Fundamentals of Linear Algebra and Optimization

Ridge Regression: Learning an Affine Function

Jean Gallier and Jocelyn Quaintance

CIS Department
University of Pennsylvania

jean@cis.upenn.edu

April 13, 2022

Ridge Regression for an Affine Function

It is easy to adapt the above method to learn an affine function $f(x) = x^\top w + b$ instead of a linear function $f(x) = x^\top w$, where $b \in \mathbb{R}$. We have the following optimization program

Program $(\mathbf{RR3})$:

$$\begin{aligned} \text{minimize} \quad & \xi^\top \xi + K w^\top w \\ \text{subject to} \quad & \\ & y - Xw - b\mathbf{1} = \xi, \end{aligned}$$

with $y, \xi, \mathbf{1} \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$. Note that in Program $(\mathbf{RR3})$ minimization is performed over ξ, w and b, but b is *not* penalized in the objective function.

Ridge Regression: Program (RR3) Solution

The objective function is *convex*.

The Lagrangian associated with this program is

$$L(\xi, w, b, \lambda) = \xi^\top \xi + K w^\top w - w^\top X^\top \lambda - \xi^\top \lambda - b\mathbf{1}^\top \lambda + \lambda^\top y.$$

Since L is *convex as a function of* ξ, b, w, it has a minimum iff $\nabla L_{\xi,b,w} = 0$.

Ridge Regression: Dual Function of (RR3)

We get

$$\lambda = 2\xi$$
$$\mathbf{1}^\top \lambda = 0$$
$$w = \frac{1}{2K} X^\top \lambda = X^\top \frac{\xi}{K}.$$

As before, if we set $\xi = K\alpha$, we obtain $\lambda = 2K\alpha$, $w = X^\top \alpha$, and

$$G(\alpha) = -K\alpha^\top (XX^\top + KI_m)\alpha + 2K\alpha^\top y.$$

Ridge Regression: Dual Program of (RR3)

Since $K > 0$ and $\lambda = 2K\alpha$, the dual to ridge regression is the following program

Program $(\mathbf{DRR3})$:

$$\begin{aligned} &\text{minimize} \quad \alpha^\top (XX^\top + KI_m)\alpha - 2\alpha^\top y \\ &\text{subject to} \\ &\qquad\qquad \mathbf{1}^\top \alpha = 0, \end{aligned}$$

where the minimization is over α.

Ridge Regression: Solution to (DRR3)

Observe that up to the factor $1/2$, this problem satisfies the conditions of a previous proposition from the first lesson of the quadratic optimization lesson with

$$\begin{aligned} A &= (XX^\top + KI_m)^{-1} \\ b &= y \\ B &= \mathbf{1}_m \\ f &= 0, \end{aligned}$$

and x renamed as α.

Ridge Regression: Solution to (DRR3)

Therefore, it has a unique solution (α, μ) (beware that $\lambda = 2K\alpha$ is **not** the λ used before, which we rename as μ), which is the unique solution of the KKT-equations

$$\begin{pmatrix} XX^\top + KI_m & \mathbf{1}_m \\ \mathbf{1}_m^\top & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \mu \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}.$$

Ridge Regression: Solution to (DRR3)

Since the solution is

$$\mu = (B^\top AB)^{-1}(B^\top Ab - f), \quad \alpha = A(b - B\mu),$$

we get

$$\mu = (\mathbf{1}^\top (XX^\top + KI_m)^{-1}\mathbf{1})^{-1}\mathbf{1}^\top (XX^\top + KI_m)^{-1} y$$
$$\alpha = (XX^\top + KI_m)^{-1}(y - \mu\mathbf{1}).$$

Ridge Regression: Solution to (DRR3)

Interestingly $b = \mu$, which is not obvious a priori.

Proposition. We have $b = \mu$.

Ridge Regression: Program (RR3) Solution

In summary the KKT-equations determine both α and μ, and so $w = X^\top \alpha$ and b as well.

Ridge Regression: Averaging Formula for b

There is also a useful expression of b as an average. We have

$$b = \bar{y} - \sum_{j=1}^{n} \overline{X^j} w_j = \bar{y} - (\overline{X^1} \cdots \overline{X^n}) w,$$

where $\bar{y}$ is the mean of y and $\overline{X^j}$ is the mean of the jth column of X.

