Regularization

February 3, 2019

1 Regularization

1. Ridge Regression
2. Lasso Regression
3. Elastic Net Regression

The 3 commonly used methods for finding sweet spot between simple models and complicated models are: - Regularization - Boosting - Bagging

1.1 Ridge Regression

- Helpful when sample size of training data is small which can lead to poor Least Square estimates that result in bad ML predictions.
- Ridge Regression can improve predictions made from new data (i.e. reduce variance) by making the predictions less sensitive to the training data.
- When the model fits the training data well (SSR is small) compared to the test data (SSR is large), it means that the line fit is overfit to the training data (has low bias and high variance).
- The main idea behind Ridge Regression is to find a new line that does not fit the training data as well by introducing a small amount of Bias in to how the new line is to fit the training data.
- By introducing the small amount of Bias, we get a significant drop in variance.
- In other words, by starting with a slightly worse fit to the training data, Ridge Regression can provide better long term predictions.
- Even when there is not enough data to find the Least Squares parameter estimates, Ridge Regression can find a solution with Cross Validation and the Ridge Regression Penalty.

1.1.1 Ridge Regression Penalty

- When Least Squares determines values for the parameters in the equation, it minimizes the sum of squared residuals (SSR).
  $y = mX + c$

where,
c is the y-intercept
m is slope

- In contrast, when Ridge Regression determines values for the parameters in the above equation, it minimizes the Ridge Regression Penalty term.

$SSR + \lambda \times m^2$.
m^2 term adds a penalty to the traditional LS method
λ determines how severe the penalty is.

- Ridge Regression is just Least Squares (SSR) + Ridge Regression Penalty ($\lambda \times m^2$).
- Without the small amount of Bias that the penalty creates, the LS fit has a large amount of Variance.
- In contrast, the Ridge regression line which has the small amount of Bias due to penalty, has less Variance.
- The predicted value of y is sensitive to the slope m. For large values of m the value of y fluctuates greatly wrt changes in x.
- Ridge regression with the added Bias term helps in identifying m more accurately in turn predicting y.
- Ridge Regression does not fit the training data as well as the LS method. In other words, Ridge Regression has more Bias than Least Squares, but in return for that small amount of Bias, the Ridge Regression line has a significant drop in Variance.
- The main idea is that by starting with a slightly worse fit, Ridge Regression provides better long term predictions.

1.1.2 Multiple linear regression

$y = w_1 x_1 + w_2 x_2 + c$

- Minimizing factor:
  $SSR + \lambda \times ({w_1}^2 + {w_2}^2)$

  where, $\lambda \times ({w_1}^2 + {w_2}^2)$ is called the Ridge Regression Penalty.

- For an equation with n parameters, Least Squares needs atleast n data points to estimate all n parameters.
- Ridge Regresson can find a solution with cross validation and the Ridge Regression penalty that favors smaller parameter values.

Lambda

- λ value can be anything from 0 to +infinity.
- When $\lambda = 0$, the Ridge regression line will minimize only SSR value, which is same as Least Squares method since $\lambda m^2 = 0$.
- When λ value is larger, the predictions for y become less and less sensitive to x.

How to determine the value of λ?

- Try a number of values for λ using k-fold cross validation to determine which one results in the lowest Variance.

Other applications

- Ridge regression can be applied to data containing both continuous and discrete variables.
- Ridge regression can be applied to Logistic Regression as well.
- Ridge regression optimizes the sum of likelihoods instead of the squared residuals because Logistic Regression is solved using Maximum Likelihood.
- Factor to minimize:
  the sum of likelihoods + $\lambda \times m^2$

1.2 Lasso Regression

- Lasso regression is similar to Ridge Regression.
- Lasso and Ridge regression make the predictions y less sensitive to the small training datasets.
- In the Lasso Regression Penalty, instead of square the parameter coefficients, absolute values are considered.

$y = w_1x_1 + w_2x_2 + c$

- Minimizing factor:
  $SSR + \lambda \times (|w_1| + |w_2|)$

where, $\lambda \times (|w_1| + |w_2|)$ is called the Ridge Regression Penalty.

- When λ = 0 then, the Lasso regression line will be the same as the Least Squares line.
- As λ increases in value, the slope gets smaller until the slope = 0.

Similarities between Lasso and Ridge regression

- Just like Ridge Regression λ can be any value from 0 to +infinity and is determined by cross validation.
- When Ridge and Lasso Regression shrink parameters, they don't have to shrink them all equally.

Dissimilarities between Lasso and Ridge regression

- The big difference between Ridge and Lasso Regression is that Ridge regression can only shrink the slope assymptotically close to 0 while Lasso regression can shrink the slope all the way to 0.

Example $y = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + c$

In the above equation, - y is the variable to predict - w_1 and w_2 are reasonable variables to predict y, and - w_3 and w_3 are the least reasonable variables to predict y.

Using Ridge Regression, the penalty term is $\lambda \times ({w_1}^2 + {w_2}^2 + {w_3}^2 + {w_4}^2)$

The larger we make the value of λ, the parameters w_1 and w_2 may shrink a little bit and w_3 and w_4 may shrink a lot, but they will never be equal to 0.

In contrast, when we increase the value of λ, w_3 and w_4 can go all the way to 0 and we are left with a model to predict y that only includes w_1 and w_2 and excludes all the bad predictors.

- Since, Lasso Regression can exclude useless variables from equations, it is a little better than Ridge Regression at reducing the Variance in models that contain a lot of useless variables.
- Ridge Regression is useful, when most of the variables are useful.

1.3 Elastic Net Regression

- When we know a lot about all the parameters in our model, it is easy to choose if we want to use Lasso Regression or Ridge Regression.
- When using models that include millions of parameters - far too many to know everything about, and whe you have millions of parameters, then some form of regularization is needed to estimate them.
- In such case, where we have a lot of variables and we dont know which variables are useful, then Elastic-Net Regression is useful.
- Starting with Least Squares, Elastic Net Regression combines the strengths of Lasso and Ridge Regression.

Elastic Net Regression Penalty:
$SSR + \lambda_1 \times (|w_1| + |w_2| + |w_3| + |w_4|) + \lambda_2 \times ({w_1}^2 + {w_2}^2 + {w_3}^2 + {w_4}^2)$

- We use cross validation on different combinations of λ_1 and λ_2 to find best values.
- When the values of both λ_1 and λ_2 are 0, we get Least Square parameter estimates.
- When $\lambda_1 > 0$ and $\lambda_2 = 0$, we get Lasso Regression parameter estimates.
- When $\lambda_1 = 0$ and $\lambda_2 > 0$, we get Ridge Regression parameter estimates.
- When $\lambda_1 > 0$ and $\lambda_2 > 0$, we get a hybrid of Ridge and Lasso Regression parameter estimates.

Advantages

- The hybrid Elastic-Net Regression is especially good at dealing with situations when there are correlations between parameters.
- This is because Lasso Regression on its own, tends to pick just one of the correlated terms and eliminates the others.
- Ridge regression tends to shrink all of the parameters for the correlated variables together.
- By combining Lasso and Ridge Regression, Elastic-Net Regression groups and shrinks the parameters associated with the correlated variables and leaves them in the equation or removes them all at once.