MACHINE LEARNING

Electrical Summer Workshops (ESW) 2022

Electrical Engineering Department - Sharif University of Technology

Instructors: *Alireza Gargoori Motlagh – Amir Mirrashid – Ali Nourian*

REGULARIZATION AND SHRINKAGE METHODS

Ridge Regression

Ridge regularization is a shrinkage method which its goal is to reduce the variance of the model through reducing the absolute value of the regression parameters (coefficients).

The RSS cost function is modified by the L_2 norm of the coefficients:

$$cost = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\rightarrow cost = \sum_{i=1}^{n} \left(y^{(i)} - \beta_0 - \beta_1 x_1^{(i)} - \beta_2 x_2^{(i)} - \cdots - \beta_p x_p^{(i)}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

So a regularization term is added to the residual term in order to limit the radius of the parameters.

λ is called the hyperparameter of the model and it controls the reduction of the coefficients' values.

Ridge Regression Solution

First, by removing the intercept (β_0) from the cost function, we can rewrite the cost as follows:

$$cost = \left|\left|y - X\beta\right|\right|_{L2}^2 + \lambda\left|\left|\beta\right|\right|_{L2}^2 = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

Again, this is a convex function of regression parameters and we can set its derivative w.r.t. $\boldsymbol{\beta}$ to zero to find the optimal solution:

$$\frac{\partial\, RSS}{\partial\, \boldsymbol{\beta}} = -2\boldsymbol{X^T}(\boldsymbol{y} - \boldsymbol{X\beta}) + 2\lambda\boldsymbol{\beta} = 0 \rightarrow (\boldsymbol{X^TX} + \lambda\boldsymbol{I})\,\boldsymbol{\beta} = \boldsymbol{X^Ty}$$

Hence the solution would be:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X^TX} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X^Ty}$$

(which is *always unique even if the features are correlated and* X is not full column-rank)

Lasso

Lasso is a shrinkage method which its goal is to reduce the variance of the model through reducing the absolute value of the regression parameters (coefficients).

The RSS cost function is modified by the L_1 norm of the coefficients:

$$cost = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

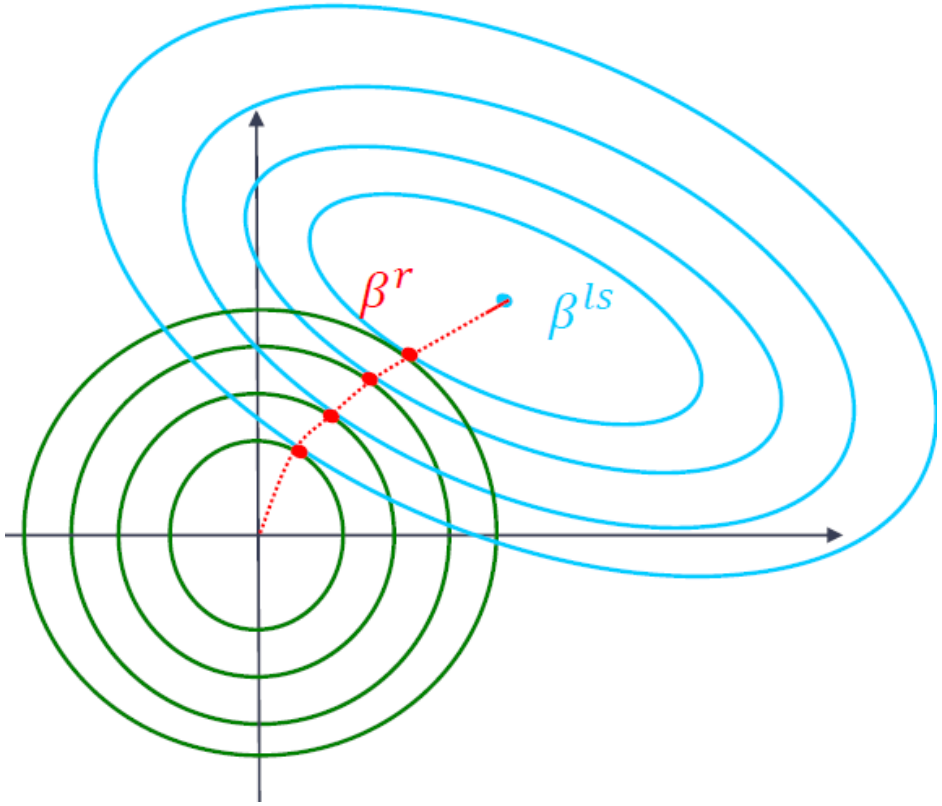
$$\rightarrow cost = \sum_{i=1}^{n} \left(y^{(i)} - \beta_0 - \beta_1 x_1^{(i)} - \beta_2 x_2^{(i)} - \cdots - \beta_p x_p^{(i)}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

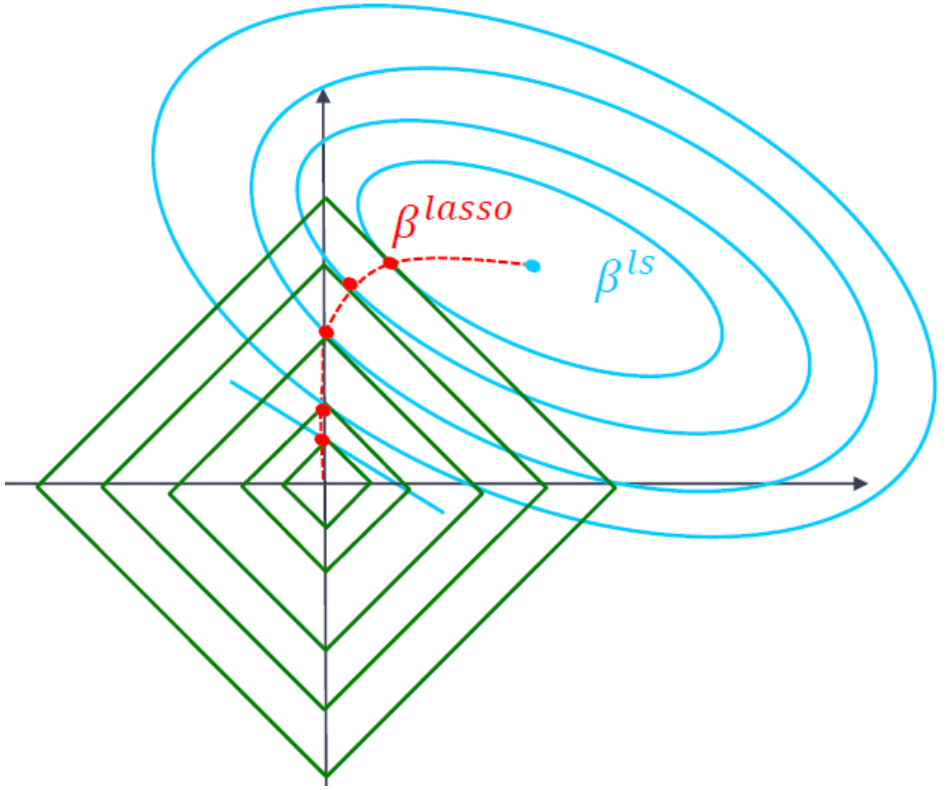
So a regularization term is added to the residual term in order to limit the radius of the parameters.

λ is called the hyperparameter of the model and it controls the reduction of the coefficients' values.

This is still a convex function of the regression parameters but with no analytical solution.

Ridge and Lasso Coefficients Path





Linear Regression vs Ridge vs Lasso

The effect of λ hyperparameter also results in U-shape curve for test error.