Equivalence of Several L1 Sparsity Problem

http://freemind.pluskid.org/machine-learning/equivalence-of-several-l1-sparsity-problem

posted on Free Mind on June 26, 2013
generated with pandoc on December 3, 2015
category: Machine Learning

tags: Sparsity, Regularization, Compressive Sensing

有这么几个 ℓ_1-norm 相关的稀疏优化问题在某种意义下是等价的。首先是如下问题：

$$x^a(\lambda) = \underset{x}{\operatorname{argmin}} \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1$$

这个形式通常来自于 regularized linear regression 等问题，其中使用 ℓ_1-norm 作为 regularizer 会偏向于得到稀疏的解，我们之前也曾简要地讨论过这个问题。同样的形式还可以解释为使用 Laplace prior 的 MAP 参数估计所得到的目标函数。

另一个问题是主要来自于 Compressive Sensing 的如下形式：

$$x^b(\epsilon) = \underset{x}{\operatorname{argmin}} \|x\|_1 \quad s.t. \|Ax - y\|^2 \le \epsilon$$

第三个是把目标函数和 constraint 的位置反过来一下的形式，这实际上是 LASSO 的原始形式：

$$x^c(t) = \underset{x}{\operatorname{argmin}} \frac{1}{2}\|Ax - y\|^2 \quad s.t. \|x\|_1 \le t$$

关于这几个问题的等价性似乎直观上来讲也是可以接受的，而且大部分论文在提及这里的等价性的时候也都只是一笔带过或者只是很模糊地提了一下在相应的参数（这里要求它们都是正实数）λ、ϵ 和 t 之间满足某个对应关系的时候它们的最优解是可以对应起来的。不知道有没有记错的是，好像若干年前还在某次组会上被提起，老师问大家有没有谁能上黑板写一下为什么等价，大家面面相觑..... ^_^bb 今天在这里简单总结一下，似乎也并不是 trivial 的。

开始之前需要一点点关于 subgradient 和 subdifferential 的了解，可以参考我们之前所给的定义[1]另外，知道如下的性质会对 subdifferential 的计算带来方便。

定理 1 (Theorem 23.8 (Convex Analysis, by R.T. Rockafellar)) 令 $f = f_1 + \ldots + f_m$，其中 $f_1, \ldots, f_m$ 为 $\mathbb{R}^n$ 中的 *proper convex function*，则对任意 x，我们有：

$$\partial f(x) \supset \partial f_1(x) + \ldots + \partial f_m(x)$$

[1] 这两个量可以针对任意函数定义也可以指针对凸函数来定义，似乎并没有像标准的 gradient 和 differential 那样有标准的定义，不过在这里并不影响我们的问题分析，因为我们处理的都是凸函数。

如果更进一步地，凸集 $ri(domf_i), i = 1, \ldots, m$ 有非空交集的话，对任意 x，我们有

$$\partial f(x) = \partial f_1(x) + \ldots + \partial f_m(x)$$

再知道一下凸函数 f 在 x^* 处取到最优值当且仅当 $0 \in \partial f(x^*)$，就可以开始我们的分析了。为了方便起见，我们把三个问题分别叫做 URLS (Unconstrained Regularized Least Square)、CS (Compressive Sensing) 和 LASSO（注意这些并不是标准的叫法，只是在这篇文章中用的临时名字)，并定义如下三个目标函数：

$$\begin{aligned} f^a(x) &= \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1 \\ f^b(x) &= \|x\|_1 + \delta(x|B_2^\epsilon) \\ f^c(x) &= \frac{1}{2}\|Ax - y\|^2 + \delta(x|B_1^t) \end{aligned}$$

其中集合 indicator 函数 $\delta(x|C)$ 定义为

$$\delta(x|C) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

而 B_1^t 和 B_2^ϵ 则分别是对应的 ℓ_1-ball 和 ℓ_2-ball 集合[2]。通过使用这种非 regular 的函数，可以在形式上把后面两个问题也写成 unconstrained 的问题，在分析上变得方便一点。

[2] 注意这里的 B_2^ϵ Ball 并不一定是中心在原点的，为了简化符号起见就没有写关于 A 和 y 的依赖性了。

首先来看一下 URLS 和 LASSO 之间的等价性，也就是 $x^a(\lambda)$ 和 $x^c(t)$ 什么时候会相等。

第一点可以明确的是，如果 $t < \|x^a(\lambda)\|_1$，那么两者最优值肯定是不想等的，因为此时 URLS 的最优解在 LASSO 问题下是 infeasible 的。反过来，如果 $t > \|x^a(\lambda)\|_1$ 也是不行的。假设不然，$x^a(\lambda) = x^c(t)$，此时 LASSO 问题的最优解在 B_1^t 的内部取到，而在内部目标函数 $f^c(x)$ 是可微的，于是 subdifferential 也退化为单点集：

$$\partial f^c(x^c) = \{A^T(Ax^c - y)\}$$

根据最优解的充要条件，$A^T(Ax^c - y) = 0$。但是这会和 $0 \in \partial f^a(x^a)$ 的条件矛盾，这里我们需要计算一下 ℓ_1-norm 的 subdifferential，因为 $\|x\|_1 = \sum_i |x_i|$，所以可以利用刚才的定理 1 来做计算。而对于每一个 component $h_i(x) = \lambda|x_i|$，根据定义可以算出：

$$\partial h_i(x) = \begin{cases} \{\alpha e_i : \alpha \in [-\lambda, \lambda]\} & x_i = 0 \\ \{\lambda \text{sign}(x_i) e_i\} & x_i \neq 0 \end{cases}$$

