IAI, TCG CREST

Avisek Gupta (avisek003@gmail.com)

Machine Learning

13 – Ridge & Lasso Regression

October 29, 2022

Penalties in Regression

Ridge Regression: Uses an ℓ_2-norm to not let the model parameters attain large magnitudes.

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 + \lambda ||\mathbf{w}||_2^2$$

Lasso Regression: Uses an ℓ_1-norm to drop weights that are close to zero.

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 + \lambda ||\mathbf{w}||_1$$

Elastic Net: Penalizes both the ℓ_2 and ℓ_1 norms.

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 + \lambda_1 ||\mathbf{w}||_2^2 + \lambda_2 ||\mathbf{w}||_1$$

Ridge Regression

The Ridge Regression objective function:

$$\min_{\mathbf{w}} \; J_{RR} = \sum_{i=1}^{n}(y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 + \lambda||\mathbf{w}||_2^2$$

Ridge Regression

The Ridge Regression objective function:

$$\min_{\mathbf{w}} \; J_{RR} = \sum_{i=1}^{n}(y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 + \lambda||\mathbf{w}||_2^2$$

The objective can be rewritten as,

$$\begin{aligned}\min_{\mathbf{w}} \; J_{RR} &= (\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\mathbf{w} - \mathbf{w}^T X^T\mathbf{y} + \mathbf{w}^T X^T X\mathbf{w} + \lambda\mathbf{w}^T\mathbf{w}\end{aligned}$$

Ridge Regression

The Ridge Regression objective function:

$$\min_{\mathbf{w}} \; J_{RR} = \sum_{i=1}^{n} (y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 + \lambda ||\mathbf{w}||_2^2$$

The objective can be rewritten as,

$$\begin{aligned}
\min_{\mathbf{w}} \; J_{RR} &= (\mathbf{y} - X\mathbf{w})^T (\mathbf{y} - X\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w} \\
&= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T X \mathbf{w} - \mathbf{w}^T X^T \mathbf{y} + \mathbf{w}^T X^T X \mathbf{w} + \lambda \mathbf{w}^T \mathbf{w}
\end{aligned}$$

Equating the gradient to zero,

$$\begin{aligned}
\nabla_{\mathbf{w}} J_{RR} &= -2X^T \mathbf{y} + 2X^T X \mathbf{w} + 2\lambda \mathbf{w} = 0 \\
\implies \mathbf{w} &= (X^T X + \lambda I)^{-1} X^T \mathbf{y}
\end{aligned}$$

Lasso Regression

The Lasso Regression objective function:

$$\min_{\mathbf{w}} J_{lasso} = \sum_{i=1}^{n}(y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 + \lambda||\mathbf{w}||_1$$

Lasso Regression

The Lasso Regression objective function:

$$\min_{\mathbf{w}} J_{lasso} = \sum_{i=1}^{n}(y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 + \lambda||\mathbf{w}||_1$$

- J_{lasso} is not differentiable, so we cannot apply Gradient Descent.

Lasso Regression

The Lasso Regression objective function:

$$\min_{\mathbf{w}} J_{lasso} = \sum_{i=1}^{n}(y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 + \lambda||\mathbf{w}||_1$$

- J_{lasso} is not differentiable, so we cannot apply Gradient Descent.

Outline of the following mathematical discussions -

1. J_{lasso} is a convex function.
2. A convex non-differentiable function can be optimized by following the direction of a *subgradient*.

Convex Sets and Functions

Convex Set: S is a convex set if $\forall \mathbf{x}_1, \mathbf{x}_2 \in S$ and $\forall \lambda \in [0, 1]$ we have $\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 \in S$.

This means that a convex set contains the line between any two points in the set.

Convex Sets and Functions

Convex Set: S is a convex set if $\forall \mathbf{x}_1, \mathbf{x}_2 \in S$ and $\forall \lambda \in [0,1]$ we have $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in S$.

This means that a convex set contains the line between any two points in the set.

