Active learning homework assignment with hands-on activities

Alejandro Strachan and Juan Carlos Verduzco
Purdue University

The following problems will help you better understand active learning and to modify workflows to fit your needs. 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 *Machine learning guided design of ceramic oxides for batteries* notebook in the following tool: https://nanohub.org/tools/citrinetools. You will need an API key to access data from citrination, follow the instructions in the tool, you can learn more about data queries in a separate module in this series.

Problem 1. Data collected. You will explore the data collected. After downloading the data into a Pandas dataframe, add a cell (Insert -> Cell below) and explore the entire data by executing "display(data)". How many data points are included? Study all the column titles to understand the collected data.

Problem 2. After adding the composition to the dataframe, display it again. What is the "composition" field of the first entry? What is included in composition, only the elements involved or elements and at.%?

Problem 3. How many data points remain after removing duplicates?

Problem 4. Before starting the active learning workflow, the notebook creates neural network and random forest models. Answer the following questions.
Question 4.1. What fraction of the data is used for training?
Question 4.2. How many hidden layers separate the input layer and the output layer in the neural network?
Question 4.3. After creating the model explain why dense_3 layer has 1830 parameters. Hint: learn more about neural networks in the module of this series.
Question 4.4. Explain the meaning of the parameter "n_estimators" when setting up a random forest using scikit-learn.

Problem 5. Before entering the active learning loop, the notebook prints the next selected experiment based on the 10 initial experiments. Compare the predicted uncertainties for MU and the other acquisition functions.

Problem 6. Go over the active learning loop and identify the code to implement each of the acquisition functions. Starting from the code cells following "5.3 Approach" modify and re-run the notebook so that the MEI acquisition function searched the oxide with the lowest (not highest). You can check the selection based on the first 10 experiments before the active learning loop. Include a snapshot of the last frame of the output movie.