AliCloud

Community

6 Top AutoML Frameworks for Machine Learning Applications (May 2019)

Alibaba Clouder September 2, 2019 48,582 1

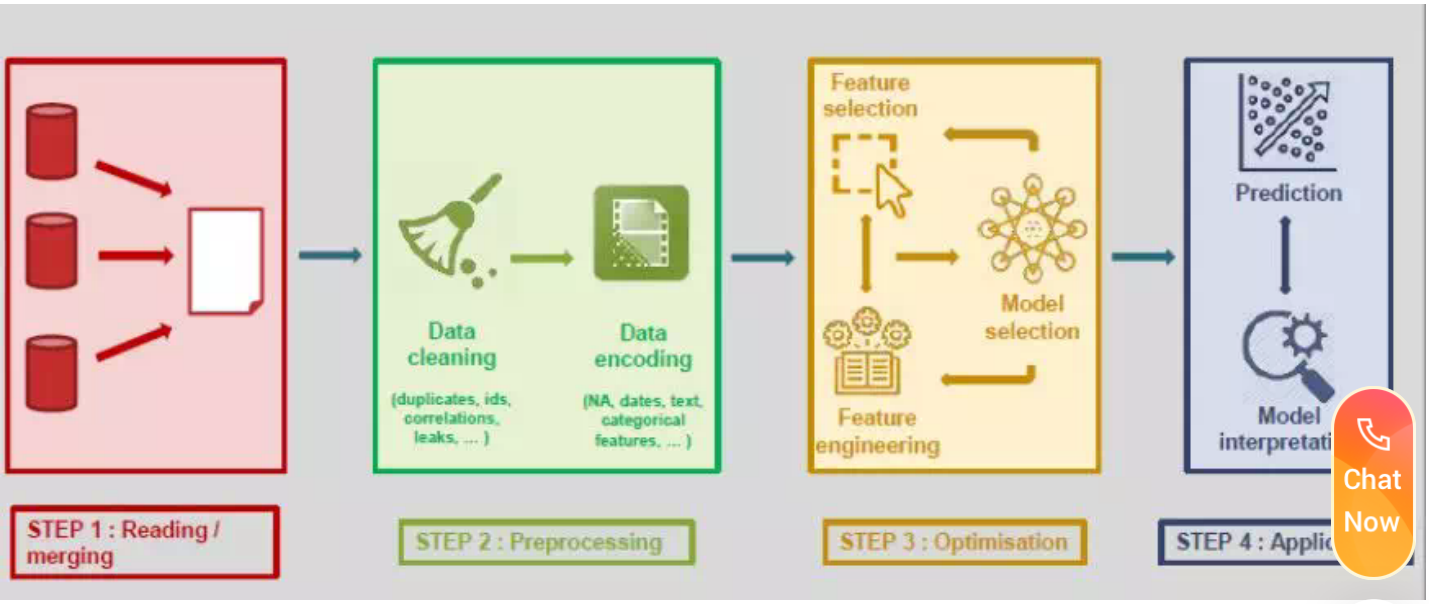
In this post, we 6 key automated machine learning (AutoML) platforms that can assist data scientists to accelerate machine learning development.

1. What Is AutoML?

1.1 Overview

AutoML (automated machine learning) refers to the automated end-to-end process of applying machine learning in real and practical scenarios.

A typical machine learning model includes the four following steps:



Chat Now

From data reading, pre-processing, optimization, and result prediction, each step is controlled and performed manually. AutoML focuses on two main aspects: data collection and prediction. Any

other intermediate steps can be easily automated. In addition, AutoML provides models that have been optimized and ready for prediction.

Currently, AutoML mainly falls into three categories: 1. AutoML for automated parameter tuning (a relatively basic type) 2. AutoML for non-deep learning, for example, AutoSKlearn. This type is mainly applied in data pre-processing, automated feature analysis, automated feature detection, automated feature selection, and automated model selection. 3. AutoML for deep learning/neural networks, including NAS and ENAS as well as Auto-Keras for frameworks.

1.2 Why Is AutoML Needed?

From the application perspective, the demand for machine learning systems has soared over the past few years. ML has been adopted in a wide range of applications. However, although it is proven that machine learning can provide better support for some enterprises, many enterprises are still struggling to implement ML model deployment.