Convex Function: For a convex set $S \subseteq \mathbb{R}^n$, a function $f : S \to \mathbb{R}$ is convex if for any two points $\mathbf{x}_1, \mathbf{x}_2 \in S$ and any $\lambda \in [0,1]$ we have,

$$f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2)$$

Convex Sets and Functions

Convex Set: S is a convex set if $\forall \mathbf{x}_1, \mathbf{x}_2 \in S$ and $\forall \lambda \in [0,1]$ we have $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in S$.

This means that a convex set contains the line between any two points in the set.

Convex Function: For a convex set $S \subseteq \mathbb{R}^n$, a function $f : S \to \mathbb{R}$ is convex if for any two points $\mathbf{x}_1, \mathbf{x}_2 \in S$ and any $\lambda \in [0,1]$ we have,

$$f(\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2)$$

Theorem: The sum of two convex functions is convex.

Taylor Series Approximations

Taylor Series Approximation: The value of a function $f(x)$ at a point a is approximated by a polynomial that has similar values in a neighborhood around a.

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^3(a)}{3!}(x-a)^3 + ...$$

Taylor Series Approximations

Taylor Series Approximation: The value of a function $f(x)$ at a point a is approximated by a polynomial that has similar values in a neighborhood around a.

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^3(a)}{3!}(x-a)^3 + \ldots$$

First-Order Taylor Series Approximation: Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable at $\bar{\mathbf{x}} \in \mathbb{R}^n$. Then,

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^T(\mathbf{x} - \bar{\mathbf{x}}) + o(||\mathbf{x} - \bar{\mathbf{x}}||), \ \forall \mathbf{x} \in \mathbb{R}^n,$$

where,

$$\lim_{\bar{\mathbf{x}} \rightarrow \mathbf{x}} \frac{o(||\mathbf{x} - \bar{\mathbf{x}}||)}{||\mathbf{x} - \bar{\mathbf{x}}||} = 0.$$

Taylor Series Approximations

Taylor Series Approximation: The value of a function $f(x)$ at a point a is approximated by a polynomial that has similar values in a neighborhood around a.

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^3(a)}{3!}(x-a)^3 + \ldots$$

First-Order Taylor Series Approximation: Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at $\bar{\mathbf{x}} \in \mathbb{R}^n$. Then,

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^T(\mathbf{x} - \bar{\mathbf{x}}) + o(||\mathbf{x} - \bar{\mathbf{x}}||), \ \forall \mathbf{x} \in \mathbb{R}^n,$$

where,

$$\lim_{\bar{\mathbf{x}} \to \mathbf{x}} \frac{o(||\mathbf{x} - \bar{\mathbf{x}}||)}{||\mathbf{x} - \bar{\mathbf{x}}||} = 0.$$

(This means when $\bar{\mathbf{x}}$ is close to $\mathbf{x}$, $f(\mathbf{x})$ can be approximated by an affine function $f(\bar{\mathbf{x}}) + \nabla f(\bar{\mathbf{x}})^T(\mathbf{x} - \bar{\mathbf{x}})$)

Directional Derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is differentiable at $\mathbf{x} \in \mathbb{R}^n$, and let $\mathbf{d} \in \mathbb{R}^n$ with $||\mathbf{d}|| = 1$. The derivative of f at $\mathbf{x}$ in direction $\mathbf{d}$ is,

$$f'(\mathbf{x}, \mathbf{d}) = \lim_{\lambda \to 0} \frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda}.$$

Directional Derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is differentiable at $\mathbf{x} \in \mathbb{R}^n$, and let $\mathbf{d} \in \mathbb{R}^n$ with $||\mathbf{d}|| = 1$. The derivative of f at $\mathbf{x}$ in direction $\mathbf{d}$ is,

$$f'(\mathbf{x}, \mathbf{d}) = \lim_{\lambda \to 0} \frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda}.$$

Claim 1: $f'(\mathbf{x}, \mathbf{d}) = \nabla f(\mathbf{x})^T \mathbf{d}$.

