CMO Assignment 3

Biswadeep Debnath BISWADEEPD@IISC.AC.IN
24528

1 LASSO Regression and KKT Conditions

1.1 Solution

Optimization Problem

$$\begin{aligned} &\min_{\beta \in \mathbb{R}^n} \ \frac{1}{2} \|X\beta - y\|_2^2 \\ &\text{subject to } \ \|\beta\|_1 \leq t \qquad (1) \end{aligned}$$

where $X \in \mathbb{R}^{50\times 15}$ and $y \in \mathbb{R}^{50}$

Lagrangian,

$$L(\beta, \lambda) = \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\left(\|\beta\|_1 - 1\right)$$

KKT conditions:

1. Stationarity condition:

$$\nabla_i L(\beta, \lambda) = x_i^T(X\beta - y) + \lambda \partial|\beta_i| = 0, \quad i \in \{1, 2, \ldots, 15\}$$

which implies

$$x_i^T(X\beta - y) + \lambda \begin{cases} \text{sign}(\beta_i), & \beta_i \neq 0 \\ u_i \in [-1, 1], & \beta_i = 0 \end{cases} = 0$$

which implies

$$\begin{cases} x_i^T(y - X\beta) = \lambda \text{sign}(\beta_i), & \text{if } \beta_i \neq 0 \\ |x_i^T(X\beta - y)| \leq \lambda, & \text{if } \beta_i = 0 \end{cases}$$

2. Complementary slackness:

$$\lambda\left(\|\beta\|_1 - t\right) = 0, \quad \lambda \geq 0$$

3. Primal feasibility:

$$\|\beta\|_1 \leq t$$

1.2 Solution

$$\min_{\beta \in \mathbb{R}^n} \quad f(\beta) + \lambda \|\beta\|_1 \quad \text{where } \lambda > 0 \text{ is a known regularization parameter.} \qquad (2)$$

Yes, optimizing (1) is equivalent to optimizing (2) as if we observe the stationary condition for (2), we have,

$$\nabla f(\beta) + \lambda \, \partial \|\beta\|_1 = 0,$$

where $\partial \|\beta\|_1$ is the subdifferential, and $\lambda > 0$.
This is equivalent to the first KKT condition for (1).
And also from the complementary slackness of (1) it is ensured that $\lambda \geq 0$.

Since (2) is an unconstrained problem, we can solve it using methods like gradient descent, which are computationally more efficient than solving the constrained optimization problem (1).
The value of λ can also be tuned easily, which is more convenient than setting a value of t for the constrained optimization case.

1.3 Solution

The number of fitted (non-zero) coefficients against λ to compare the sparsity of β_* is plotted in Figure 1.

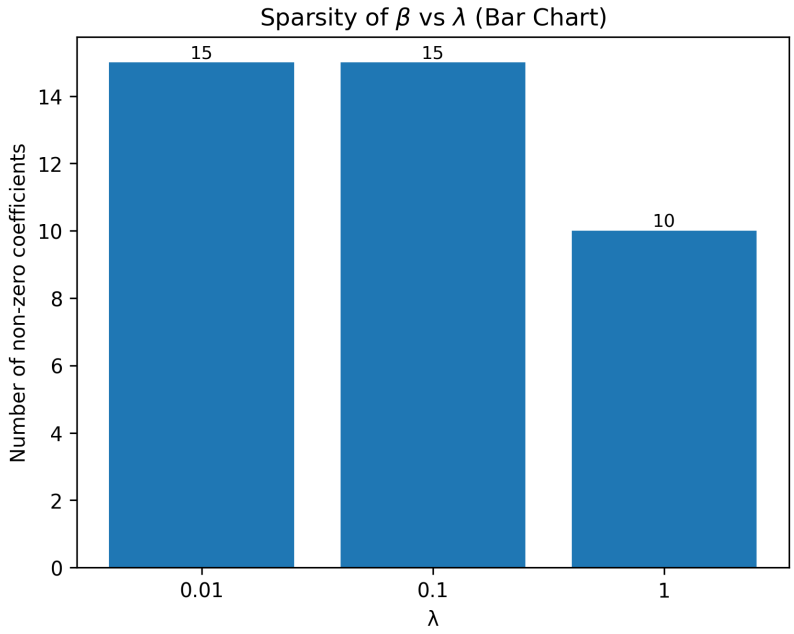

Figure 1: Number of non-zero coefficients(β_i) vs λ.

