Adding New Machine Learning Models to DFFML

Google Summer of Code

Yash Lamba
Cluster Innovation Centre, University of Delhi
Mentor:- John Andersen, Open Source Security Engineer at Intel Corporation

python SOFTWARE FOUNDATION

Abstract

Data Flow Facilitator for Machine Learning (DFFML) provides APIs for dataset generation and storage, and model definition using any machine learning framework, from high level down to low level use is supported.

The goal of DFFML is to build a community driven library of plugins for dataset generation and model definition. So that we as developers and researchers can quickly and easily plug and play various pieces of data with various model implementations.

During the community bonding period, the proposed work was modified to achieve optimized result from the summer. The finalized work was:

1. Adding Linear Regression Model from scratch
2. Adding Linear Regression and other proposed models using scikit-learn
3. Adding tests for the added models
4. Documenting the models

Accomplished Tasks

1. Added Linear Regression model from scratch with tests:

Simple Linear Regression model implemented from scratch. This was successfully completed with tests and documentation, and was also released on PyPI.

2. Added scikit models with dynamic support :

Initially, it was planned to add certain number of models from scikit but as we completed one model (Multiple Linear Regression with scikit), we decided to extend this and make a base for all scikit models and make other model classes dynamic. This was successful and now adding scikit models to DFFML is as easy as appending the model name to a python dictionary. The tests are complete and the documentation material is ready but we are still figuring out a more understandable way of documenting this before release.

Future Work

The project was started just before GSoC'19 and it has come a long way since. I plan on contributing significantly to the project after GSoC'19. Few of the planned stuff:

1. Adding more scikit models
2. Working on more machine learning libraries and add models
3. Construct DFFML Web UI from scratch which was conceptulized during summer and much more.

Example Integration

Dataset for Linear Regression

Years of Experience	Expertise	Trust Factor	Salary
0	01	0.2	10
1	03	0.4	20
2	05	0.6	30
3	07	0.8	40
4	09	1.0	50
5	11	1.2	60

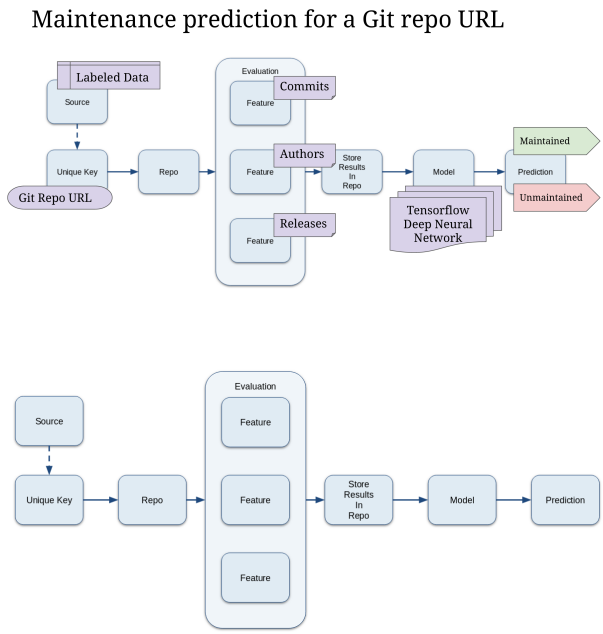
Available Models

Type	Model	Entrypoint
Regression	LinearRegression	scikitlr
Classification	KNeighborsClassifier	scikitknn
	AdaBoostClassifier	scikitadaboost
	GaussianProcessClassifier	scikitgpc
	DecisionTreeClassifier	scikitdtc
	RandomForestClassifier	scikitrfc
	QuadraticDiscriminantAnalysis	scikitqda
	MLPClassifier	scikitmlp
	GaussianNB	scikitgnb

Using DFFML

```
$ cat > train.csv << EOF
Years,Expertise,Trust,Salary
0,1.0,0.2,10
1,3,0.4,20
2,5,0.6,30
3,7,0.8,40
EOF
$ cat > test.csv << EOF
Years,Expertise,Trust,Salary
4,9,1.0,50
5,11,1.2,60
EOF
$ dffml train \
  -model scikitlr \
  -features def:Years:int:1 def:Expertise:int:1 def:Trust:float:1 \
  -model-predict Salary \
  -sources f=csv \
  -source-filename train.csv \
  -source-readonly \
  -log debug
$ dffml accuracy \
  -model scikitlr \
  -features def:Years:int:1 def:Expertise:int:1 def:Trust:float:1 \
  -model-predict Salary \
  -sources f=csv \
  -source-filename test.csv \
  -source-readonly \
  -log debug
1.0
$ echo -e 'Years,Expertise,Trust\n6,13,1.4\n' | \
 dffml predict all \
  -model scikitlr \
  -features def:Years:int:1 def:Expertise:int:1 def:Trust:float:1 \
  -model-predict Salary \
  -sources f=csv \
  -source-filename /dev/stdin \
  -source-readonly \
  -log debug
[
  {
    "extra": {},
    "features": {
      "Expertise": 13,
      "Trust": 1.4,
      "Years": 6
    },
    "last_updated": "2019-09-18T19:04:18Z",
    "prediction": {
      "confidence": 1.0,
      "value": 70.00000000000001
    },
    "src_url": 0
  }
]
```

DFFML Architecture



References:

1. github.com/intel/dffml
2. intel.github.io/dffml/master
3. summerofcode.withgoogle.com/projects/#5733002032709632

Cluster Innovation Centre University of Delhi