Sparsity learning

Kaikai Zhao
Email: kkai_zhao@yeah.net

First draft: January 12, 2022 Last update: March 18, 2022

1 ℓ_0 norm regularization

2 Lasso

2.1 Solve a lasso problem

First we solve ℓ_1 the regularization problem in the section and then make the regularization as a contraint in the next section. In other words, to project a given vector onto an ℓ_1 ball if the constraint is $\|x\|_1 \leq \lambda$.
I will use Lagrangian method and derive to get a closed-form solution, refer to Anderson Ang's slides.

2.2 Group lasso regression with ADMM

Refer to Section 20.9 of my optimization notes, page 61.

2.3 Overlap group lasso regression with ADMM

Refer to the details in section 6.4.2 and 7.2 in Stephen Boyd, et al. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, 2011.

3 Projection onto unit ℓ_1 ball

The projection of $\mathbf{x}$ onto a unit ℓ_1 ball, i.e, $\|\mathbf{x}\|_1 \leq 1$, can be transformed into the following problem in which $d = 1$. Clearly, $\theta = 0$ if $\|\mathbf{x}\|_1 \leq 1$. Otherwise, we can get θ via solving the following problem.

Problem 1 (Clipping and summation)**.** *Given a vector $\mathbf{x} \in \mathbb{R}^n$, find a nonnegative real number θ such that*

$$\sum_{i=1}^{n} \max\{|x_i| - \theta, 0\} = d \tag{1}$$

where $d \in \mathbb{R}^+$. Then the clipped $\mathbf{x}$ will be given by

$$\begin{cases} x_i - \theta, & \text{if } x_i > \theta \\ 0, & \text{if } -\theta \leq x_i \leq \theta \\ x_i + \theta, & \text{if } x_i < -\theta. \end{cases}$$

which is called soft-thresholding operator, denoted as $S_d(x)$.

Algorithm 1 The Computation of θ in Problem 1

Input: $\mathbf{x} \in \mathbb{R}^n, d \in \mathbb{R}^+$ **Output:** θ
Initialization: $A = 0, \theta = 0, k = 0$
Sort the elements of $\mathbf{x}$ in descending order of their absolute values to get a new x, denoted as $\mathbf{x}'$.
while $A \leq d$ and $k < n$ **do**
  $A := A + |\mathbf{x}'_k|$
  $k := k + 1$
end while
if $A > d$ **then**
  $\theta = \frac{A-d}{k}$
else
  $\theta = 0$
end if

Lemma 1. *If the clipped* $\mathbf{x}$ *in Problem 1 has* k *nonnegative elements, then these* k *elements must correspond to the* k *largest absolute values of* $\mathbf{x}$*.*

Proof. Suppose there exists an element of $\mathbf{x}$, $|x_k| > |x_i|, i = 0, \ldots, k$ where $|x_0| \geq |x_1| \geq \ldots \geq |x_{k-1}|$. Since $|x_k| - d > 0$, this implies that there would be $k + 1$ nonnegative elements after clipping. This contradicts the assumption that the clipped vector has k nonnegative elements. Therefore, the k nonnegative elements in the clipped vector must correspond to the k largest absolute values of the original vector. □

Lemma 2. *The solution to Problem 1 is unique.*

Proof. Suppose there exists two solutions θ' and θ'', there are two cases to consider. The first case is the clipped $\mathbf{x}'$ and $\mathbf{x}''$ share the same number of nonnegative elements. By Lemma 1, the nonnegative elements of $\mathbf{x}'$ and $\mathbf{x}''$ correspond to the first k elements of the original vector whose elements have been sorted in descending absolute-value order. Then $\sum_{i=0}^{k-1}(|x_i| - \theta') = \sum_{i=0}^{k-1}(|x_i| - \theta'')$. So $k\theta' = k\theta''$, namely $\theta' = \theta''$.
The second case is that $\mathbf{x}'$ and $\mathbf{x}''$ share different numbers of nonnegative elements. Suppose $\theta' < \theta''$, it is easy to get that the numbers of nonnegative elements in $\mathbf{x}'$ and $\mathbf{x}''$ satisfy $k' > k''$. We have the following

$$\begin{aligned}
\sum_{i=0}^{k'-1}(|x_i| - \theta') - \sum_{i=0}^{k''-1}(|x_i| - \theta'') &= \sum_{i=k'}^{k''-1}|x_i| - k'\theta' + k''\theta'' \\
&= \sum_{i=k'}^{k''-1}|x_i| - (k' - k'')\theta'' + k'(\theta'' - \theta') \\
&= \sum_{i=k'}^{k''-1}\underbrace{(|x_i| - \theta'')}_{>0} + k'\underbrace{(\theta'' - \theta')}_{>0} > 0
\end{aligned}$$

which contradicts the requirement that $\sum_{i=0}^{k'-1}(|x_i| - \theta') - \sum_{i=0}^{k''-1}(|x_i| - \theta'') = d - d = 0$. Hence, $\theta' = \theta''$. □

