Abstract – MS in Statistics

Many experimental real-world, large-scale data sets are composed of highly correlated features, where the correlation structure is unknown prior to data collection. A common goal in empirical research to model the mean behavior of a feature by conditioning on a linear model of the remaining feature space; in this instance a high degree of correlation in the feature space can lead to an unstable model. Additionally, one may find that some of the data features are completely unrelated to a feature of interest, adding no information to the model. Thus, a researcher would like a regularization procedure that accurately identifies the unrelated features, and sets their model coefficients to 0, resulting in a sparse model.

The Ridge regularization was developed to yield statistical models that are stable in the presence of a highly correlated feature space. Lasso regularization was introduced to accurately identify sparse statistical models. The Elastic Net combines both regularization terms to yield a statistical model with both properties. It has been shown that as the dimension of the feature space grows, and the number of observations grow, the Lasso and the Elastic Net are both inconsistent estimators of the true set of nonzero model coefficients. It can be demonstrated that when the signal to noise ratio is low, the Lasso yields highly erroneous models.

Consistency of estimation is a desirable property of any statistical modeling procedure. It has been shown consistency is possible by

1. Assuming additional regularization constraints.
2. Weighting the coefficients in the L1 penalty of the Lasso regularization term.

These results are not only theoretically interesting, but demonstrably useful when modeling real world data-sets of moderate scale and with a moderate correlation structure. Yet, to date, an easy-to-use package has not been developed or implemented in Python.

In this presentation I introduce regularized statistical linear models, demonstrate the issues that arise in real data sets leading to the need of regularization. I also demonstrate the inconsistency of the Elastic Net, the consistency of the Adaptive Lasso and Adaptive Elastic Net, and provide an easy to use Python package that allows the user to implement an AENet and ALasso on their own. My presentation and associated code lives in a Jupyter notebook that will be made available to attendees. If you plan on attending my talk, please follow the link below to register so that I can send out a link to the notebook prior to my talk.