Machine Learning in Material Simulations

Zhenyao Fang

The Objective: The incorporation of machine learning techniques into material simulations has been a center of attraction in recent years. By extracting features either from crystal structures or from existing material properties in database, these techniques significantly accelerate the predictions of various functional properties of materials. The goal of this workshop activity is to extract material properties from databases, construct machine learning models for regression tasks, and understand the effect of hyperparameters in those models.

The Problem: This activity is based on two typical tasks in machine learning: regression and classification. In regression tasks, the machine learning models will learn to predict a continuous target, such as the formation energy or the band gap. When calculating those properties in materials with large supercells (such as in disordered materials), traditional first-principles calculations are computationally expensive, motivating the usage of cluster expansion techniques, which expand the lattice into clusters of different sizes and linearly fit the target quantity to the number of each cluster in the lattice. Here we will use a typical configurationally disordered material, face-centered tetragonal AuCu, as an example to illustrate how to construct those linear regression models and how to evaluate their performance on training sets and testing sets.

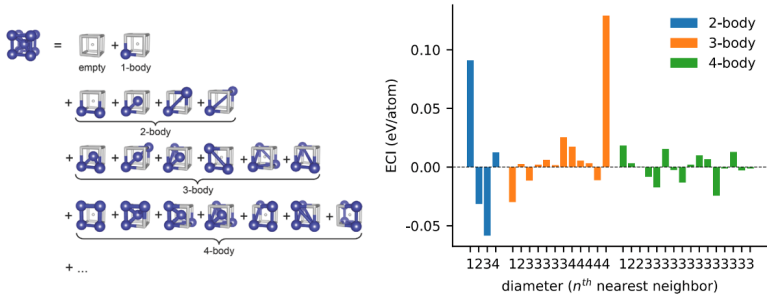

Figure 1 – The cluster expansion methods to predict material properties. The crystal structure is expanded into clusters of different sizes, such as 1-body, 2-body, 3-body interactions. Each cluster correspond to an energy value, called the effective cluster interaction (ECI). In our tasks, this will be the weights in the linear regression models and will help determine the total energy of the whole crystal.

Different from regression tasks, classification tasks focus on discrete target quantities, such as the magnetic or the topological classification, and the space group of a given material. Classification tasks are usually achieved through logistic regression methods. The core of logistic regression is the logistic function $\sigma(t) = \frac{e^t}{1+e^t}$, which outputs a value between 0 and 1. In logistic regression methods, we first use a regression model with the output being the logit t, and feed the logit into the logistic function to obtain the probability of the given material belonging to a certain class; the logistic function will amplify the difference of the probability of different classes. Here we will classify whether a material in the given database is a metal or a non-metal, using the existing material samples in the Materials Project database. We will discuss the effect of the hyperparameters used in the logistic regression methods and explore the importance of each feature of the sample.

The Experiments: With the above introduction, we will now explore more details on incorporating machine learning models into material simulations.

(1) Cluster Expansion Methods for Formation Energies: The first experiment we will do is to use the dataset of configurationally disordered face-centered cubic AuCu as an example and apply various linear regression techniques to predict the formation energies.

We will explore (1) the effect of L1 and L2 regularization; (2) the effect of mixing ratios used in the regularization schemes; (3) the physical meaning of the regression weights.

Questions: What is the performance of the linear regression model? Why do we need a validation set? What are the limitations of the cluster expansion methods?

(2) Logistic Regression for Classification: The second experiment we will do is a data mining project. We will retrieve material properties from existing databases and perform logistic regression to classify them as metals or non-metals.

Questions: Can you think of other properties and retrieve them from the database? Are those properties better descriptors in this task?

Further Reading:

[1] Andreas C. Müller and Sarah Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, 2016, O'Reilly Media.

[2] Keith T. Butler, Felipe Oviedo, Pieremanuele Canepa. Machine Learning in Materials Science. 2022, American Chemical Society.