DEPARTMENT OF STATISTICS

STATS 760 A Survey of Modern Applied Statistics

Assignment 3 2017. Due May 15.

Tetko et al. (2001) and Huuskonen (2000) investigated a set of compounds with corresponding experimental solubility values using complex sets of descriptors. They used linear regression and neural network models to estimate the relationship between chemical structure and solubility. For this assignment, we will use 1267 compounds and a set of more understandable descriptors that fall into one of three groups: 208 binary "fingerprints" that indicate the presence or absence of a particular chemical sub-structure, 16 count descriptors (such as the number of bonds or the number of Bromine atoms) and 4 continuous descriptors (such as molecular weight or surface area).

The data are in a text file Solubility.txt on the web page.

1. Read the data into R. Some of the non-binary variables are quite skewed so you may wish to transform them to make them more symmetric. The caret package has a function BoxCoxTransform that you may find useful.
2. Fit a linear regression as a benchmark. Calculate a bootstrap prediction error.
3. Fit a ridge regression and a lasso, finding a suitable value of lambda in each case. Calculate bootstrap predict errors (including the process of finding the lambda.) Do these improve over linear regression?
4. Fit boosted trees and random forests. Again, calculate prediction errors.
5. Compare all your results, taking care to explain how you got the various estimates of PE (particularly for part 4.)
6. Which variables are the most important in predicting the solubility?
7. Draw a partial dependence plot for each of the four continuous descriptors. How do these variables relate to solubility?

Tetko, I., Tanchuk, V., Kasheva, T., and Villa, A. (2001). Estimation of aqueous solubility of chemical compounds using E-state indices. *Journal of Chemical Information and Computer Sciences*, 41(6), 1488-1493.

Huuskonen, J. (2000). Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *Journal of Chemical Information and Computer Sciences*, 40(3), 773-777.