 = 0.55$ as your initial guess. Plot the evolution of n_0^m as a function of the EM iterations (i.e., for $m = 0, ..., M$).