Logistic Regression: Filtering Spam Emails

Signal Data Science

The CSDMC2010 SPAM Corpus

You'll be looking at the data from the CSDMC2010 SPAM Corpus. The data is in `spam-emails.csv`, with `spam-emails-key.txt` giving the correct classification as spam or not-spam.

- Use the `tm` package to construct a `DocumentTermMatrix` from the data, where each row represents a document, each column represents a word, and each entry contains the word frequency for the associated word in the associated document.
- Use the `caret` package to find the optimal values of (λ, α) for regularized logistic regression in the classification of the documents.
- Using the optimal hyperparameters you found, train a regularized logistic regression model on the whole dataset. Look at the ROC curve.

Some things to keep in mind (read these before you begin):

- For a reference about text preprocessing with `tm`, look here or here.
- Not every single email in the dataset successfully made it into `spam-emails.csv`. You'll want to do an `inner_join()` (from `dplyr`) on the dataset and the classification key, retaining only the rows which are successfully matched.
- Remove columns corresponding to words that show up fewer than 10 times in total throughout the entire corpus.
- The matrix of documents and word frequencies is very sparse – don't scale it! If you do so, it will make it non-sparse, which is bad!