Regressing Towards Simpler Prediction Systems

Tushar Chandra
Google, Inc.
tushar@google.com

Abstract

This talk will focus on our experience in managing the complexity of Sibyl, a large scale machine learning system that is widely used within Google. We believe that a large fraction of the challenges faced by Sibyl are inherent to large scale production machine learning and that other production systems are likely to encounter them as well[1]. Thus, these challenges present interesting opportunities for future research.

The Sibyl system is complex for a number of reasons. We have learnt that a complete end-to-end machine learning solution has to have subsystems to address a variety of different needs: data ingestion, data analysis, data verification, experimentation, model analysis, model serving, configuration, data transformations, support for different kinds of loss functions and modeling, machine learning algorithm implementations, etc. Machine learning algorithms themselves constitute a relatively small fraction of the overall system.

Each subsystem consists of a number of distinct components to support the variety of product needs. For example, Sibyl supports more than 5 different model serving systems, each with its own idiosyncrasies and challenges. In addition, Sibyl configuration contains more lines of code than the core Sibyl learner itself. Finally existing solutions for some of the challenges don't feel adequate and we believe these challenges present opportunities for future research.

Though the overall system is complex, our users need to be able to deploy solutions quickly. This is because a machine learning deployment is typically an iterative process of model improvements. At each iteration, our users experiment with new features, find those that improve the model's prediction capability, and then "launch" a new model with those improved features. A user may go through 10 or more such productive launches. Not only is speed of iteration crucial to our users, but they are often willing to sacrifice the improved prediction quality of a high quality but cumbersome system for the speed of iteration of a lower quality but nimble system.

In this talk I will give an example of how simplification drives systems design and sometimes the design of novel algorithms.

Categories and Subject Descriptors
I.2.6 Computing Methodologies, ARTIFICIAL INTELLIGENCE, Learning: Parameter learning

Keyword
Machine Learning

REFERENCE

[1] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young. Machine learning: The high interest credit card of technical debt. In SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop), 2014.

WSDM'15, February 2–6, 2015, Shanghai, China.
ACM 978-1-4503-3317-7/15/02.
http://dx.doi.org/10.1145/2684822.2697048