Chemistry 213/313 Homework Assignment 1 Due: April 4, 2025

In this homework assignment, you will gain experience with basic supervised machine learning methods. To submit the assignment, export each notebook to HTML and upload the resulting file to Canvas. Your code should be well commented through Python comments and markdown cells. If you need to install Python, the Anaconda download is recommended. You can also use Google Colab, a cloud-based platform.

You are encouraged to use large language models (LLMs) such as http://phoenixai.uchicago.edu, but make sure that you understand the code that they produce. Some useful Python application programming interfaces (APIs) with which to familiarize yourself are pandas, scikit-learn, NumPy, and Matplotlib. Each of the linked pages provides a quick introduction to the API in addition to complete documentation. You may also find Python's built-in help() function handy: calling help(*obj*) on object *obj* displays its docstring. You can of course also query an LLM about about the code.

1. Construct a linear model to predict solubility using the dataset discussed in White, Chapter 2: https://github.com/whitead/dmol-book/raw/main/data/curated-solubility-dataset.csv. To this end, develop a Jupyter notebook that does the following.

    (a) Gets the data, loads it into a pandas dataframe, and also saves it locally.

    (b) Does some exploratory data analysis, such as that done by White in Chapter 2 (e.g., making histograms and scatter plots of the features and predictive target); you should also compute the correlations of the features with each other and with the target. Note that some of White's plotting code is deprecated—you should update it. Which features do you expect to be most important based on this analysis?

    (c) Makes an 80:20 train:test split and normalizes the features as in White Chapter 3. Why is it useful to normalize the features?

    (d) Uses scikit-learn to make a linear model without regularization and plots the resulting predictions against the actual values. Which features do you interpret to be the most important based on this model?

    (e) Uses scikit-learn to make a linear model with L1 regularization and plots the resulting predictions against the actual values. You should vary the strength of the regularization to examine the tradeoff between accuracy and interpretability. Which features do you interpret to be the most important based on this model?

2. Construct a classifier to distinguish white from red wine based on its chemical properties. In this case, the data are in two separate datasets:
http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv
http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv
More information about the dataset and an alternative way of downloading it are available at
https://archive.ics.uci.edu/dataset/186/wine+quality
Develop a Jupyter notebook that does the following.

(a) Gets the data, labels it according to whether the wine is white or red, concatenates it in a single pandas dataframe, and saves the resulting dataset locally. Comment on whether there is class imbalance. Why in general is this important to check?

(b) Does some exploratory data analysis. You should think about what plots and statistics might be most useful given that the goal is classification rather than regression.

(c) Makes an 80:20 train:test split and normalizes the features.

(d) Uses scikit-learn to construct perceptron and logistic regression models, and prints the confusion matrix in each case. How do the performances of the two models compare?

(e) Uses scikit-learn to build a random forest model with varying depth and numbers of trees. How do the accuracy and interpretability vary with these hyperparameters? Which features are the most important? In addition to trying numerical methods (see the scikit-learn documentation for some options), you may wish to plot a few of the contributing decision trees for a shallow depth.