Engine of Automatic Machine Learning

Project proposal for CSPC 490

September 4, 2018

Felipe Pires ([felipe.pires@yale.edu](mailto:felipe.pires@yale.edu)), author
Kyle Jensen ([kyle.jensen@yale.edu](mailto:kyle.jensen@yale.edu)), advisor

Summary

I propose, as a final project for my B.Sc. in Computer Science, the development of a program capable of solving generic machine learning problems without the need for human intervention (ie. to find the most appropriate model to train, to tune hyperparameters etc). Given a simple description of a machine learning problem and an appropriate dataset to apply it on, the program should be able to train a model to solve it as well as it can.

Motivation

Over the past couple of years, the adoption of Machine Learning by companies outside of the high-tech sector has continued to underwhelm the expectations of analysts.\(^1\) According to the McKinsey Global Institute, one leading cause of the lag in adoption is the difficulty in hiring data scientists, professionals with background in Computer Science and Statistics who are able to use Machine Learning tools to drive growth inside the companies. Indeed, there is said to be a fast-growing shortage for this type of professional in the US. A report published last year by IBM Analytics estimated three million jobs in the US for “data professionals” by 2020.\(^2\)

Despite this widespread difficulty that industries have found in applying Machine Learning, relatively little attention has been paid in the applied AI community to the

---

\(^1\) “Artificial Intelligence, The Next Digital Frontier - McKinsey.”

\(^2\) “IBM Predicts Demand For Data Scientists Will Soar 28% By 2020.” 13 May. 2017,
problem of making the tools it develops more accessible to laymen. This suggests a dissonance between the industry needs from the AI community and the community is focusing on.

We speculate that this dissonance is aggravated by the significant influence that high-tech companies have recently had in driving innovation in the sector. Companies like Google and Facebook are big patrons of AI research, dedicating hundreds of millions of dollars to fund studies both internally and through academia. ³ ⁴ The fact that Google and Facebook have can easily apply the machine learning technologies that they fund or develop themselves causes little attention to be paid to the needs of other, low-tech companies. To make matters worse, it is much easier for high-tech companies to attract and keep top talent in data science, leaving less powerful companies having to pay way above market average.⁵

Solution

We propose addressing this issue by creating an easy-to-use library that gives users with elementary knowledge of programming the ability to easily solve a set of key machine learning problems, just by providing the dataset for training and little else.

High-level Specification

The product of this project will be a Python library that exposes a small number of functions, for training new models, running previously trained models, and (time permitting) for cleaning datasets and detecting anomalies in them.

We will be leveraging existing open-source ML libraries to create different models to solve different types of problems. Over the past few years, there has been a surge in the number of high-quality ML frameworks being open-sourced by tech companies and teams from academic. The most popular example is perhaps Tensorflow, a

---


⁵ Where’s the source for this?
framework for Deep Learning which powered AI applications inside Google for years, before being open-sourced in late 2015. It is today typical for big tech companies and startups to have open-source AI frameworks of their own.

The program will be written in the Python programming language. Python is the most popular language for data science work today and the majority of popular Machine Learning frameworks have a stable API for it, including Tensorflow and XGBoost.6

Project Deliverables
Here is a rough list of possible project deliverables, which may be worked on in a random order.

**Use XGBoost for classification tasks**
Implement a script that builds on top of XGBoost (https://github.com/dmlc/xgboost) to solve machine learning classification problems.

**Use SKLearn for clustering tasks**
Scikit-learn (http://scikit-learn.org) is a popular Python framework for running miscellaneous machine learning tasks. It implements several types of clustering algorithms, which we can use to solve clustering tasks.7

**Use Prophet for forecasting tasks**
Prophet (https://facebook.github.io/prophet) is a Python framework for modeling of time series, which we can use to solve forecasting tasks.

**Automatic hyperparameter search**
Use a framework that does hyperparameter search to automatically find good hyperparameters for the models we build without the need for human intervention.

---

6 “The Most Popular Language For Machine Learning and Data Science Is.”

7 “The 5 Clustering Algorithms Data Scientists Need to Know.” 5 Feb. 2018,
**Ability to store/load models in the cloud**

Use-cases of machine learning in the industry typically involve training a model based on the big data that the company or a partner company generates, and then keeping using this trained model to generate insights over a period of time. This means that an important feature of a easy-to-use ML library is the ability to serialize models and save them, so they can be later deserialized and run again for new data.