Regularization and condition number

Ka Hung Wong

January 6, 2018

1 Regularization and Condition Number

1.1 Introduction of Regularization

The residual sum of squares (RSS) with regularization on the coefficients of fitting is given by $J = ||y - X\beta||^2 + ||\beta||^2$. If we use L2 norm, we have:

$$J = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

Solving for the coefficients, β by first taking the derivative to the RSS function:

$$\begin{aligned}\frac{\partial J}{\partial \beta} &= 2(y - X\beta)^T - \frac{\partial X\beta}{\partial \beta} + 2\lambda\beta^T I \\ &= -2(y - X\beta)^T X + 2\lambda\beta^T I\end{aligned}$$

Then, we set the above equation to zero, we get:

$$\begin{aligned}\bar{\beta}^T &= y^T X(\lambda I + X^T X)^{-1} \\ \bar{\beta} &= (X^T X + \alpha I)^{-1} X^T y\end{aligned}$$

1.2 Condition Number

In real world application, there is some error in X and y, which can come from the inability to represent real numbers with finite precision, measurement, etc. Therefore, it is nice to estimate the relative error in estimated coefficients, β, due to the error in X and y.

First, we assuming there is no error in X but error in $X^T y$. Let $A = X^T X$ such A is a symmetric matrix. The condition number of A, which measures the maximum ratio of the error in estimated β due to the error in $X^T y$, is defined as:

$$\begin{aligned}k(A) &= \sup_{\Delta b} \frac{\frac{||A^{-1}\Delta b||}{||A^{-1}b||}}{\frac{||\Delta b||}{||b||}} \\ &= \sup_{\Delta b} \frac{||A^{-1}\Delta b||}{||\Delta b||} \sup_{b} \frac{||b||}{||A^{-1}b||} \\ &= \sup_{\Delta b} \frac{||A^{-1}\Delta b||}{||\Delta b||} \sup_{A^{-1}b} \frac{||AA^{-1}b||}{||A^{-1}b||} \\ &= ||A^{-1}||||A|| \\ &\geq ||A^{-1}A|| \\ &= 1\end{aligned} \tag{1}$$

Assuming there is no error on $A = X^T X$, and that $\Delta b \leq b$, we have

$$\begin{aligned}k(A) &= \sup_{\Delta b} \frac{||A^{-1}\Delta b||}{||A^{-1}b||} \sup_{b} \frac{||b||}{||\Delta b||} \\ &\geq \sup_{\Delta b} \frac{||A^{-1}\Delta b||}{||A^{-1}b||} \\ &= \frac{||\Delta\beta||}{||\beta||}\end{aligned}$$

This suggests the error in the estimated β is bounded by the product of the condition number of matrix A and the error in X^Ty. If the condition number is large, the error in the estimated β can be large even there is a relatively small error in X^Ty. Now we consider the case where there is no error in X^Ty but error in A, then we have

$$\begin{aligned} A\beta &= b \\ (A+\Delta A)(\beta+\Delta\beta) &= b \end{aligned} \tag{2}$$

Combining both equations, we have:

$$\begin{aligned} \Delta A(\beta+\Delta\beta) &= -A\Delta\beta \\ A^{-1}\Delta A(\beta+\Delta\beta) &= -\Delta\beta \\ ||A^{-1}||||\Delta A||||(\beta+\Delta\beta)|| &\geq ||\Delta\beta|| \\ ||A^{-1}||||\Delta A|| &\geq \frac{||\Delta\beta||}{||(\beta+\Delta\beta)||} \\ ||A^{-1}||||\Delta A|| &\geq \frac{||\Delta\beta||}{||(\beta+\Delta\beta)||} \\ k(A)\frac{||\Delta A||}{||A||} &\geq \frac{||\Delta\beta||}{||(\beta+\Delta\beta)||} \end{aligned} \tag{3}$$

This also suggests there is a relationship between the condition number and the error in estimated β assuming there is no error in X^Ty. In summary, condition number calculation is a tool for estimating the bound of the error in estimation based on the data. The calculation of condition number is related with the eigenvalues when the 2-norm is considered. Since the matrix is normal, $A^T = A$. The matrix has unitary decomposition.