Proof: From the first order Taylor series approximation of f at $\mathbf{x}$,

$$\begin{aligned} f(\mathbf{x} + \lambda \mathbf{d}) &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\lambda \mathbf{d}) + o(||\lambda \mathbf{d}||) \\ \implies \frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda} &= \nabla f(\mathbf{x})^T \mathbf{d} + o(\lambda ||\mathbf{d}||) \\ \lim_{\lambda \to 0} \frac{f(\mathbf{x} + \lambda \mathbf{d}) - f(\mathbf{x})}{\lambda} &= \nabla f(\mathbf{x})^T d. \end{aligned}$$

On Convex functions

Theorem 1: Let $f : \mathbb{R}^n \to \mathbb{R}$, and S be a convex subset of $\mathbb{R}^n$. Then f is convex **iff** for any $\mathbf{x}, \mathbf{y} \in S$ we have $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$.

On Convex functions

Theorem 1: Let $f : \mathbb{R}^n \to \mathbb{R}$, and S be a convex subset of $\mathbb{R}^n$. Then f is convex **iff** for any $\mathbf{x}, \mathbf{y} \in S$ we have $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$.

Proof: $[\implies]$ Assume f is convex, and let $\mathbf{z} = \lambda\mathbf{y} + (1-\lambda)\mathbf{x}$ for some $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0, 1]$. Then,

$$f(\mathbf{z}) = f(\lambda\mathbf{y} + (1-\lambda)\mathbf{x}) \leq \lambda f(\mathbf{y}) + (1-\lambda)f(\mathbf{x})$$

$$\implies f(\mathbf{x} + \lambda(\mathbf{y}-\mathbf{x})) - f(\mathbf{x}) \leq \lambda f(\mathbf{y}) + (1-\lambda)f(\mathbf{x}) - f(\mathbf{x}) = \lambda f(\mathbf{y}) - \lambda f(\mathbf{x})$$

$$\implies \frac{f(\mathbf{x} + \lambda(\mathbf{y}-\mathbf{x})) - f(\mathbf{x})}{\lambda} \leq f(\mathbf{y}) - f(\mathbf{x}).$$

[Using Claim 1] $$\implies \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}).$$

On Convex functions

Theorem 1: Let $f : \mathbb{R}^n \to \mathbb{R}$, and S be a convex subset of $\mathbb{R}^n$. Then f is convex **iff** for any $\mathbf{x}, \mathbf{y} \in S$ we have $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$.

Proof: [$\Longleftarrow$] Let $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$ for any $\mathbf{x}, \mathbf{y} \in S$. Let $\mathbf{z} = \lambda\mathbf{y} + (1-\lambda)\mathbf{x}$. Then,

$$f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^T(\mathbf{y} - \mathbf{z}) \tag{1}$$

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^T(\mathbf{x} - \mathbf{z}) \tag{2}$$

λ times eqn.(1) added to $(1-\lambda)$ times eqn.(2) gives,

$$\begin{aligned}
&\lambda f(\mathbf{y}) + (1-\lambda) f(\mathbf{x}) \\
&\geq \lambda f(\mathbf{z}) + \lambda \nabla f(\mathbf{z})^T(\mathbf{y} - \mathbf{z}) + (1-\lambda) f(\mathbf{z}) + (1-\lambda)\nabla f(\mathbf{z})^T(\mathbf{x} - \mathbf{z}) \\
&= f(\mathbf{z}) + \nabla f(\mathbf{z})^T(\lambda\mathbf{y} - \lambda\mathbf{z}) + \nabla f(\mathbf{z})^T((1-\lambda)\mathbf{x} - (1-\lambda)\mathbf{z}) \\
&= f(\mathbf{z}) + \nabla f(\mathbf{z})^T(\lambda\mathbf{y} + (1-\lambda)\mathbf{x} - \mathbf{z}) \\
&= f(\mathbf{z}) = f(\lambda\mathbf{y} + (1-\lambda)\mathbf{x}).
\end{aligned}$$

Subdifferentials

Theorem 1: Let $f : \mathbb{R}^n \to \mathbb{R}$, and S be a convex subset of $\mathbb{R}^n$. Then f is convex **iff** for any $\mathbf{x}, \mathbf{y} \in S$ we have $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$.

