Putting Machine Learning into Production Systems

DATA VALIDATION AND SOFTWARE ENGINEERING FOR MACHINE LEARNING

ADRIAN COLYER

This time around with The Morning Paper I've chosen two papers that address different aspects of putting machine learning into production systems. In "Data Validation for Machine Learning," Breck et al. share details of the pipelines used at Google to validate petabytes of production data every day. With so many moving parts it's important to be able to detect and investigate changes in data distributions before they can impact model performance. As a bonus, the data-validation library at the core of Google's approach has also been made available in open source so that you can experiment with it, too (https://github.com/tensforflow/data-validation).

"Software Engineering for Machine Learning: A Case Study" shares lessons learned at Microsoft as machine learning started to pervade more and more of the company's systems, moving from specialized machine-learning products to simply being an integral part of many products and services. This means that software-engineering processes and practices on those projects have had to adapt. This paper demonstrates once again the importance of a rock-solid data pipeline, as well as some of the unique challenges that machine learning presents to development projects.

For the complete column, go to
https://queue.acm.org/TheMorningPaper/41

Adrian Colyer *is a venture partner with Accel in London, where it's his job to help find and build great technology companies across Europe and Israel. (If you're working on an interesting technology-related business, he would love to hear from you at acolyer@accel.com.) Prior to joining Accel, he spent more than 20 years in technical roles, including CTO at Pivotal, VMware, and SpringSource.*

CONTENTS