Report from the Workshop on Common Model Infrastructure, ACM KDD 2018

Chaitanya Baru
UC San Diego
baru@sdsc.edu

Abstract

Here we provide a summary of the Workshop on Common Model Infrastructure which was organized at KDD 2018 to focus on model lifecycle management and the tools and infrastructure needed to support development, testing, discovery, sharing, reuse, and reproducibility of machine learning, data mining, and data analytics models.

This workshop was organized by Chaitan Baru, UC San Diego; Amol Deshpande, University of Maryland; Robert Grossman, University of Chicago; Bill Howe, University of Washington; Jun Huan, Baidu; Vandana Janeja, University of Maryland Baltimore County; and, Arun Kumar, UC San Diego. The workshop organizers would like to acknowledge the strong support and encouragement provided by KDD Workshop Chairs, Jun Huan and Nitesh Chawla.

1 Motivation

The *Workshop on Common Model Infrastructure* was organized at KDD 2018 to focus on *model life-cycle management* and the tools and infrastructure needed to support development, testing, discovery, sharing, reuse, and reproducibility of machine learning, data mining, and data analytics models.

The workshop provided a forum for discussion of emerging issues, challenges, and solutions in this area, focusing on the principles, services and infrastructure needed to help data scientists of every ilk—from scientific researchers to industry analysts, and other practitioners—share data analytics models, reproduce analysis results, support transfer learning, and reuse pre-constructed models.

The continuing, rapid accumulation of large amounts of data and increasing use of machine learning, data mining, and data analytics techniques across a broad range of applications creates the need for systematic approaches for managing the increasingly complex of the modeling processes, given the large numbers of data-driven models being generated. However, current modeling practices are rather ad hoc, often depending upon the experience and expertise of individual data scientists and types of pre-processing used, which may be specific to domains. Different application domains/disciplines may use similar models and modeling tools, yet, sharing is limited; modeling results often have poor reproducibility; information on when/how a model works, and when it may fail, is oftentimes not clearly recorded; model provenance and the original intent behind the knowledge discovery process is not well-recorded; and, furthermore, many predictive analytics algorithms are not transparent to end-users. The CMI workshop was designed as a forum to discuss these and related issues.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

2 Workshop Format

CMI 2018 was a half-day workshop held on August 20th at KDD 2018, London, UK (`https://cmi2018.sdsc.edu/`). The workshop included three invited talks and two long lightning talks and four short lightning talks, selected from a set of submitted papers. Two of invited talks were from industry and one was from academia; four of the six lightning talks were from industry and two were from academia. The workshop ended with a group discussion, with all the speakers serving as an ad hoc panel. There were about 60 attendees at the workshop at any given time.

3 Invited Talks

The invited talk, *"What is the code version control equivalent for ML and data science workflows?"* by Clemens Mewald, Product Lead of TensorFlow Extended (TFX) from the Research & Machine Intelligence group at Google discussed how Google AI is addressing the challenges of managing the new paradigm of software development workflows being introduced in machine learning projects, which require code, data, and all derived artifacts (including models) to be indexed, updated, and shared. Traditionally, software developers primarily dealt with code that was compiled or interpreted, usually with deterministic behavior. With ML, products now rely on behaviors and patterns, often expressed as predictions, that are a function of code and evolving data plus models, leading to dynamic behaviors. The talk discussed the specific challenges faced by researchers and software engineers, and how systems like Google's TensorFlow Extended and TensorFlow Hub are addressing these issues.

In the invited talk, *"Why it is Important to Understand the Differences Between Deploying Analytic Models and Developing Analytic Models?"*, Professor Bob Grossman from the University of Chicago highlighted the two cultures in data science and analytics — one that is focused on *developing* analytic models, and the other that focuses on *deploying* analytic models into operational systems. The talk described the typical lifecycle of analytic models and provided an overview of some of the approaches that have been developed for managing and deploying analytic models and workflows, which included an overview of languages for analytic models, such as PMML, and for analytic workflows, such as PFA. It was posited that there is the emerging discipline of AnalyticOps that has borrowed some of the techniques of DevOps.

The third invited talk on, *"Beaker: A collaborative platform for rapid and reproducible research"* was delivered by Marc Milstone, leader of the Aristo Engineering team at the Allen Institute for Artificial Intelligence. The talk described Beaker, a robust experimentation platform for computational researchers that helps streamline the reproducible training, analysis and dissemination of machine learning results. Beaker was designed collaboratively with researchers at the Allen Institute for AI, focusing on ease-of-use for AI research workflows. The goal of the Beaker system is to provide a tool to help researchers deal with the details of organizing results and scaling experiments, so that they may spend more time on discovery and validation of new ideas that advance their field, rather than on the various ancillary details of running computational experiments. Beaker utilizes systems like Docker and Kubernetes to help reduce the "cognitive overhead" of infrastructure.

4 Lightning Talks

Papers for all the talks are available here `https://cmi2018.sdsc.edu/#accepted`.

One of the "long talks" was on *"Context: The Missing Piece in the Machine Learning Lifecycle"*, presented by Rolando Garcia, UC Berkeley, on behalf of co-authors, Vikram Sreekanti, Neeraja Yadwadkar, Daniel Crankshaw, Joseph E. Gonzalez, and Joseph M. Hellerstein. The ML lifecycle is characterized as consisting of three phases—Pipeline Development, Training, and Inferencing—and they identify that the crucial missing

piece within and across every phase is context, viz., "all the information surrounding the use of data in an organization." In current practice, the transitions between stages and teams are usually ad hoc and unstructured, which means that no single individual or system has a global, end-to-end view of the ML Lifecycle, leading to issues like irreproducibility, over-fitting, and missed opportunities for improved productivity, performance, and robustness.