By Theorem 1, if a function f is convex and differentiable, then $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$.

- However, the function $||x||$ we use in Lasso Regression is not differentiable.
- For convex functions whose derivatives are undefined at some points of their domains, we can use subdifferentials.

Subdifferentials

Subdifferential: The subdifferential ∂f is the set [a,b] of all subderivatives g of a function f at point $\mathbf{x}_0$,

$$\partial f(\mathbf{x}) = \{g : f(\mathbf{x}) \geq f(\mathbf{x}_0) + g(\mathbf{x} - \mathbf{x}_0),\ \forall \mathbf{x} \in S\},$$

where,

$$a = \lim_{\mathbf{x} \to \mathbf{x}_0^-} \frac{f(\mathbf{x}) - f(\mathbf{x}_0)}{\mathbf{x} - \mathbf{x}_0}$$

$$b = \lim_{\mathbf{x} \to \mathbf{x}_0^+} \frac{f(\mathbf{x}) - f(\mathbf{x}_0)}{\mathbf{x} - \mathbf{x}_0}$$

Subdifferentials

Some properties of subdifferentials:

- A convex function is differentiable at $\mathbf{x}_0$ iff the subdifferential has only one point, the derivative at $\mathbf{x}_0$.
- $\mathbf{x}_0$ is the global minima of a convex function f iff 0 is contained in the subdifferential.
- Moreau-Rockafeller Theorem: If f and g are both convex, then the subdifferential of $f + g$ is $\partial(f + g) = \partial f + \partial g$.

Subdifferentials

Some properties of subdifferentials:

- A convex function is differentiable at $\mathbf{x}_0$ iff the subdifferential has only one point, the derivative at $\mathbf{x}_0$.
- $\mathbf{x}_0$ is the global minima of a convex function f iff 0 is contained in the subdifferential.
- Moreau-Rockafeller Theorem: If f and g are both convex, then the subdifferential of $f + g$ is $\partial(f + g) = \partial f + \partial g$.

The function $f(x) = |x|$ is convex but non-differentiable at $x = 0$.

Its subdifferential at the origin is the interval $[-1, 1]$. It has the origin, so $x = 0$ is the global minimum.

For $x < 0$, the subdifferential has -1, and for $x > 0$, the subdifferential has $+1$.

Lasso Regression

$$\min_{\mathbf{w}} J_{lasso} = J_{OLS} + J_{\ell_1} = \left\{ \sum_{i=1}^{n} (y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 \right\} + \{\lambda ||\mathbf{w}||_1\}$$

The derivative of J_{OLS},

$$\begin{aligned} \frac{\partial}{\partial w_j} J_{OLS} &= -2 \sum_{i=1}^{n} (y^{(i)} - \sum_{k \neq j}^{d} w_k x_k^{(i)}) x_j^{(i)} + 2 w_j \sum_{i=1}^{n} (x_j^{(i)})^2. \\ &= -\rho_j + w_j z_j \end{aligned}$$

Lasso Regression

$$\min_{\mathbf{w}} J_{lasso} = J_{OLS} + J_{\ell_1} = \left\{ \sum_{i=1}^{n} (y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 \right\} + \{\lambda ||\mathbf{w}||_1\}$$

We can write, $J_{\ell_1} = \lambda ||\mathbf{w}||_1 = \lambda \sum_{j=0}^{d} |w_j| = \lambda |w_j| + \lambda \sum_{k \neq j}^{d} |w_k|$.