其中 e_i 是第 i 维上的单位向量。根据定理 1 和最优解相等的假设，我们有

$$\begin{aligned} \partial f^a(x^a) &= A^T(Ax^a - y) + \sum_i \partial h_i(x^a) \\ &= A^T(Ax^c - y) + \sum_i \partial h_i(x^a) \\ &= \sum_i \partial h_i(x^a) \end{aligned}$$

所以为了要求 $0 \in \partial f^a(x^a)$，而 $\lambda \neq 0$ 的情况下，就只能 x^a 的每一个分量都必须等于零了。小结一下就是：除非两个问题的最优解同时是 $x^c = x^a = 0$ 这种特殊情况，否则在 $t > \|x^a(\lambda)\|_1$ 的时候也会导出矛盾。

因此就只剩下 $t = \|x^a(\lambda)\|_1$ 的情况了。很容易看到这种情况下 x^c 必须等于 x^a，因为如果不是这样的话，根据 x^c 是 LASSO 问题的最优解 (而 x^a 不是)，我们可以得到

$$\|Ax^c - y\|^2 < \|Ax^a - y\|^2$$

另一方面，根据 x^c 的 feasibility，我们有 $\|x^c\|_1 \leq t = \|x^a\|_1$。这样一来，我们实际上有

$$f^a(x^c) < f^a(x^a)$$

因此 x^a 不可能是 URLS 问题的最优解，得到矛盾。这个是充分性，必要性同样可以通过分析 subdifferential 的结构来得到。因此结论是，在参数满足关系 $t = \|x^a(\lambda)\|_1$ 的情况下，LASSO 问题和 URLS 问题是等价的。需要注意的是这虽然是一个一一对应（因为接下来我们要导出逆向的映射)，但是并不是有一个显式的易于计算的映射可以用于在实际问题中直接得出一个问题关于另一个问题的等价形式——因为我们可以看到要得到映射的结果需要先把其中一个问题的最优解算出来。

反过来的映射，我们还是通过 $0 \in \partial f^a(x^a)$ 这个条件，根据刚才我们得到的 ℓ_1-norm 的 subdifferential 的形式，可以再把这个条件显式地写出来：

$$\begin{aligned} \left(A^T(y - Ax^a)\right)_i &= \lambda \text{sign}(x_i^a), \quad \text{if } x_i^a \neq 0 \\ \left|\left(A^T(y - Ax^a)\right)_i\right| &\leq \lambda, \quad \text{if } x_i^a = 0 \end{aligned}$$

所以除非是 $x^a = 0$ 这个解的话，我们必须有 $\lambda = \max_i |(A^T(y - Ax^a))_i|$。此时如果两个最优解相等的话，$\lambda$ 和 t 的关系自然就变成了 $\lambda = \max_i |(A^T(y - Ax^c(t)))_i|$。这是必要条件。

反过来如果我们直接令 $\lambda = \max_i |(A^T(y - Ax^c(t)))_i|$ 的话，显然将 $x = x^c(t)$ 带入是可以满足上面的条件的，根据凸优化解的唯一性，也就有 $x^a(\lambda) = x^c(t)$。

小结一下，URLS 问题和 LASSO 问题之间有一个参数的一一对应关系使得对应的问题等价，虽然在实际中一般没法直接计算，不过形式上两者之间的相互映射如下：

$$\begin{aligned} t &= \|x^a(\lambda)\|_1 \\ \lambda &= \max_i |(A^T(y - Ax^c(t)))_i| \end{aligned}$$

用同样的方法可以分析 URLS 和 CS 之间的关系。具体来说，我们可以验证，当 $\epsilon = \|Ax^a(\lambda) - y\|^2$ 时，会有 $x^a(\lambda) = x^b(\epsilon)$。而当两个最优解相等的时候，CS 问题如果排除了零解这种情况，那么最优解肯定出现在 B_2^ϵ 边界上，于是反过来也能得到

$$\epsilon = \|Ax^b(\epsilon) - y\|^2 = \|Ax^a(\lambda) - y\|^2$$

逆向的从 ϵ 到 λ 的映射则和之前的分析完全一样了，就不重复写一遍了。

一般来说，从 CS 或者 LASSO 的形式往 URLS 的形式转化的场合比较多一点，因为后面两者问题的 constraint 的具体值的 “物理” 意义通常可以比较容易诠释，而且这类的 formulation 会更多地直接从具体问题中抽象出来。但是另一方面 URLS 的优化算法方面的研究似乎又是最多的，所以在实际中通常需要把后两种问题转化成第一种，但是实际中通常 λ 等价情况下的具体值是没法事先算出来的——如果先算出原始问题的最优解再带入上面的公式，则问题都已经解决了后续的步骤就毫无意义了。所以有时候会直接采用经验或者纯 empirically 地使用 cross validation 选择最优的 λ 之类的；另外一种可能就是可以使用像 LARS 之类的 Homotopy Method，解出随着 λ 变化所能得到的一系列 $x^a(\lambda)$ 解，然后再在得到的一系列解中去找和原始问题的对应关系。