Regularization

Iván Andrés Trujilllo Abella

ivantrujillo1229@gmail.com

OLS

...

- $$\underset{n\times 1}{y} \textbf{ target variable}$$
- $$\underset{n\times D}{X} \textbf{ feature matrix}$$
- $$\underset{D\times 1}{\Theta} \textbf{ parameters vector}$$

The problems is expressed as:

...

$$J(\Theta) = \|X\Theta - y\|^2 \tag{1}$$

Loss function

...

$$J(\Theta) = \|X\Theta - y\|^2 \tag{2}$$

Now we need rewrite the function:

...

$$\begin{aligned} J(\Theta) &= (X\Theta - y)^T(X\Theta - y) \\ &= \Theta^T X^T X\Theta - \Theta^T X^T y - y^T X\Theta + y^T y \\ &= \Theta^T X^T X\Theta - 2\Theta^T X^T y + y^T y \end{aligned} \tag{3}$$

Note that $y^T X\Theta$ is an escalar and the tranpose of a scalar is itself then.

Minimizing

...

$$\frac{\partial J(\Theta)}{\partial \Theta} \tag{4}$$

...

- $$\frac{\partial \left(\Theta^T X^T X^\Theta\right)}{\partial \Theta} = 2X^T X\Theta$$
- $$\frac{\partial \left(-2\Theta^T X^T y\right)}{\partial \Theta} = -2X^T y$$

(See the appendix)

...

$$\frac{\partial J(\Theta)}{\partial \Theta} = 2X^T X\Theta - 2X^T y = 0 \tag{5}$$

$$X^T X\Theta = X^T y \tag{6}$$

...

$$\Theta^* = \left(X^T X\right)^{-1} X^T y \tag{7}$$

Regularization

Ridge

...

$$J(\Theta) = \|A\Theta - y\|^2 + \lambda\|\Theta\|^2 \tag{8}$$

Regularization is a constrain to solution, the strengh of the constrin is measured by λ and the way that affect is $\|\Theta\|^2$.

Ridge

...

$$
\begin{aligned}
J(\Theta) &= \|A\Theta - y\|^2 + \lambda\|\Theta\|^2 \\
&= (A\Theta - y)^T (A\Theta - y) + \lambda\Theta^T\Theta \\
&= \Theta X^T X\Theta - 2\Theta^T X^T y + y^T y + \lambda\Theta^T\Theta
\end{aligned} \tag{9}
$$

Remember that $y^T X\Theta$ is an scalar and its transpose is: $\Theta^T X^T y$

Derivatives

...

$$\frac{J(\Theta)}{\partial\Theta} = 2X^TX\Theta - 2X^Ty + 2\lambda\Theta \tag{10}$$

...

- $$\frac{\partial\left(2\Theta^TX^TX\right)}{\partial\Theta} = 2X^TX\Theta \tag{11}$$
- $$\frac{\partial\left(\Theta^Tx^Ty\right)}{\partial\Theta} = X^Ty \tag{12}$$
- $$\frac{\partial\left(\lambda\Theta^T\Theta\right)}{\partial\Theta} = 2\lambda\Theta \tag{13}$$

...

...

$$\frac{J(\Theta)}{\partial\Theta} = 2X^T X\Theta - 2X^T y + 2\lambda\Theta = 0$$
$$= X^T X\Theta + \lambda\Theta = X^T y$$
$$= (X^T X + \lambda I)\Theta = X^T y \tag{14}$$

...

$$\Theta^* = \left(X^T X + \lambda I\right)^{-1} X^T y \tag{15}$$

is closed-form solution for **ridge** penalty.

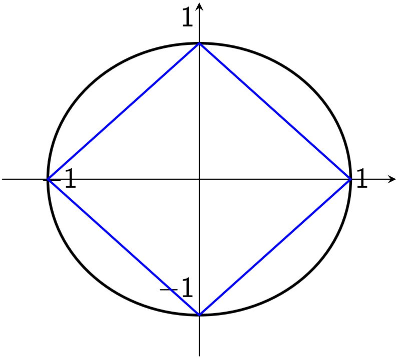
Choose λ

How must be select λ? uses cross-validation.

...

- if λ is bigger then the magnitude the coefficients is low.

Norms



In this figure are shown the **isosurfaces** norms $l1 = 1$ and $l2 = 1$. In general the norms could be written as:

...

$$\|x\|_p = \left(\sum |x_i|^p\right)^{\frac{1}{p}} \tag{16}$$

Sparsity

Could be very important reach a corner solution (feature elimination)
Least Absolute Shrinkage and Selection Operator (LASSO).

lasso

$$J(\Theta) = \|X\Theta - y\| + \lambda\|\Theta\|_1 \tag{17}$$

How solve if absolute values are not differentiable?.

Apendix

...

