The classical linear regression model is good. Why do we need regularization?

Grace Zhang
Oct 28, 2018 · 5 min read

Motivation

The linear regression model is probably the simplest and the most commonly used prediction model. There are a lot of advantages of using a linear regression model. The most important one is that under the assumption of i.i.d normal distribution of error terms, the OLS (Ordinary Least Squares) estimators of the linear regression model are unbiased (based on *Gauss-Markov Theorem*), thus yield useful inferences.

But there are cases that the classical linear regression model doesn't handle well:

1. When there is **multicollinearity. Multicollinearity** is the phenomenon that one (or more) of the independent variable(s) can be expressed as the linear combination of other independent variables. In fact, this problem almost exists everywhere in the real world.
2. When the number of independent variables is larger than the number of observations. When this happens, the OLS estimates are **not valid** mainly because there are infinite solutions to our estimators.

……

Hence, we need better alternatives to solve these problems.

. . .

Introduction to Regularization

There are several ways to solve the above-mentioned problems, like feature selection, regularization, dimensionality reduction, etc. Today, I will only focus on

Regularization is a process of introducing additional information in order to solve an ill-posed problem or to prevent overfitting (Wikipedia). One way to regularize is to add a constraint to the loss function:

Regularized Loss = Loss Function + Constraint

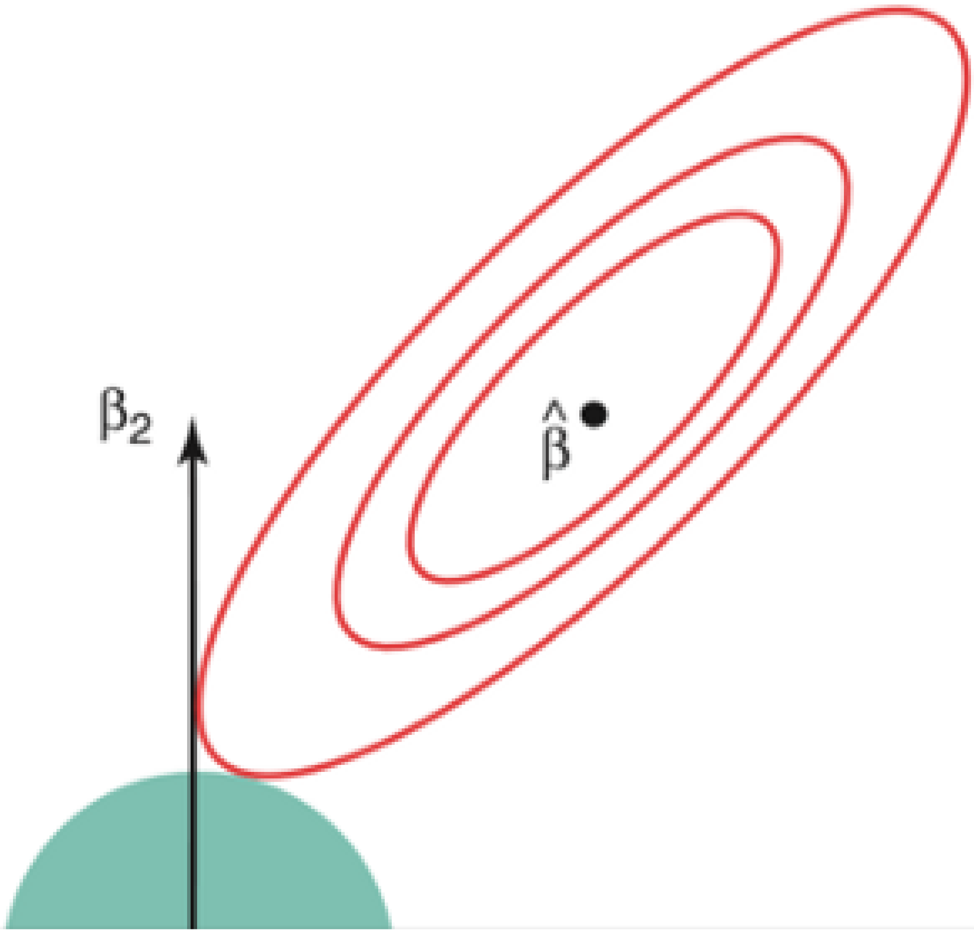
There are multiple different forms of constraints that we could use to regularize. The three most popular ones are Ridge Regression, Lasso, and Elastic Net.

