Computational Chemistry 2, CHEM-E4225, Exercise 5 25.11.2022

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

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/CC2-2022-examples

To see what is in this dir type ls -l /home/kari/CC2-2022-examples (ls is the list command)

you can copy the example files to your own directory: cp /home/kari/CC2-2022-examples/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. To run a python program type python validation-vdw-polyfit.py 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 R2 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 the coordinates with ase gui)

3) Run the previous problem with Gradient Boost. python 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 pKa_Shap_Model_plain.py the data it need is in pKa_Data.csv This will optimize three used ML models. What are the optimal parameters in this fit. Which mode is the best. You may need Shap model type: pip install shap.

You can also start the Jupyter notebook. First you need to install it with pip install notebook you need also the shap package. Load it with pip install shap Then start the notebook jupyter notebook. It will start browser from which you can start the xx.ipynb programs. There are pKa_Shap_Model.ipynb, poly-fit.ipynb and others.

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.

The sklearn library can be started with 'module load python'

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