Homework 3

Note: As a solution to each of the exercises, upload your python codes, discussion narrative and image files in Canvas.

1. QM9 is a comprehensive dataset that provides geometric, energetic, electronic and thermodynamic properties for a subset of GDB-17 database, comprising 134,000 stable organic molecules with up to nine heavy atoms. All molecules are modeled using density functional theory (B3LYP/ 6-31G(2df,p) based DFT). This dataset is described in :

Wu et al., “MoleculeNet: a benchmark for molecular machine learning” *Chemical Science* (2018), **9**, 513.

Use the scikit-learn methods to perform principal components analysis of the QM9 and determine which basic original features are well described by the first couple principal components. The dataset can be downloaded using the following link: http://moleculenet.ai/datasets-1

A. For ease of manipulation, read only the data for the first 150 organic molecules from QM9. Drop the string based column (mol_id). Drop columns corresponding to A, B, C, u0, h298, g298, u0_atom, h298_atom, and g298_atom.

B. Print information about the data contained in the dataset, such as the header, description or summary.

C. Graph the dataset using a pair plot representation.

D. Use the **StandardScaler** method to perform **standardization** of the QM9 dataset to a **normal distribution**. Write the unscaled and scaled QM9 datasets to csv files.

E. Use the **PCA method** to describe the QM9 data. Start by using the cumulative explained variance to determine the number of principal components needed to describe the data. Plot the cumulative explained variance ratio as a function of the number of components and determine the number of components needed to describe at least 0.90 of the QM9 data.

F. Plot the principal components scores for the first two principal components.

G. Determine the loadings on the first 2 principal components and print them. What do the values tell you regarding how well the PCs describe the 10 original features?

H. Select in turn each of the following 4 original features: mu (dipole moment), cv (heat capacity), u298 (internal energy at 298 K), and gap. Perform a **simple linear regression** to determine how well each of these features is reproduced by the first principal component, then the second principal component, and then the third principal component.

- Write the respective R^2 statistics and Residual Standard Error (RSE) in a file
- Illustrate the fitted line in a graph along with the scatter plot of these original features versus the scores along each PCs
- what do you learn from this analysis?

G. Repeat E to H without first scaling the QM9 data. Just move the mean to zero. What do you learn from these results compared to the ones for the scaled QM9 data? Which set-up should be used when performing PCA on QM9?