Materials descriptors for data science: homework assignment with hands-on activities

Alejandro Strachan, Zachary McClure, Juan Carlos Verduzco
Purdue University

The following problems will help you better understand the use of descriptors in materials modeling. Before starting with the assignment make sure you go over the accompanying lecture and hands-on tutorial. For the assignments below, you will work with the following nanoHUB tool: https://nanohub.org/tools/featureselect.

The following problems build on the first notebook: *Role of periodic table descriptors and properties in models for melting temperature*

Problem 1. Data exploration. After uploading the data into a Pandas dataframe, the dataframe is displayed in the second code cell. How many data points are included? Study all the column titles to understand the collected data.

Problem 2. What are the four properties included as inputs in the first models developed in Section 2?

Problem 3. The data is divided into training and testing sets. What percent is used for training?

Problem 4. In Section 3, one of the periodic table descriptors added as a descriptor is the electronegativity of each atom in the oxide. Why do you think electronegativities could help predict melting temperature? Hint: think about the relationship between electronegative and the strength of ionic bonding and the relationship between bonding strength and melting.

Problem 5. The function "RandomForestRegressor" sets up the random forest model, how many trees are being used?

Problem 6. Compare the average MAE over the ten random forests between the first model (with four input descriptors) to the value obtained after adding stiffness and Lindemann melting law. How much is the average MAE reduced?

The following problems build on the second notebook: *Quantifying descriptors: Pearson correlations*

Problem 7. Study the first set of Pearson coefficients (Section 4.1). List the top three input properties in terms of their correlation with the melting temperature?

Problem 8. Which of the properties in Section 4.1 shows the highest negative correlation with melting temperature?

Problem 8. Study the properties ranked by their correlation to the melting temperature in Section 5. Discuss why stiffness and melting temperature are correlated.

Problem 9. With knowledge of Pearson correlations, we will go back to the first notebook and add shear modulus to the initial set of four descriptors to explore how much the accuracy can be improved. Before re-running, note the average MAE using the original set of four descriptors.
Add G_VRH to the list of properties:

```
# Create an array with the input properties
#   These are the properties in our data file:
#      IPF is packing fraction, density, space group (crystal structure info), and molar volume
input_properties = np.array(np.column_stack((oxide_melting.IPF,oxide_melting.Density,oxide_melting['spacegroup.number'],
                                             oxide_melting.Molar_Volume, oxide_melting.G_VRH
                                             )))
```

And re-run the cell and the next two. Report the percent decrease in MAE when shear stiffness is added.