Project: Molecular Energy Prediction

May 3, 2022

The goal of the challenge is to model the inter-atomic potential energy surface of small organic molecules [1]. We denote by $\mathbf{r} = \{r_1, r_2, ..., r_N\}$ the positions of atoms in a molecule in 3D space, and denote by $E(\mathbf{r})$ the atomization energy of the configuration $\mathbf{r}$. One should define a unified model for all the molecules to predict $E(\mathbf{r})$, based on the geometric information represented by $\mathbf{r}$ and some extra information about the atoms. This amounts to a high dimensional regression problem.

The most challenging component of this problem is to respect the symmetry constraint. In other words, we define the translational, rotational, and permutational operations of $\mathbf{r}$ as:

$$T_b(\mathbf{r}) = \mathbf{r} + b, \quad T_U(\mathbf{r}) = U\mathbf{r}, \quad T_\sigma(\mathbf{r}) = \{r_{\sigma(1),\sigma(2),\cdots,\sigma(N)}\}$$

The symmetry constraint implies that

$$E(T_b(\mathbf{r})) = E(T_U(\mathbf{r})) = E(T_\sigma(\mathbf{r})) = E(\mathbf{r})$$

for all possible translation b, rotation U and permutation σ (acting on the N particles in the 3D domain).

Data description

We use a subset of QM7-X, which contains 8466 structures of molecules, with various number of particles (atoms). QM7-X is an extension of the dataset QM7, and there are many works to address the same regression problem for QM7, c.f. `quantum-machine.org`. Further details about QM7-X can be found in [1].

The data source contains 2 folders: atoms and energies. In the folder of atoms, the training and test data are further separated. The file is ordered by the id of each molecule configuration, written in the extended xyz format. You may use the python package ase to read the coordinates $\mathbf{r}$ in each xyz file. This file also contains the type of each atom. In the folder of energies, there is one csv file, with two columns: id and energy. The energy is real-valued and it is the atomization energy of the molecules in the training set. In this challenge, the goal is produce another csv file in the same format, computed from the test data.

Metric description

The metric used in this challenge to decide the winner is the root mean square error. Assume you model predicts $\tilde{E}(\mathbf{r}_{id})$ on D test configurations $\{\mathbf{r}_{id}\}_{id \le D}$, then the error is

$$\sqrt{\frac{1}{D}\sum_{id=1}^{D}(E(\mathbf{r}_{id}) - \tilde{E}(\mathbf{r}_{id}))^2}$$

Group work and Final presentation/ report

The project is to be realized in group (e.g. en binome). To validate the project, each group should present their work in 5 minutes, followed by a 5 minute discussion. Each group should also write a report which contains the following elements:

- The first part is about how you explore and pre-process the training and test data.
- The second part contains all the details on how you address the problem. It should be concise by highlighting your main ideas (question, motivation, solution). It should also contain enough details in order to reproduce your results (method description, hyper-parameters, etc).
- The third part should present a detailed analysis of your results. This analysis can be quantitative or qualitative.
- Finally, you write a conclusion with a brief summary of the main idea and your results, including your insights and perspectives on the topic.

References

[1] Johannes Hoja, Leonardo Medrano Sandonas, Brian G Ernst, Alvaro Vazquez-Mayagoitia, Robert A DiStasio Jr, and Alexandre Tkatchenko. Qm7-x, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Scientific data*, 8(1):1–11, 2021.