Topics for Literature Seminar

Evan Curtin
(Dated: September 12, 2016)

I. EVOLUTIONARY ALGORITHMS IN COMPUTATIONAL CHEMISTRY

This technique borrows many ideas from evolution in the natural world. They are useful for solving problems that lie in a high dimensional, complex space. They work by randomly generating a population with various 'traits' and assessing the individual's "fitness," which is essentially a user-defined scoring mechanism. The most fit solutions are allowed to "mate" by mixing their attributes with other individuals in the population. This process is then iterated on until a satisfactory solution is found. The population may also be subject to mutations and cross-linking, in direct analogy to those processes in biology. This method has been applied to protein folding, characterizing biological NMR samples, and and kinetics of high temperature reactions. It shows promise for various global optimization problems, which are commonplace in chemistry.

II. NEURAL NETWORKS IN CHEMISTRY

Again drawing inspiration from biology, neural networks seek to emulate the human brain in order to tackle problems which are not amenable to traditional computation. It is commonly used in image recognition and other pattern recognition problems. This technique can be used to try to answer questions of the form "is this molecule a suitable drug to treat a certain disease?". It accomplishes this by setting up a network of "neurons" which, during the learning process, can be given different weights. Once trained, the neural network can "understand" rules such as "carbonyls have peaks at 1700cm^{-1} in IR." Problems can be encountered in this technique if one uses either too many or too few nodes. Too few will not allow the network to learn detailed information about a problem, while too many will simply categorize the entire training set individually.

III. THE SUPPORT VECTOR MACHINE ALGORITHM IN CHEMISTRY

A support vector is a way to classify sets of data. This technique is performed by maximizing the distance between the two closest points in separate classes in high dimensional space. It can be used for classifying many classes or regression of functions with large number of variables. It's commonly used in bioactivity studies as well as toxicology. It's performed by using a kernel function to map the data to a high dimensional space, and generating a hyperplane which splits the data. New entries are categorized by checking which side of the hyperplane they lie on.

IV. REFERENCES

1) Carwright, H. (2007). Reviews in Computational Chemistry. (K. B. Lipkowitz & T. R. Cundari, Eds.)Reviews in Computational Chemistry (Vol. 25). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/9780470189078

2) Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. Wiley Interdisciplinary Reviews: Computational Molecular Science, 4(5), 468481. http://doi.org/10.1002/wcms.1183