1. LSE:

For convenience, we consider two dimensional data (x, y).

Ideally, we would expect that $A\vec{w} = \vec{b}$:

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ & & \vdots & & \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}}_{n \times (m+1)} \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix}}_{(m+1) \times 1} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1}$$

Then $\vec{w}$ can be calculated by $\vec{w} = A^{-1}\vec{b}$, however, A is not guaranteed to be invertible. In fact, A doesn't even need to be a square matrix. Therefore, we change our

goal from finding $\vec{w}$ such that $A\vec{w} = \vec{b}$ to minimizing $\|A\vec{w} - \vec{b}\|^2$.

$$\|A\vec{w} - \vec{b}\|^2 = (A\vec{w} - \vec{b})^T (A\vec{w} - \vec{b})$$

$$= (\vec{w}^T A^T - \vec{b}^T)(A\vec{w} - \vec{b})$$

both are scalar

$$= \vec{w}^T A^T A\vec{w} - \vec{w}^T A^T \vec{b} - \vec{b}^T A\vec{w} + \vec{b}^T \vec{b}$$

$$= \vec{w}^T A^T A\vec{w} - 2\vec{w}^T A^T \vec{b} + \vec{b}^T \vec{b}$$

$$\frac{\partial(\vec{w}^T A^T A\vec{w})}{\partial \vec{w}} = \begin{bmatrix} \frac{\partial}{\partial w_0} \\ \vdots \\ \frac{\partial}{\partial w_m} \end{bmatrix} [w_0 \; w_1 \cdots w_m] \begin{bmatrix} a_{0,0} & a_{0,1} \cdots & a_{0,m} \\ a_{1,0} & a_{1,1} \cdots & a_{1,m} \\ \vdots & \vdots & \vdots \\ a_{m,0} & a_{m,1} \cdots & a_{m,m} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial w_0} \\ \vdots \\ \frac{\partial}{\partial w_m} \end{bmatrix} \underbrace{\begin{bmatrix} w_0(a_{0,0}\cdot w_0 + a_{0,1}\cdot w_1 + \cdots + a_{0,m}\cdot w_m) \\ + w_1(a_{1,0}\cdot w_0 + a_{1,1}\cdot w_1 + \cdots + a_{1,m}\cdot w_m) \\ \vdots \\ + w_m(a_{m,0}\cdot w_0 + a_{m,1}\cdot w_1 + \cdots + a_{m,m}\cdot w_m) \end{bmatrix}}_{1\times 1 \text{ (scalar)}}$$

$$= \underbrace{\begin{bmatrix} (a_{0,0}w_0 + a_{0,1}w_1 + \cdots + a_{0,m}w_m) + (w_0a_{0,0} + w_1a_{1,0} + \cdots + w_ma_{m,0}) \\ (a_{1,0}w_0 + a_{1,1}w_1 + \cdots + a_{1,m}w_m) + (w_0a_{0,1} + w_1a_{1,1} + \cdots + w_ma_{m,1}) \\ \vdots \\ (a_{m,0}w_0 + a_{m,1}w_1 + \cdots + a_{m,m}w_m) + (w_ma_{0,m} + w_1a_{1,m} + \cdots + w_ma_{m,m}) \end{bmatrix}}_{(m+1)\times 1}$$

$$= (A^TA)\vec{w} + (A^TA)^T\vec{w}$$

$$= 2(A^TA)\vec{w} \ \#$$

Similarly, $\frac{\partial(\vec{w}^T A^T \vec{b})}{\partial \vec{w}} = \begin{bmatrix} \frac{\partial}{\partial w_0} \\ \vdots \\ \frac{\partial}{\partial w_m} \end{bmatrix} [w_0\ w_1\ \cdots\ w_m]\ (A^T\vec{b})$

$$= A^T\vec{b}\ \#$$

$\therefore f(\vec{w}) = \vec{w}^T A^T A\vec{w} - 2\vec{w}^T A^T\vec{b} + \vec{b}^T\vec{b}$

$$\frac{\partial f(\vec{w})}{\partial \vec{w}} = 2A^TA\vec{w} - 2A^T\vec{b} = 0$$

$$\Rightarrow \boxed{\vec{w} = (A^TA)^{-1}A^T\vec{b}\ \#}$$

Hence $f(\vec{w})$ attains its minimum at $\vec{w} = (A^TA)^{-1}A^T\vec{b}$. Note that we only have $\det(A^TA) \geq 0$, it's still possible that (A^TA) is not invertible.

