

# Sparsity and Some Basics of L1 Regularization

<http://freemind.pluskid.org/machine-learning/sparsity-and-some-basics-of-l1-regularization>

Sparsity 是当今机器学习领域中的一个重要话题。John Lafferty 和 Larry Wasserman 在 2006 年的一篇评论中提到：

posted on [Free Mind](#) on May 24, 2012  
generated with pandoc on December 3, 2015  
category: Machine Learning

“Some current challenges ... are high dimensional data, sparsity, semi-supervised learning, the relation between computation and risk, and structured prediction.”

—John Lafferty and Larry Wasserman. Challenges in statistical machine learning. Statistica Sinica. Volume 16, Number 2, pp. 307-323, 2006.

tags: Sparsity, Regularization

Sparsity 的最重要的“客户”大概要属 high dimensional data 了吧。现在的机器学习问题中，具有非常高维度的数据随处可见。例如，在文档或图片分类中常用的 bag of words 模型里，如果词典的大小是一百万，那么每个文档将由一百万维的向量来表示。高维度带来的一个问题就是计算量：在一百万维的空间中，即使计算向量的内积这样的基本操作也会是非常费力的。不过，如果向量是稀疏的话（事实上在 bag of words 模型中文档向量通常都是非常稀疏的），例如两个向量分别只有  $L_1$  和  $L_2$  个非零元素，那么计算内积可以只使用  $\min(L_1, L_2)$  次乘法完成。因此稀疏性对于解决高维度数据的计算量问题是非常有效的。

当然高维度带来的问题不止是在计算量上。例如在许多生物相关的问题中，数据的维度非常高，但是由于收集数据需要昂贵的实验，因此可用的训练数据却相当少，这样的问题通常称为 “small  $n$ , large  $p$  problem”——我们一般用  $n$  表示数据点的个数，用  $p$  表示变量的个数，即数据维度。当  $p \gg n$  的时候，不做任何其他假设或者限制的话，学习问题基本上是无法进行的。因为如果用上所有变量的话， $p$  越大，通常会导致模型越复杂，但是反过来  $n$  有很小，于是就会出现很严重的 overfitting 问题。例如，最简单的线性回归模型：

$$f(\mathbf{x}) = \sum_{j=1}^p w^j x^j = \mathbf{w}^T \mathbf{x}$$

使用 square loss 来进行学习的话，就变成最小化如下的问题

$$J(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 = \frac{1}{n} \|\mathbf{y} - X\mathbf{w}\|^2$$

这里  $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$  是数据矩阵，而  $\mathbf{y} = (y_1, \dots, y_n)^T$  是由标签组成的列向量。该问题具有解析解

$$\hat{\mathbf{w}} = \left( X^T X \right)^{-1} X^T \mathbf{y} \quad (1)$$

然而，如果  $p > n$  的话，矩阵  $X^T X$  将会不是满秩的，而这个解也没法算出来。或者更确切地说，将会有无穷多个解。也就是说，我们的数据不足以确定一个解，如果我们从所有可行解里随机选一个的话，很可能并不是真正好的解，总而言之，我们 **overfitting** 了。

解决 **overfitting** 最常用的办法就是 **regularization**，例如著名的 **ridge regression** 就是添加一个  $\ell_2$  **regularizer**：

$$J_R(\mathbf{w}) = \frac{1}{n} \|\mathbf{y} - X\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|^2$$

直观地来看，添加这个 **regularizer** 会使得模型的解偏向于 **norm** 较小的  $\mathbf{w}$ 。从凸优化的角度来说，最小化上面这个  $J(\mathbf{w})$  等价于如下问题：

$$\min_{\mathbf{w}} \frac{1}{n} \|\mathbf{y} - X\mathbf{w}\|^2, \quad s.t. \|\mathbf{w}\| \leq C$$

其中  $C$  是和  $\lambda$  一一对应的是个常数。也就是说，我们通过限制  $\mathbf{w}$  的 **norm** 的大小实现了对模型空间的限制，从而在一定程度上（取决于  $\lambda$  的大小）避免了 **overfitting**。不过 **ridge regression** 并不具有产生稀疏解的能力，得到的系数  $\mathbf{w}$  仍然需要数据中的所有特征才能计算预测结果，从计算量上来说并没有得到改观。

