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

# Exercise Sheet 03

Published: May 25, 2022
Due: June 1, 2022
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. **Stochastic Block Model**

   (a) Consider a network with adjacency matrix 1P

   $$A = \begin{bmatrix} & 1 & 1 & & & & \\ 1 & & 1 & & & & \\ 1 & 1 & & 1 & & & \\ & & 1 & & 1 & 1 & \\ & & & 1 & & 1 & \\ & & & 1 & 1 & & 1 \\ & & & & & 1 & \end{bmatrix}$$

   and a stochastic block model with block assignment vector $\vec{z} = (0, 0, 0, 1, 1, 1, 1)$. Give a maximum likelihood estimate for the the stochastic block matrix $\mathbf{M}$. Is this the vector $\vec{z}$ that maximizes the likelihood function for $B = 2$?

   (b) Consider the so-called *micro-canonical stochastic block model*, i.e. a generalization of the $G(n, m)$ model in the spirit of the stochastic block matrix, where the block matrix entries $M_{kl}$ give the number of edges randomly generated between nodes in block $k$ and block $l$ rather than link probabilities. Give an expression for the microstate probability of this model. 2P

2. **Simulated Annealing**

   (a) Use the toy example network with six nodes from the lecture and calculate the maximum likelihood of a stochastic block model with block assignment vector $\vec{z} = (0, 1, 2, 3, 4, 5)$. Compute the maximum likelihood estimates of the stochastic block matrix entries and interpret the resulting matrix $\widehat{\mathbf{M}}$. 2P

   (b) In the discussion so far, we have fixed the number of blocks $B$ for the inference of the block assignment vector, i.e. we had to specify the *number* of communities to be detected in advance. Addressing this isse, adapt the simulated annealing algorithm from the practice session, such that the number of blocks $B$ can change during the optimization process. One approach to achieve this is to start with a block assignment vector that assigns all nodes to a single community. In each step you can randomly split or merge communities and accept the change depending on the current temperature and the difference in likelihood. The result will be an optimization algorithm that detects both the number of block $B$ and the optimal block assignment vector. Apply this algorithm to the toy example and report your finding. interpret the result in the context of task 2 (a). 3P

   (c) Investigate $k$-means clustering, a simple algorithm to detect clusters in Euclidean data. Explain how the detection of the optimal block number $B$ in the stochastic block model is related to the detection of the optimal cluster number $k$ in $k$-means. What is the minimal value of the loss function for $k$-means clustering for $k = n$, i.e. if the number of clusters corresponds to the number of data points. Explain how this is related to your result from 2(a) and 2(b). 2P