Deploying Machine Learning at Web Scale

Christoph Schmitz

1&1 Mail & Media Development & Technology GmbH

Presentation Abstract

1&1 uses machine learning on some of the largest German web portals with practical challenges which are underrepresented in the academic literature. In our presentation, we will discuss these and some of our solutions in practice.

Data. Data quality is a major concern when integrating data sources within the company. The hardest problems in our production environment are gradual degradations in data quality. The root causes are hard to find, often occuring several steps upstream in the data pipeline. Organizational constraints can impede the collection of good quality data. Data sets from questionnaires can be skewed and thus require considerable preprocessing to be usable.

Modeling. Machine learning tools are usually targeted at an exploratory, interactive work flow. Building and maintaining hundreds of models at the same time leads to other requirements, though. We treat models like code, using versioning, continuous integration, and deployment strategies from software development. Much of the training work flow is automated, allowing a small team of data scientists to maintain models for a large number of target groups.

Constraints. In our applications, constraints are important when assessing the quality of models. One major example is the joint distribution of target variables with the age and gender of customers. Thus, measuring and visualizing these additional constraints is part of our modeling work flow. We are also looking into including these constraints directly in the training of models itself.

Processing. To be able to efficiently score more than 300 models, we use custom planning logic to split data flows into a minimal number of MapReduce jobs. Again, the common machine learning tools are not made for this. We will discuss challenges and solutions embedding Weka into a Hadoop application, e. g., schema handling, missing values, and dealing with errors.

Keeping Up. Since big data technology evolves at a breakneck pace, we need to trade off missing the latest features or the newest frameworks against the considerable cost of updating dozens of machines. While vendors promise hassle-free rolling upgrades, in practice upgrades are much more involved and entail considerable risk, effort, and organizational overhead.

 In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org