不过，特别是在像生物或者医学等通常需要和人交互的领域，稀疏的解除了计算量上的好处之外，更重要的是更具有“可解释性”。比如说，一个病如果依赖于 5 个变量的话，将会更易于医生理解、描述和总结规律，但是如果依赖于 5000 个变量的话，基本上就超出人肉可处理的范围了。

在这里引入稀疏性的方法是用  $\ell_1$  **regularization** 代替  $\ell_2$  **regularization**，得到如下的目标函数

$$J_L(\mathbf{w}) = \frac{1}{n} \|\mathbf{y} - X\mathbf{w}\|^2 + \lambda \|\mathbf{w}\|_1 \quad (2)$$

该问题通常被称为 **LASSO** (**least absolute shrinkage and selection operator**)。LASSO 仍然是一个 **convex optimization** 问题，不过不再具有解析解。它的优良性质是能产生稀疏性，导致  $\mathbf{w}$  中许多项变成零。

可是，为什么它能产生稀疏性呢？这也是一直让我挺感兴趣的一个问题，事实上在之前申请学校的时候一次电话面试中我也被问到了这个问题。我当时的回答是背后的理论我并不是很清楚，但是我知道一个直观上的理解。下面我们就先来看一下这个直观上的理解。

首先，和 ridge regression 类似，上面形式的 LASSO 问题也等价于如下形式：

$$\min_{\mathbf{w}} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2, \quad \text{s.t.} \|\mathbf{w}\|_1 \leq C$$

也就是说，我们将模型空间限制在  $\mathbf{w}$  的一个  $\ell_1$ -ball 中。为了便于可视化，我们考虑两维的情况，在  $(w^1, w^2)$  平面上可以画出目标函数的等高线，而约束条件则成为平面上半径为  $C$  的一个 norm ball。等高线与 norm ball 首次相交的地方就是最优解。如图 1 所示：

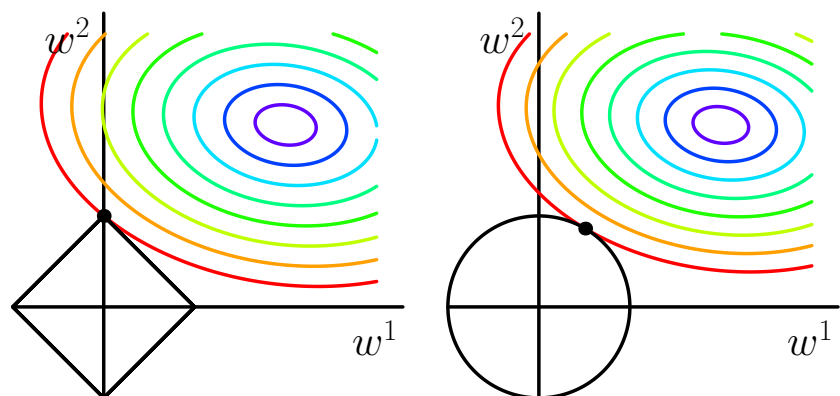


Figure 1: .....

(a)  $\ell_1$ -ball meets quadratic function.  $\ell_1$ -ball has corners. It's very likely that the meet-point is at one of the corners.

(b)  $\ell_2$ -ball meets quadratic function.  $\ell_2$ -ball has no corner. It is very unlikely that the meet-point is on any of axes."