Theorem 1. *Problem 1 can be solved by Algorithm 1.*

Proof. By Lemma 1, we sort elements of $\mathbf{x}$ in descending order by absolute values and get a new vector $\mathbf{x}'$. Suppose there are k nonnegative elements after clipping. By Lemma 2, we assume that

the solution is θ. Then we have

$$\sum_{i=0}^{k-1}(|x_i|-\theta)=d \implies \sum_{i=0}^{k-1}|x_i|-d=k\theta \implies \theta=\frac{\sum_{i=0}^{k-1}|x_i|-d}{k}.$$

Since θ is supposed to be a positive real number, if $\sum_{i=0}^{k-1}|x_i|\leq d$ for all k, then $\theta=0$. Otherwise, $\theta=\frac{\sum_{i=0}^{k-1}|x_i|-d}{k}$.

□

Proposition 1. *The proximal operator with $\|\cdot\|$ has a closed form solution:*

$$\operatorname*{argmin}_{\mathbf{x}}\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2+\lambda\|\mathbf{x}\|=\max(\|\mathbf{y}\|-\lambda,0)\frac{\mathbf{y}}{\|\mathbf{y}\|}, \tag{2}$$

where $\lambda>0$ and $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$.

Proof. Recall the definition of the subdifferential of a function $f(\mathbf{x})$ with regard to $\mathbf{x}$:

$$\partial f(\mathbf{x})=\{\mathbf{v}\in\mathbb{R}^n|\ f(\mathbf{y})\geq f(\mathbf{x})+\mathbf{v}^{\mathrm{T}}(\mathbf{y}-\mathbf{x}),\ \forall\mathbf{x},\mathbf{y}\in\mathrm{dom}(f)\}$$

Given $f(\mathbf{x})=\|\mathbf{x}\|_2$, its gradient at $\mathbf{x}\neq\mathbf{0}$ is $\nabla f(\mathbf{x})=\mathbf{x}/\|\mathbf{x}\|_2$. Its subdifferential at $\mathbf{x}=\mathbf{0}$ is

$$\partial f(\mathbf{0})=\{\mathbf{v}\in\mathbb{R}^n|\|\mathbf{y}\|_2\geq\mathbf{v}^T\mathbf{y},\ \forall\mathbf{y}\in\mathrm{dom}(f)\}\Rightarrow\partial f(\mathbf{0})=\{\mathbf{v}\in\mathbb{R}^n|\|\mathbf{v}\|_2\leq 1\}.$$

Taking derivatives of $\ell_\lambda(\mathbf{x})$ w.r.t $\mathbf{x}$, according to the first-order optimality condition, we get

$$\partial\ell_\lambda(\mathbf{x})=\mathbf{x}-\mathbf{y}+\lambda\partial\|\mathbf{x}\|_2\ni\mathbf{0} \tag{3}$$

We need to consider two cases: $\mathbf{x}\neq\mathbf{0}$ and $\mathbf{x}=\mathbf{0}$.

- Case 1: When $\mathbf{x}\neq\mathbf{0}$, $\partial\|\mathbf{x}\|_2=\mathbf{x}/\|\mathbf{x}\|_2$. Thus,

$$\begin{aligned}\mathbf{x}-\mathbf{y}+\lambda\frac{\mathbf{x}}{\|\mathbf{x}\|_2}=\mathbf{0}\Longleftrightarrow&(1+\frac{\lambda}{\|\mathbf{x}\|_2})\mathbf{x}=\mathbf{y}\ (\mathbf{x}\text{ and }\mathbf{y}\text{ share the same direction})\\ \Longleftrightarrow&(1+\lambda/\|\mathbf{x}\|_2)\|\mathbf{x}\|_2=\|\mathbf{y}\|_2\Longleftrightarrow\|\mathbf{x}\|_2+\lambda=\|\mathbf{y}\|_2\\ \Longleftrightarrow&\|\mathbf{x}\|_2=\|\mathbf{y}\|_2-\lambda\ (\text{this is the amplitude of }\mathbf{x})\\ \Longleftrightarrow&\mathbf{x}=(\|\mathbf{y}\|_2-\lambda)\frac{\mathbf{y}}{\|\mathbf{y}\|_2}=(1-\frac{\lambda}{\|\mathbf{y}\|_2})\mathbf{y}\ (\text{with }\|\mathbf{y}\|_2>\lambda),\end{aligned}$$