Ridge Regression: Affine Case Reduction

It can be shown that solving the Dual $(\mathbf{DRR3})$ for α and obtaining $w = X^\top \alpha$ is **equivalent** to solving our previous ridge regression Problem $(\mathbf{RR2})$ applied to the centered data $\widehat{y} = y - \overline{y}\mathbf{1}_m$ and $\widehat{X} = X - \overline{X}$, where $\overline{X}$ is the $m \times n$ matrix whose jth column is $\overline{X^j}\mathbf{1}_m$, the vector whose coordinates are all equal to the mean $\overline{X^j}$ of the jth column X^j of X.

Ridge Regression: Program (RR6)

Program $(\mathbf{RR6})$ *is equivalent to ridge regression without an intercept term applied to the centered data* $\widehat{y} = y - \overline{y}\mathbf{1}$ *and* $\widehat{X} = X - \overline{X}$,

Program $(\mathbf{RR6})$:

$$\begin{aligned} \text{minimize} \quad & \xi^\top \xi + K w^\top w \\ \text{subject to} \quad & \\ & \widehat{y} - \widehat{X} w = \xi, \end{aligned}$$

minimizing over ξ and w.

Ridge Regression: Program (RR6) Solution

If $\widehat{w}$ is the optimal solution of this program given by

$$\widehat{w} = \widehat{X}^\top (\widehat{X}\widehat{X}^\top + KI_m)^{-1}\widehat{y}, \tag{$*_{w_6}$}$$

then b is given by

$$b = \overline{y} - (\overline{X^1} \ \cdots \ \overline{X^n})\widehat{w}.$$

Ridge Regression: Learning an Affine Function

In practice Program $(\mathbf{RR6})$ involving the centered data appears to be the preferred one.

Ridge Regression: Illustrated Example

Example. Consider the data set (X, y_1) with

$$X = \begin{pmatrix} -10 & 11 \\ -6 & 5 \\ -2 & 4 \\ 0 & 0 \\ 1 & 2 \\ 2 & -5 \\ 6 & -4 \\ 10 & -6 \end{pmatrix}, \quad y_1 = \begin{pmatrix} 0 \\ -2.5 \\ 0.5 \\ -2 \\ 2.5 \\ -4.2 \\ 1 \\ 4 \end{pmatrix}$$

as illustrated in Figure 1.

Ridge Regression: Illustrated Example

We find that $\bar{y} = -0.0875$ and $(\overline{X^1}, \overline{X^2}) = (0.125, 0.875)$. For the value $K = 5$, we obtain

$$w = \begin{pmatrix} 0.9207 \\ 0.8677 \end{pmatrix}, \quad b = -0.9618,$$

for $K = 0.1$, we obtain

$$w = \begin{pmatrix} 1.1651 \\ 1.1341 \end{pmatrix}, \quad b = -1.2255,$$

and for $K = 0.01$,

$$w = \begin{pmatrix} 1.1709 \\ 1.1405 \end{pmatrix}, \quad b = -1.2318.$$

See Figure 2.

Ridge Regression: Illustrated Example

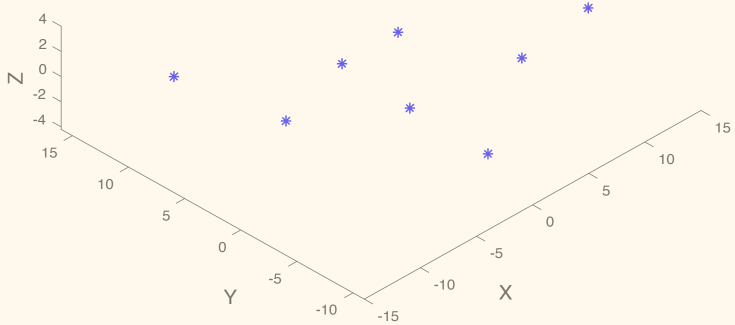

Figure 1: The data set (X, y_1).

Ridge Regression: Illustrated Example

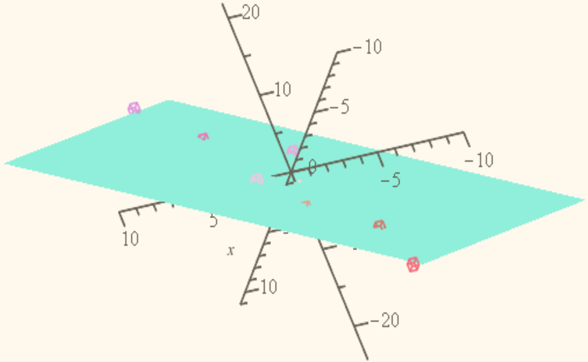

Figure 2: The graph of the plane $f(x, y) = 1.1709x + 1.1405y - 1.2318$ as an approximate fit to the data (X, y_1).

