Module 14

Ridge Regression. Lasso Regression, Elastic Net Regression Solutions

Problem 1: 10 points

Verify the formula
$$(X^\top X + KI_n)^{-1}X^\top = X^\top(XX^\top + KI_m)^{-1},$$
where X is a real $m \times n$ matrix and $K > 0$. You may assume without proof that both $X^\top X + KI_n$ and $XX^\top + KI_m$ are invertible (because they are symmetric positive definite).

Solution. From the equation
$$X^\top XX^\top + KX^\top = X^\top XX^\top + KX^\top,$$
we get
$$(X^\top X + KI_n)X^\top = X^\top(XX^\top + KI_m),$$
and since both $X^\top X + KI_n$ and $XX^\top + KI_m$ are invertible, by multiplying both sides of the above equation on the left by $(X^\top X + KI_n)^{-1}$ and on the right by $(XX^\top + KI_m)^{-1}$, we get
$$X^\top(XX^\top + KI_m)^{-1} = (X^\top X + KI_n)^{-1}X^\top,$$
as claimed.

Problem 2: 40 points

Recall that elastic net regression is the following optimization problem:

Program (**elastic net**):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\xi^\top\xi + \frac{1}{2}Kw^\top w + \tau\mathbf{1}_n^\top\epsilon \\
&\text{subject to} \\
&\qquad\qquad y - Xw - b\mathbf{1}_m = \xi \\
&\qquad\qquad\qquad w \leq \epsilon \\
&\qquad\qquad\quad\ - w \leq \epsilon,
\end{aligned}$$

with X an $m \times n$ matrix, $y, \xi \in \mathbb{R}^m$, $w, \epsilon \in \mathbb{R}^n$, $b \in \mathbb{R}$, where $K > 0$ and $\tau \geq 0$ are two constants controlling the influence of the ℓ^2-regularization and the ℓ^1-regularization.

The Lagrangian associated with this optimization problem is

$$
\begin{aligned}
L(\xi, w, \epsilon, b, \lambda, \alpha_+, \alpha_-) = {} & \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y - b\mathbf{1}_m^\top \lambda \\
& + \epsilon^\top(\tau \mathbf{1}_n - \alpha_+ - \alpha_-) + w^\top(\alpha_+ - \alpha_- - X^\top \lambda) + \frac{1}{2}K w^\top w,
\end{aligned}
$$

with $\lambda \in \mathbb{R}^m$ and $\alpha_+, \alpha_- \in \mathbb{R}^n_+$.

(1) (5 points) Prove that the gradient $\nabla L_{\xi,w,\epsilon,b}$ of the above Lagrangian is given by

$$
\begin{pmatrix} \xi - \lambda \\ Kw + (\alpha_+ - \alpha_- - X^\top \lambda) \\ \tau \mathbf{1}_n - \alpha_+ - \alpha_- \\ -\mathbf{1}_m^\top \lambda \end{pmatrix}.
$$

(2) (10 points) By setting the gradient $\nabla L_{\xi,w,\epsilon,b}$ to zero we obtain the equations

$$
\begin{aligned}
\xi &= \lambda \\
Kw &= -(\alpha_+ - \alpha_- - X^\top \lambda) \qquad (*_w) \\
\alpha_+ + \alpha_- - \tau \mathbf{1}_n &= 0 \\
\mathbf{1}_m^\top \lambda &= 0.
\end{aligned}
$$

We find that $(*_w)$ determines w.

It is more convenient to write $\lambda = \lambda_+ - \lambda_-$, with $\lambda_+, \lambda_- \in \mathbb{R}^m_+$ (recall that $\alpha_+, \alpha_- \in \mathbb{R}^n_+$), and to rescale our variables by defining $\beta_+, \beta_-, \mu_+, \mu_-$ such that

$$
\alpha_+ = K\beta_+, \;\; \alpha_- = K\beta_-, \;\; \lambda_+ = K\mu_+, \;\; \lambda_- = K\mu_-.
$$

We also let $\mu = \mu_+ - \mu_-$ so that $\lambda = K\mu$.

