Machine Learning for Complex Networks
SoSe 2023

Prof. Dr. Ingo Scholtes
Chair of Informatics XV
University of Würzburg

Exercise Sheet 04

Published: May 24, 2023
Due: June 1, 2023
Total points: 10

Please upload your solutions to WueCampus as a scanned document (image format or pdf), a typesetted PDF document, and/or as a `jupyter` notebook.

1. **MDL with the Degreee corrected SBM**

In Lecture 04, you learned about the Degree-Corrected Stochastic Block Model (DCSBM). In the practice session, you saw an example of inferring communities using maximum likelihood estimation with the DCSBM in a Jupyter Notebook. However, not specifying the number of communities can lead to trivial community assignments, as seen in Exercise 03 and Lecture 05. In Lecture 05's practice session, you learned about using the Minimum Description Length (MDL) principle to detect communities in a toy network with the standard Stochastic Block Model (SBM). Now, we will explore detecting communities in the Zachary Karate Club network using MDL with the DCSBM.

(a) For the undirected DCSBM with N nodes and B blocks, while the MDL equation takes the familiar form 4P

$$\Sigma = S - L$$

where Σ is the description length, S is the entropy of the DCSBM, and L is the information necessary to describe the model with its parameters, S and L are modified to account for the degree distribution of the network.

S can be expressed as

$$S \approx -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs} \ln\left(\frac{e_{rs}}{n_r n_s}\right)$$

where $E = \frac{1}{2}\sum_{rs} e_{rs}$ is the total number of edges, e_{rs} is the number of edges between blocks r and s, and n_r and n_s are the number of nodes belonging to blocks r and s, respectively. N_k is the total number of nodes with degree k, and the term $-\sum_k N_k \ln k!$ accounts for the degree distribution of the network.

L can be expressed as

$$L \approx Eh\left(\frac{B\left(B+1\right)}{2E}\right) + N \ln B - N\sum_k p_k \ln p_k$$

where $h\left(x\right) = \left(1+x\right)\ln\left(1+x\right) - x\ln(x)$ and p_k is the fraction of nodes with degree k. As you may have noticed, term $-N\sum_k p_k \ln p_k$ modifies L to account for the degree distribution. Use this MDL equation to partition the Karate Club network into communities. The network can be obtained via Netzschleuder under the name `karate`. Plot the communities and the degree distributions of the nodes in each community. Compare these communities to communities inferred using MDL with the standard SBM.

(b) The above approximate expression for the entropy S only holds when 1P

$$e_{rs}\frac{\langle k^2\rangle_r - \langle k\rangle_r}{\langle k\rangle_r^2}\frac{\langle k^2\rangle_s - \langle k\rangle_s}{\langle k\rangle_s^2} \ll n_r n_s$$

where $\langle k^l\rangle_r = \sum_{i\in r} k_i^l/n_r$. Show that this condition holds for the communities you inferred.

(c) How is the entropy S related to the number of microstates Σ of the DCSBM? What is the probability $P(G)$ of each microstate G? 1P

For the interested reader, you can find more information on using MDL with the SBM and DCSBM (and their directed variants) at `https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.110.148701`.

2. Kullback-Leibler Divergence

For two discrete probability mass functions Q and P defined on the same sample space Ω with events i, the Kullback-Leibler divergence (which is also called relative entropy) from Q to P is defined as:

$$D_{\mathsf{KL}}(P\|Q) := -\sum_{i\in\Omega} P(i)\cdot\log\frac{Q(i)}{P(i)}$$

(a) Consider a dice X with six faces and two different probability mass functions 2P

$$Q(X=1)=\frac{1}{6}, Q(X=2)=\frac{1}{6}, Q(X=3)=\frac{1}{6}, Q(X=4)=\frac{1}{6}, Q(X=5)=\frac{1}{6}, Q(X=6)=\frac{1}{6}$$

and

$$P(X=1)=\frac{1}{3}, P(X=2)=\frac{1}{3}, P(X=3)=\frac{1}{12}, P(X=4)=\frac{1}{12}, P(X=5)=\frac{1}{12}, P(6)=\frac{1}{12}$$

Use `python` to compute two sequences of 1000 dice rolls with probabilities according to P and Q respectively. Use a binary Huffman code to encode the sequence and compute the number of bits required per symbol.

(b) Calculate the difference between the bits required for both sequences and compare it with the Kullback-Leibler divergence from Q to P. Interpret and explain your findings. 2P