

# Homework 1

Due on 02/14/2021

In this exercise, we will predict solubility of compounds using their chemical structures. The training data are in the file “solubility\_train.csv” and the test data are in “solubility\_test.csv”. Among the 228 predictors, 208 are binary variables that indicate the presence or absence of a particular chemical substructure, 16 are count features, such as the number of bonds or the number of bromine atoms, and 4 are continuous features, such as molecular weight or surface area. The response is in the column “Solubility” (the last column).

- (a) Fit a linear model using least squares on the training data and calculate the mean squared error using the test data.
- (b) Fit a ridge regression model on the training data, with  $\lambda$  chosen by cross-validation. Report the test error.
- (c) Fit a lasso model on the training data, with  $\lambda$  chosen by cross-validation. Report the test error and the number of non-zero coefficient estimates in your model.
- (d) Fit a principle component regression model on the training data, with  $M$  chosen by cross-validation. Report the test error and the value of  $M$  selected by cross-validation.
- (e) Which model will you choose for predicting solubility?