It is observed that as the regularization parameter λ increases, the number of non-zero coefficients decreases. For example, at $\lambda = 1$, the beta coefficients tend to be smaller since the optimization problem minimizes $f(\beta) + \lambda\|\beta\|_1$.
As λ increases, the penalty term $\lambda\|\beta\|_1$ forces the coefficients β closer to zero, thereby increasing sparsity. This behavior is consistent with the KKT condition for $\beta_i = 0$:

$$|x_i^T(y - X\beta)| \leq \lambda.$$

Hence, larger values of λ impose stronger shrinkage on the coefficients, increasing the number of coefficients set exactly to zero.

1.4 Solution

The first KKT condition (stationarity) of problem (1) is verified using the values of β obtained by solving problem (2) for three different regularization parameters $\lambda \in \{0.01, 0.1, 1\}$. For each β_i, the stationary condition

$$\begin{cases} x_i^T(y - X\beta) = \lambda \operatorname{sign}(\beta_i), & \text{if } \beta_i \neq 0 \\ |x_i^T(X\beta - y)| \leq \lambda, & \text{if } \beta_i = 0 \end{cases}$$

is checked element-wise for all i with a tolerance of 10^{-4}.
Using this tolerance, all the stationary conditions are satisfied for all tested β values. The tolerance can be changed in the code to check the values of LHS and RHS in the above equations.
Regarding the KKT conditions of complementary slackness and primal feasibility, since they involve the parameter t as a constraint bound, these have not been verified in this analysis.

1.5 Solution

After duplicating a feature and running the same experiment, the beta values for the original feature and the duplicated one split approximately in half.

Table 1: Beta Values Before Duplication for Feature x_{15}

Lambda (λ)	Beta (β_{x15})
0.01	2.6549611080910425
0.1	2.6494964554451954
1	2.5886902914454053

Table 2: Beta Values After Duplication for Features x_{15} and $x_{15\_duplicate}$

Lambda (λ)	Beta (β_{x15})	Beta ($\beta_{x15\_duplicate}$)
0.01	1.32748050541599	1.3274805736853825
0.1	1.3247478158837094	1.3247477116199664
1	1.2943448786978824	1.294344858576933

Analyzing the KKT subgradient condition, since β is non-zero in this case, we have:

$$x_{15}^T(y - X\beta) = x_{15\_duplicate}^T(y - X\beta) \quad \text{since } x_{15} = x_{15\_duplicate}$$

and also

$$x_{15}^T(y - X\beta) = \lambda \operatorname{sign}(\beta_{x_{15}})$$

$$x_{15\_duplicate}^T(y - X\beta) = \lambda \operatorname{sign}(\beta_{x_{15\_duplicate}})$$

which implies

$$\operatorname{sign}(\beta_{x_{15}}) = \operatorname{sign}(\beta_{x_{15\_duplicate}})$$

This means that both coefficients have the same sign and splitting the coefficient approximately in half still satisfies the condition since the sign remains unchanged.

Question 2

2.1 Solution

$$\min_{\beta} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1$$

Transforming the primal problem using $z = X\beta$, we have

$$\min_{\beta,z} \frac{1}{2}\|z - y\|_2^2 + \lambda\|\beta\|_1 \quad \text{subject to} \quad z = X\beta$$

The Lagrangian formulation is

$$L(\beta, z, \mu) = \frac{1}{2}\|z - y\|_2^2 + \lambda\|\beta\|_1 + \mu^T(z - X\beta) \tag{3}$$

The dual function is given by

$$g(\mu) = \min_{\beta,z} \frac{1}{2}\|z - y\|_2^2 + \mu^T z + \lambda\|\beta\|_1 - \mu^T X\beta$$

We split this into two parts:

$$\underbrace{\min_{z} \frac{1}{2}\|z - y\|_2^2 + \mu^T z}_{\text{I}} + \underbrace{\min_{\beta} \lambda\|\beta\|_1 - (X^T\mu)^T\beta}_{\text{II}}$$

Considering part I:

$$\min_{z} \frac{1}{2}\|z\|_2^2 + \frac{1}{2}\|y\|_2^2 - z^T y + \mu^T z = \min_{z} \frac{1}{2}\|z\|_2^2 - z^T y + \mu^T z$$