  where the second last "$\Longleftrightarrow$" indicates $\|\mathbf{y}\|_2>\lambda$ since $\mathbf{x}\neq\mathbf{0}$. Note that the "$\Longleftarrow$" of the second "$\Longleftrightarrow$" follows from the last "$\Longleftrightarrow$", i.e., $\mathbf{x}=(\|\mathbf{y}\|_2-\lambda)\frac{\mathbf{y}}{\|\mathbf{y}\|_2}$.

- Case 2: When $\mathbf{x}=\mathbf{0}$, after substituting it into (3), we have

$$\mathbf{x}=\mathbf{0}\Longleftrightarrow-\mathbf{y}+\lambda\mathbf{v}\ni\mathbf{0}\Longleftrightarrow\mathbf{y}\in\lambda\mathbf{v}\Longleftrightarrow\|\mathbf{y}\|_2\leq\lambda,$$

  where $\mathbf{v}\in\partial f(\mathbf{0})$. The last "$\Longleftrightarrow$" follows from $\|\mathbf{v}\|_2\leq 1$.

Combining the above two cases, we get the solution to the group lasso problem

$$\mathbf{x}=\begin{cases}(1-\frac{\lambda}{\|\mathbf{y}\|_2})\mathbf{y}, & \text{if }\|\mathbf{y}\|_2>\lambda\\ \mathbf{0}, & \text{if }\|\mathbf{y}\|_2\leq\lambda.\end{cases}$$

For notational simplicity,

$$\mathbf{x}=(1-\frac{\lambda}{\|\mathbf{y}\|_2})_+\mathbf{y}=\max\{1-\frac{\lambda}{\|\mathbf{y}\|_2},0\}\mathbf{y}$$

□

Remark 1. We can also show the sufficiency of $\|\mathbf{y}\|_2 \leq \lambda$ for $\mathbf{x} = \mathbf{0}$ to be a solution by contradiction. Specifically, given $\|\mathbf{y}\| \leq \lambda$, we suppose $\mathbf{x} \neq \mathbf{0}$, then $\partial f(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$ holds. According to the first-order optimality condition,

$$\mathbf{x} - \mathbf{y} + \lambda \mathbf{x}/\|\mathbf{x}\|_2 \ni \mathbf{0} \Longrightarrow \mathbf{x}(1 + \lambda/\|\mathbf{x}\|_2) \ni \mathbf{y} \Longrightarrow \|\mathbf{x}\|_2 + \lambda = \|\mathbf{y}\|_2 \leq \lambda \Longrightarrow \|\mathbf{x}\|_2 \leq 0$$

which contradicts $\mathbf{x} \neq \mathbf{0}$. Hence, $\mathbf{x} = \mathbf{0}$ must hold.

4 My thoughts

Problem 2. *Given y, solve the following problem*

$$\min_x \ \frac{1}{2}\|x - y\|_2^2 + \sum_{i=1}^{g} \epsilon_i \mathbf{1}_{\|x_{G_i}\|_2} \tag{4}$$

where $x = \cup_{i=1}^{g} x_{G_i}, G_i \cap G_j = \emptyset, \forall i, j$.

At the first glimpse, this problem is easy to solve, but it may be not as easy as you imagined. Take a try. Can you find a closed-form solution for the above problem? If so, try to show its correctness.

How about the overlapping case? Try to find an algorithm for this case and prove its correctness or convergence.

Can we think of both x and ϵ as parameters? I think so. Then the resultant ϵ_i will indicate the importance of the corresponding group. Specifically, the greater ϵ_i is, the greater the penalty imposed on that group is, which implys the less important that group is. In this way, we can find out which groups are critical for our task. In some sense, this means we can find out which inputs determine the output and which do not. Even for multiple heterogeneous inputs, we can find out the causalities between some groups of the inputs and outputs. This is very helpful for our decision tasks.

If there are 10 parameters in our model, then there are at most $2^{10} = 1024$ groups and exploring 1024 groups is not a strenuous task for a computer. But if there are 20 parameters, there will be $1024 \times 1024 > 10^6$ groups to explore. How can we speed up? Try to propose a more efficient algorithm. First sort them by their L-2 norms, then suppose there are only k groups playing key roles...