University of Pittsburgh | School of Computing and Information

Proposal Defense

Doctor of Philosophy in Computer Science

"**Relational Machine Learning**" by **Alireza Samadian**

Date: October 15, 2020
Time: 3:00 PM – 4:00 PM
Place: Virtual at https://pitt.zoom.us/j/97709184393

Committee:

- Dr. Kirk Pruhs (advisor), Professor, School of Computing and Information, University of Pittsburgh
- Dr. Panos Chrysanthis, Professor, School of Computing and Information, University of Pittsburgh
- Dr. Adriana Kovashka, Assistant Professor, School of Computing and Information, University of Pittsburgh
- Dr. Benjamin Moseley, Associate Professor, Tepper School of Business, Carnegie Mellon University

Abstract:
Many of the learning tasks faced by data scientists involve relational data. Most commonly, the relational data is stored in tables in a relational database. The data is usually stored in a normalized form to prevent repetition, and it may have both numerical and categorical values. Yet most of the standard algorithms for standard machine learning problems are not designed to accept relational data as input. The standard practice to address this issue is to join the relational data to create the type of geometric input that standard learning algorithms expect. Unfortunately, this standard practice has exponential worst-case time and space complexity. This issue with the standard techniques leads us to consider what we call the Relational Learning Question: ``Which standard learning algorithms can be efficiently implemented on relational data, and for those that cannot, is there an alternative algorithm that can be efficiently implemented on relational data and that has similar performance guarantees to the standard algorithm?'' In this proposal, we start by explaining some preliminary results on training linear SVM on relational data. After explaining two main approaches for solving linear SVM, we discuss relational algorithms for clustering methods such as k-centers and DBSCAN and getting different guarantees for linear SVM and logistic regression as future work.

One of the main optimization algorithms for training machine learning models is gradient descent. We prove that it is #P-Hard to approximate the gradient of linear SVM up to any constant factor, and similar proof can be applied for logistic regression. In fact, it can be shown that some simpler problems such as counting the number of points lying on one side of a hyperplane are also #P-Hard. Therefore, instead of trying to directly find a relational implementation of gradient descent, we have investigated two different approaches: (1) Extracting a manageably small (potentially weighted) sample from the data set, and then directly solving (a weighted version of) the problem on the (weighted) sample. (2) Introducing a relational algorithm for those instances of SVM that have some stability properties. Both approaches will be discussed in detail as preliminary results.