Medical Data Science, WS 2023/2024
Prof. Dr. Nico Pfeifer
Chair for Methods in Medical Informatics
University of Tuebingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

2023-11-20

Assignment 3

Deadline: Tuesday, December 5, 7:59 p.m.

This problem set is worth 50 points. You should submit in groups of two people. Submit your solutions digitally by uploading to the ILIAS webpage. 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, 18 Points)

(a) (3P) Explain the *multitask learning* approach of the lecture and discuss whether multitask learning with many source domains or dualtask learning lead to better results. What can you say about the position of two really similar tasks?

(b) (1P) What is the difference between the *Major Histocompatibility Complex (MHC)* and human leukocyte antigens (HLA) molecules in the immune system? What is the main purpose of *MHC I*?

(c) (1P) Describe what leveraging is. Heckerman et al. might help.

(d) (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.

(e) (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?

(f) (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.

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

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

Problem 2 (P, 18 Points)

In this exercise, you will implement a kernel-based multitask learning approach in R or Python to predict HLA class I peptide binding. For the peptide kernel use the linseq approach presented in the fifth lecture. Use ten-fold cross-validation to estimate performances (running over different values of the cost parameter C of the SVM). The data can be found in the assignment folder (BindingData.csv). here. The binding class label is in the last column.

(a) (3P) Implement the Dirac, uniform, multitask and peptide kernel functions.

(b) (7P) Build SVM models using the Dirac kernel, the uniform kernel, and a multitask kernel (consisting of the two former kernels) in combination with the peptide kernel in the cross-validations with $C \in \{10^{-4}, 10^{-3}, ..., 10^{2}\}$. For R, use the **ksvm()** function of the **kernlab** package.

(c) (4P) What happens if you add another Dirac kernel that is based on the supertypes (supertype.csv) from LANL (where available) to the multitask kernel?

(d) (2P) Generate a ROC curve that shows the performances of the different approaches. For R you can use the ROCR package. Calculate AUCs to compare the different approaches and comment on your findings.

(e) (2P) Compare AUCs to accuracy, and discuss possible discrepancies.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Problem 3 (T, 14 Points)

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

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

- (2P) Use the node degrees as labels. Draw the two graphs with the corresponding labels.
- (8P) Perform the one dimensional Weisfeiler-Lehman test for graph isomorphism.
- (4P) 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.