Ridge Regression: Illustrated Example

We conclude that the points (X_i, y_i) (where X_i is the ith row of X) almost lie on the plane of equation

$$x + y - z - 1 = 0,$$

and that f is almost the function given by $f(x, y) = 1.1x + 1.1y - 1.2$. See Figures 3 and 4.

Ridge Regression: Illustrated Example

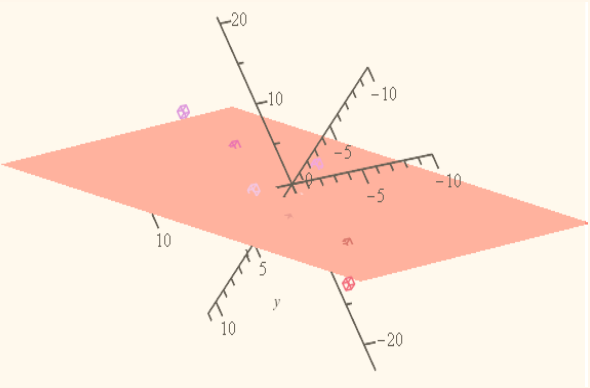

Figure 3: The graph of the plane $f(x, y) = 1.1x + 1.1y - 1.2$ as an approximate fit to the data (X, y_1).

Ridge Regression: Illustrated Example

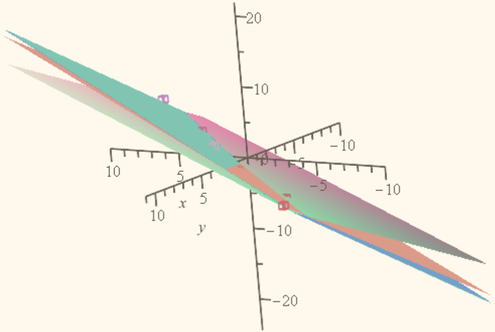

Figure 4: A comparison of how the graphs of the planes corresponding to $K = 1, 0.1, 0.01$ and the salmon plane of equation $f(x, y) = 1.1x + 1.1y - 1.2$ approximate the data (X, y_1).

Ridge Regression: Illustrated Example

If we change y_1 to

$$y_2 = \begin{pmatrix} 0 & -2 & 1 & -1 & 2 & -4 & 1 & 3 \end{pmatrix}^\top,$$

as evidenced by Figure 5, the exact solution is

$$w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad b = -1,$$

and for $K = 0.01$, we find that

$$w = \begin{pmatrix} 0.9999 \\ 0.9999 \end{pmatrix}, \quad b = -0.9999.$$

Ridge Regression: Illustrated Example

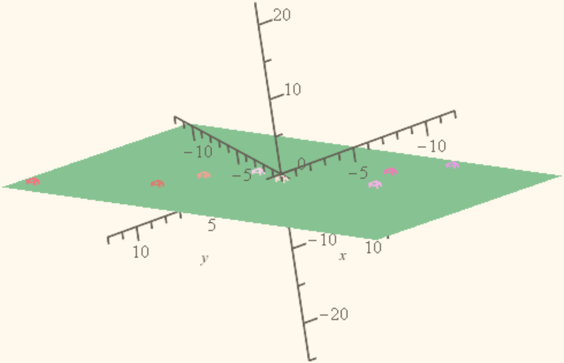

Figure 5: The data (X, y_2) is contained within the graph of the plane $f(x, y) = x + y - 1$.

Ridge Regression: Learning an Affine Function

We can see how the choice of K affects the quality of the solution (w, b) by computing the norm $\|\xi\|_2$ of the error vector $\xi = \widehat{y} - \widehat{X}w$. We notice that the smaller K is, the smaller is this norm.

As a least squares problem, the solution is given in terms of the pseudo-inverse $[X\,\mathbf{1}]^+$ of $[X\,\mathbf{1}]$ by

$$\begin{pmatrix} w \\ b \end{pmatrix} = [X\,\mathbf{1}]^+ y.$$