Project 3: Regularization in Least Squares

University of Colorado Boulder
Department of Applied Mathematics

1 Project Summary

Regularized least squares is a family of methods that adds terms to the least squares formulations to constrain the solution. Regularization can helps users incorporate goals that we want the resulting approximating function to have such as the size of the coefficients, sparsity (how many are non-zero), smoothness of the model, etc. In this project, you will explore the least squares regularization techniques used in ridge regression and Tikhonov regularization. How do the different techniques effect the resulting least squares problem in terms of normal equations and other linear algebra techniques? In this project you will explore how these techniques can be used to minimize the effects of noise and other factors on the resulting regression lines. Possible options for the independent direction of this project include generalizations of Tikhonov, L_1 regularization (e.g. LASSO), continuous regularized least squares, and smooth spline approximation (Reproducing Kernel Hilbert Spaces).

2 Project Background

A key numerical tool for many data scientists and statisticians is regression models. At this point, the primary algorithm we have explored for building these models is the least squares model, which aims to solve for a model $\boldsymbol{x}$ that minimizes the loss function or constraint

$$\underset{\boldsymbol{x}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2, \tag{1}$$

Where $\boldsymbol{A}$ is an $n \times m$ matrix, $\boldsymbol{x}$ is $m \times 1$ column vector, and $\boldsymbol{b}$ is an $n \times 1$ column vector.

In order for least squares approximations to produce a useful result several conditions must be met. First, the resulting linear system needs to have a unique solution. Second, enough information about the problems must be known in order to chose a appropriate approximating function and avoid over-fitting. In addition to these requirements, it may be preferable for $\boldsymbol{x}$ to be sparse (have few non-zero entries) to allow for simpler interpretation of $\boldsymbol{x}$.

Addressing these issues is non-trivial and an active area of research. In this project, you will explore some of the most common ways of trying to avoid these issues. These approaches reformulate the least squares problem in order to avoid the issues always aiming for greater numerical stability and sometimes sparsity. The result is a less accurate approximation than one would expect with the amount of given data.

2.1 Ridge Regression

The first method you will explore that modifies the least squares problem is *ridge regression*. In some cases the $\boldsymbol{A}$ matrix that represents the data we are using to build our regression to will be ill-posed. This can be caused by noise in measurements or the inclusion of variables that have no real correlation with the output being predicted. When these conditions occur it becomes possible that an exact solution to the original system does not exist and the solution to the problem $\boldsymbol{x}$ will begin to exhibit undesirable properties, such as large oscillatory regression coefficients as the least squares problem attempts to compensate for the abnormalities in the data. One way to fight the growth in regression coeffcients is to change the loss function of the least squares problem in equation 1 to include a penalization term on the magnitude of $\boldsymbol{x}$ as shown below

$$\underset{\boldsymbol{x}}{\operatorname{argmin}} \ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \gamma\|\boldsymbol{x}\|_2^2 \tag{2}$$

where γ is a constant chosen to optimally penalize the l_2-norm of $\boldsymbol{x}$.

As with the method of least squares, the loss function for ridge regression cannot be solved directly to determine the minimizer $\boldsymbol{x}$. Instead we must determine a matrix $\boldsymbol{E}$ such that $\boldsymbol{x}^* = Eb$. It can be shown that the estimator that provides the optimal solution to the matrix system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ under the constraint set forth by equation 2 is

$$E_{ridge} = (\boldsymbol{A}^T\boldsymbol{A} + \gamma\boldsymbol{I})^{-1}\boldsymbol{A}^T.$$

It is the form of the ridge estimator that gives ridge regression its name. If $\gamma >> 0$ the values along the main diagonal of the estimator become very large forming a sort of "ridge" in the matrix.

2.1.1 Questions to Investigate

1. Deriving the Ridge Estimator

   (a) Given that $\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^T\boldsymbol{x}$, rewrite equation (2) in terms of matrix products. *Hint: just like the least squares case you should get a quadratic.*

   (b) Using methods from calculus and properties from linear algebra, determine the point $\boldsymbol{x}^*$ that minimizes equation (2).

2. Exploring the Ridge Estimator

   (a) i. Sample the line $y = 3x + 2$ 20 times and add some Gaussian noise.

      ii. Select 10 of these data points at random and use them to build a line of best fit ($y = mx + b$) for the data using ridge regression, $\gamma = 0$. We will call the other 10 points the validation set.

      iii. Calculate the sum of squared errors of your model on the other 10 points.

      iv. Increase γ by 0.1 and build another regression to test against the validation set. Does it perform worse or better?

      v. Repeat this process for different values of γ, what is the optimal value? Which model most accurately fits $y = 3x + 2$?

   (b) Repeat the same process you perfromed in 2a but this time sample the line $y = x^2$ and fit the model $y = ax^5 + bx^4 + cx^3 + dx^2 + ex + f$. How does Ridge regression help improve this model?

