Homework 0

Instructions: Use this PDF file as a template to develop your homework. Submit your homework on time as a ZIP file to **Canvas** containing:

(1) A single PDF or Word document with your answers to the questions.

(2) Your code within a Jupyter Notebook.

Homework can be completed in any format, including handwritten answers. ***Late submissions will not be accepted and will receive 0 credit.***

1 Processing Chemical Compound Dataset Using Python (50 points)

Fetch the Delaney dataset (`https://raw.githubusercontent.com/deepchem/deepchem/master/datasets/delaney-processed.csv`), which contains solubility data for a set of chemical compounds, and complete the following five tasks. This problem requires the use of Python, along with the Pandas and RDKit packages.

Hint: i) You may refer to the provided `homework0_reference.ipynb` file or other internet resources to help you solve these questions; ii) You may use Google Colab (`https://colab.research.google.com/`) if you don't have a local Python environment. iii) If you're unsure how to use a specific functionality of a Python package, refer to its official documentation website for detailed explanations and examples.

1.1 Please calculate the total number of chemical compounds present in the dataset. **(10 points)**
The number of compounds: 1128.

1.2 Please provide the minimum and maximum molecular weights of chemical compounds in the dataset. **(10 points)**
Min MW: 16.04.
Max MW: 780.95

1.3 Please calculate the number of chemical compounds with at least two rings in the dataset. **(10 points)**
The number of compounds with at least two rings: 425.

1.4 Please select any chemical compound with at least two rings and provide its Compound ID along with its SMILES string. **(10 points)**
Compound 1:
Compound ID: Amigdalin
SMILES: OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)C(O)C3O

Compound 2:
Compound ID: Fenfuram
SMILES: Cc1occc1C(=O)Nc2ccccc2

1.5 Please visualize your selected chemical compound using RDkit package. **(10 points)**

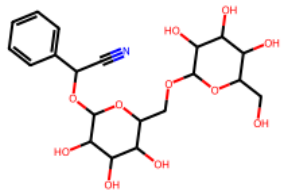

Figure 1: Compound 1 Amigdalin

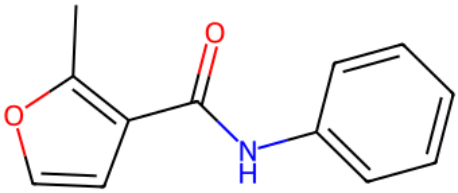

Figure 2: Compound 2 Fenfuram

2 Predicting Compound Solubility Using Linear Regression (50 points)

Please use your fetched Delaney dataset to complete the following four tasks. This problem requires using Python and the scikit-learn package.

Hint: i) You may refer to the provided `homework0_reference.ipynb` file or other internet resources to help you solve these questions; ii) You may use Google Colab (`https://colab.research.google.com/`) if you don't have a local Python environment. iii) If you're unsure how to use a specific functionality of a Python package, refer to its official documentation website for detailed explanations and examples.

2.1 Please perform a 70:30 train-test split on the dataset. Fit a linear regression model using "Molecular Weight" as the predictor to estimate the compound's log solubility on the training set. Provide the slope and intercept of the fitted linear regression model. **(10 points)**
Slope: -0.011
Intercept: -0.766

2.2 Please report the mean squared error (MSE) for both the training and test sets. **(10 points)**
MSE of training: 1.52
MSE of test: 1.52

2.3 Please visualize the distribution of chemical compounds in the molecular weight vs. log solubility space for both the training and test sets, and overlay the fitted linear regression line on the plots. **(10 points)**

