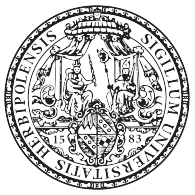

Machine Learning for Complex Networks
SoSe 2022

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

Exercise Sheet 01

Published: May 4, 2022
Due: May 12, 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. **Computing connected components in graphs**

   (a) Implement Tarjan's algorithm to compute the (strongly) connected components in a (directed) network. Apply your algorithm to a directed example network that has multiple strongly connected components. Explain how the `low_link` and `dfs_num` counters in your example are used to assign nodes to strongly connected components. 2P

   (b) For a Laplacian matrix $\mathcal{L}$ of an undirected graph G with n nodes consider the sequence 1P

   $$\lambda_1 = 0 \leq \lambda_2 \leq \ldots \leq \lambda_n$$

   of eigenvalues in ascending order. Generate undirected networks with $n = 20$ nodes and different numbers of connected components from one to 50. Calculate the eigenvalue sequences of the corresponding Laplacian. What do you observe? Can you explain your observation?

   (c) Repeat the experiment above using the sorted sequence of eigenvalues of the adjacency matrix. 1P

   (d) Use your finding from the previous tasks to implement a `python` function that uses the Laplacian matrix to calculate the number of connected components in an undirected graph. Test your function in an example network. What is the computational complexity of your function? 1P

2. **Density-based Clustering with DBSCAN**

   Cluster detection, i.e. identifying groups of objects more similar to each other than to objects in other groups, is an important unsupervised machine learning task for collections of data points in a Euclidean space. Considering that similarities between points in Euclidean space can be represented by links, graph-based algorithms have been successfully applied to this problem. A well-known example is DBSCAN, a density-based clustering algorithm that uses connected components in a graph to identify clusters based on contiguous regions with high point density. Consider the following paper (available on WueCampus) and answer the questions below:

   M Ester, HP Kriegel, J Sander, X Xu: **A density-based algorithm for discovering clusters in large spatial databases with noise**, In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226-231, August 1996

   (a) Give a pseudocode implementation of DBSCAN and explain the following aspects of the algorithm: 2P
   - Explain how the parameter ϵ is used to represent the data points in terms of a graph.
   - Explain how the parameter ϵ influences the categorization of nodes as core, border, and noise nodes.
   - Explain how we can apply Tarjan's algorithm to detect clusters.
   - Discuss how the choice of the parameter δ influences the number of detected clusters.

   (b) Implement the algorithm in python and test it using synthetic data generated by the function `make_moons` available in `sklearn.datasets`. 2P

   (c) Investigate for which values of δ the algorithm returns a reasonable cluster structure and compare the performance to k-means clustering (e.g. using the implementation included in `sklearn`). 1P