Automatic Feature Assembling for Event Prediction

Project proposal for CPSC 490

Felipe Pires (felipe.pires@yale.edu)

Spring 2019

A well-established application of data science is in trying to predict the future. Banks use data science to assess borrowers’ risk of default; big retail uses it to forecast upcoming demand; healthcare, to calculate a patient’s chances of developing lifestyle diseases etc. Today, state-of-the-art workflow to answer many such questions look very similar, regardless of the industry: (1) data scientists use programming to take historical data and process them into features, then (2) feed them to use case-agnostic machine learning frameworks, like TensorFlow or Keras, which learn to generalize the present to predict the future. The models that these frameworks create learn to identify risky borrowers, unhealthy patients, trends in customer demand etc. While the second half of this pipeline has seen many advances over the years, with ML frameworks becoming increasingly sophisticated and performant, the work of turning historical data into features has remained practically the same: somewhat repetitive, yet still done by data scientists on a case-by-case basis.

We propose creating a framework that partly automates the assembly of features for event-prediction modeling. Given a set of historical records and a description of features to generate, the framework should be able to assemble these features without the help of a data scientist. The benefits of abstracting away the low-level details of manipulating historical data using programming, we believe, goes beyond
the time that will be saved by data scientists. Manipulating at a higher level will allow reasoning about prediction at a higher level. Feature selection, for instance, the work of discovering which features are best suited to solve a certain problem, will become much easier.

Work will begin by automating the creation of “lag features”, which we found to be standard practice in building event-prediction models.¹

As much as we already have an idea of which types of features to automate, part of the effort in putting this framework together will be to bring to light common feature engineering strategies. For this, we will explore solutions to famous data science problems available online, through websites like Kaggle and TopCoder.