Linear Regression: Kaggle Africa Soil Challenge

Signal Data Science

Note: The R script used in the presentation is located in the `lectures/africa-soil` folder of the curriculum repository. Feel free to refer to it as you work.

After finishing the regularization assignment using the speed dating dataset, you'll be working on the Kaggle Africa Soil Property Prediction Challenge. Download the data. If you haven't made a Kaggle account, do so.

- Download the data from the Kaggle website. Load both the train and test sets into R. Use `read_csv()` from the `readr` package to load the data instead of `read.csv()` (the former is substantially faster).

The premise here is that we can predict amounts of soil organic carbon, pH values, calcium, phosphorus and sand content in a soil sample soil's absorbance at many electromagnetic wave numbers.[1] We'll use regularized linear regression for this. For simplicity, restrict your attention to the wave number features.

First, we'll focus on predicting calcium levels.

- As in the final lines of the R script used in the presentation, explore how the graph of the coefficients varies with α, using `cv.glmnet()` with `alpha` set to 1, 0.1, 0.05, 0.01, 0.001 and 0.
- Use the `caret` package to tabulate cross validated RMSE as a function of (α, λ). Make sure that your grid of values of α includes values close to zero.

Next, we can expand consideration to all 5 target variables.

- Use the `caret` package to find the best values of the hyperparameters (α, λ) for each of the 5 target variables. Generate predictions for the test sets and upload them to Kaggle per the instructions.

The Kaggle competition features an advanced machine learning algorithm called Bayesian Additive Regression Trees (BART)[2]. The predictive power of regularized linear regression won't be as good as the BART algorithm, so your

[1] Wave number is equal to inverse frequency.
[2] Chipman *et al.*, BART: Bayesian additive regression trees.

position on the leaderboard won't be high, but it's striking that you're able to get as good predictive power as you are without needing to know anything about chemistry, and after just ~10 days of doing data science!

- Plot histograms of P and $\log(P+1)$. How much of an improvement in predictive power do you get from fitting a linear model for the latter instead of the former? (Don't forget to transform the predictions back into the non-logarithmic scale.)