Then by the definition of the subdifferential,

$$\partial_{w_j} J_{\ell_1} = \partial_{w_j} \lambda |w_j| = \begin{cases} \{-\lambda\} & w_j < 0 \\ [-\lambda, \lambda] & w_j = 0 \\ \{\lambda\} & w_j > 0 \end{cases}$$

Lasso Regression

$$\min_{\mathbf{w}} J_{lasso} = J_{OLS} + J_{\ell_1} = \left\{ \sum_{i=1}^{n} (y^{(i)} - \sum_{j=0}^{d} w_j x_j^{(i)})^2 \right\} + \{\lambda ||\mathbf{w}||_1\}$$

Therefore equating the subdifferential of J_{lasso} to zero,

$$\partial_{w_j} J_{lasso} = 0 = -\rho_j + w_j z_j + \partial_{w_j} \lambda |w_j|$$

$$\implies 0 = \begin{cases} \{-\rho_j + w_j z_j - \lambda\} & w_j < 0 \\ [-\rho_j - \lambda, -\rho_j + \lambda] & w_j = 0 \\ \{-\rho_j + w_j z_j + \lambda\} & w_j > 0 \end{cases}$$

$w_j = 0$ will be the global minima if $0 \in [-\rho_j - \lambda, -\rho_j + \lambda]$

$$\implies -\rho_j - \lambda \leq 0 \text{ and } -\rho + \lambda \geq 0 \implies -\lambda \leq \rho_j \leq \lambda.$$

We can define a soft-thresholding function $\frac{1}{z_j} S(\rho_j, \lambda)$.

Lasso Regression

Therefore equating the subdifferential of J_{lasso} to zero,

$$\partial_{w_j} J_{lasso} = 0 = -\rho_j + w_j z_j + \partial_{w_j} \lambda |w_j|$$

$$\implies 0 = \begin{cases} \{-\rho_j + w_j z_j - \lambda\} & w_j < 0 \\ [-\rho_j - \lambda, -\rho_j + \lambda] & w_j = 0 \\ \{-\rho_j + w_j z_j + \lambda\} & w_j > 0 \end{cases}$$

$w_j = 0$ will be the global minima if $0 \in [-\rho_j - \lambda, -\rho_j + \lambda]$,

$$\implies -\rho_j - \lambda \leq 0 \text{ and } -\rho_j + \lambda \geq 0 \implies -\lambda \leq \rho_j \leq \lambda.$$

We can define a soft-thresholding function $\frac{1}{z_j} S(\rho_j, \lambda)$.

$$\frac{1}{z_j} S(\rho_j, \lambda) = \begin{cases} w_j = \frac{\rho_j + \lambda}{z_j} & \rho_j < -\lambda \\ w_j = 0 & -\lambda \leq \rho_j \leq \lambda \\ w_j = \frac{\rho_j - \lambda}{z_j} & \rho_j > \lambda \end{cases}$$

Lasso Regression Optimization

Coordinate Descent Algorithm to Optimize the Lasso Regression model:

`for` $j = 0, 1, ..., d$

(i) Compute $\rho_j = \sum_{i=1}^{n} x_j^{(i)} \{y^{(i)} - \sum_{k \neq j}^{d} w_k x_k^{(i)}\}$

(ii) Compute $z_j = \sum_{i=1}^{n} (x_j^{(i)})^2$

(iii) Set $w_j = \frac{1}{z_j} S(\rho_j, \lambda)$

References

[1] Xavier Bourret Sicotte. Lasso regression: derivation of the coordinate descent update rule. Data Blog. 2018. https://xavierbourretsicotte.github.io/lasso_derivation.html

[2] Harvard AM 221, 2016, Lecture 8. https://people.seas.harvard.edu/~yaron/AM221-S16/lecture_notes/AM221_lecture8.pdf

[3] Gareth James, Daniela Witten, Trevor Hastie, Rob Tibshirani. Introduction to Statistical Learning with R, Second Edition. 2021. https://www.statlearning.com/