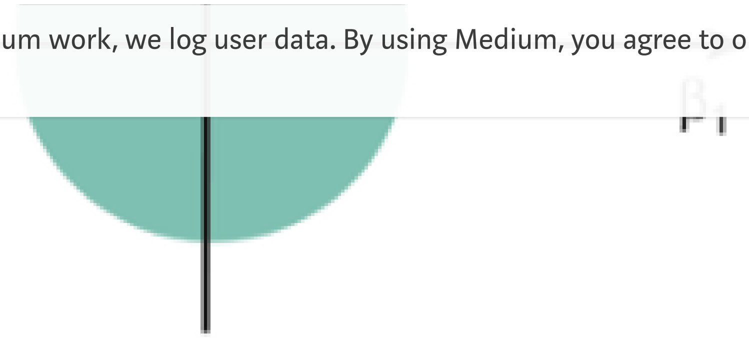
Ridge Regression

Ridge regression is also called **L2 regularization**. It adds a constraint that is a linear function of the squared coefficients.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

picture from Towards Data Science





To minimize the regularized loss function, we need to choose λ to minimize the sum of the area of the circle and the area of the ellipsoid chosen by the tangency.

Note that when λ tends to zero, the regularized loss function becomes the OLS loss function.

When λ tends to infinity, we get an intercept-only model (because in this case, the ridge regression coefficients tend to zero). Now we have smaller variance but larger bias.

A critique of ridge regression is that all the variables tend to end up in the model. The model only shrinks the coefficients.

Lasso

Lasso is also known as **L1 regularization**. It penalizes the model by the absolute weight coefficients.

Lasso works in the following way: it forces the sum of the absolute value of the coefficients to be less than a constant, which forces some of the coefficients to be zero and results in a simpler model. This is because comparing to the L2 regularization, the ellipsoid has the tendency to touch the diamond-shaped constraint on the corner.

Lasso performs better than ridge regression in the sense that it helps a lot with feature selection.

Elastic Net

Elastic Net is the combination of the L1 regularization and L2 regularization. It can both shrink the coefficients as well as eliminate some of the insignificant coefficients.

• • •

Tuning Parameter λ

As we discussed above, choosing the right tuning parameter λ is quite important in regularization. For instance, ridge regression mainly solves the problem of multicollinearity. Therefore we want the tuning parameter λ to be relatively small in order to control the bias being introduced to the model. Lasso works well in terms of feature selection thus we want to balance between fitting the model and shrinking the coefficients.

Statistically, there are multiple available options for tuning parameter selection, including CV (Cross Validation), AIC (Akaike's Information Criterion), and BIC (Bayesian Information Criterion). Below is a list of researches you can refer to if you would like to know more:

Regularization and variable selection via the elastic net

LASSO Tuning Parameter Selection

Luckily, we have useful Python and R packages that could help us figure out the appropriate tuning parameters. glmnet (R) and scikit-learn (Python) are powerful packages that can help you solve the problem.

The End

We always want to fit a model that does well on training set (low bias) and predicts well on the unseen data (low variance). Although regularization does introduce bias to the model, the tradeoff is that it lowers the variance of the model. In practice, whether we should choose the OLS linear regression model or the regularized model varies case by case. You can always use the linear regression model as the base model, compare it with the regularized model, and find the best fit.

This article is inspired by the Linear Regression Analysis and Machine Learning classes I take in University of San Francisco.

I hope you enjoy this article. Please let me know if you have any questions, comments, suggestions, etc. Thanks for reading :)

Reference

Linear Regression Analysis at University of San Francisco by Jeff Hamrick

Machine Learning at University of San Francisco by Brian Spiering

Penalized Regression Essentials: Ridge, Lasso & Elastic Net

Lasso, Ridge and Elastic Net Regularization by Jayesh Bapu Ahire

Regularization in Machine Learning by Prashant Gupta

Machine Learning Linear Regression Data Science Beginners Guide