Computational Chemistry 2, CHEM-E4225, Exercise 6 30.11.2023

In this exercise you need **python which contain the sklearn library**. It can be started with 'module load python' all .py programs can be run with command python prog.py

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

There are a lot of example file in /home/kari/CC2-2023-examples

To see what is in this dir type ls -l /home/kari/CC2-2023-examples (ls is the list command)

you can copy the example files to your own directory: cp /home/kari/CC2-2023-examples/h2o.inp .
(there is a dot at the end it is your working directory)

1) Use the pKa-loop.py program to analyse the size convergence of the data. Which of the 3 methods is best. Does this conclusion depend on the error criterion.

2) Use the UMAP-numbers.ipynb or UMAP-numbers.py program to analyse the handwritten number recognition. It will have two data set data_id=41082 and mnist_784 they are rather large so the ML will take a bit of time. See the scattering of the numbers, like 3 and 8, which of the data sets are cleaner.

To run the program you type jupyter notebook UMAP-numbers.ipynb
you need jupyter notebook and UMAP install them with pip install notebook pip install umap-learn

3) Use the UMAP-numbers.ipynb or ….py program to analyse the PCA effect of numbers. Use 5, 15 and 25 components in the PCA analysis. See the singular values and the reconstructed numbers fit. Are the number recognizable with 5 PCA components.

4) Run the shap-ncnt.py program and see the shap analysis. The first page is the violin plot. Look the dCNN data. Will the low and high values produce different SHAP values. Look the mu, Eg and dCNN effect to the SHAP value. What is the average energy value (SHAP base value). This needs ncnt-gga.csv data.

you need rdkit install it with pip install rdkit

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 mylly2 are included.

In the first time make your own directory in /home/kari/CC2-2022-results
mkdir /home/kari/CC2-2022-results/ossi (ossi should be your own name)

At end of exercise copy the results to your result dir: cp *out /home/kari/CC2-2022-results/ossi

Extra: should work

5) Use PCA-feature-single-29.11.py to analyse the molecular clustering. To run this

```
python  PCA-feature-single-29.11.py  -i claisen-smiles-cyclic.csv -t SMILES -m UMAP -p claisen
```

and

```
python  PCA-feature-single-29.11.py  -i claisen-smiles-cyclic.csv -t SMILES -m PCA -p claisen
```

You need the claisen-smiles-cyclic.csv, fingerprints-claisen.csv and submolecules-claisen.csv data. The program is not very nice. Look what is SMILES representation of molecules. Try to understand the fingerprint data (NOT how they have been done). See the PCA and UMAP analysis.

Here you no not need rdkit and to install it is difficult in mylly2. At the moment it does not work.