ridgeRegression

October 27, 2019

1 Ridge Regression

Our goal is to learn about Ridge Regression. This technique is closely related to standard linear regression, but adds an extra twist. The twist involves including an additional term in our minimization of sums of squares that encourages coefficients to shrink.

The idea behind ridge regression is that influencing coefficients to shrink towards zero will also discourage overfitting to our training data. If we can prevent overfitting than we can better generalize our model to yet uncollected data.

1.1 Mechanics

We saw in previous lectures that finding optimal coefficients in a linear regression is the same as minimizing the sum squares error. More concretely, give a set of observations $(x,y)_1, (x,y)_2, \cdots, (x,y)_N$ we can write our probabilistic model as

$$p(y|x) \sim N(\beta_0 + \beta' x, \sigma^2)$$

where β_0 is an intercept.

The optimal β parameters are found by minimizing the following function of β

$$\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta' x_i \right)^2 \right\}$$

Ridge regression adds an additional term to the above optimization problem

$$\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta' x_i \right)^2 \right\} + \lambda \sum_{m=1}^{M} \beta_m^2$$

The λ parameter is a choice on part of the investigator. Typically, λ is chosen by cross-validation.

This additional term penalizes coefficients related to covariates. Ridge regression does **not** penalize the size of the intercept.

1.2 Optimal β

We can solve the above by taking partial derivatives and finding the β that zeros out these derivatives.

$$f(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta' x_i\right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2$$

$$f(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta' x_i\right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2 \tag{1}$$

1.3 Example data

[ ]:

```
```