Since this is a quadratic function, it can be transformed as

$$\max_{\mu} -\frac{1}{2}\|\mu\|_2^2 + y^T\mu$$

Considering part II:

$$\min_{\beta} \ \lambda\|\beta\|_1 - (X^T\mu)^T\beta$$

Let $v = X^T\mu$. We analyze this element-wise. For the j-th component:

$$f(\beta_j) = \lambda|\beta_j| - v_j\beta_j$$

Case 1: $|v_j| \leq \lambda$

$$\begin{aligned} \text{If } \beta_j \geq 0: \quad & f(\beta_j) = (\lambda - v_j)\beta_j \\ \text{If } \beta_j \leq 0: \quad & f(\beta_j) = -(\lambda + v_j)\beta_j \end{aligned}$$

If $\lambda - v_j \geq 0$ (which holds), the minimum over $\beta_j \geq 0$ is at $\beta_j = 0$.
If $-(\lambda + v_j) \leq 0$ (since $-\lambda \leq v_j$), the minimum over $\beta_j \leq 0$ is also at $\beta_j = 0$.
Thus, when $|v_j| \leq \lambda$, the minimum occurs at $\beta_j = 0$.

Case 2: $|v_j| > \lambda$
For $\beta_j \geq 0$:

$$f(\beta_j) = (\lambda - v_j)\beta_j$$

If $v_j > \lambda$, then $\lambda - v_j < 0$, so $f(\beta_j) \to -\infty$ as $\beta_j \to +\infty$.
For $\beta_j \leq 0$:

$$f(\beta_j) = -(\lambda + v_j)\beta_j$$

If $v_j < -\lambda$, then $\lambda + v_j < 0$, so $f(\beta_j) \to -\infty$ as $\beta_j \to -\infty$.
Thus, if $|v_j| > \lambda$, the function is unbounded below and the minimum is $-\infty$.

Summary with infinity norm notation:
Replacing μ with u for notation

$$\min_{\beta} \lambda\|\beta\|_1 - (X^T u)^T\beta = \begin{cases} 0, & \text{if } \left\|\frac{X^T u}{\lambda}\right\|_\infty \leq 1 \\ -\infty, & \text{if } \left\|\frac{X^T u}{\lambda}\right\|_\infty > 1 \end{cases}$$

Therefore, the dual problem is

$$\max_{u} \; -\frac{1}{2}\|u\|_2^2 + y^T u \quad \text{subject to} \quad \|X^T u\|_\infty \leq \lambda$$

2.2 Solution

We have the primal problem:

$$\min_{\beta} \frac{1}{2}\|z - y\|_2^2 + \lambda\|\beta\|_1 \quad \text{s.t.} \quad z = X\beta$$

We can prove strong duality in this problem using Slater's condition (Boyd and Vandenberghe (2004)).

Slater's condition for this convex problem states that if there exists a $\beta \in \mathbb{R}^d$ such that

$$g_i(\beta) < 0, \quad i = 1, \ldots, m, \quad \text{and} \quad h_j(\beta) = 0, \quad j = 1, \ldots, r,$$

then strong duality holds.
In our problem, there are no inequality constraints, so we only consider equality constraints:

$$X^T X\beta = X^T z \implies \beta = (X^T X)^{-1} X^T z,$$

assuming $X^T X$ is invertible (i.e., X has full rank).
Thus, the assumption that X has full rank ensures the existence of such β satisfying the constraint.
Therefore, by Slater's condition, strong duality holds for this problem.

2.3 Solution

Implemention is in code.

2.4 Solution

We have the Lagrangian from the primal problem where we use $z = XB$

$$\mathcal{L}(B, z, u) = \frac{1}{2}\|z - y\|_2^2 + \lambda\|B\|_1 + u^T(z - XB)$$

Taking the gradient of $\mathcal{L}$ with respect to z, we get

$$\nabla_z \mathcal{L} = z - y + u = 0$$

which implies

$$u = y - z$$

Substituting $z = XB$, we obtain

$$u^* = y - XB^*$$

Results from the code where the difference in norms is presented in Table 3.