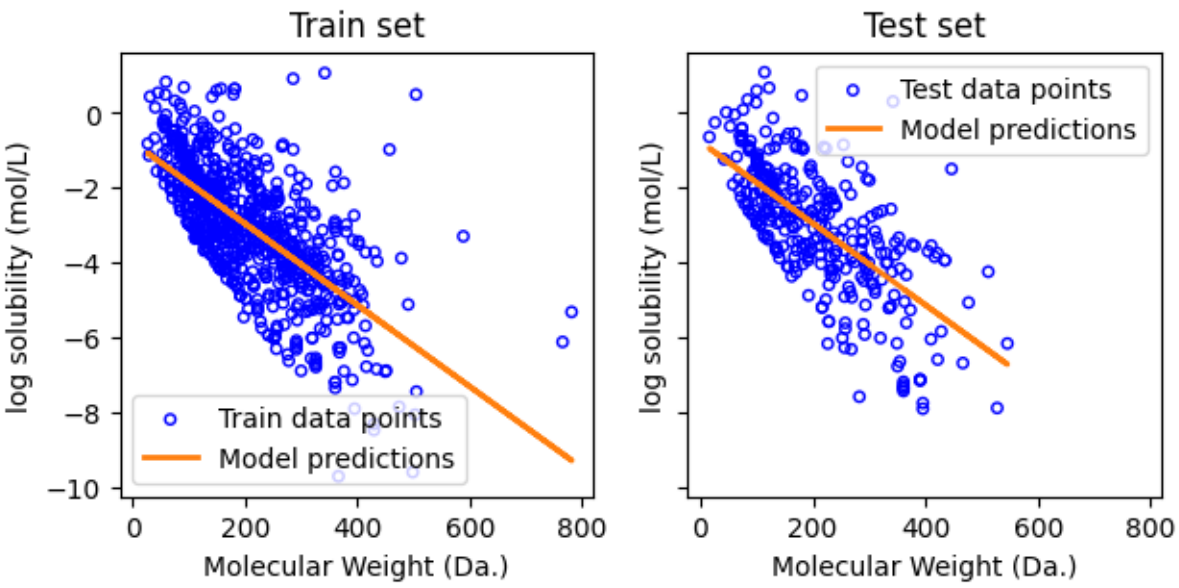

Figure 3: log Solubility vs. Molecular weight

2.4 Please fit another linear regression model using the property "Polar Surface Area" as the predictor. Report the mean squared error (MSE) for both the training and test sets. Visualize the distribution of chemical compounds in the polar surface area vs. log solubility space for both the training and test sets, and overlay the fitted linear regression line on the plots. Please briefly explain whether molecular weight or polar surface area is a better feature for prediction of log solubility. **(20 points)**

Slope: 0.0090
Intercept: -3.31
MSE of training: 2.67
MSE of test: 2.87

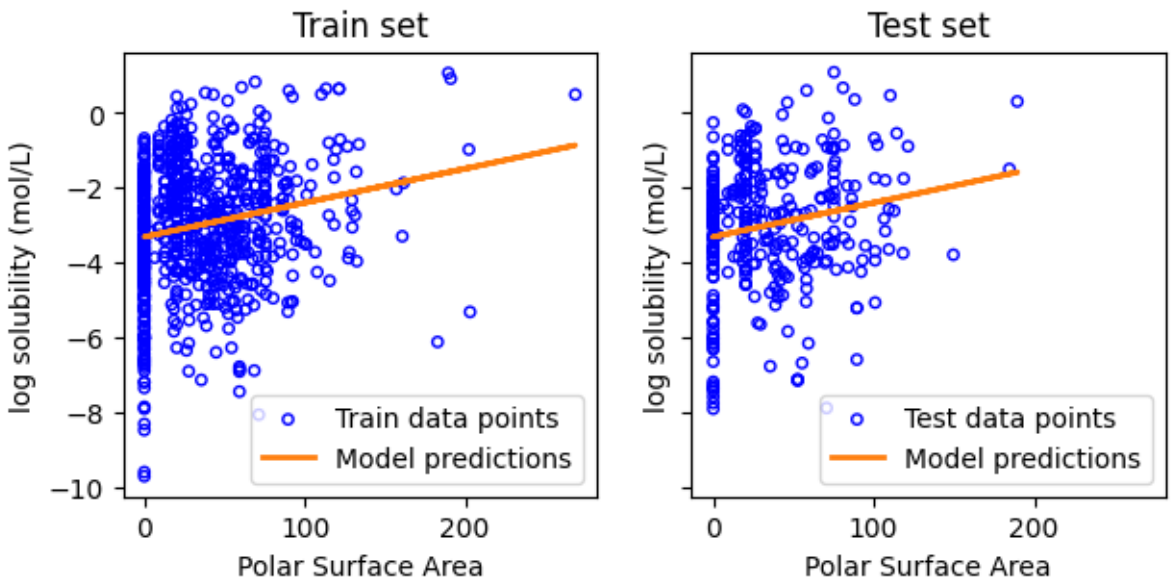

Figure 4: log Solubility vs. Polar Surface Area

Molecular weight is a better feature for log solubility prediction. The R2 score of the linear model using molecular weight is 0.45 on training data and 0.49 on test data. The R2 score of the linear model using polar surface area is 0.04 on training data and 0.03 on test data. A higher R2 score indicates better fitting curve. Therefore, molecular weight is a better feature for the prediction of log solubility.