Theoretically, one goal of AI is to replace a portion of manpower. Specifically, a large part of the AI design work can also be implemented by using proper algorithms. Take parameter tuning for example: Algorithms like Bayes, NAS, and evolutionary programming can be used in the parameter tuning process to replace manpower by allowing more computing power.

To deploy AI models, an enterprise first needs to have a team of experienced data scientists, who expect high salaries. Even if an enterprise does have an excellent team, usually more experience rather than AI knowledge is needed to decide which model best fits the enterprise. The success of machine learning in a variety of applications leads to an increasingly higher demand for machine learning systems, which are supposed to be easy-to-use even for non-experts. AutoML tends to automate as many steps as possible in ML pipelines and retain good model performance with minimum manpower.

AutoML has three major advantages:

- Improve efficiency by automatically running repetitive tasks. This allows data scientists to focus more on problems instead of models.
- Automated ML pipelines also help avoid potential errors caused by manual work.
- AutoML is a big step toward the democratization of machine learning and allows everyone to use ML features.

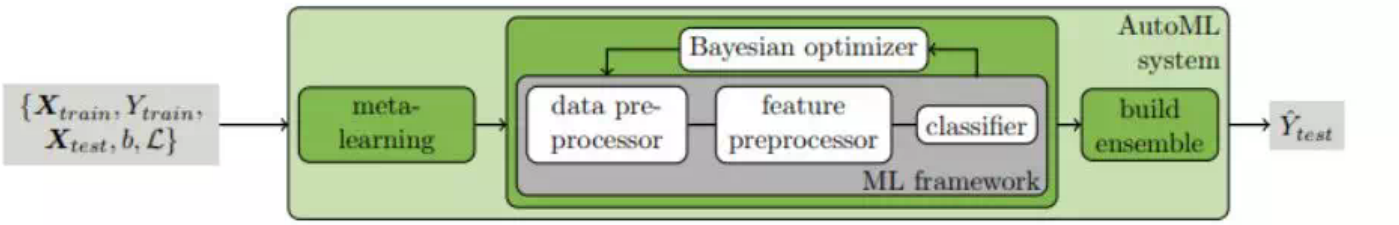
2. Common AutoML Frameworks (as of May 2019)

AutoML has a history of many years. Since last year, many excellent AutoML frameworks have emerged. This article only briefly describes several common frameworks. Subsequent articles will give more information about the use and performance of these frameworks.

2.1 Auto-SKLearn

Since SKLearn is usually a choice for getting started, let's first talk about AutoSKLearn.

Auto-SKLearn is an automated machine learning software package built on scikit-learn. Auto-SKLearn frees a machine learning user from algorithm selection and hyper-parameter tuning. It includes feature engineering methods such as One-Hot, digital feature standardization, and PCA. The model uses SKLearn estimators to process classification and regression problems.



Auto-sklearn pipeline

Auto-SKLearn creates a pipeline and uses Bayes search to optimize that channel. In the ML framework, two components are added for hyperparameter tuning by means of Bayesian reasoning: Meta learning is used to initialize optimizers using Bayes and evaluate the auto collection construction of the configuration during the optimization process.

Auto-SKLearn performs well in medium and small datasets, but it cannot produce modern deep learning systems with the most advanced performance in large datasets.

Demo

The following example shows how to use Auto-SKLearn to fit a simple regression model.

```
import sklearn.model_selection
import sklearn.datasets
import sklearn.metrics

import autosklearn.regression

def main():
X, y = sklearn.datasets.load_boston(return_X_y=True)
feature_types = (['numerical'] * 3) + ['categorical'] + (['numerical'] * 9)
X_train, X_test, y_train, y_test = \
sklearn.model_selection.train_test_split(X, y, random_state=1)

automl = autosklearn.regression.AutoSklearnRegressor(
time_left_for_this_task=120,
```

```
per_run_time_limit=30,
tmp_folder='/tmp/autosklearn_regression_example_tmp',
output_folder='/tmp/autosklearn_regression_example_out',
)
automl.fit(X_train, y_train, dataset_name='boston',
feat_type=feature_types)

print(automl.show_models())
predictions = automl.predict(X_test)
print("R2 score:", sklearn.metrics.r2_score(y_test, predictions))


if __name__ == '__main__':
main()
```

Resource link:

GitHub: https://github.com/automl/auto-sklearn

2.2 MLBox

MLBox is a powerful Automated Machine Learning python library. According to the official document, it provides the following features:

