L_1 regularized Least Squares

$$\frac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_1$$

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft: June 16, 2018
Last update : October 28, 2019

Overview

L_1 regularized least square problem

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^{m \times 1}$, find $\mathbf{x} \in \mathbb{R}^{n \times 1}$ by solving

$$(\mathcal{P}_1): \quad \min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

A regularized least square problem

- $\frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$: data fitting term
- $\lambda \|\mathbf{x}\|_1$: regularizer
- $\lambda \geq 0$: regularization parameter
- $\|\mathbf{x}\|_1$: L_1 norm of vector $\mathbf{x}$, not differentiable

The regularizer $\lambda \|\mathbf{x}\|_1$ is a sparsity inducing regularizer, it promote the sparsity of solution $\mathbf{x}$ obtained by solving $\mathcal{P}_1$

L_2 regularized least square problem

The classical L_2 regularized least square

$$(\mathcal{P}_2): \quad \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$$

- a.k.a. Tikhonov regularization
- $\|\mathbf{x}\|_2$: L_2 norm of vector $\mathbf{x}$, it is differentiable
- The regularizer $\lambda\|\mathbf{x}\|_2^2$ forces solution $\mathbf{x}$ to have small size (in L_2 norm)

L_1 norm and L_2 norm

For a vector $\mathbf{x} \in \mathbb{R}^{n \times 1}$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$$

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| = |x_1| + |x_2| + ... + |x_n|$$

- $\|\mathbf{x}\|_2$ is differentiable. We have $\frac{\partial \|\mathbf{x}\|_2}{\partial \mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ and $\frac{\partial \|\mathbf{x}\|_2^2}{\partial \mathbf{x}} = 2\mathbf{x}$
- $\|\mathbf{x}\|_1$ is not-differentiable : absolute value is non-differentiable
- We use sub-gradient to handle the non-differentiable $\|\mathbf{x}\|_1$

L_2 norm has squares, L_1 don't

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x_1^2 + x_2^2 + ... + x_n^2} = \text{ sum of squares}$$

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| = |x_1| + |x_2| + ... + |x_n| = \text{ sum of magnitudes}$$

Consider the squaring action on a component x_i :

- If x_i is large (says $x_i > 1$) : x_i^2 becomes very big, so x_i has more contribution in the sum.
- If x_i is small (says $x_i < 1$) : x_i^2 becomes very small, so x_i has less contribution in the sum.
- So L_2 norm will pay more attention on large components

L_1 norm does not square the magnitude in both cases, so L_1 norm pay will more attention on small components than L_2 norm $\implies$ it forces small components to zero more quickly than L_2 norm.

L_1 norm regularization induces sparsity

$$(\mathcal{P}_2): \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2, \quad (\mathcal{P}_1): \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

For $(\mathcal{P}_2)$, large components in $\mathbf{x}$ will get more attention, they will be compressed more during the minimization, so the sol.

- does not have large component
- is allowed to have many small non-zero components

For $(\mathcal{P}_1)$, small components in $\mathbf{x}$ get more attention (compared to L_2), so the sol.

- does not contain many small non-zero components as the L_2 case, small components in $\mathbf{x}$ will become zero
- is allowed to have a few large components
- i.e., the sol. obtained by solving $(\mathcal{P}_1)$ will be sparse

L_1 norm is less sensitive to outlier

Suppose component x_1 in $\mathbf{x}$ is an outlier.

Fact : outlier has extreme magnitude (far away from the normal range).

A way to handle outlier : discard it.
i.e. pay no attention to outlier : assume we know x_1 is outlier, we can just ignore x_1 as it does not provide useful but harmful information.

L_2 norm has square operation : it will be sensitive to x_1.
L_1 norm has no square operation : it is less sensitive to x_1
$\Longrightarrow$ it is more robust to outlier

L_1 norm and L_0 norm

For a vector $\mathbf{x} \in \mathbb{R}^{n \times 1}$

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| = |x_1| + |x_2| + ... + |x_n|$$

$$\|\mathbf{x}\|_0 = \#\{\, i \mid x_i \neq 0 \,\} = \text{ number of nonzeros in } \mathbf{x}$$

- Other name of L_0 norm : cardinality
- L_0 norm can be treated as the limit of L_p norm as $p \to 0$

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} x_i^p \right)^p$$

- L_0 norm is not a "norm" : L_p norm is not a norm when $p < 1$
- Like L_1 norm, L_0 norm is also non-differentiable : L_p norm is non-differentiable when $p \leq 1$

L_1 and L_0 regularized least squares

$$(\mathcal{P}_1): \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1, \quad (\mathcal{P}_0): \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_0$$

- L_0 regularization select the component of $\mathbf{x}$
- $(\mathcal{P}_0)$ is NP-hard : see p.8-10 here.

