Week 11: Regularized Regression

11/07/2019

Jake Campbell

Bias vs. Variance

- In statistical modeling, we have to think about a trade-off between bias and variance
  - Low Variance: consistent results, but not as accurate on average
  - Low Bias: Accurate on average, but inconsistent
- Adding more variables decreases the bias, but can lead to higher variance

Parsimony

- In general, larger models are going to perform better than smaller ones (at least with the data on hand!)
  - We have more info to make predictions with
- Of course, the improvement brought about by larger models may not justify a loss of model understanding
  - Does jumping 1% in adj. R squared justify adding 5 new variables?
- The idea of a parsimonious model is that it explains the model well, with the minimum number of predictors
  - Sounds great! But is this a concept we should put all of our faith into?

Data Dredging

- Data dredging is an abuse of data mining to confirm some sort of bias
  - Might be done on purpose or not
  - We are casting our net with the sole purpose to find some significant variable
- Researchers that go in with no hypothesis and don't do pre-analysis are very likely to be fishing in their data
- You can go in blind, but by using pre-analysis steps (graphing, inspecting data, looking at correlations, etc.), you can begin to generate initial hypothesis
- If our sole goal is to get a parsimonious model, we aren't being good researchers
  - A general issue with some forms of stepwise regression

What Should We Do?

- We'd like a good, relatively parsimonious model, but don't want to succumb to poor research tactics
- What if we take a larger model, and introduce bias into it to lower the variance and get more consistent results?

Ridge Regression

- Like in regular OLS regression, aiming to minimize the sum of squared residuals, BUT with a constraint on the coefficient
- This constraint is the multiple of a value λ and the sum of our squared coefficients
- Known as L2 regularization
- We are trying to minimize the error, but also minimize the size of our coefficients

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j}^{m} \beta^2 j$$

Feature Selection: What is it?

- When we talk about features of a model, we are talking about the predictors
- So far, we have looked at features independently and gauged their relationship with the response ourselves
  - Do x and y have some sort of relationship graphically?
  - Is x significant while taking other variables into account?
  - Does it even matter if x is significant?
- This can become difficult when we're dealing with thousands or millions of variables
- As much as we want to be in control of model development, sometimes it's impossible

LASSO Regression

- Least Absolute Shrinkage and Selection Operator
- Like in regular OLS regression, aiming to minimize the sum of squared residuals, BUT with a constraint on the coefficient
  - This constraint is the multiple of a value λ and the sum of the absolute value of the coefficients
- Known as L1 regularization
- Final coefficients can be minimized to `0`, making LASSO a feature selection tool

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j}^{m} \mid \beta j \mid$$

Ridge and LASSO in R

- Can use the `glmnet` package to perform both lasso and ridge regression
- `cv.glmnet` performs cross-validated regularized regression
  - Identifies the λ that has the lowest average error across several held out test sets
- Set alpha to `0` for ridge regression, `1` for LASSO regression
- We can identify the λ that minimizes the error as well as the largest λ one standard error away
  - More shrinkage takes effect, but it isn't that different in terms of performance than the model with the minimum lambda
- Specify the argument `family = "binomial"` to perform regularized logistic regression

Additional Notes

- The same assumptions apply as in linear regression, but you'll need to plot the results manually
- In terms of predictions, we have to specify `newx`, which is the matrix we'll be making predictions on
- `s` is the value of lambda for the model that is used to make the predictions