Prediction of *ab initio*-quality molecular spectra using machine learning

Eric Berquist

December 11, 2016

Machine learning is one of the primary tools used to process and interpret "big data" which is intractable using manual or brute-force interpretation. One area in which machine learning can be applied is regression modeling, where reference (training) data with a known relationship between some independent and dependent variables is used to predict dependent variables for unknown data. A simple form of regression is linear regression: $f(x) = mx + b$, where x is the independent variable and $f(x)$ is the dependent variable. With the advent of large databases containing diverse chemical information, traditional methods of hand-picking a few model systems for high-level calculations to understand macroscopic properties are becoming less applicable.

I will use machine learning techniques to build regression models capable of predicting *ab initio* calculated molecular spectra, using only atomic coordinates as input. My focus is on extending the work of Alán Aspuru-Guzik and O. Anatole von Lilienfeld from scalar quantities ($\epsilon_{\text{HOMO}}, \epsilon_{\text{LUMO}}, \Delta\epsilon, \langle r^2 \rangle, \mu_{\text{norm}}, \alpha_{\text{iso}}, U_0, E_{\text{ZPVE}}$ and others) to quantities requiring multidimensional representations, such as molecular spectra. Work has already been done to find optimal representations of molecular data in models (the input x), such as the Coulomb matrix, the "bag of bonds", or connectivity fingerprints as different forms of molecular graphs.

The first test will use a subset of the existing GDB-9 database, which consists of the "subset of all neutral molecules with up to nine atoms (CONF), not counting hydrogen" (about 134K) with thermodynamic properties calculated at the B3LYP/6-31G(2df,p) level. In order to calculate the zero-point vibrational energy, E_{ZPVE}, (harmonic) vibrational frequencies are required, so they are already present in the database. Using this publicly available data and the open-source Python library scikit-learn, I will see if current approaches to regression modeling can be extended to higher dimensional quantum chemical data.

Future work may take one of three clear directions. The first direction is use of more accurate quantum chemical models, such as modern density functionals known to perform well for thermochemistry (e.g. ωB97M-V/def2-TZVP). It has not been tested if using higher-quality methods reduces the size of the training set required for quantitative predictions. The second direction is to test other regression models, although there is already some indication that introducing non-linearity does not improve results. The third direction is to predict other kinds of molecular spectra; of particular interest are non-linear optical properties (hyperpolarizabilities) and magnetic spectroscopies (NMR and EPR). A fourth, more difficult direction is the prediction of experimental (rather than calculated) spectra. This is important for spectroscopies that are difficult and/or expensive to predict using current quantum chemical techniques, such as optical or specific rotation. However, this entails the existence or more likely the creation of experimental databases, which is outside the scope of this proposal.