可以看到， $\ell_1$ -ball 与  $\ell_2$ -ball 的不同就在于他在和每个坐标轴相交的地方都有“角”出现，而目标函数的测地线除非位置摆得非常好，大部分时候都会在角的地方相交。注意到在角的位置为产生稀疏性，例如图中的相交点就有  $w^1 = 0$ ，而更高维的时候（想象一下三维的  $\ell_1$ -ball 是什么样的？）除了角点以外，还有很多边的轮廓也是既有很大的概率成为第一次相交的地方，又会产生稀疏性。

相比之下， $\ell_2$ -ball 就没有这样的性质，因为没有角，所以第一次相交的地方出现在具有稀疏性的位置的概率就变得非常小了。这就从直观上来解释了为什么  $\ell_1$  regularization 能产生稀疏性，而  $\ell_2$  regularization 不行的原因了。

不过，如果只限于 intuitive 的解释的话，就不那么好玩了，但是背后完整的理论又不是那么容易能够搞清楚的，既然这次的标题是 Basics，我们就先来看一个简单的特殊情况好了。

接下来我们考虑 orthonormal design 的情况： $(1/n)\mathbf{X}^T\mathbf{X} = \mathbf{I}$ 。然后看看 LASSO 的解具体是什么样子的。注意 orthonormal design 实际上是要求特征之间相互正交。这可以通过对数据进行 PCA 以及模长 normalize 来实现。

注意到 LASSO 的目标函数 (2) 是 convex 的，根据 KKT 条件，在最优

解的地方要求  $\text{gradient } \nabla J_L(\mathbf{w}) = 0$ 。不过这里有一小问题： $\ell_1$ -norm 不是光滑的，不存在  $\text{gradient}$ ，所以我们需要用一点 **subgradient** 的东西。

**定义 1 (subgradient; subdifferential)** 对于在  $p$  维欧氏空间中的凸开子集  $U$  上定义的实值函数  $f: U \rightarrow \mathbb{R}$ ，一个向量  $p$  维向量  $\mathbf{v}$  称为  $f$  在一点  $\mathbf{x}_0 \in U$  处的 *subgradient*，如果对于任意  $\mathbf{x} \in U$ ，满足

$$f(\mathbf{x}) - f(\mathbf{x}_0) \geq \mathbf{v} \cdot (\mathbf{x} - \mathbf{x}_0)$$

由在点  $\mathbf{x}_0$  处的所有 *subgradient* 所组成的集合称为  $\mathbf{x}_0$  处的 *subdifferential*，记为  $\partial f(\mathbf{x}_0)$ 。

注意 *subgradient* 和 *subdifferential* 只是对凸函数定义的。例如一维的情况， $f(x) = |x|$ ，在  $x = 0$  处的 *subdifferential* 就是  $[-1, +1]$  这个区间(集合)。注意在  $f$  的 *gradient* 存在的点，*subdifferential* 将是由 *gradient* 构成的一个单点集合。这样就将 *gradient* 的概念加以推广了。这个推广有一个很好的性质。

**性质 1 (condition for global minimizer)** 点  $\mathbf{x}_0$  是凸函数  $f$  的一个全局最小值点，当且仅当  $0 \in \partial f(\mathbf{x}_0)$ 。

证明很简单，将  $0 \in \partial f(\mathbf{x}_0)$  带入定义 1 中的那个式子立即就可以得到。有了这个工具之后，就可以对 LASSO 的最优解进行分析了。在此之前，我们先看一下原始的 least square 问题的最优解 (1) 现在变成了什么样子，由于 orthonormal design，我们有

$$\hat{\mathbf{w}} = \frac{1}{n} X^T \mathbf{y} \quad (3)$$

然后我们再来看 LASSO，假设  $\bar{\mathbf{w}} = (\bar{w}^1, \dots, \bar{w}^p)^T$  是  $J_L(\mathbf{w})$  的全局最优值点。考虑第  $j$  个变量  $\bar{w}^j$ ，有两种情况。

**gradient 存在**，此时  $\bar{w}^j \neq 0$

由于 *gradient* 在最小值点必须要等于零，我们有

$$\left. \frac{\partial J_L(\mathbf{w})}{\partial w^j} \right|_{\bar{w}^j} = 0$$

亦即

$$-\frac{2}{n} \left( X^T \mathbf{y} - X^T X \bar{\mathbf{w}} \right)_j + \lambda \text{sign}(\bar{w}^j) = 0$$

