LASSO Regression

DR. JASMEET SINGH,
ASSISTANT PROFESSOR,
CSED, TIET

LASSO Regression

- **L**east **A**bsolute **S**election and **S**hrinkage **O**perator (LASSO) regression performs ‘**L1 regularization**‘, i.e. it adds a factor of sum of absolute values of coefficients in the optimization objective.

- Thus, ridge regression optimizes the following:

$$Cost\ (Objective) Function = Mean\ Square\ Error + \lambda(\text{sum of absolute values of coefficients})$$

i.e., $J = \frac{1}{2n}\sum_{i=1}^{n}((y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \cdots \dots \dots \dots .. -\beta_k x_{ik})^2 + \lambda \sum_{j=0}^{k} |\beta_j|)$

where n are the total number of training examples; k are the number of features; β_j represents regression coefficients of j^th input variable and λ is the regularization parameter.

LASSO Regression using Gradient Descent

- Since absolute functions are not differentiable at all points (for instance |β| is not differentiable at β=0), thus in order to compute partial derivative of J w.r.t β we use two techniques **Coordinate Descent** and **Soft thresholding rule**.
- To solve for each coefficient βj, coordinate descent is often used. The idea is to update one coefficient at a time while holding the others fixed.
- For the j-th coefficient, we isolate the corresponding term in the objective function, which leads to solving:

$$J(\beta_0, \beta_1, \beta_2, \dots \dots \beta_k) = \frac{1}{2n}\left(\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j}\beta_k x_{ik} - \beta_j x_{ij}\right)^2 + \lambda|\beta_j| + \lambda\sum_{j\neq k}|\beta_k|\right)$$

LASSO Regression using Gradient Descent Contd….

- Concentrating on j$^{\text{th}}$ coefficient only, the equation reduces to:

$$J(\,\beta_j) = \frac{1}{2n}\left(\sum_{i=1}^{n}(r_j - \beta_j x_{ij})^2 + \lambda|\beta_j|\right)$$

$$where\ \boldsymbol{r_j} = \boldsymbol{y_i} - \sum_{\boldsymbol{k \neq j}} \boldsymbol{\beta_k x_{ik}}\ \ and\ is\ called\ as\ \boldsymbol{partial\ residual\ or\ error}\ as\ it\ is\ the$$

$$difference\ between\ the\ actiual\ and\ predicted\ value\ (except\ predicted\ value\ of\ the\ jth\ variable)$$

- The quadratic part can be expanded and simplified as:

$$J(\,\beta_j) = \frac{1}{2n}\left(\sum_{i=1}^{n} {r_j}^2 - 2r_j\beta_j x_{ij} + {x_{ij}}^2{\beta_j}^2 + \lambda|\beta_j|\right)$$

LASSO Regression using Gradient Descent Contd….

- Notice that the first term $\sum r_j^2$ does not depend on βj, so we can ignore it for optimization purposes. The remaining terms form the expression to be minimized:

$$J(\beta_j) = \frac{1}{2n}\left(\sum_{i=1}^{n} -2r_j\beta_j x_{ij} + {x_{ij}}^2{\beta_j}^2 + \lambda|\beta_j|\right)$$

$$J(\beta_j) = \frac{1}{2n}\left(-2c_{ij}\beta_j + d_{ij}{\beta_j}^2 + \lambda|\beta_j|\right)$$

Where $\boldsymbol{c_{ij}} = \sum_{i=1}^{n} \boldsymbol{r_j x_{ij}}$ is called **correlation (or cosine similarity)** between the jth feature values and the partial residual

and $\boldsymbol{d_{ij}} = \sum_{i=1}^{n} \boldsymbol{x_{ij}}^{\boldsymbol{2}}$ *is called* ***deviation***

LASSO Regression using Gradient Descent Contd....

Case I: when $\boldsymbol{\beta_j > 0}$ $\boldsymbol{|\beta_j| = \beta_j}$

$$J(\beta_j) = \frac{1}{2n}\left(-2c_{ij}\beta_j + d_{ij}{\beta_j}^2 + \lambda\beta_j\right)$$

$$\frac{\partial J(\beta_j)}{\partial \beta_j} = \frac{1}{2n}\left(-2c_{ij} + 2d_{ij}\beta_j + \lambda\right)$$

For maxima/minima, substitute

$$\frac{\partial J(\beta_j)}{\partial \beta_j} = 0$$

$$\boldsymbol{\beta_j = \frac{c_{ij}}{d_{ij}} - \frac{\lambda}{2d_{ij}} \dots\dots\dots.(1)}$$

$$\frac{\partial^2 J(\beta_j)}{\partial \beta_j^2} = 2d_{ij} = positive\ (being\ sum\ of\ sqaured\ quantity)$$

There cost is minimum for the given β_j

LASSO Regression using Gradient Descent Contd....

