Optimization for large-scale machine learning: large data and large model

Anna Choromanska

Courant Institute of Mathematical Sciences, New York University

10/20/2015 at 11:00 am
Core A (Room 301)

Abstract

The talk will focus on selected challenges in modern large-scale machine learning in two settings: i) large data setting and ii) large model (deep learning) setting. The first part of the talk will focus on the case when the learning algorithm needs to be scaled to large data. The multi-class classification problem will be addressed, where the number of classes (k) is extremely large, with the goal of obtaining train and test time complexity logarithmic in the number of classes. A reduction of this problem to a set of binary classification problems organized in a tree structure will be discussed. A top-down online tree construction approach for constructing logarithmic depth trees will be demonstrated, which is based on a new objective function. Under favorable conditions, the new approach leads to logarithmic depth trees that have leaves with low label entropy. Discussed approach comes with theoretical guarantees following from convex analysis, though the underlying problem is inherently non-convex. The second part of the talk focuses on the theoretical analysis of more challenging non-convex learning setting, deep learning with multilayer networks. Despite the success of convex methods, deep learning methods, where the objective is inherently highly non-convex, have enjoyed a resurgence of interest in the last few years and they achieve state-of-the-art performance. In the second part of the talk we move to the world of non-convex optimization where recent findings suggest that we might eventually be able to describe these approaches theoretically. The connection between the highly non-convex loss function of a simple model of the fully-connected feed-forward neural network and the Hamiltonian of the spherical spin-glass model will be established. It will be shown that under certain assumptions i) for large-size networks, most local minima are equivalent and yield similar performance on a test set, (ii) the probability of finding a bad local minimum, i.e. with high value of loss, is non-zero for small-size networks and decreases quickly with network size, (iii) struggling to find the global minimum on the

training set (as opposed to one of the many good local ones) is not useful in practice and may lead to overfitting. Discussion of open problems concludes the talk.

Bio

Anna Choromanska is a Post-Doctoral Associate in the Computer Science Department at Courant Institute of Mathematical Sciences, New York University. She is working in the Computational and Biological Learning Lab, which is a part of Computational Intelligence, Learning, Vision, and Robotics Lab, of prof. Yann LeCun. She graduated with her PhD from Columbia University, Department of Electrical Engineering, where she was the The Fu Foundation School of Engineering and Applied Science Presidential Fellowship holder. She was advised by prof. Tony Jebara. She completed her MSc with distinctions in the Department of Electronics and Information Technology, Warsaw University of Technology with double specialization, Electronics and Computer Engineering and Electronics and Informatics in Medicine. She was working with various industrial institutions, including AT&T Research Laboratories, IBM T.J. Watson Research Center and Microsoft Research New York. Her research interests are in machine learning, optimization and statistics with applications in biomedicine and neurobiology. She also holds a music degree from Mieczyslaw Karlowicz Music School in Warsaw, Department of Piano Play. She is an avid salsa dancer performing with the Ache Performance Group. Her other hobbies is painting and photography.

Faculty Host: Dimitris Metaxas/George Moustakides