Linear Regression: Kaggle Africa Soil Challenge

Signal Data Science

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.

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 `linreg-africa-soil-exploration.R`, explore how the graph of the coefficients varies with `alpha`, 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 `alpha` and `lambda`. Make sure that your grid of values of `alpha` includes values of `alpha` close to zero.

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

- Run the `caret` package to find the best values of the hyperparameters `alpha` and `lambda` 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 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!

[1] inv

[2] Chipman *et al.*, BART: Bayesian additive regression trees.