Case II: **when $\boldsymbol{\beta_j < 0 \ \ |\beta_j| = -\beta_j}$**

$$J(\beta_j) = \frac{1}{2n}\left(-2c_{ij}\beta_j + d_{ij}{\beta_j}^2 - \lambda\beta_j\right)$$

$$\frac{\partial J(\beta_j)}{\partial \beta_j} = \frac{1}{2n}\left(-2c_{ij} + 2d_{ij}\beta_j - \lambda\right)$$

For maxima/minima, substitute

$$\frac{\partial J(\beta_j)}{\partial \beta_j} = 0$$

$$\boldsymbol{\beta_j = \frac{c_{ij}}{d_{ij}} + \frac{\lambda}{2d_{ij}} \dots \dots \dots (1)}$$

$$\frac{\partial^2 J(\beta_j)}{\partial \beta_j^2} = 2d_{ij} = positive\ (being\ sum\ of\ sqaured\ quantity)$$

There cost is minimum for the given β_j

LASSO Regression using Gradient Descent Contd….

- But since $|\beta j|$ is not defined at 0, so thresholding is done on other variable i.e., c_{ij} which is called **soft thresholding.**
- If $|c_{ij}| > \lambda/2 \; i.e., c_{ij} > \lambda/2 \; or \; c_{ij} < -\lambda/2$

$$(cosine\ simialrity\ between\ the\ feature\ and\ the\ partial\ residual\ is\ high;\ so\ feature\ is\ important\ and\ its\ coefficient\ \beta_j\ must\ be\ updated)$$

When $c_{ij} > \frac{\lambda}{2}$ *then* β_j is positive (from both equation 1 and 2).

$$\text{Hence } \boldsymbol{\beta_j} = \frac{\boldsymbol{c_{ij}}}{\boldsymbol{d_{ij}}} - \frac{\boldsymbol{\lambda}}{\boldsymbol{2d_{ij}}}$$

When $c_{ij} < -\frac{\lambda}{2}$ *then* β_j is negative (from both equation 1 and 2).

$$\text{Hence } \boldsymbol{\beta_j} = \frac{\boldsymbol{c_{ij}}}{\boldsymbol{d_{ij}}} + \frac{\boldsymbol{\lambda}}{\boldsymbol{2d_{ij}}}$$

- If $|c_{ij}| \le \frac{\lambda}{2} i.e., c_{ij} \le \frac{\lambda}{2} or \; c_{ij} \ge -\frac{\lambda}{2}$

$$(cosine\ simialrity\ between\ the\ feature\ and\ the\ partial\ residual\ is\ not\ high;\ so\ feature\ is\ not\ important\ and\ its\ coefficient\ \beta_j\ must\ be\ updated\ i.e., \boldsymbol{\beta_j} = 0)$$

LASSO Regression using Gradient Descent Contd….

- Therefore, the regression coefficients are updated as follows in LASSO regression for fixed number of iterations:

$$\beta_j = \begin{cases} \dfrac{c_{ij}}{d_{ij}} + \dfrac{\lambda}{2d_{ij}} & if\ c_{ij} < -\dfrac{\lambda}{2} \\ 0 & if\ -\dfrac{\lambda}{2} \le c_{ij} \le \dfrac{\lambda}{2} \\ \dfrac{c_{ij}}{d_{ij}} - \dfrac{\lambda}{2d_{ij}} & if\ c_{ij} > \dfrac{\lambda}{2} \end{cases}$$

where, $c_{ij} = \sum_{i=1}^{n} x_{ij} \times (y_i - \sum_{k \neq j} \beta_k\, x_{ik})$ and $d_{ij} = \sum_{1=1}^{n} x_{ij}^2$ for j=0,1,2….k