Prove that

$$
\begin{aligned}
w &= -(\beta_+ - \beta_- - X^\top \mu) \\
&= \begin{pmatrix} -I_n & I_n & X^\top & -X^\top \end{pmatrix} \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix}.
\end{aligned}
$$

Use the above result to prove that

$$
\frac{1}{2} w^\top w = \frac{1}{2} \begin{pmatrix} \beta_+^\top & \beta_-^\top & \mu_+^\top & \mu_-^\top \end{pmatrix} Q \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix},
$$

with Q the symmetric positive semidefinite matrix

$$Q = \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top & -XX^\top \\ X & -X & -XX^\top & XX^\top \end{pmatrix}.$$

(3) (10 points) Prove that the dual function is given by

$$\begin{aligned} G(\mu, \beta_+, \beta_-) &= \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y + w^\top(\alpha_+ - \alpha_- - X^\top \lambda) + \frac{1}{2} K w^\top w \\ &= -\frac{1}{2}K^2 \mu^\top \mu - \frac{1}{2} K w^\top w + K y^\top \mu. \end{aligned}$$

Hint. Use $(*_w)$.

(4) (15 points) Prove that

$$\frac{1}{2}\mu^\top \mu = \frac{1}{2}\begin{pmatrix} \mu_+^\top & \mu_-^\top \end{pmatrix} \begin{pmatrix} I_m & -I_m \\ -I_m & I_m \end{pmatrix} \begin{pmatrix} \mu_+ \\ \mu_- \end{pmatrix}.$$

Using (2) to rewrite $\frac{1}{2}w^\top w$, (4) to rewrite $\frac{1}{2}\mu^\top \mu$, and (3), prove that

$$G(\beta_+, \beta_-, \mu_+, \mu_-) = -\frac{1}{2} K \begin{pmatrix} \beta_+^\top & \beta_-^\top & \mu_+^\top & \mu_-^\top \end{pmatrix} P \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix} - K q^\top \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix}$$

with

$$\begin{aligned} P &= Q + K \begin{pmatrix} 0_{n,n} & 0_{n,n} & 0_{n,m} & 0_{n,m} \\ 0_{n,n} & 0_{n,n} & 0_{n,m} & 0_{n,m} \\ 0_{m,n} & 0_{m,n} & I_m & -I_m \\ 0_{m,n} & 0_{m,n} & -I_m & I_m \end{pmatrix} \\ &= \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top + KI_m & -XX^\top - KI_m \\ X & -X & -XX^\top - KI_m & XX^\top + KI_m \end{pmatrix}, \end{aligned}$$

and

$$q = \begin{pmatrix} 0_n \\ 0_n \\ -y \\ y \end{pmatrix}.$$

Solution. (1) The gradient is the unique vector $\nabla L_{\xi,w,\epsilon,b} \in \mathbb{R}^{m+2n+1}$ such that

$$dL_{\xi,w,\epsilon,b}(\xi_1, w_1, \epsilon_1, b_1) = \nabla L_{\xi,w,\epsilon,b} \cdot \begin{pmatrix} \xi_1 \\ w_1 \\ \epsilon_1 \\ b \end{pmatrix}$$

for all $\xi, \xi_1 \in \mathbb{R}^m$, $w, w_1, \epsilon, \epsilon_1 \in \mathbb{R}^n$, $b, b_1 \in \mathbb{R}$. From earlier results (Week 10), the derivative $df_u(v)$ of the function $f(x) = \frac{1}{2}x^\top x$ (where $x, u, v \in \mathbb{R}^n$) at u is

$$df_u(v) = u^\top v,$$

and the derivative $dg_u(v)$ of the function $g(x) = w^\top x$ (where $x, u, w \in \mathbb{R}^n$) at u is

$$dg_u(v) = w^\top v.$$

Applying the above to the Lagrangian

$$\begin{aligned} L(\xi, w, \epsilon, b, \lambda, \alpha_+, \alpha_-) &= \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y - b\mathbf{1}_m^\top \lambda \\ &\quad + \epsilon^\top(\tau \mathbf{1}_n - \alpha_+ - \alpha_-) + w^\top(\alpha_+ - \alpha_- - X^\top \lambda) + \frac{1}{2}K w^\top w, \end{aligned}$$

we find that the derivative $dL_{\xi,w,\epsilon,b}(\xi_1, w_1, \epsilon_1, b_1)$ is given by

$$\begin{aligned} dL_{\xi,w,\epsilon,b}(\xi_1, w_1, \epsilon_1, b_1) &= (\xi - \lambda)^\top \xi_1 + (Kw + (\alpha_+ - \alpha_- - X^\top \lambda))^\top w_1 \\ &\quad + (\tau \mathbf{1}_n - \alpha_+ - \alpha_-)^\top \epsilon_1 - (\mathbf{1}_m^\top \lambda) b_1. \end{aligned}$$