Relation between $(\mathcal{P}_0)$ and $(\mathcal{P}_1)$

- $(\mathcal{P}_1)$ can be used as an approximation of $(\mathcal{P}_0)$: solving $(\mathcal{P}_1)$ provides an approximate sol. for $(\mathcal{P}_0)$
- In fact $(\mathcal{P}_1)$ is a convex relaxation of $(\mathcal{P}_0)$: the cost function of $(\mathcal{P}_0)$ is non-convex while that of $(\mathcal{P}_1)$ is convex
- Under some condition on matrix $\mathbf{A}$, the sol. of $(\mathcal{P}_1)$ is a sol. of $(\mathcal{P}_0)$ – this is the theoretical foundation of compressive sensing

Therefore we can solve $(\mathcal{P}_0)$ via solving $(\mathcal{P}_1)$.

Plot of L_0, L_1, L_2 norm

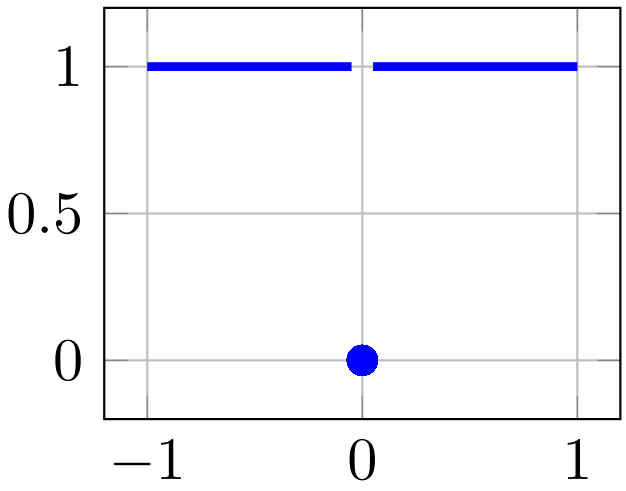

Figure: L_0

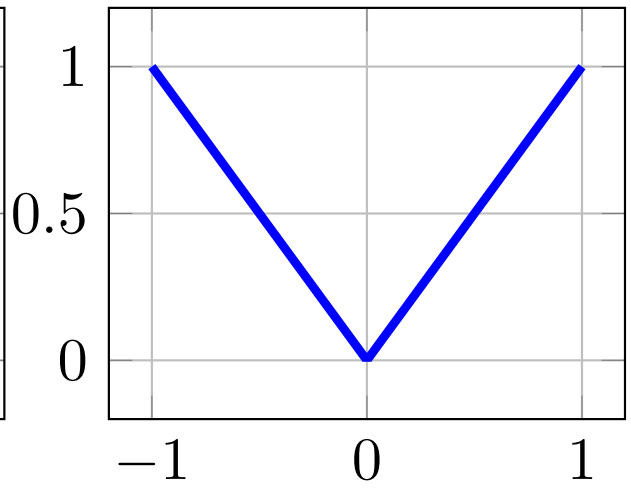

Figure: L_1

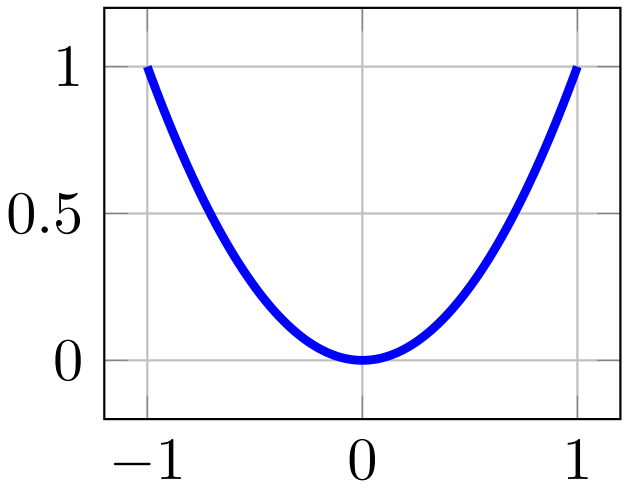

Figure: L_2

- L_1 norm is a convex envelope of L_0 norm on [-1, 1]
- L_2 norm is also a convex envelope of L_0 norm on [-1, 1]

On the gradient on scalar x

- Gradient of L_0 : zero everywhere, undefined at $x = 0$
- Gradient of L_1 : $+1$ if $x > 0$, -1 if $x < 0$ and undefined if $x = 0$
- Gradient of L_2 : $2x$

Last page - summary

- L_1 regularized least squares : given $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^{m \times 1}$, find $\mathbf{x} \in \mathbb{R}^{n \times 1}$ by solving

$$\min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

- L_1 norm compared with L_2 norm and L_0 norm
- L_1 norm promotes sparsity in solution $\mathbf{x}$
- L_1 norm is less sensitive to outlier

Next document : how to solve L_1 regularized least squares

End of document