matsML: a machine-learning toolkit for materials science

Huan Tran
Georgia Institute of Technology
huan.tran@mse.gatech.edu

August 27, 2021

Contents

1 Overview

matsML is a python-based machine-learning (ML) toolkit for some generic problems in materials science. Being initiated to support some of my (Huan Tran) non-polymer ML works, matsML was designed to be portable and self-contained, providing necessary materials to follow the workflows and reproduce the results reported. Given this objective, actual scripts used for my works can be found in some examples of matsML while others are for tutorial purposes and beyond. matsML is free at `https://github.com/huantd/matsml.git`.

A typical workflow of materials informatics includes preparing/generating/collecting suitable data, featurizing (or fingerprinting) the data, learning the featurized data to make models, using the developed models to make predictions, inverting the models to solve inverse problems, and more. This toolkit *does not* aim at providing complete solutions to any of these steps. However, demonstrations for most of the typical workflow can be found within matsML, specifically in `examples/ex1_pcm-molecs` and others.

Most of the computed data referred to in this toolkit are from my works, others are open reported data. In cases of experimental data that are subjected to copyright and ownership, suitable freely available alternatives are provided for demonstration purpose. Questions, requests, and comments are welcome at `huan.tran@mse.gatech.edu`.

2 Technical details

2.1 Code and installation

matsML is freely available at https://github.com/huantd/matsml.git. Dependencies needed for MATSML are

1. numpy (https://numpy.org/)
2. pandas (https://pandas.pydata.org/)
3. scikit-learn (https://scikit-learn.org/stable/)
4. keras (https://keras.io/)
5. tensorflow (https://www.tensorflow.org/)
6. tensorflow_probability (https://www.tensorflow.org/probability)
7. ase (https://wiki.fysik.dtu.dk/ase/)
8. dscribe (https://singroup.github.io/dscribe/latest/index.html)
9. matplotlib (https://matplotlib.org/)

Users are suggested to (1) create a devoted environment for this toolkit, e.g., using conda with python3, (2) install the dependencies, (3) obtain matsml by

```
git clone https://github.com/huantd/matsml.git
```

and then (4) install the source code of matsml by entering the matsml folder and issuing

```
python setup.py install
```

2.2 Functionalities

1. Some datasets (see Table XX for a summary) provided at www.matsml.org and used for the toolkit
2. Fingerprinting molecules with projected Coulomb matrix and crystals with projected Ewald sum matrix
3. Building and training ML models using Kernel Ridge Regression (with scikit-learn), Gaussian Process Regression ((with scikit-learn)), fully connected Neural Network (with TensorFlow), and probability Neural Network (with TensorFlow-Probability).
4. Load trained models and make predictions on new cases

3 Details

3.1 Data

Two classes of data used in matsml (and obtained from www.matsml.org) are raw data and featured data. Generally, raw data are similar to those that can be obtained from various databases, and they need to be featured before learning. For some reasons, raw data are not available for some examples, and in these cases, featured data are provided for learning step.

3.1.1 Raw data

Raw data that will be featurized are atomic structures of materials, and they should be summarized in a csv-format file with the following structure

```
file_name,prop
CF4-00001.xyz,-.25466963E+02
CF4-00002.xyz,-.25357728E+02
CF4-00003.xyz,-.25463676E+02
CF4-00004.xyz,-.25312495E+02
```

3.1.2 Featurized data