Introduction

This workshop will be focused on the application of machine learning techniques in materials science and intended to provide some hands-on experience. We will introduce a couple of supervised and unsupervised machine learning methods and apply those methods on some database for demonstration. We will provide the materials a couple of days before this workshop, and we expect students to follow the steps in the workshop. We will be using *Jupyter Notebook* for this purpose. It will be beneficial if students already install *Jupyter Notebook* before this workshop in their respective system (Windows preferred). This workshop is planned for 3 exercises, and we will try to cover as many materials as possible.

Installation

- Search for *Anaconda Navigator*, download, and install it.
- Open *Anaconda Navigator* in your system and look for *Jupyter Notebook* in it.
- Click on *Launch*.

Exercise 1

- Objective/Background: To understand some basic data processing and modeling techniques on a hypothetical database.
- Database: *materialsDataset.txt*
- It has 5 columns viz., *cityName*, *peopleWorkingInMaterialsScience*, *budgetInMillions*, *numberOfPapersPublished*, and *efficient*.
- The first 4 columns is exactly what it sounds like. The last column is a result of various factors like whether conditions, students safety, tuition fee, sports culture and so on. In other words, it gives says if students are likely to join that university or not.
- We will understand the data by gathering information through plots and clusters.
- We will apply *Gaussian Naïve Bayes* and *Decision Tree* model to this database.

Exercise 2

- Objective/Background: We will apply regression models to a real thermodynamic database. The system under consideration is *Cr-Mo* binary alloy system.
- Database: *eta2Dataset.txt*

- It has 9 columns, but we will just use eta1, eta2, eta3, eta4, and u0 in this workshop.
- eta's are dimensionless numbers and represent normalization of energy parameters i.e. $\frac{e}{RT}$. In a Body Centered Cubic (BCC) system, we have 2 pairs, one triangle, and one tetrahedron, each has its energy values and thus we have 4 eta values. u0 is composition of a thermodynamic alloy system under consideration.
- We will apply *Linear Regression* and *K-Nearest Neighbor* model to this database.

Exercise 3

- Objective/Background: To understand the formation of carbon nanostructures in the circumstellar environments, we will use unsupervised machine learning techniques. We have generated a lot of molecular dynamics simulation data for multilayered graphene system. On these data we will use unsupervised machine learning techniques to identify structures with configurations close to the C_{60} (Buckminsterfullerene) molecule.
- Database: *absoluteTotalAveragedData1.txt*
- It has 3 columns viz., *abs_diff_pe*, *abs_diff_rings*, and *abs_diff_rdf*. They are abbreviated for absolute difference in potential energy, absolute difference in the number of carbons of carbon rings (5 and 6 membered carbon rings), and absolute difference in the pair separation distance of the center of mass of 5 and 6 membered carbon rings. These values are calculated with respect to that of C_{60} molecule.
- We will apply *Silhouette analysis*, *K-Means clustering* to this database.