Chem361: Machine Learning in Chemistry
Spring 2025, UW-Madison
Homework 2

Due date: 5pm, March 4, 2025

Instructions: This homework covers "Chapter 4: Predicting Chemical Outcomes – Supervised Learning for Classification and Regression". Use this PDF file as a template to complete your homework. Submit your work on time as a ZIP file to **Canvas**, including:

(1) A single PDF or Word document containing your answers (handwritten or typed).

(2) Your code in a Jupyter Notebook.

You may refer to the provided `homework2_reference.ipynb` file or online resources to help you solve the questions; If needed, you can use Google Colab (https://colab.research.google.com/) instead of a local Python environment. For Python package functionality, consult the official documentation for detailed explanations and examples.

There are 4 homework assignments; the 3 highest scores will each contribute 20% to your final grade ***Late homework submissions will not be accepted.***

1 Linear Regression (35 points)

Fetch the Delaney solubility dataset (https://raw.githubusercontent.com/deepchem/deepchem/master/datasets/delaney-processed.csv), to complete the tasks in this homework. For each chemical compound, use input feature vectors consisting of four molecular properties: **"Molecular Weight", "Number of H-Bond Donors", "Number of Rings", and "Number of Rotatable Bonds"**. Train a linear regression model to predict the "measured log solubility in mols per litre" based on these features.

1.1 Randomly split the dataset to training and validation set using a 90:10 ratio on four features. Train a linear regression model and report the mean squared error and coefficient of determination for both the training and validation sets. **(10 points)**

1.2 If selecting three out of four input features for training a linear regression model, use cross-validation to determine the optimal feature combination: (a). Enumerate all possible combinations of three features. (b). For each combination, perform 5-fold cross-validation. (c). Report the MSE values for both the training and validation sets. (d). Select the best three-feature combination and justify your choice. **(25 points)**

2 Logistic Regression (25 points)

Please use feature vectors containing the molecular properties of **"Molecular Weight", "Number of H-Bond Donors", "Number of Rings" and "Number of Rotatable Bonds"** to train a logistic regression model for

binary classification of compounds. Compounds are labelled as "soluble" if their "measured log solubility in mols per litre" is equal or greater than −2 and as "non-soluble" otherwise.

2.1 Please report the the number of soluble and non-soluble compounds in the dataset. **(10 points)**

2.2 Randomly split the dataset into training and validation sets using a 90:10 ratio. Train a logistic regression model on the training set to classify "soluble" v.s. "non-soluble" compounds. Determine the number of true positive and true negative compounds on the validation set. Does the model perform better at correctly predicting the "soluble" or "non-soluble" class? Please justify your choice **(15 points)**

3 K-Nearest Neighbour (20 points)

Please use the feature vectors containing molecular properties of **"Molecular Weight", "Number of H-Bond Donors", "Number of Rings" and "Number of Rotatable Bonds"** to build the k-nearest neighbour model for binary classification of compounds. Compounds are labelled as "soluble" if their "measured log solubility in mols per litre" is equal or greater than −2 and as "non-soluble" otherwise.

3.1 Please hold out the first compound, **Amigdalin**, for testing. Construct the **KNeighborsClassifier** with $k = 5$ using the remaining compounds in the dataset. Please report the predicted class ("soluble" vs. "non-soluble") for **Amigdalin**. (Hint: Data normalization is critical).**(10 points)**

3.2 Continuing from Q3.1, please report the compound IDs, the "measured log solubility in mols per litre", and their solubility class ("soluble" vs. "non-soluble") for the $k = 5$ nearest neighbours of **Amigdalin**. **(10 points)**

4 Decision Tree (20 points)

Please use the feature vectors containing molecular properties of **"Molecular Weight", "Number of H-Bond Donors", "Number of Rings" and "Number of Rotatable Bonds"** to build a decision tree for binary classification of compounds. Compounds are labelled as "soluble" if their "measured log solubility in mols per litre" is equal or greater than −2 and as "non-soluble" otherwise.

4.1 Please hold out the first compound, **Amigdalin**, for testing. Build a decision tree using the remaining compounds in the dataset with max_depth= 3. Please plot the tree, and identify which feature is used for the first split. **(15 points)**

4.2 Please predict the class ("soluble" vs. "non-soluble") of **Amigdalin**. **(5 points)**