The second "long talk" on *"Fighting Redundancy and Model Decay with Embeddings"* was presented by Dan Shiebler from the Twitter Cortex group, on behalf of co-authors Luca Belli, Jay Baxter, Hanchen Xiong, Abhishek Tayal. Twitter faces the issue that topics of interest are perpetually changing and evolving on the Internet at fairly rapid pace. Thus, models must keep up with the pace of change in the actual contents of the data streams. This talk detailed the commoditized tools and pipelines that Twitter has developed, and is developing, to regularly generate high quality, up-to-date embeddings and share them broadly across the company. Every day, hundreds of millions of new Tweets containing over 40 languages of ever-shifting vernacular flow through Twitter. Models that attempt to extract insight from this firehose of information must face the torrential covariate shift that is endemic to the Twitter platform. While regularly-retrained algorithms can maintain performance in the face of this shift, fixed model features that fail to represent new trends and tokens can quickly become stale, resulting in performance degradation. To mitigate this problem, Twitter employs learned features, or embedding models, that can efficiently represent the most relevant aspects of a data distribution. Sharing these embedding models across teams can also reduce redundancy and multiplicatively increase cross-team modeling productivity.

The remaining four presentation were "short talks". First was a talk on *"Recommender systems for machine learning pipelines"*, presented by Raymond Wright on behalf of the co-authors Insorge Silva and Ilknur Kaynar-Kabul, all from the SAS Institute Inc, about enhancing the existing SAS Model Studio interactive workbench to recommend one or more pipelines to use for a new, unseen dataset based on metafeatures derived from the dataset, since novice users often struggle with how to build or select a pipeline. The pipelines typically include feature generation, feature selection, model building, ensembling, and selection of a champion model, and represents a reproducible machine learning process. Once a general-purpose recommender model has been built using data from a variety of domains, the hope is that one can then generate specialized recommender rules for domains such as insurance, credit risk, etc. The work is inspired by recent work in automated machine learning, such as AUTO-SKLEARN, with the variation that here they aim to select from among a finite set of existing pipelines that have been found generally useful among users of the workbench.

Next, was a paper on *"Building a Reproducible Machine Learning Pipeline"* by Peter Sugimara and Florian Hartl from Tala Co, a startup company. The objective here is to automate and abstract the process from idea to deploying a machine learning model in production—which includes many steps—from machine learning practitioners, in order to improve modeling speed and quality, and realize key benefits like reproducibility of results. The presentation described the framework, which is comprised of four main components—data, feature, scoring, and evaluation layers—which are themselves comprised of well—defined transformations. This enables exact replication of models, and the reuse of transformations across different models. As a result, the platform can dramatically increase the speed of both offline and online experimentation while also ensuring model reproducibility.

The third presentation was on *"Knowledge Aggregation via Epsilon Model Spaces"* by Neel Guha who presented work done while he was a student at Stanford University. This work tackles scenarios where the machine learning task is divided over multiple agents, where each agent learns a different task and/or learns from a different dataset, which is a situation that can occur in many practical applications. The Epsilon Model Spaces (EMS), framework that was presented learns a global model by aggregating local learnings performed by each agent. This approach forgoes sharing of data between agents, makes no assumptions on the distribution of data across agents, and requires minimal communication between agents. Experiments were performed on the MNIST dataset, providing a validation of the methods used for both shallow and deep aggregate models. EMS is among the first to lay out a general methodology for "combining" distinct models. The EMS approach could help in future by allowing the development of "libraries" of models that could act as building blocks when

learning new models.

The fourth talk was by Microsoft on *"An Open Platform for Model Discoverability"*, presented by Yasin Hajizadeh, on behalf of co-authors Vani Mandava and Amit Arora. Recognizing that the need to rapidly experiment, share and enable collaboration at scale between a fast-growing community of machine learning experts and domain experts has created an acute need for a platform for management, discoverability, and sharing of machine learning models, Microsoft has developed a novel cloud-based AI Gallery that enables data scientists and domain experts to seamlessly share and collaborate on machine learning solutions. While there are siloed systems that address this issue, there is no single comprehensive platform that addresses all these challenges. The Azure AI Gallery, `https://gallery.azure.ai/`models, provides a gallery of models, projects and solution templates for data scientists, application developers and model managers that works with local and cloud environments. The gallery currently has several ONNX models from `https://onnx.ai` and is expected to grow as more models are contributed.

5 Summary

The first community workshop on *"Common Model Infrastructure"*—informally referred to as the *ModelCommons*—held at KDD 2018, London, UK was, indeed, a success. There was significant industry presence among the speakers at the workshop as well as in the audience, indicating a strong industry interest in the topic, as well as the relevance of this topic area to practical applications. The workshop concluded with a panel of all presenters engaging in some question and answers with the audience, but also in a group discussion. First, it was felt that this is an important topic that should be continued in the future, including with a workshop at the next KDD Conference in Anchorage, Alaska in 2019. Indeed, the next meeting on this topic is sponsored by Baidu and will occur at the NIPS Expo on December 2, in conjunction with the NIPS 2018 Conference in Montreal, Canada . There was also a discussion of the appropriate venue for publication of full-length papers from this workshop. One idea was to request for a special issue of the new ACM Transactions on Data Science.

There was an important discussion on whether "Common Model Infrastructure" was sufficiently descriptive of the range of issues that fall under the "ML lifecycle". There was agreement that a more descriptive term is needed that indicates coverage of the broad set of issues—some of which may be research while others may be infrastructural in nature—but the group did not discuss what the new description should be. Future workshop may well adopt a different name for this set of topics.