Thesis Review by Opponent

Explainability of classification of graph represented data

Master Thesis by Bc. Bibiána Lajčinová, 2020

I read Ms. Lajčinová‘s thesis with interest, as it touches two topics of current interest in machine learning community – learning from graph structures and providing explanation of black box model verdicts. Both problems are related, as graph-based models require specialized explanation techniques. At the same time, both of these problem areas are not yet sufficiently enough understood, hence any advance has the potential of great impact.

The thesis starts by introductory chapter which gives only very brief overview over basics of machine learning (ML), Bayesian probability theory and graph theory. The second chapter extends the intoduction to ML by covering Logistic regression, Bayesian logistic regression and Neural networks (NN), including their history, details of relevant learning concepts (backpropagation, gradient descent, model testing) and a overview of activation functions as well as introduction to interpretability and explainability. This chapter is very well written and covers its subject in a easy-to read manner, while covering all reasonable details to enable the reader to get good understanding of the concepts even if the reader does not have this specific knowledge beforehand. The third chapter first introduces the recent GNN Explainer algorithm suitable for explaining graph neural network verdicts. Then it introduces the concept of Predictive projection, useful to reduce the feature space relevant for explanations; in particular two approaches are discussed, one based on a combination of stochastic gradient with Langevin dynamics, the other based on Deep Ensembles. The chapter culminates in definition of original GNN Explorer+, which takes use of ideas both from GNN Explainer and the discussed Predictive projection techniques. Chapter 4 introduces a framework for testing the results of explanation on synthetic data. Chapter 5 follows similar testing methodology on a real world problem example – credit scoring of bank clients based on a set of features describing each client combined with information about related clients. Example explanations are given and brief conclusions about which factors affect client credibility the most.

The key result presented in the thesis – GNN Explainer+ - is a non-trivial result, which takes use of Bayesian learning tools previously unused in this explanation context. This result as well as the fact that I did not notice any technical mistakes, well justifies my recommendation to accept this thesis as Master thesis. That said, I need to point out numerous flaws that the thesis suffers from.

There seems to be a logical gap between Chapter 2 and 3. Chapter 3 discusses GNN Explained, the tool defined on top of graph neural networks (GNN); but Chapter 2 introduces only the standard non-structural NNs. No definition of GNNs is provided, nor any much needed discussion of their principal differences to NNs. It is very confusing for the reader to see what role exactly are the GNNs supposed to play here. Further discussion seems to refer to graph explanations throughout, but even the key applied experiement concluding the thesis actually does not use of GNNs, instead maps the input data into non-structural (vector) form and then applies standard NN. It is no less difficult to follow the graph line of enquiry in section 3.3; the relation of uncertainty estimation on NNs to graph explainability should be described and explained more clearly and understandably. Introduction of Shapley values is too brief, does not give understanding how Shapley values work.

Structure of the thesis is at times confusing. Some concepts of high importance are introduced in sub-sub-subsections instead of at more prominent level, e.g., the important and general problem of

uncertainty estimation is introduced only within sub-sub-section 3.3.2 called “ways of generating \theta”.

Mathematical symbols are often used in multiple incompatible definitions thoughout the text which is very confusing (“p” once stands for index of NN instance, another time it stands for client feature. “\Phi” once stands for trained model to enter GNN explanation, at another place it stands for Shapley value, number of classess is sometimes marked by “C”, sometimes by “K”, “D” on page 43 refers to dataset, while on page 42 it refers to number of NNs) and shows lack of effort invested in providing a unified thesis text. Some terms are referred to in multiple ways (“vrchol” vs “vertex”). Sometimes symbols get orphaned – on page 41 “\beta^(j)” is claimed to refer to samples of projected coefficient, but is in fact nowhere used in the related method description. Overall, the clarity of explanation seems to decline from chapter 3 onwords, and the number of minor notation and description mistakes seems to increase. Logical duplicities occur, even some completely unnecessary mistakes (e.g., I do not understand what does Fig 5.2 actually bring beyond the fact that a client can have relations to other clients? While Fig 5.2 seems omissable, it is even more striking that it is at odds with text, where on page 51 I read “clients in dataset have 1 to 6 relations” while Fig 5.2 displays 7 edges). The closer the reader reads through to the end of the thesis, the more the thesis seems to have been finished in rush.

Some constants would deserve further explanation to allow reader understant the reasoning behind their particular value. What exactly is the role of the constant 10 in the limit on the number od Nns in step 2 in algorithm on page 42? On page 51 why exactly 1398 samples has been selected out of 5000? Why on page 53 in the experiments the 4-fold cross-validation is used, while all previous discussions considered 5-fold cross-validation? Can the choice of 15% FPR on page 53 be reasoned about? Figures 4.2, 4.4 are not referenced from text, although one can guess which paragraphs they are supposed to accompany. Figure 5.3 would be easier to follow if it contained graphical legend. How have the particular 2-D projections been selected for Fig 5.4? Why the experiment in section 5.1 regarding Shapley values has been defined to yield exactly 20 features? Parameter K in formula (3.27) would deserve deeper discussion to let the reader understand its role and the influence of its setting to the performance of the algorithm.

The key omission in the experimental Chapters is the comparison of GNN Explainer to GNN Explainer+. Is not one of the key messages of the thesis that GNN Explainer+ is better than GNN Explainer? No experimental verification is given. Experimental section and related discussion has more flaws. It is noted that GNN Explainer+ is more computationally complex than GNN Explainer, it is even explained that each of the changed components makes it slower (see section 3.5.1). But there is no enumeration of how much slower, nor an experiment illustrating the time complexity is provided.

To summarize, the key result presented in the thesis is very good, meaningful and stands on good theoretical grounds. The thesis as a whole suffers from many structural flaws, illogical omissions, overloading of mathematical symbols and irregular text quality. I do recommend this thesis to be accepted as Master thesis. I recommend to rate the thesis by mark **C** – good.

In Prague, 18 July 2020
RNDr. Petr Somol, Ph.D.