- Fast reading and distributed data preprocessing/cleaning/formatting
- Highly robust feature selection and leak detection as well as accurate hyper-parameter optimization
- State-of-the art predictive models for classification and regression (Deep Learning, Sta LightGBM,...)
- Prediction with model interpretation
  MLBox has been tested on Kaggle and shows good performance. (See Kaggle "Two Sigma Connect：Rental ListingInquiries"| Rank：85/2488)
- Pipeline

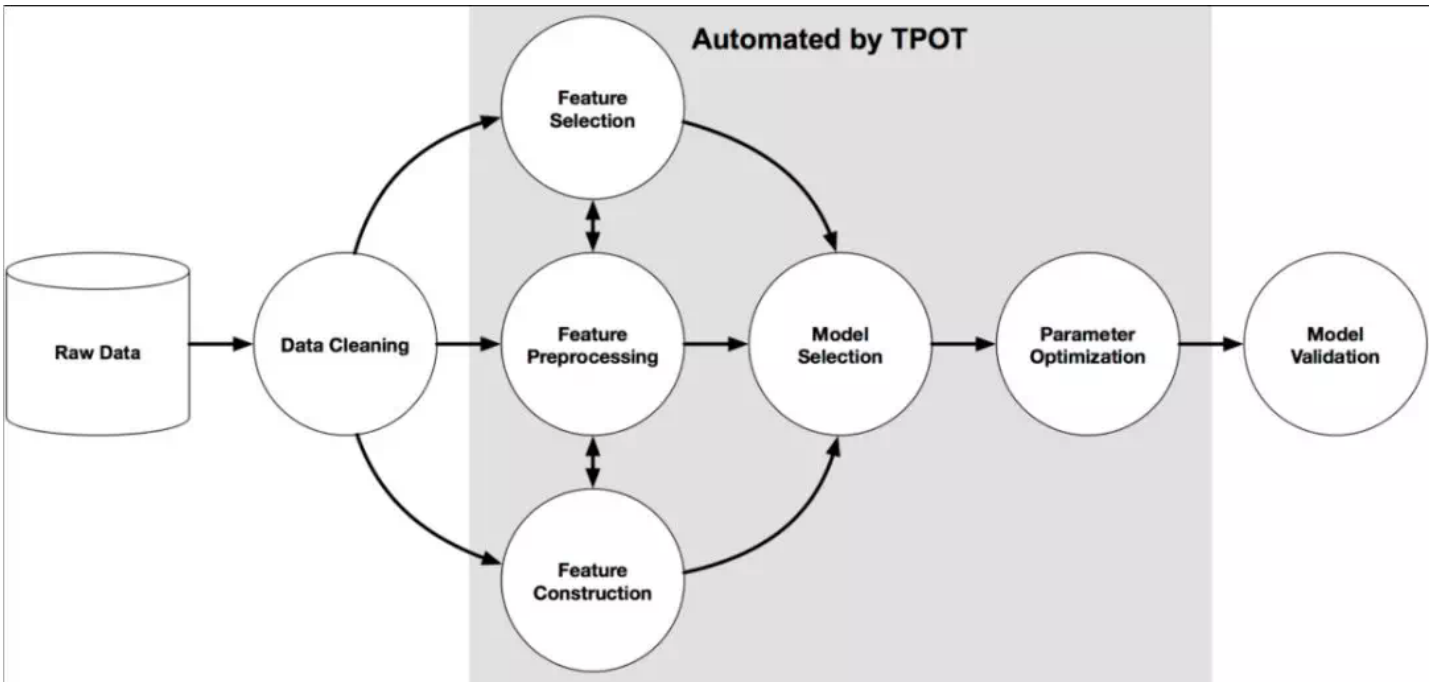
MLBox architecture

MLBox main package contains 3 sub-packages:

- Pre-processing: reading and pre-processing data
- Optimization: testing or optimizing a wide range of learners
- Prediction: predicting the target on a test dataset

2.3 TPOT

TPOT is a tree-based pipeline optimization tool that uses genetic algorithms to optimize machine learning pipelines. TPOT is built on top of scikit-learn and uses its own regressor and classifier methods. TPOT explore thousands of possible pipelines and finds the one that best fit the data.



Resource link:

GitHub: https://epistasislab.github.io/tpot/

2.4 H2O AutoML

H2O is an open source and distributed in-memory machine learning platform developed by H2O.ai. H2O supports both R and Python. It supports the most widely used statistical and machine learning algorithms including gradient boosted machines, generalized linear models, deep learning and more.

