Molecular Structure Prediction Using Infrared Spectra

Michael Chen, Sophia Chen, Yanbing Zhu
{misch, schen10, yanbingz}@stanford.edu
CS 229, Stanford University

Summary

The molecular identification of a sample is a difficult task. Experimentalists often resort to spectroscopic methods to help identify the molecular composition of a given sample. The integration of spectral clues to help narrow down possibilities is not systematic and often time consuming. We developed several machine learning approaches to **predict the molecular structure of organic compounds using infrared (IR) spectra**.

Dataset

From the NIST Chemistry Webbook, we scraped 3D molecular structure files and IR spectra for various organic molecules (~1600). The collected IR spectra contained intensity measurements (either transmittance or absorbance) for a range of incident light frequencies with varying resolutions. The 3D molecule files for a given molecule detailed the atoms and their connectivities, which we used to label each molecule with its constituent functional groups.

Features

To standardize the IR spectra, all intensities were converted to **absorbances**. Additionally we featurized the spectra to fit a common range and resolution, requiring that we linearly interpolate between the raw spectral points. Spectral intensities were normalized by dividing by the spectrum's maximum intensity.

The functional groups we used were **alkanes, alkenes, alkynes, alcohols, amines, nitriles, aromatics, alkyl halides, esters, ketones, aldehydes, carboxylic acids, and acyl halides**.

Models

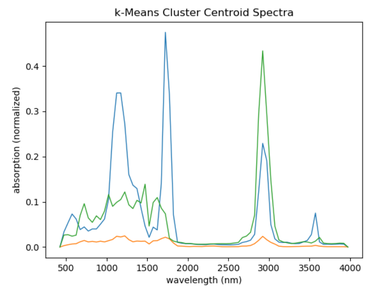
Logistic Regression

We applied a One-vs-Rest logistic regression classifier using a stochastic gradient descent solver to classify carbonyls, alkenes, and alcohols.

The SGD parameter update is given by: $\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta\left(x^{(i)}\right)\right)x_j^{(i)}$

K-means

Our K-means model was able to cluster carbonyls, alkenes, and alcohols. The other functional groups had peaks that were too similar to other functional groups, and k-means was unable to distinctly cluster the remaining organic groups.

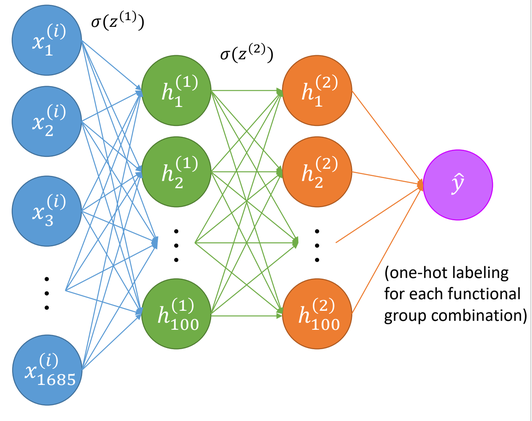


PCA

With PCA, we were able to lower our error after reducing to three components to represent the three functional groups.

Neural Network

We constructed a **feedforward** neural network with two hidden layers to classify different functional groups. Our activation function for the hidden layers was the sigmoid function. We were able to classify 13 different functional groups.



Results/Discussion

Model	Training Error	Testing Error	Training/Testing Points
Logistic Regression	41.54%	42.14%	1176/208 samples
K-means	N/A	36.05%	508 samples
K-means w/PCA	N/A	35.83%	508 samples
Neural Network	13.28%	23.84%	1516/169 samples

In attempting to classify a molecule's functional groups based on its IR spectrum, we conducted a survey of methods and found our **neural network** gave the highest performance. We expected this result given the nonlinear nature of the decision boundary, as typified by the overlapping of characteristic peaks amongst different functional groups.

Initially, we hypothesized that using **PCA** for dimensionality reduction would additionally help improve accuracies, given that certain frequencies appear to not be specific to any functional group. However in applying PCA we found there to be only slight accuracy improvements.

We also hypothesized that **k-means** would be able to cluster the molecules based on the characteristic peaks of each functional group. However, because characteristic peaks are often located in similar spots, k-means was only able to distinguish between functional groups with very different locations.

Future Work

Actual sample composition determination usually involves integrating clues from multiple sources, and analogously we would extend our models to account for a combination of spectral features in addition to IR (e.g. UV-Vis, NMR, Mass-spec, etc.). Ideally, we would extend the problem of classifying a molecule based on common functional groups to the regression problem of determining the molecule's exact elemental composition and 3D structure

References: [1] P.J. Linstrom and W.G. Mallard, Eds., **NIST Chemistry WebBook, NIST Standard Reference Database Number 69**, National Institute of Standards and Technology, Gaithersburg MD, 20899
[2] Madden, M. G., T. Howley, A machine learning application for classification of chemical spectra, in: Proc. of 28th SGAI International Conference, Cambridge, UK, (2008)