根据 orthonormal design 性质以及 least square 问题在 orthonormal design 时的解 (3) 化简得到

$$\bar{w}^j = \hat{w}^j - \frac{\lambda}{2} \text{sign}(\bar{w}^j)$$

从这个式子也可以明显看出  $\bar{w}^j$  和  $\hat{w}^j$  是同号的，于是  $\text{sign}(\bar{w}^j)$  等于  $\text{sign}(\hat{w}^j)$ ，所以上面的式子变为

$$\bar{w}^j = \hat{w}^j - \frac{\lambda}{2} \text{sign}(\hat{w}^j) = \text{sign}(\hat{w}^j) \left( |\hat{w}^j| - \frac{\lambda}{2} \right)$$

再用一次  $\text{sign}(\hat{w}^j) = \text{sign}(\bar{w}^j)$ ，两边同时乘以  $\text{sign}(\bar{w}^j)$ ，可以得到

$$|\hat{w}^j| - \frac{\lambda}{2} = |\bar{w}^j| \geq 0$$

于是刚才的式子可以进一步写为

$$\bar{w}^j = \text{sign}(\hat{w}^j) \left( |\hat{w}^j| - \frac{\lambda}{2} \right)_+ \quad (4)$$

这里  $(x)_+ = \max\{x, 0\}$  表示  $x$  的正部。

**gradient 不存在，此时  $\bar{w}^j = 0$**

根据 **subgradient** 在最小值点处的性质的性质，此时比有

$$\begin{aligned} 0 = \bar{w}^j \in \partial J_L(\bar{\mathbf{w}}) &= \left\{ -\frac{2}{n} \left( X^T \mathbf{y} - X^T X \bar{\mathbf{w}} \right)_j + \lambda e : e \in [-1, 1] \right\} \\ &= \left\{ 2\bar{w}^j - 2\hat{w}^j + \lambda e : e \in [-1, 1] \right\} \end{aligned}$$

亦即存在  $e_0 \in [-1, 1]$  使得

$$0 = 2\bar{w}^j - 2\hat{w}^j + \lambda e_0 = -2\hat{w}^j + \lambda e_0$$

于是

$$|\hat{w}^j| = \frac{\lambda}{2} |e_0| \leq \frac{\lambda}{2}$$

又因为  $\bar{w}^j = 0$ ，所以这个时候式子也可以统一为 (4) 的形式。如此一来，在 **orthonormal design** 的情况下，LASSO 的最优解就可以写为 (4)，可以用图 2 形象地表达出来。

图上画了原始的 **least square** 解，LASSO 的解以及 **ridge regression** 的解，用上面同样的方法（不过由于 **ridge regularizer** 是 **smooth** 的，所以过程却简单得多）可以得知 **ridge regression** 的解是如下形式

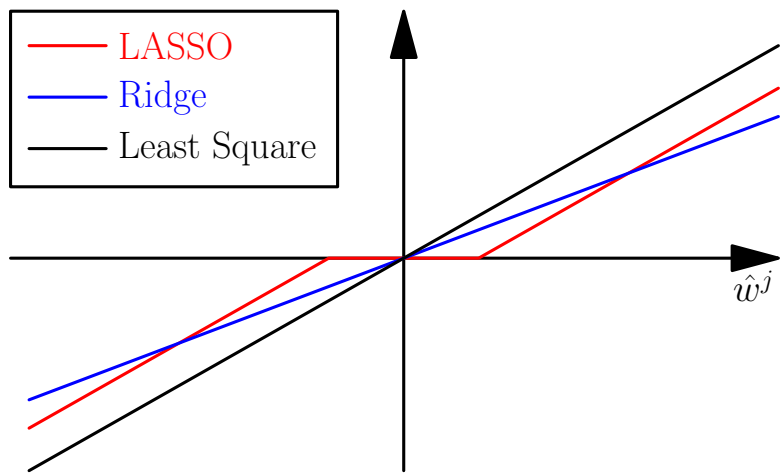


Figure 2: .....

$$\frac{n}{1+n\lambda}\hat{w}^j$$

可以 ridge regression 只是做了一个全局缩放，而 LASSO 则是做了一个 soft thresholding：将绝对值小于  $\lambda/2$  的那些系数直接变成零了，这也就更加令人信服地解释了 LASSO 为何能够产生稀疏解了。