H2O includes an Automated Machine Learning module and uses its own algorithms to create pipelines. It uses exhaustive search for feature engineering methods and model hyper-parameters to optimize pipelines.

H2O automates some complex data science and machine learning tasks, such as feature engineering, model validation, model adjustment, model selection, and model deployment. In addition, it provides automatic visualization and machine learning interpretation (MLI).

Resource link: https://h2o-release.s3.amazonaws.com/h2o/master/3888/docs-website/h2o-docs/downloading.html

2.5 Auto-Keras

Auto-Keras is an open source software library for automated machine learning (AutoML) developed by DATA Lab. Built on top of the deep learning framework Keras, Auto-Keras provides functions to automatically search for architecture and hyper-parameters of deep learning models.

Auto-Keras follows the classic Scikit-Learn API design and therefore is easy to use. The current version provides the function to automatically search for hyper-parameters during deep learning.

In Auto-Keras, the trend is to simplify ML by using automatic Neural Architecture Search (NAS) algorithms. NAS basically uses a set of algorithms that automatically adjust models to replace deep learning engineers/practitioners.

Resource link: https://github.com/keras-team/autokeras

2.6 TransmogrifAI

Now let's look at a killer library - TransmogrifAI released by Salesforce in 2018.

Einstein, a flagship ML platform of Salesforce, is also powered by TransmogrifAI. TransmogrifAI is an end-to-end AutoML library for structured data written in Scala that runs on top of Apache Spark. It automates feature analysis, feature selection, feature validation, model selection and more. TransmogrifAI is especially useful in the following scenarios:

- Quickly train quality machine learning models with minimal manual adjustment
- Construct modular, reusable, and strongly-typed machine learning workflows

3. Thoughts on AutoML

The essence of AutoML is to automate repetitive tasks such as pipeline creation and hyper-parameter tuning so that data scientists can spend more time on business problems on hand in practical scenarios. AutoML also allows everyone instead a small group of people to use the

machine learning technology. Data scientists can accelerate ML development by using AutoML to implement really efficient machine learning.

Whether AutoML will be a success depends on its usage and progress in the machine learning field. Obviously, AutoML is an important part of machine learning in the future.

Appendix

- https://ml.informatik.uni-freiburg.de/papers/15-NIPS-auto-sklearn-preprint.pdf
- Benchmarking Automatic Machine Learning Frameworks

Artificial Intelligence | Deep Learning | Machine Learning | AutoML | Automated Machine Learning

Read previous post:

MVP #FridayFive: August 2019 Edition 5

Read next post:

Alibaba Open-Source and Lightweight Deep Learning Inference Engine - Mobile Neural Network (MNN)

Introduction to EMR DataScience

Alibaba EMR - August 19, 2020

Shortening Machine Learning Development Cycle with AutoML

Alibaba Clouder - November 29, 2018

What Is Machine Learning?

Alibaba Clouder - December 30, 2020

Building a High-Level Frontend Machine Learning Framework Based on the tfjs-node

Alibaba F(x) Team - December 10, 2020

Comments

5232790373409361

May 25, 2020 at 2:24 am

Great summary. Thanks for the detailed analysis. Our team evaluated all of the common AutoML frameworks and selected Auger, www.auger.ai. In addition to the benefits of AutoML referenced above - hyperparameter tuning, algorithm selection, Python APIs, etc. - Auger provides an automation API on top of AutoML called A2ML that automates the retraining and data pipeline management process. It's ideal for developers who aren't Data Scientists. Happy to share details from our analysis with anyone interested. -Dan Turchin (dan@peoplereign.io)

0 0

Write your comment...

Post

Chat Now

More Posts by Alibaba ...

See All >

- MyBatis with a More Fluent Experience
- Alibaba Cloud Sustainability Report 2021
- Comparing CNI Models in Container Service for Kubernetes — Alibaba Cloud Series Part 1
- [Infographic] 5 Steps to Accelerate Your Digitalization in Asia
- Attackers Use the Vulnerability of ShowDoc to Spread Botnets
- Powerful: MyBatis and Three Streaming Query Methods
- What is the Difference between Spring Boot and Spring?
- What are the Differences and Functions of the Redo Log, Undo Log, and Binlog in MySQL?
- On the In-Depth Cluster Scheduling and Management
- Alibaba Technological Practices: Experiences in Cloud Resource Scheduling