Consequently, the gradient $\nabla L_{\xi,w,\epsilon,b}$ is given by

$$\nabla L_{\xi,w,\epsilon,b} = \begin{pmatrix} \xi - \lambda \\ Kw + (\alpha_+ - \alpha_- - X^\top \lambda) \\ \tau \mathbf{1}_n - \alpha_+ - \alpha_- \\ -\mathbf{1}_m^\top \lambda \end{pmatrix}.$$

(2) By setting the gradient $\nabla L_{\xi,w,\epsilon,b}$ to zero we obtain the equations

$$\begin{aligned} \xi &= \lambda \\ Kw &= -(\alpha_+ - \alpha_- - X^\top \lambda) \qquad (*_w) \\ \alpha_+ + \alpha_- - \tau \mathbf{1}_n &= 0 \\ \mathbf{1}_m^\top \lambda &= 0. \end{aligned}$$

Since

$$Kw = -(\alpha_+ - \alpha_- - X^\top \lambda)$$

and

$$\alpha_+ = K\beta_+,\ \ \alpha_- = K\beta_-,\ \ \lambda = K\mu,\ \ \mu = \mu_+ - \mu_-,$$

we obtain

$$\begin{aligned} w &= -(\alpha_+ - \alpha_- - X^\top \lambda)/K \\ &= -(\beta_+ - \beta_- - X^\top \mu) = -\beta_+ + \beta_- + X^\top \mu_+ - X^\top \mu_- \\ &= \begin{pmatrix} -I_n & I_n & X^\top & -X^\top \end{pmatrix} \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix}. \end{aligned}$$

As a consequence,

$$\frac{1}{2} w^\top w = \frac{1}{2} \begin{pmatrix} \beta_+^\top & \beta_-^\top & \mu_+^\top & \mu_-^\top \end{pmatrix} \begin{pmatrix} -I_n & I_n & X^\top & -X^\top \end{pmatrix}^\top \begin{pmatrix} -I_n & I_n & X^\top & -X^\top \end{pmatrix} \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix}.$$

But

$$\begin{aligned} \begin{pmatrix} -I_n & I_n & X^\top & -X^\top \end{pmatrix}^\top \begin{pmatrix} -I_n & I_n & X^\top & -X^\top \end{pmatrix} &= \begin{pmatrix} -I_n \\ I_n \\ X \\ -X \end{pmatrix} \begin{pmatrix} -I_n & I_n & X^\top & -X^\top \end{pmatrix} \\ &= \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top & -XX^\top \\ X & -X & -XX^\top & XX^\top \end{pmatrix}, \end{aligned}$$

so

$$\frac{1}{2} w^\top w = \frac{1}{2} \begin{pmatrix} \beta_+^\top & \beta_-^\top & \mu_+^\top & \mu_-^\top \end{pmatrix} Q \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix},$$

with Q the symmetric positive semidefinite matrix

$$Q = \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top & -XX^\top \\ X & -X & -XX^\top & XX^\top \end{pmatrix}.$$

(3) The value of the dual function $G(\mu, \beta_+, \beta_-)$ corresponds to the minimum value of the Lagrangian

$$
\begin{aligned}
L(\xi, w, \epsilon, b, \lambda, \alpha_+, \alpha_-) = \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y - b\mathbf{1}_m^\top \lambda \\
+ \epsilon^\top(\tau \mathbf{1}_n - \alpha_+ - \alpha_-) + w^\top(\alpha_+ - \alpha_- - X^\top \lambda) + \frac{1}{2}Kw^\top w
\end{aligned}
$$

when $\nabla L_{\xi, w, \epsilon, b} = 0$, namely

$$
\begin{aligned}
\xi &= \lambda \\
Kw &= -(\alpha_+ - \alpha_- - X^\top \lambda) \qquad (*_w) \\
\alpha_+ + \alpha_- - \tau \mathbf{1}_n &= 0 \\
\mathbf{1}_m^\top \lambda &= 0.
\end{aligned}
$$