Table 3: L2 Norms of $u^* - (y - X\beta^*)$ for Different λ Values

λ	$\|u^* - (y - X\beta^*)\|_2$
0.01	1.9774314995749002e-06
0.1	2.8575889480641164e-05
1	4.097589234641439e-05

This suggests the relation between u^* and β^* is satisfied.

2.5 Solution

$$\nabla_\beta L(\beta, z, \mu) = \lambda\partial\|\beta\|_1 - X^T\mu = 0$$

Analyzing element-wise:

$$|X_j^T \mu^*| < \lambda \implies \beta_j^* = 0$$

$$|X_j^T \mu^*| = \lambda \implies \beta_j^* \neq 0$$

Now also $u = y - X\beta$ is a kind of residual which means this tells what the model hasn't explained. If for a feature j we have $X_j^T \mu$ tells how much the residual or inexplanability of the model is aligned with the feature and if the feature influences the residual upto the threshold λ then the β_j's are non-zero. Since in data there might be much features which doesn't align with this resiudal much and are less than the threshold which introduces sparsity in β_js by setting them to 0. In a way it creates a threshold for the features that if they reach this threshold only then the betas would be non-zero. Also, if we set λ to a high value it means its more likely that not much features would reach this threshold and therefore we also observed in Question 1 that increasing λ makes the betas zeros.

Question 3

3.1 Solution

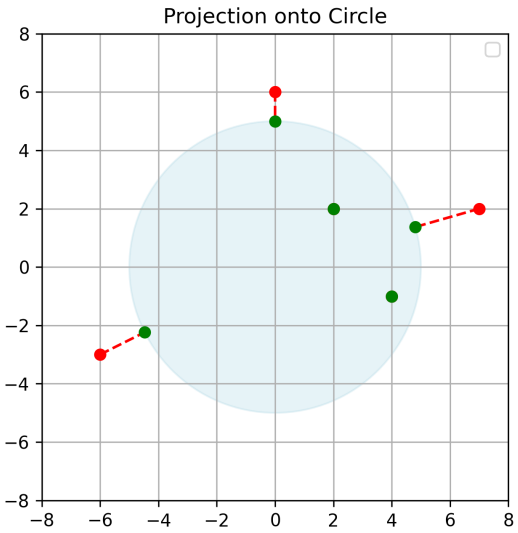

Figure 2: Projected points into the circle. The red color denotes the original point and green color the projected point. If the point is inside initially then its green.

The projected points are shown in the plots in Figure 2 and Figure 3 .

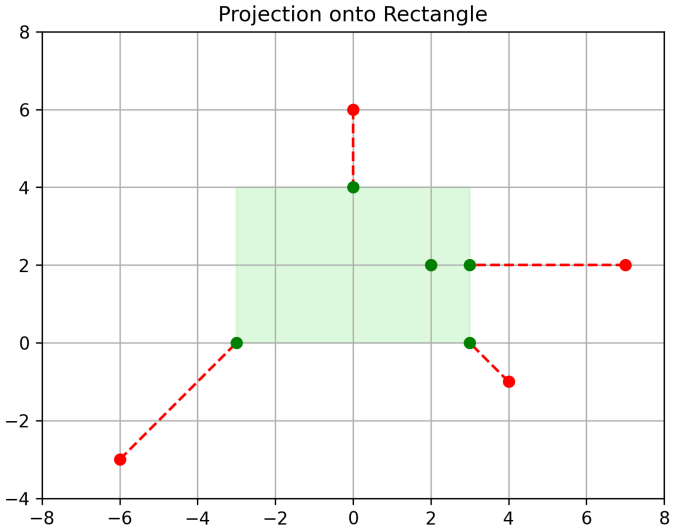

Figure 3: Projected points into the rectangle. The red color denotes the original point and green color the projected point. If the point is inside initially then its green.

3.2 Solution

In this solution, the method of alternating projections (Boyd et al. (2003)) is used to find the shortest distance between the sets and then calculate the perpendicular bisector as a hyperplane.

3.3 Solution

Each of the row of $A^T x \leq b$ is a capacity constraint or constraint in this problem. And y here assigns non-negative weights to these constraints. This y_i values for each constraint say i, acts like a penalty for violating the constraint i. Higher the value means it is more responsible for the infeasibility.

References

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike-20&path=ASIN/0521833787`.

Stephen Boyd, Jon Dattorro, et al. Alternating projections. *EE392o, Stanford University*, 2003.