The Progression of Regression Analysis: Least Squares, Ridge Regression, Lasso, and Random Lasso

Created by, James Matthew Hamilton

KENNESAW STATE UNIVERSITY
College of Computing and Software Engineering

Abstract

Random Lasso is the state-of-the-art method of regression for predicting gene regulatory networks in bioinformatics. The Complexity of lasso regression requires a deeper understanding of its predecessors in order to make it a malleable tool of bioinformatic research. Several refinements occurred during the development of regression analysis. Now a vast array of methods exist, but it is often the case that one regression method will not fit all tasks. Looking into obscure branches of regression may hold clues needed for the development of more personalized regression methods.

Research Question

What statistical challenges culminated in the development of the modern regression analysis used in bioinformatics?

Least Squares

Least squares regression derives an equation that attempts to minimize the sum of the square residuals, i.e. sum of all errors. Least squares regression predicts the values of y, the dependent variable, based on the value of x, the independent variable. Switching the x and y variables may seem inconsequential, but this in fact can radically alter the line of best fit. One way to tackle this uncertainty is to do a bit of math with both x and y as the independent variable.

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Ridge Regression

Ridge regression is well suited for machine learning, because an adjustable lambda value is used to impose some bias on the least squares method. The least squares methods is often used for two variables, but ridge regression can effectively be used for millions of variables.

$$\beta^{ridge} = \frac{1}{n} \sum_{i=1}^{n} y_i + \lambda \cdot |slope_1| + |slope_2| + ... + |slope_n|$$
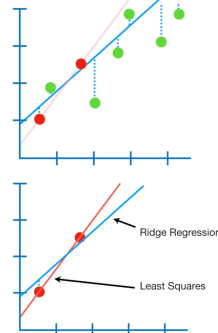
$\alpha = 1 \rightarrow Lasso$
$\alpha = 0 \rightarrow Ridge$
$\alpha = (0,1) \rightarrow Elasticnet$
$\lambda \rightarrow$ Tuning parameter
$\beta \rightarrow$ Parameter estimates

Ridge Regression

Least Squares

Lasso Regression

Lasso's equation is identical to ridge regression, except that it squares the slope. Since lasso regression can reduce useless variables to zero, it is a little better than ridge regression for at reducing the variance in models that contain a lot of useless variables. This is helpful in gene regulatory networks where many genes have no effect and should be reduced to zero.

$$\beta^{lasso} = \frac{1}{n} \sum_{i=1}^{n} y_i + \lambda \cdot slope_1^2 + slope_2^2 + ... + slope_n^2$$

Problems Caused by Variance in the Independent Variable

A line of best fit derived from the least squares method. The equation is linear. Given a value for y we should be able to predict x within reason.

Two lines of best fit derived from flipping the dependent and independent variable in the least squares method. The shallow slope is obtained when the independent variable is set to the x-axis. The steeper slope is obtained when the independent variable is set to the y-axis.

Elastic Net Regression

Elastic net regression capitalizes on the strengths of both ridge regression and lasso regression. When data with millions of variables is used, it is often that case that we do not know if there are useless variables that should be shrunk to zero. Two adjustable lambda values exist. If either lambda is set is zero we no longer have electric net regression — we would have either ridge regression or lasso regression.

$$\beta^{net} = \frac{1}{n} \sum_{i=1}^{n} y_i + \lambda_1 \cdot |slope| ... + \lambda_2 \cdot slope^2 ...$$
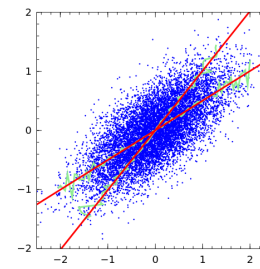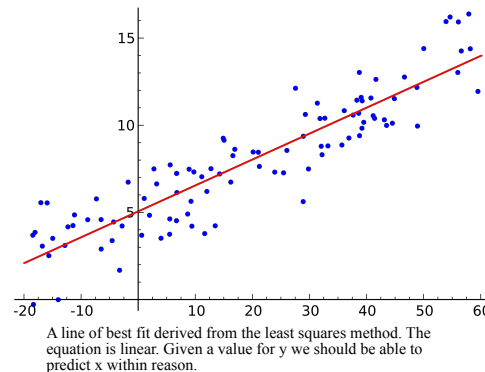
Random Lasso

Random lasso tends to select or remove highly correlated variables more efficiently than the elastic-net regression. Random lasso also alleviates two possible limitations of lasso:

I. When a model includes several highly correlated variables, all of which are related to some extent to the response variable, lasso tends to pick only one or a few of them and shrinks the rest to 0. This may not be a desirable feature.

II. When p > n, lasso can identify at most only n variables before it saturates. This again may not be a desirable feature for many practical problems, particularly microarray studies, for it is unlikely that only such a small number of genes are involved in the development of a complex disease.

Random lasso is too complex to be written as an equation, and is more of an algorithm.

[†] Sijian Wang, Bin Nan, Saharon Rosset and Ji Zhu *Random Lasso* University of Wisconsin, University of Michigan, Tel Aviv University