Computational Chemistry 2, CHEM-E4225, Exercise 5 14.5.2021

In this exercise you need **python/3.8 which contain the sklearn library**. It can be started with 'module load python/3.8'

More information of the sklearn manual pages. Search sklearn.ensambe (Warning these are very technical)

There are a lot of example file in /home/kari/CC2-2021-example

To see what is in this dir type ls -l /home/kari/CC2-2021-example (ls is the list command)

you can copy the example files to your own directory: cp /home/kari/CC2-2021-example/h2o.inp . (there is a dot at the end it is your working directory)

1) Use the validation-vdw-polyfit.py program to test the polynomial fit to noisy van der Waal equation. The program will plot the data, full fit and one test fit, to continue close the window (from the upper right red button). Look the R^2 values of the of the full data fit and the validation fits. Which order polynome gives the best prediction. You can also increase the max order of the polynome (now 7). Set the randomness off (rnscale = 0.0). What happens?

2) Run the RandomForest (RF) predictor of Hydrogen binding energy on Pt CNT system: python rf-Pt-CNT-new.py. You need the Pt-CNT data (rf-data-Pt-tube-CNT.dat). Take a look of the data. There are 7 descriptors: HCCn C-C distance between the CH and it's neighbour C's, HCPt shortest C-Pt distance between the CH and Pt's, CH neighbours shortest distance to Pt's. What are the Training and test set R2 values. See also the Cross Validation (CV) values. Figure 1: look the prediction and the 3 descriptor panels. Do you think you could predict the results with least square type method. The Pt-CNT system coordinates are in the Pt-tube-CNT1212-last.xyz (see also exer 6)

3) Run the Gradient Boost predictor, gb-Pt-CNT-new.py. (it uses the same rf-data..). Is this better than RF? Look the Cross validation and values at low energy.

4) Run the RandomForest predictor of Hydrogen binding energy on H2-CNT system: python rf-CNT-H2-new.py. You need the H2-CNT data (rf-data-H2-CNT.dat). Take a look of the data. There are 7 descriptors: HCCn C-C distance between the CH and it's neighbour C's, dHH H-H distance, HHx,y,z the components of the H-H vector. What can you say of the importance of the features. What are the Training, test set and CV R2 values. Look the prediction and the 3 descriptor panels. Do you think you could predict the results with least square type methods. An example of the system coordinates is in the CNT1212-H2-56-lb.xyz. Is this more difficult system than the previous one.

5) Make a new (sub)directory and copy the awk-Pt-CNT-loop.addH, Pt-tube-CNT1212-add.xyz and opt-Pt-tube-CNT1212-diag.inp to it. Take a look of the awk file and run it by typing ./awk-Pt-CNT-loop.addH see the files. This script will make the input and coordinate files for 10 first systems. Take a look of few of the coord files. (The CP2K input is not interesting). You see that the awk script has also submission command (it is commented out). This is a relatively simple example of High

Throughput computation. With few lines one can make input for 100's of calculations and send them.

6) Optimize the ML method. There is an example file Surface_analysis-plain.py and the data file surf-data-smol.dat Study the program and run it. This will optimize the RF model. What are the optimal parameters in this fit. The data set is rather small but compare the ML model predictions to BEP model. The data contain barriers of several simple reaction on various surfaces.

python/3.8 which contain the sklearn library can be started with 'module load python/3.8'

To see geometries you can use ase, module load ase, ase-gui …

The instructions of Wihuri are included.

In the first time make your own directory in /home/kari/CC2-2021-results
mkdir /home/kari/CC2-2021-results/ossi (ossi should be your own name)
At end of exercise copy the results to your result dir: cp *out /home/kari/CC2-2021-results/ossi

Orca input library: https://sites.google.com/site/orcainputlibrary/home