Medical Data Science, WS 2021/2022
Prof. Dr. Nico Pfeifer
Chair for Methods in Medical Informatics
University of Tuebingen

2021-23-11

Assignment 3

Deadline: Tuesday, December 7, 7:59 p.m.

This problem set is worth 25 points. You should submit in groups of two people. Submit your solutions digitally by uploading to the ILIAS webpage (none of the other students can see the files you upload). Just upload a zipped folder containing all necessary files and name the folder by your last names. The folder should be named according to the following scheme:

cs[MDS][Assignment 3]_lastname1_lastname2

Problem 1 (T, 17 Points)

(a) (3P) Where do we encounter graphs in the biomedical field and why can it be interesting to compare two graphs and take similarities between them into account? Think about a biomedical example (that you don't know from the lecture) where graphs can help to gain further insight.

(b) (4P) Why is complexity a problem in the computation of graph similarity? How is this problem handled currently? Name and briefly explain two possibilities of comparing two graphs. Which one leads to better results?

(c) (2P) What is a graph isomorphism? Give an example for a graph isomorphism by drawing two graphs with four nodes each. (You can draw the graphs by hand and insert an image of it.) Explain in your own words why the graphs are isomorphic.

(d) (2P) What is a graphlet? Describe one similarity and one difference of graphlet kernels to previously discussed kernels.

(e) (2P) Name and explain two different Weisfeiler-Lehman kernels. Which one leads to better results?

(f) (2Ps) Why do integrative analyses for cancer make sense? In which other scenarios could you imagine integrative analyses?

(g) (2Ps) Describe the assumptions that are used in a Gaussian latent variable model and relate them to iCluster.

Problem 2 (Exciting news, 5 Points)

Last year there were some exciting news in the biomedical machine learning field! Whatch the following video and briefly summarize it. You can also use information from this recent nature news article: nature.
Pay special attention to the following questions:
What was the big news?
Why is this topic of such huge interest and why can't it be solved using a deterministic approach?
What is (probably) the underlying machine learning approach?
How can this achievement probably improve biomedical research?
Do you think the "protein folding problem" (see reference article) is solved given this new approach?

Problem 3 (T, 10 Points)

Weisfeiler-Lehman test for graph isomorphism.
You have two unlabeled graphs defined by the following adjacency matrices:

$$G_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \text{ and } G_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

- (1P) Use the node degrees as labels. Draw the two graphs with the corresponding labels.
- (6P) Perform the one dimensional Weisfeiler-Lehman test for graph isomorphism.
- (3P) There exist graphs for which this test fails. Show two such graphs by a sketch. Explain how the algorithm fails in your example (e.g. continuous loop, false positive/negative,...). You don't need to provide a formal proof.

Problem 4 (P, 18 Points)

Evaluate the performance of different graph kernels using GraKeL: link to website and the MUTAG data set from here: website with benchmark data sets.

- Compute the graphlet kernel using sampling for the graphlets of size 3 (1000 samples). Perform a 10-fold cross-validation for the binary classification problem using the kernel with an SVM. What is the accuracy for the best $C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$?
  How many samples would you need such that the deviation from the real distribution is less than 0.05 with probability larger than 0.9?
- Compute the Weisfeiler-Lehman subtree kernel for 4 iterations. Perform a 10-fold cross-validation for the binary classification problem using the kernel with an SVM. What is the accuracy for the best $C \in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$?