As a result, we can add the regularized term $\lambda I (\lambda > 0)$ to (A^TA), and then $\det(A^TA + \lambda I) > 0$, which means invertible.

$\therefore$ $\boxed{\vec{w} = (A^TA + \lambda I)^{-1} A^T \vec{b}}$ This is called rLSE

Next we'll illustrate on LU decomposition and how to apply it to find the inverse matrix.

$$A = LU = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{2,1} & 1 & 0 & & 0 \\ l_{3,1} & l_{3,2} & 1 & & 0 \\ \vdots & \vdots & & & \vdots \\ l_{n,1} & l_{n,2} & \cdots & & 1 \end{bmatrix} \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ 0 & u_{2,2} & & u_{2,n} \\ 0 & 0 & & \vdots \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & u_{n,n} \end{bmatrix}$$

A: $n \times n$, L: $n \times n$, U: $n \times n$

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & & & \vdots \\ \vdots & & & \\ a_{n,1} & \cdots & \cdots & a_{n,n} \end{bmatrix}$$

$$\Rightarrow a_{1,1} = u_{1,1},\ a_{1,2} = u_{1,2},\ \cdots\ a_{1,n} = u_{1,n}$$

$$a_{2,1} = l_{2,1}\cdot u_{1,1},\ a_{3,1} = l_{3,1}\cdot u_{1,1},\ \cdots\ a_{n,1} = l_{n,1}\cdot u_{1,1}$$

$$\Rightarrow l_{2,1} = \frac{a_{2,1}}{u_{1,1}},\ l_{3,1} = \frac{a_{3,1}}{u_{1,1}},\ \cdots\ l_{n,1} = \frac{a_{n,1}}{u_{1,1}}$$

$$a_{2,2} = l_{2,1}\cdot u_{1,2} + u_{2,2},\ a_{2,3} = l_{2,1}\cdot u_{1,3} + u_{2,3}\ \cdots$$

$$\Rightarrow u_{2,2} = a_{2,2} - l_{2,1}\cdot u_{1,2},\ u_{2,3} = a_{2,3} - l_{2,1}\cdot u_{1,3}$$

$$a_{3,2} = l_{3,1}\cdot u_{1,2} + l_{3,2}\cdot u_{2,2} \Rightarrow l_{3,2} = \frac{a_{3,2} - l_{3,1}\cdot u_{1,2}}{u_{2,2}}$$

$$\vdots$$

Suppose that we already have $A = LU$.

Define $A^{-1} = [x_1\ x_2\ \cdots\ x_n]$ where $x_1, x_2, \cdots, x_n$ are column vectors. Then $A(A^{-1}) = [Ax_1\ Ax_2\ \cdots\ Ax_n] = I = [e_1\ e_2\ \cdots\ e_n]$

$\Rightarrow LUx_1 = e_1,\ LUx_2 = e_2 \cdots LUx_n = e_n$

We solve for $\begin{cases} Ux_1 = y_1,\ Ly_1 = e_1 \\ Ux_2 = y_2,\ Ly_2 = e_2 \\ \vdots \\ Ux_n = y_n,\ Ly_n = e_n \end{cases}$

Then A^{-1} can be calculated #

2. Steepest descent method:

The formula of steepest descent, a.k.a gradient descent, can be written as:

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

where f is the loss function.

Assume that f is Lipschitz continuous with constant $L > 0$. Then $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$ for any x, y.

