Lasso Regression

Christian Thieme

5/13/2021

Using Lasso Regression for Variable Selection

Lasso regression, like ridge regression, penalizes variables in a model and shrinks them. However, lasso regression, unlike ridge regression, will shrink some of the coefficients to 0. With this in mind, we can use lasso regression to perform variable selection. Coefficients that have not been shrunk to 0, are those that should be included in the model.

We'll demonstrate this with a well known baseball dataset called Hitters in which the goal is to predict a player's salary.

```
Hitters <- na.omit(Hitters)
Hitters <- Hitters %>% dplyr::select(-NewLeague)
glimpse(Hitters)
```

```
## Rows: 263
## Columns: 19
## $ AtBat    <int> 315, 479, 496, 321, 594, 185, 298, 323, 401, 574, 202, 418...
## $ Hits     <int> 81, 130, 141, 87, 169, 37, 73, 81, 92, 159, 53, 113, 60, 4...
## $ HmRun    <int> 7, 18, 20, 10, 4, 1, 0, 6, 17, 21, 4, 13, 0, 7, 20, 2, 8, ...
## $ Runs     <int> 24, 66, 65, 39, 74, 23, 24, 26, 49, 107, 31, 48, 30, 29, 8...
## $ RBI      <int> 38, 72, 78, 42, 51, 8, 24, 32, 66, 75, 26, 61, 11, 27, 75,...
## $ Walks    <int> 39, 76, 37, 30, 35, 21, 7, 8, 65, 59, 27, 47, 22, 30, 73, ...
## $ Years    <int> 14, 3, 11, 2, 11, 2, 3, 2, 13, 10, 9, 4, 6, 13, 15, 5, 8, ...
## $ CAtBat   <int> 3449, 1624, 5628, 396, 4408, 214, 509, 341, 5206, 4631, 18...
## $ CHits    <int> 835, 457, 1575, 101, 1133, 42, 108, 86, 1332, 1300, 467, 3...
## $ CHmRun   <int> 69, 63, 225, 12, 19, 1, 0, 6, 253, 90, 15, 41, 4, 36, 177,...
## $ CRuns    <int> 321, 224, 828, 48, 501, 30, 41, 32, 784, 702, 192, 205, 30...
## $ CRBI     <int> 414, 266, 838, 46, 336, 9, 37, 34, 890, 504, 186, 204, 103...
## $ CWalks   <int> 375, 263, 354, 33, 194, 24, 12, 8, 866, 488, 161, 203, 207...
## $ League   <fct> N, A, N, N, A, N, A, N, A, A, N, N, A, N, N, A, N, N, A, N...
## $ Division <fct> W, W, E, E, W, E, W, W, E, E, W, E, E, E, W, W, W, E, W, W...
## $ PutOuts  <int> 632, 880, 200, 805, 282, 76, 121, 143, 0, 238, 304, 211, 1...
## $ Assists  <int> 43, 82, 11, 40, 421, 127, 283, 290, 0, 445, 45, 11, 151, 4...
## $ Errors   <int> 10, 14, 3, 4, 25, 7, 9, 19, 0, 22, 11, 7, 6, 8, 10, 16, 2,...
## $ Salary   <dbl> 475.000, 480.000, 500.000, 91.500, 750.000, 70.000, 100.00...
```

In taking a glimpse at the dataset, you can see that there are 19 features that we can use to predict our target variable `Salary`. How do we know which of these variables are actually useful in predicting salary? This is where Lasso Regression comes in. In using lasso regression we'll need to adjust our dataframe to a matrix.

```
x <- model.matrix(Salary ~ ., Hitters)[,-1]
y <- Hitters$Salary
```

Next, we can train a lasso model. The `glmnet` function works for both ridge and lasso regression. Alpha is set to 0 for ridge regression and is set to 1 for lasso regression. The optimal lambda can be found through a grid search. Here I show the value found for this particular dataset. Next we can run `predict`, with type 'coefficients' to see the coefficients.

```
model <- glmnet(x,y, alpha = 1, lambda = 2.436791)
coefficients <- predict(model, type = "coefficients", s = 2.436791)
coefficients
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                        1
## (Intercept)  129.3771309
## AtBat         -1.6336069
## Hits           5.8538016
## HmRun          .
## Runs           .
## RBI            .
## Walks          4.8802498
## Years         -9.7441004
## CAtBat         .
## CHits          .
## CHmRun         0.5836731
## CRuns          0.6967581
## CRBI           0.3696566
## CWalks        -0.5672148
## LeagueN       32.7904226
## DivisionW   -119.0549258
## PutOuts        0.2750933
## Assists        0.1878598
## Errors        -2.1368193
```

In viewing the output above, we can see that 5 variables have been forced to 0 (the dots). The coefficients in this model that have not been forced to 0 are those that should be included in the model.

Here we have shown that lasso regression can be used for more than just regression. We can use it as a means of variable selectio as well.