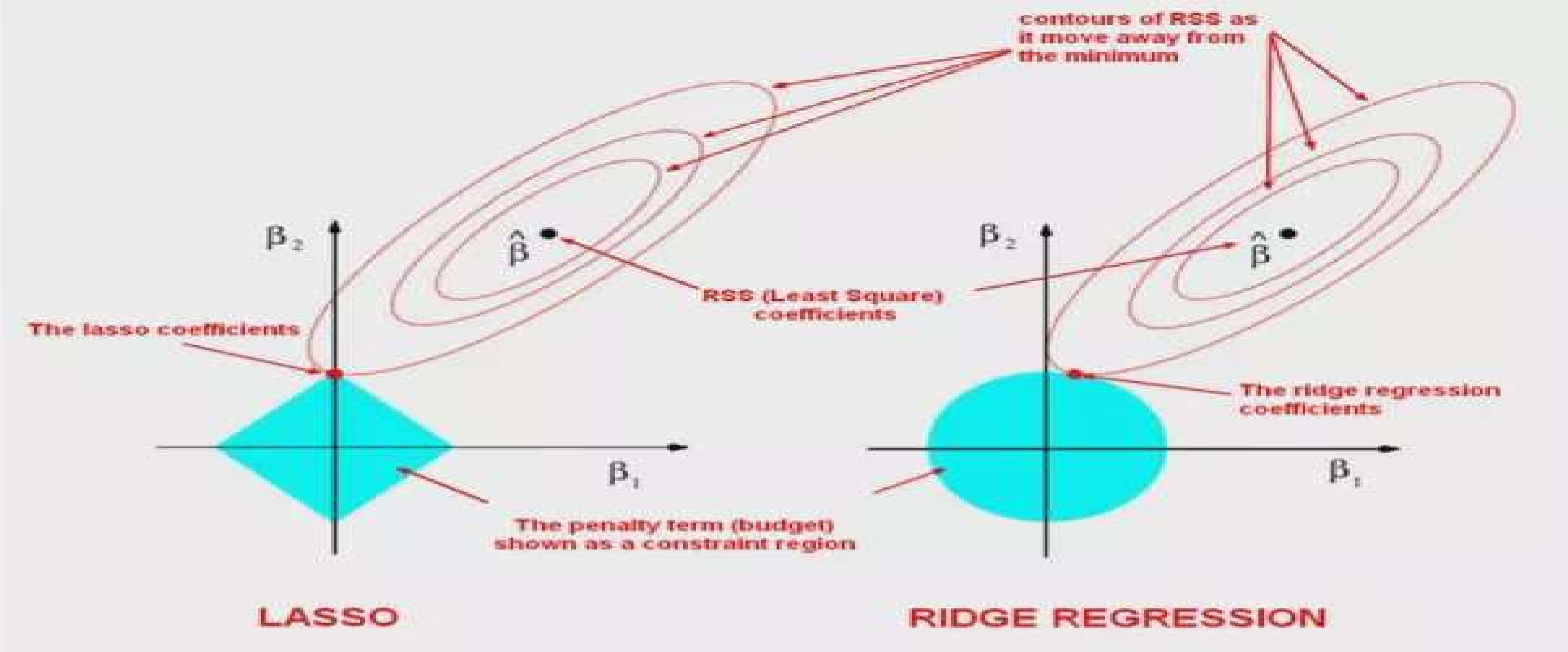
Ridge vs LASSO

1. Key Difference

- **Ridge:**
  - It includes all (or none) of the features in the model.
  - Thus, the major advantage of ridge regression is coefficient shrinkage and reducing model complexity.

- **LASSO:**
  - Along with shrinking coefficients, lasso performs feature selection as well. (Remember the '*selection*' in the lasso full-form)
  - Some of the coefficients become exactly zero, which is equivalent to the particular feature being excluded from the model.

Ridge vs. LASSO



Ridge vs LASSO

2. Typical Use Cases

- **Ridge:**
  - It is majorly used to *prevent overfitting*.
  - Since it includes all the features, it is not very useful in case of exorbitantly high features, say in millions, as it will pose computational challenges.
- **LASSO:**
  - Since it provides *sparse solutions*, it is generally the model of choice (or some variant of this concept) for modelling cases where the #features are in millions or more.
  - In such a case, getting a sparse solution is of great computational advantage as the features with zero coefficients can simply be ignored.

Ridge vs LASSO

3. Presence of Highly Correlated Features

- **Ridge:**
  - ➢ It generally works well even in presence of highly correlated features as it will include all of them in the model but the coefficients will be distributed among them depending on the correlation.
- **LASSO:**
  - ➢ It arbitrarily selects any one feature among the highly correlated ones and reduced the coefficients of the rest to zero.
  - ➢ Also, the chosen variable changes randomly with change in model parameters. This generally doesn’t work that well as compared to ridge regression.