Logistic Regression: U.S. National Elections

Signal Data Science

For some additional practice with logistic regression, we'll be looking at American election data from the National Election Study from 1948 through 2002.

A note on `caret`

In the following, you'll be using the `caret` package to obtain cross-validated estimates of α and λ for regularized logistic regression. Keep the following points in mind:

- Use 5-fold cross validation without any repeats.
- In order to tell `train()` that you want to perform two-class classification instead of standard regression, set `classProbs=TRUE` in the control parameters.
- In addition, in order to use area under the ROC curve as your metric of model quality, set `summaryFunction=twoClassSummary` in the control parameters and pass in `metric="ROC"` to `train()`.

Getting started

The dataset is located in the `nat-elections` directory as `elections-cleaned.dta`. Information about the data is located in `nes-glossary.txt`.

- `.dta` files are Stata data files which R cannot natively read. Load the `foreign` package and use `read.dta()` to load the dataset into R.

For your convenience, we've already cleaned the dataset by imputing missing values and properly renaming factor levels. In addition, we've restricted consideration to years with presdential elections and selected a subset of the original variables.

Exploring the data

Before you do any data analysis, it's typically a good idea to do some basic exploratory visualizations to build intuition around the dataset.

- Use `mosaicplot()` to make a couple mosaic plots from the cleaned and simplified dataset. For example, try `mosaicplot(table(df$income, df$presvote))`. Can you find any counterintuitive results?
- Considering the entire time period from 1948–2000, it there any relationship between what region a voter lives in and which presidential party they support? Is this relationship any different if you restrict to looking data from smaller timespans (*e.g.*, a single election year or 2 consecutive elections)? You can just look at a couple mosaic plots to answer this.

Analysis with logistic regression

We need to expand out the factor columns into a set of binary indicator variables in order to fit linear models.

Earlier, you learned that a factor with k levels should be expanded out into a set of $k-1$ indicator variables, because k indicator variables (one for each level) would suffer from multicollinearity. When using regularized models, we however *do* want to use k indicator variables. Here's why: intuitively, the multicollinearity arises from our being able to write one of the k indicator variables as a linear combination of the others. However, when we *regularize*, we constrain the magnitudes of the model's coefficients and effectively overcome this problem. As such, adding in k instead of $k-1$ indicator variables for factors can improve the performance of a regularized model.

Thankfully, we don't need to write our own function to perform this expansion.

- Use `dummy.data.frame()` from the `dummies` package to create a *new* data frame with the factors expanded out into indicator variables. When calling `dummy.data.frame()`, set `sep="_"` to make the resulting column names more readable.

We're now ready to use regularized logistic regression to explore the dataset. As described in the regularization assignment, use `caret`'s `train()` function to search for the correct values of α and λ. It typically gives good initial results to search over $\alpha \in \{0, 0.1, \ldots, 1\}$ and $\lambda \in \{2^{-4}, 2^{-3}, \ldots, 2^{1}\}$; if you want further improvements, you can perform a finer grid search over a smaller range of values.

Note: In the following, `scale()`ing various subsets of the data might introduce `NAs` into the data frame because a constant column cannot be scaled (as it has standard deviation 0). Be sure to check and correct for this.

For each of the following questions, you should interpret the nonzero regression coefficients and calculate the area under the ROC curve. They are relatively open-ended; feel free to do any additional analysis which interests you or to explore questions which aren't listed here.

- Predict support for George H. W. Bush in the 1992 election. (Restrict consideration to people who actually voted!) Can you improve your model by adding interaction terms?
- Predict party support for different years and look at how the coefficients of the models change over time. Again, consider adding some interaction terms to your model. Which demographic variables have increased or decreased the most in importance over time?
- For voters who didn't vote, predict how they *would have* voted. If you aggregate these predictions by election year, how do they appear to change over time?