$$\begin{aligned} k(A) &= ||A||||A^{-1}|| \\ &= ||UDU^*||||U^{-1*}D^{-1}U^{-1}|| \\ &= ||D||||D^{-1}|| \\ &= |\lambda_{max}|||D^{-1}|| \\ &= \frac{|\lambda_{max}|}{|\lambda_{min}|} \end{aligned}$$

Now we can prove that the L2 regularization can reduce the error in estimated β. Since A is X^TX and real, so it is positive definite. Now we consider the scaled identity matrix,

$$\begin{aligned} B &= A+\alpha I \\ &= UDU^T + U\alpha IU^T \\ &= U(D+\alpha I)U^T \end{aligned} \tag{4}$$

The eigenvalue of A is positive. Then, the condition number of B is:

$$\begin{aligned} k(B) &= \frac{\lambda_{max}+\alpha}{\lambda_{min}+\alpha} \\ &\leq \frac{\lambda_{max}}{\lambda_{min}} \\ &= k(A) \end{aligned} \tag{5}$$

This suggests the L2 regularization reduces the condition number of the system.

2 Elastic Net

The cost function of the elastic net is: $J = ||y - X\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2$. The cost function of the elastic net can transform to a lasso regression model by merging the sum of errors observed in predicted β and L2 regularization.

$$\begin{aligned}
||\widetilde{y} - \widetilde{X}\widetilde{\beta}||_2^2 &= ||\begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} \beta||_2^2 \\
&= ||y - X\beta||_2^2 + || - \sqrt{\lambda_2}\beta I||_2^2 \\
&= ||y - X\beta||_2^2 + \lambda_2 ||\beta||_2^2
\end{aligned}$$

where $\widetilde{y} = (y, 0)$, $\widetilde{X} = (X, \sqrt{(}\lambda_2)I)$, and $\widetilde{\beta} = \sqrt{(}1+\lambda_2)\beta$. This is known as the normal version of the elastic net, however, there is an improved version of elastic net is presented in Section 13.5.3.3 of Machine learning. It is:

$$J = \beta^T \left(\frac{X^T X + \lambda_2 I}{1 + \lambda_2} \right) \beta - 2y^T X \beta + \lambda_1 ||\beta||_1 \tag{6}$$

Starting from this formulation, we can add the weight term, and then we have:

$$\beta^T \left(\frac{X^T W X + \lambda_2 I}{1 + \lambda_2} \right) \beta - 2y^T W X \beta + \lambda_1 ||\beta||_1 \tag{7}$$

2.1 Gradient

Taking sub-derivative on the improved cost function, we have:

$$\begin{aligned}
\partial_\beta J &= 2\beta^T \left(\frac{X^T W X + \lambda_2 I}{1 + \lambda_2} \right) - 2y^T W X + \partial \lambda_1 ||\beta||_1 \\
&= 2\beta^T \left(\frac{X^T W X + \lambda_2 I}{1 + \lambda_2} \right) - 2y^T W X + \begin{cases} \lambda_1 \text{ if } \beta > 0 \\ -\lambda_1 \text{ if } \beta < 0 \\ [-\lambda_1, \lambda_1] \text{ if } \beta = 0 \end{cases}
\end{aligned} \tag{8}$$

We can use this to optimize the cost function, which is known as sub-gradient descent.

On the other hand, the cost function can be optimized by coordinate descent. The first step is to take a derivative on each component of the coefficients. First, we write the partial derivative as:

$$\partial_{\beta_k} J = \beta_i D_{ij} \beta_j - 2E_i \beta_i + \partial \lambda_1 ||\beta||_1$$

by letting $D = \frac{X^T W X + \lambda_2 I}{1 + \lambda_2}$ and $E = y^T W X$. Then, we have:

$$\partial_{\beta_k} J = 2 \sum_i \beta_i D_{ij} - N_k + \partial_{\beta_k} \lambda_1 ||\beta||_1$$

By rearranging the term, we get:

$$\beta_k^* = S\left(\alpha_k, \widetilde{\lambda_k}\right) \tag{9}$$

where

$$\begin{aligned}
\alpha_k &= \frac{E_k - \beta D_k + \beta_k D_{kk}}{D_{kk}} \\
\widetilde{\lambda_k} &= \frac{\lambda_1}{2D_{kk}}
\end{aligned}$$