Consequently,

$$
G(\mu, \beta_+, \beta_-) = \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y + w^\top(\alpha_+ - \alpha_- - X^\top \lambda) + \frac{1}{2}Kw^\top w.
$$

Using $(*_w)$ and the fact that $\xi = K\mu$, $\lambda = K\mu$, $\alpha_+ = K\beta_+$, $\alpha_- = K\beta_-$, we find that the dual function is given by

$$
\begin{aligned}
G(\mu, \beta_+, \beta_-) &= \frac{1}{2}\xi^\top \xi - \xi^\top \lambda + \lambda^\top y + w^\top(\alpha_+ - \alpha_- - X^\top \lambda) + \frac{1}{2}Kw^\top w \\
&= \frac{1}{2}\xi^\top \xi - K\xi^\top \mu + K\mu^\top y + Kw^\top(\beta_+ - \beta_- - X^\top \mu) + \frac{1}{2}Kw^\top w \\
&= \frac{1}{2}K^2\mu^\top \mu - K^2\mu^\top \mu + Ky^\top \mu - Kw^\top w + \frac{1}{2}Kw^\top w \\
&= -\frac{1}{2}K^2\mu^\top \mu - \frac{1}{2}Kw^\top w + Ky^\top \mu.
\end{aligned}
$$

(4) The equation $\mu = \mu_+ - \mu_-$ can be written in matrix form as

$$
\mu = \begin{pmatrix} I_m & -I_m \end{pmatrix} \begin{pmatrix} \mu_+ \\ \mu_- \end{pmatrix},
$$

so

$$
\frac{1}{2}\mu^\top \mu = \frac{1}{2} \begin{pmatrix} \mu_+^\top & \mu_-^\top \end{pmatrix} \begin{pmatrix} I_m & -I_m \\ -I_m & I_m \end{pmatrix} \begin{pmatrix} \mu_+ \\ \mu_- \end{pmatrix}.
$$

Since

$$
G(\mu, \beta_+, \beta_-) = -\frac{1}{2}K^2\mu^\top \mu - \frac{1}{2}Kw^\top w + Ky^\top \mu = -\frac{1}{2}Kw^\top w - \frac{1}{2}K^2\mu^\top \mu + Ky^\top \mu,
$$

using (2) to rewrite $\frac{1}{2}w^\top w$, (4) to rewrite $\frac{1}{2}\mu^\top\mu$, and (3), we obtain

$$G(\beta_+, \beta_-, \mu_+, \mu_-) = -\frac{1}{2}K \begin{pmatrix} \beta_+^\top & \beta_-^\top & \mu_+^\top & \mu_-^\top \end{pmatrix} P \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix} - Kq^\top \begin{pmatrix} \beta_+ \\ \beta_- \\ \mu_+ \\ \mu_- \end{pmatrix}$$

with

$$\begin{aligned} P &= Q + K \begin{pmatrix} 0_{n,n} & 0_{n,n} & 0_{n,m} & 0_{n,m} \\ 0_{n,n} & 0_{n,n} & 0_{n,m} & 0_{n,m} \\ 0_{m,n} & 0_{m,n} & I_m & -I_m \\ 0_{m,n} & 0_{m,n} & -I_m & I_m \end{pmatrix} \\ &= \begin{pmatrix} I_n & -I_n & -X^\top & X^\top \\ -I_n & I_n & X^\top & -X^\top \\ -X & X & XX^\top + KI_m & -XX^\top - KI_m \\ X & -X & -XX^\top - KI_m & XX^\top + KI_m \end{pmatrix}, \end{aligned}$$

and

$$q = \begin{pmatrix} 0_n \\ 0_n \\ -y \\ y \end{pmatrix}.$$

Total: 50 points