Data challenges with modularization and code submission

Lessons learned

Full text

Summary

Motivated by the shortcomings of traditional data challenges, we have developed a unique concept and platform, called Rapid Analytics and Model Prototyping (RAMP), based on **modularization and code submission**. Open code submission allows participants to **build on each other's ideas**, provides the organizers with a **fully functioning prototype**, and makes it possible to build **complex machine learning workflows** while keeping the contributions simple. We will start this presentation by describing the **context and motivation**, the guiding **design principles**, and some of the **technical details** (front and backend) of the platform. We will then walk you through some of the most interesting workflows and applications (e.g., **anomaly detection in particle physics** detectors, **classifying molecular spectra** for safe drug administration, **spatio-temporal time series** prediction in climate science). In the last third of the talk we will present a preliminary analysis of the RAMPs that touches on both the technical (machine learning) aspects of the tool and on the **management of crowdsourcing** data analytics.

Extended abstract

A data challenge is a recently developed unconventional dissemination and communication tool. Well-organized challenges are immensely useful for **generating visibility** in the data science community about novel application domains. Our HiggsML challenge attracted almost **2000 participants** and generated **30000 visits** on our main site. On the other hand, they are **not adapted to solving complex and open-ended data science problems** in realistic environments. Neither the challenge host nor the organizers have **automatic access to the software** that generated the solutions or to the **participants** who are typically spread around the world. They also emphasize competition between individuals or small teams, so they are diametrically opposed to the open source scheme where collaboration is the main driver.

Motivated by the shortcomings of data challenges, the Paris-Saclay Center for Data Science (CDS) has developed a unique platform called **Rapid Analytics and Model Prototyping (RAMP).** RAMPs are **data challenges with modularization and code submission**. They are combining distributed open-source development and data challenges. Code submission solves three of the main bottlenecks of classical (prediction submission) data challenges. First, at the end of the challenge, the **organizers have access to a fully functioning prototype** which can be easily transferred into a production pipeline. Second, opening the code for free consultation among participants means that that they can **build on each other's ideas**. Third, submitting code means that we can develop **complex workflows that would be impossible in a classical challenge due to data leakage** (e.g., time series forecasting).

In the last 24 months we have organized **twelve single day hackathon-style RAMPs**. In each RAMP, we invite typically 30-50 researchers and students around a scientific data science problem. In the morning of the RAMP, the **data provider explains the problem and the data set**, then the participants tackle the problem during the day, **guided by coaches from the core CDS team**. Our scientific problems range from particle and astrophysics, through climate science (El Niño and Arctic sea ice forecasting) and biodiversity, to health care. Following our local success, we were invited to several events outside Saclay (the two last Climate Informatics workshops, to the Paris School of Economics, and to the Epidemium hackathon). At each RAMP, thanks to the collaborative format, we improve the baseline performance significantly, sometimes spectacularly.

The RAMP platform was originally designed for a **collaborative prototyping** tool that makes efficient use of the time of data scientist in solving the data analytics segment of high-impact domain science or business problems, but it became clear from the beginning that the tool has great value for **hands-on training of data scientists**. About half of our participants attended the RAMPs for learning data science. We have been using RAMPs in **professional training** and in **several M.Sc. programs** at Université Paris-Saclay. From the next school-year on, all of the roughly **500 M.Sc-level data science students** of UPSaclay will go through a RAMP-based Data Camp course at Polytechnique, Telecom, UPSud, UVSQ, ENS, CentraleSupelec, and Mines.