CS181 Spring 2016 Practical 1

Robert J. Johnson — Dinesh Malav — Matthew McKenna

February 13, 2016

Abstract

The Harvard Clean Energy Project has been investigating the features and characteristics of organic photovoltaic molecules in an effort to create carbon-based solar cells. Density Functional Theory (DFT) has been traditionally used to estimate the energy difference from the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). DFT calculations are computationally expensive, and thus we have developed an [insert method here] algorithm to predict the HOMO-LUMO gap based on the chemical structure of an organic photovoltaic molecule. Using [insert method here] the HOMO-LUMO gaps of our test data were predicted with a RMSE of [insert final score here].

1 Technical Approach

The approach that our team took was an iterative one. Very early on in the process, our team established that we had relatively little domain knowledge of organic chemistry. Thus, we attempted to implement various proven techniques and algorithms such as Ridge Regression, Random Forest Regression, Bayesian Linear Regression, Neural Networks, and Support Vector Machines. However, these were not able to beat the RMSE baselines set by the example OLS (RMSE = 0.29846) and Random Forest Regression (RMSE = 0.27207) established by the teaching staff. A different technique was needed.

It became abundantly clear that feature engineering was the key to improving the model. Using the RDKit package in Python, we were able to extract some basic information regarding the SMILEs molecules. These were basic yet important pieces of information regarding a molecule: the total number of atoms, the total number of bonds, and the molecular weight of the molecule. Undoubtedly, knowledge of specific types of bonds and chemical rings may have improved the analysis here, but for the time being, these factors were determined to be enough to improve on the model.

2 Results

The model that we ultimately chose to tune was the Random Forest Regressor. Our training data contained 256 features, of which we were not familiar with either their function or how they impacted the HOMO-LUMO gap of the molecule. Accordingly, a chief concern was preventing over-fitting.

3 Discussion