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A} \tag{18}$$

Think the following:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1,m} \\ a_{21} & a_{21} & \dots & a_{2,m} \\ \vdots & \vdots & \dots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,m} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} a_{1i} x_i \\ \sum_{i=1}^{m} a_{2i} x_i \\ \vdots \\ \sum_{i=1}^{m} a_{ni} x_i \end{bmatrix}$$

If apply we derivate with respect to $\boldsymbol{x}$

$$\begin{bmatrix} \frac{\partial f_1()}{\partial x_1} & \frac{\partial f_1()}{\partial x_2} & \cdots & \frac{\partial f_1()}{\partial x_m} \\ \frac{\partial f_2()}{\partial x_1} & \frac{\partial f_2()}{\partial x_2} & \cdots & \frac{\partial f_2()}{\partial x_m} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_n()}{\partial x_1} & \frac{\partial f_n()}{\partial x_2} & \cdots & \frac{\partial f_n()}{\partial x_m} \end{bmatrix}$$

Notice that

$$\frac{\partial \left(\sum_{i=1}^{n} a_{ki} x_i \right)}{\partial x_k} = a_{ki}$$

Therefore for k row and i column:

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A} \tag{19}$$

.

...

$$\frac{\partial\left(\boldsymbol{x}^T\boldsymbol{A}x\right)}{\partial\boldsymbol{A}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x} \tag{20}$$

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} a_{11}x_1 + a_{12}x_3 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{bmatrix}$$

that could be expressed as:

$$\begin{aligned} & x_1(a_{11}x_1 + a_{12}x_1 + a_{13}x_3)+ \\ & x_2(a_{21}x_1 + a_{22}x_2 + a_{23}x_3)+ \\ & x_3(a_{31}x_1 + a_{32}x_2 + a_{33}x_3) \end{aligned} \tag{21}$$

Note the following:
$\frac{\partial(f(x_1,\ldots,x_n))}{x_1} = 2a_{11}x_1 + \sum_{i=1,i\neq 1}^{3} a_{1i}x_i + \sum_{i=1,i\neq 1}^{3} a_{i1}x_i$
Therefore is easily extended:

$$\frac{\partial f}{\partial x_j} = 2a_{jj}x_j + \sum_{i=1,i\neq j}^{n} a_{ji}x_i + \sum_{i=1,i\neq j}^{n} a_{ij}x_i \tag{22}$$

The before equation we have that:

$$\left(\sum_{i=1,i\neq j}^{n} a_{ji}x_i + a_{jj}x_j\right) + \left(\sum_{i=1,i\neq j}^{n} a_{ij}x_i + a_{jj}x_j\right) \tag{23}$$

This can be rewritten as:

$$\left(\sum_{i=1}^{n} a_{ji} + \sum_{i=1}^{n} a_{ij}\right) x_i = \sum_{i=1}^{n} (a_{ji} + a_{ij}) x_i \tag{24}$$

if $\vec{a}_j$ is the $j - row$ row of $\boldsymbol{A}$ matrix then $\vec{a}_j\vec{x} = \sum_{i=1}^{n} a_{ji}x_i$ therefore for the $j - th$ row of $\boldsymbol{A}^T$ $\vec{a}_j\vec{x} = \sum_{i=1}^{n} a_{ij}x_i$. for m variables therefore we have that:

$$(A + A^T)_{ji} = [a_{ji} + a_{ij}] \tag{25}$$

Therefore:

$$\frac{\partial\left(\boldsymbol{x}^T \boldsymbol{A} x\right)}{\partial \boldsymbol{A}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x} \tag{26}$$

Notice when the matrix $\boldsymbol{A}$ is symmetric then:

$$\frac{\partial\left(\boldsymbol{x}^T \boldsymbol{A} x\right)}{\partial \boldsymbol{A}} = 2\boldsymbol{A}\boldsymbol{x} \tag{27}$$

give that A is symmetric also can be written as: $2\boldsymbol{x}^T\boldsymbol{A}$.

Remember that:

$$\frac{\partial(\boldsymbol{x}^T \boldsymbol{z})}{\partial \boldsymbol{y}} = \boldsymbol{z} \tag{28}$$

Now $\boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{y}$

$$\frac{\partial \left(\boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{y}\right)}{\partial \boldsymbol{x}} = \boldsymbol{A}\boldsymbol{y} \tag{29}$$

from $\boldsymbol{x}^T \left(\boldsymbol{A}^T \boldsymbol{y}\right)$ the term $\left(\boldsymbol{A}^T \boldsymbol{y}\right)$ is a vector then $\boldsymbol{z} = \left(\boldsymbol{A}^T y\right)$ the last derivative is $\frac{\partial\left(\boldsymbol{x}^T \boldsymbol{z}\right)}{\partial \boldsymbol{x}} = \boldsymbol{z} = \boldsymbol{A}^T \boldsymbol{y}$.