We can perform a quadratic expansion of f around $f(x_t)$ and obtain the following inequality:

$$f(x_{t+1}) \le f(x_t) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{1}{2} \nabla^2 f(x_t) \|x_{t+1} - x_t\|^2$$

$$\le f(x_t) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{1}{2} L \|x_{t+1} - x_t\|^2$$

$$= f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{1}{2} L \eta^2 \|\nabla f(x_t)\|^2$$

$$= f(x_t) - (1 - \frac{1}{2} L \eta) \eta \|\nabla f(x_t)\|^2$$

Note that $\|x_{t+1} - x_t\|$ has to be small enough, which implies that η also has to be small enough.

Choose $\eta \le \frac{1}{L}$, then $-(1 - \frac{1}{2} L \eta) = \frac{1}{2} L \eta - 1$

$$\le \frac{1}{2} L (\frac{1}{L}) - 1 = -\frac{1}{2}$$

$$\therefore \boxed{f(x_{t+1}) \le f(x_t) - \frac{1}{2} \eta \|\nabla f(x_t)\|^2} \quad (**)$$

↳ positive unless $\nabla f(x) = 0$

Thus the sequence $\{f(x_0), f(x_1), \cdots\}$ is indeed decreasing.

Assume that f is convex and $f(x)$ attains its minimum at $x = x^*$, then we have:

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) \quad \text{for any } x$$

$$\Leftrightarrow f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$$

Then (**) becomes $f(x_{t+1}) \leq f(x^*) + \nabla f(x_t)^T (x_t - x^*) - \frac{1}{2}\eta \|\nabla f(x_t)\|^2$

$$\Rightarrow f(x_{t+1}) - f(x^*) \leq \frac{1}{2\eta} (2\eta \nabla f(x_t)^T (x_t - x^*) - \eta^2 \|\nabla f(x_t)\|^2)$$

$$= \frac{1}{2\eta} (\|x_t - x^*\|^2 - \|x_t - \eta \nabla f(x_t) - x^*\|^2)$$

$$= \frac{1}{2\eta} (\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)$$

This implies that the sequence $\{f(x_0), f(x_1) \cdots\}$ is bounded.

In the end, consider $g(\vec{w}) = \|\vec{b} - A\vec{w}\|^2$, a.k.a LSE loss.

Hence $\frac{\partial g}{\partial \vec{w}} = 2A^TA\vec{w} - 2A^T\vec{b}$ $\Rightarrow$ gradient

$$= 2A^T(A\vec{w} - \vec{b})$$

On the other hand, for the regularized term in L1-norm, the gradient of it can be written as the sign function. $\text{sign}(W_i) = \begin{cases} 1, & \text{if } W_i > 0 \\ -1, & \text{if } W_i < 0 \\ 0, & \text{if } W_i = 0 \end{cases}$

Thus the gradient in total is $2A^T(A\vec{w} - \vec{b}) + \lambda \cdot \text{sign}(\vec{w})$ #

3. Newton's method:

We have the following equation (from Taylor expansion):

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + O((x - x_0)^2)$$

If we want to find x such that $f(x) = 0$, and $|x - x_0|$ is small enough, then

$$0 = f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

$$\Rightarrow x \approx x_0 - \frac{f(x_0)}{f'(x_0)}$$

Then we derive the formula of Newton's method:

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$$

It will converge to the root of the original equation, and the error is up to your tolerance.

If we want to apply Newton's method to an optimization problem, we may need to solve $f(x) = 0$:

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}$$

$\Rightarrow x_{t+1} = x_t - (\nabla^2 f(x_t))^{-1} \nabla f(x_t)$ in multi-dimensional case.

From (1) LSE and (2) Steepest descent method, we know that $\nabla f(\vec{w}) = 2A^T(A\vec{w} - \vec{b})$ for f: square error.

$\nabla^2 f(\vec{w}) = 2A^TA$ #