2.2 Tikhonov's Regularization

Ridge regression is a method that is able to create more accurate approximations than the standard least squares approach when the corresponding linear system is under determined or the underlying physics being modeled is poorly understood. It does this by introducing an acceptable amount of bias into the problem. By bias we mean that we penalize the magnitude of the $\boldsymbol{x}$ vector, we bias our solution to have a smaller magnitude. Unfortunately, ridge regression does not solve all of the original issues with least squares and introduces some new issues that one would like to avoid. For example, the ridge normalization effects each of the variables in the regression equally. In other words, the ridge model will shrink all of the coefficients of $\boldsymbol{x}$ no matter how important they are. The next technique you will investigate allows for more control over how the regularization will affect the solution. It is called *Tikhonov Regularization.*

Note that the minimization problem for ridge regression can be written as

$$\underset{\boldsymbol{x}}{\operatorname{argmin}} \; \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + ||\sigma \boldsymbol{I}\boldsymbol{x}||_2^2,$$

where $\sigma = \sqrt{\gamma}$. Tikhonov Regularization considers a more general form of this minimization problem. Specifically, it poses a more general minimization problem of the form

$$\underset{\boldsymbol{x}}{\operatorname{argmin}} \; \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \|\boldsymbol{\Gamma}\boldsymbol{x}\|_2^2.$$

where $\boldsymbol{\Gamma}$ is a *weight matrix* that is chose so that $\boldsymbol{x}$ will satisfy properties desired by the user or so that certain vectors $\hat{\boldsymbol{x}}$ can be eleminated from the possible solutions. Typically, the Tikhonov Regularizatio problem is recast as

$$\underset{\boldsymbol{x}}{\operatorname{argmin}} \; \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \lambda^2\|\boldsymbol{L}\boldsymbol{x}\|_2^2. \tag{3}$$

where $\boldsymbol{L}$ is a unitary matrix and the constant λ is a weighting constant. For the remainder of this document, we consider the minimization problem in equation (3) when we refer to Tikhonov regularization.

Like ridge regression, Tikhonov regularization discourages the elements of the solution $\boldsymbol{x}$ from being oscillatory and large. Additional feature of Tikhonov regularization is that it is able to enforce smoothness in the derivative of the solution $\boldsymbol{x}$. Here the derivative is mean in a finite difference sense. In other words, the derivative is approximated via weighted averages. One of the most common finite difference equations for approximating the first derivative of a function is the centered difference. The centered difference approximation of $f'(t)$ is given by

$$\frac{df}{dt}(t) \approx \frac{f(t+\Delta t) - f(t-\Delta t)}{2\Delta t}.$$

The truncation error in this approximation is $\mathcal{O}((\Delta t)^2)$.

Remark 2.1. Finite difference approximations can be constructed by adding together linear combinations of Taylor series. The truncation error is of the order of the first term in the expansion of the remainder of the series that is not included in the approximation.

For a vector function $\boldsymbol{x}$ of length n with j^{th} entry x_j and assuming $\Delta t = 1$, the derivative of $\boldsymbol{x}$ can be approximated using the matrix product

$$\boldsymbol{x}' = \boldsymbol{D}\boldsymbol{x} = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \dots & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{4}$$

The matrix $\boldsymbol{D}$ is an $(n-2) \times n$ tridiagonal matrix.

In Tikhonov regularization, this derivative matrix is used in the regularization term as follows:

$$\underset{\boldsymbol{x}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda^2 \|\boldsymbol{D}\boldsymbol{x}\|_2^2. \tag{5}$$

With the new loss function, the least squares estimator can be derived, and the final solution to the problem can be solved. To determine an optimal value for λ the most simple technique that is used is cross-validation. Cross validation is a fancy way of saying guess and check, increasing the magnitude of λ until the error in the model is minimized on a data set that is separate from the data that was used to build the model.

2.2.1 Questions to Consider

1. Show that $\boldsymbol{D}$ is unitary.

2. Using the methods you used to derive the ridge estimator, derive the estimator needed to solve the the Tikhonov regularization problem.

3. Sample the curve $y = sin(x) + sin(5x)$ and add gaussian noise. Use Tikhonov Regularization to calculate a regression for a data set with the smoothing method explained above. How does the model fit? How does it compare to the least squares solution. How do the derivatives compare.

4. Other finite difference methods exist to approximate the first derivative of a discrete signal. What happens if you use those methods to penalize this model? One option is the forward finite difference formula which has a truncation error of $\mathcal{O}(\Delta h)$. The forward difference is given as

$$\frac{df}{dt}(t) \approx \frac{f(t + \Delta t) - f(t)}{Deltat}.$$

3 Software Expectations

You may use software for this project to solve the systems of equations resulting from the ridge and Tikhonov estimators. By this we mean you do not need to code up your own version of gaussian elimination. Though it is easy to build your own solver using standard linear algebra packages.

If you would like to dive deeper into these methods (possibly for the independent direction), you can explore the python library cvxpy to perform your regressions. Using this library is not required nor expected. Documentation on this library can be found at https://www.cvxpy.org/.

4 Independent Directions

Possible next steps for this project include, but are not limited to:

1. Explore LASSO and elastic net constraints for the least squares problem. These methods hope to adjust the loss function to favor non-zero coefficients only for the most variables/terms to promote sparsity in the final system.

2. Explore the applications of least squares in system identification techniques. One of these techniques, the Sparse Identification of Non-linear Dynamics algorithm (SINDY), employs regularized least squares to fit differential equations to discrete data sets by approximating derivatives with finite difference methods. This extension is very cool and but will require additional research by the students.

3. The application of regularization in continuous least squares models. These methods seek to build least squares approximations to a function over a given interval instead of a predetermined discreet set of points.

5 Helpful Sources

1. https://epubs.siam.org/doi/abs/10.1137/S0895479897326432

2. https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b